CSEP 546: Machine Learning Autumn 2026

Jamie Morgenstern



Course Website

https://courses.cs.washington.edu/courses/csep546/26au/

Everything you need to know is there:

- Lecture slides
- Links to places to get help (Ed Discussion, staff email)
- Homeworks and due dates
- Practice quiz generators
- Textbook and other references
- Office hours
- Grading, FAQ, etc.

Autumn 2026 logistics

Mondays, 6:30–9:20 PM Pacific

CSE2 G10

October 5, 12, 19, 26

November 2, 9, 16, 23, 30; December 7

No Monday registrar holidays in this schedule

Regular instruction ends December 11

No regular lecture December 14

Recordings: Panopto in Canvas; questions: Ed

Course Staff - Instructor

Jamie Morgenstern
Instructor, CSEP 546
Autumn 2026



Course Staff - Teaching Assistant

- *Bernie Zhu*
- *Teaching assistant, Autumn 2026*
- *Course staff: csep546-staff@cs.washington.edu*
- *Course questions: Ed Discussion in Canvas*
- *Office hours: see the course website*

CSE546: Machine Learning

CSE546: Machine Learning

What this class is:

- **Fundamentals of ML:** bias/variance tradeoff, overfitting, optimization and computational tradeoffs, supervised learning (e.g., linear, boosting, deep learning), unsupervised models (e.g. k-means, EM, PCA)
- **Preparation for further learning:** the field is fast-moving, you will be able to apply the basics and teach yourself the latest

CSE546: Machine Learning

What this class is:

- **Fundamentals of ML:** bias/variance tradeoff, overfitting, optimization and computational tradeoffs, supervised learning (e.g., linear, boosting, deep learning), unsupervised models (e.g. k-means, EM, PCA)
- **Preparation for further learning:** the field is fast-moving, you will be able to apply the basics and teach yourself the latest

What this class is not:

- **Survey course:** laundry list of algorithms, how to win Kaggle
- **An easy course:** familiarity with intro linear algebra and probability are assumed, homework will be time-consuming

Prerequisites

- Formally:
  - MATH 308, CSE 312, STAT 390 or equivalent
- Familiarity with:
  - Linear algebra
    - linear dependence, rank, linear equations, SVD
  - Multivariate calculus
  - Probability and statistics
    - Distributions, marginalization, moments, conditional expectation
  - Algorithms
    - Basic data structures, complexity
- “Can I learn these topics concurrently?”
- Use HW0 to judge skills
- **See website for review materials!**

Grading

- *Homework determines your provisional grade*
- *HW0: 10%; HW1–3: 20% each; HW4: 30%*
- *Document collaboration and AI use; understand your submitted work*

- *Quiz performance sets the highest grade you can earn*
- *Eight in-person quizzes; three questions each*
- *Drop the two lowest quizzes, including absences*
- *Best six quizzes count: 18 points total*

Quiz grade eligibility

- *8 quizzes × 3 questions = 24 points offered*
- *Drop 2 whole quizzes, including zeros for absences*
- *Best 6 × 3 questions = 18 counted points*
- *15/18: eligible for 3.5–4.0; 13/18: 3.0–3.4*
- *11/18: eligible for 2.5–2.9; 9/18: 2.0–2.4*
- *Final grade = min(homework grade, quiz grade ceiling)*
- *Practice generators on the course website*

Independent understanding

- *Quizzes are the expected route*
- *Complete independently: no notes, devices or AI*
- *Only a best-six score below 9/18 permits a final*
- *The optional final covers the quarter's essential concepts*
- *Its result replaces the quiz result, even if lower*
- *At 9/18 or above, no final to improve quiz eligibility*
- *Approved accommodations and required exceptions apply*

Homework

- HW 0 is out (**Due Monday, October 12, 2026 at 11:59 PM Pacific**)
  - Short *review*
  - Work individually, treat as barometer for readiness
- HW 1,2,3,4
  - They are not easy or short. Start early.
- Submit to Gradescope
- Regrade requests on Gradescope
- **There is no credit for late work**
  - **5 total late days**
  - **at most 2 days used per assignment.**

Homework

- HW 0 is out (**Due Monday, October 12, 2026 at 11:59 PM Pacific**)
  - Short *review*
  - Work individually, treat as barometer for readiness
- HW 1,2,3,4
  - They are not easy or short. Start early.
- Submit to Gradescope
- Regrade requests on Gradescope
- **There is no credit for late work, 5 late days**

Homework

- HW 0 is out (**Due Monday, October 12, 2026 at 11:59 PM Pacific**)
  - Short *review*
  - Work individually, treat as barometer for readiness
- HW 1,2,3,4
  - They are not easy or short. Start early.
- Submit to Gradescope
- Regrade requests on Gradescope

1. All code must be written in Python
2. All written work must be typeset (e.g., LaTeX)

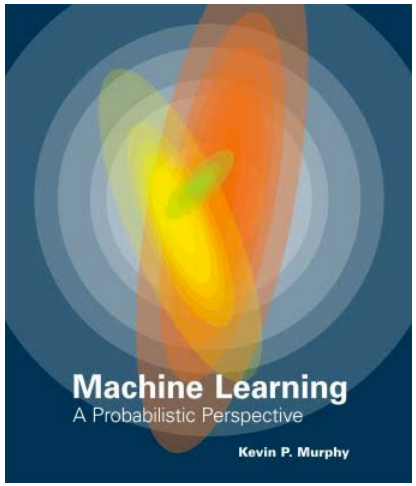
See course website for tutorials and references.

Communication Channels

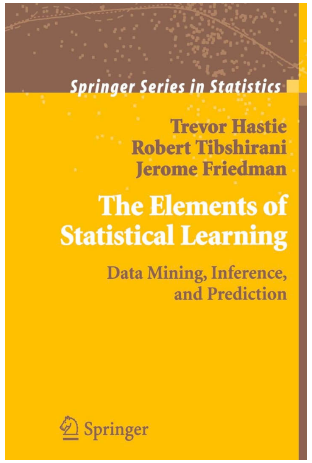
- Course questions and announcements: Ed Discussion
- Access Ed and Panopto recordings through Canvas
- Personal concerns: csep546-staff@cs.washington.edu
- Homework submission/regrades: see assignment instructions
- Office hours will be announced on Ed
- Anonymous feedback: feedback.cs.washington.edu

Textbooks

- Required Textbook:
  - ***Machine Learning: a Probabilistic Perspective*; Kevin Murphy**



- Optional Books (free PDF):
  - *The Elements of Statistical Learning: Data Mining, Inference, and Prediction;* Trevor Hastie, Robert Tibshirani, Jerome Friedman



Addcodes

- Email: Elle Brown (ellean@cs.washington.edu) for addcodes

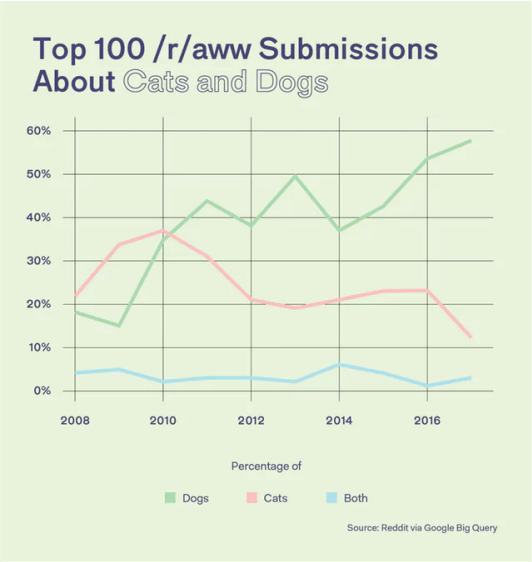
Enjoy!

- ML is becoming ubiquitous in science, engineering and beyond
- It's one of the hottest topics in industry today
- This class should give you the basic foundation for applying ML and developing new methods
- The fun begins…

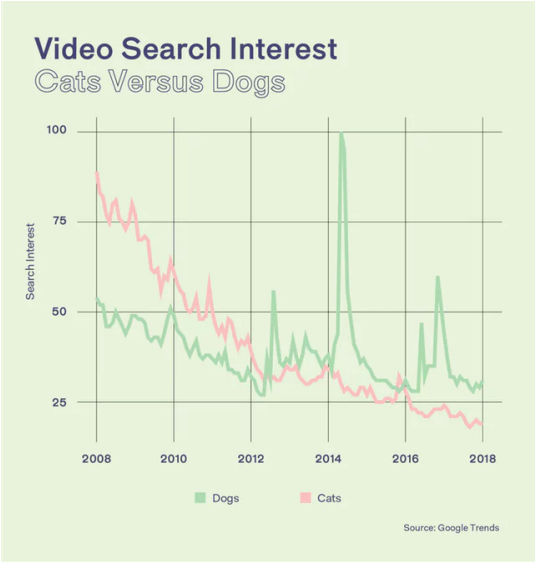
Traditional algorithms

Social media mentions of Cats vs. Dogs

Reddit



Google



Twitter?

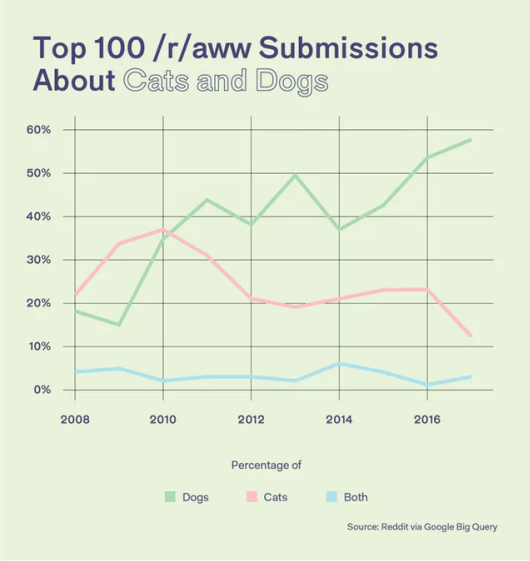
Write a program that sorts tweets into those containing "**cat**", "**dog**", or ***other***

Graphics courtesy of https://theoutline.com/post/3128/dogs-cats-internet-popularity?zd=1

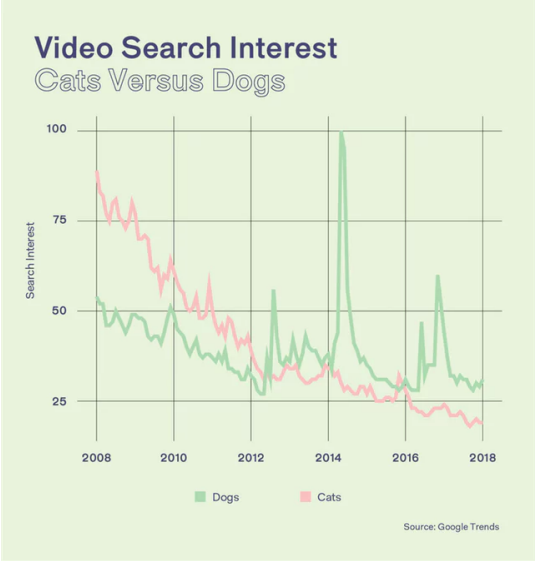
Traditional algorithms

Social media mentions of Cats vs. Dogs

Reddit



Google



Twitter?

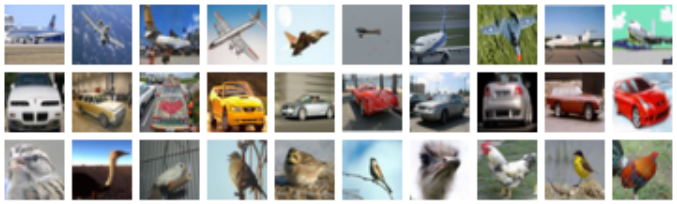
```
cats = []
dogs = []
other = []
for tweet in tweets:
    if "cat" in tweet:
        cats.append(tweet)
    elseif "dog" in tweet:
        dogs.append(tweet)
    else:
        other.append(tweet)
return cats, dogs, other
```

Write a program that sorts tweets into those containing "**cat**", "**dog**", or ***other***

Graphics courtesy of https://theoutline.com/post/3128/dogs-cats-internet-popularity?zd=1

Machine learning algorithms

Write a program that sorts image
into those containing "**birds**",
"**airplanes**", or ***other.***



airplane

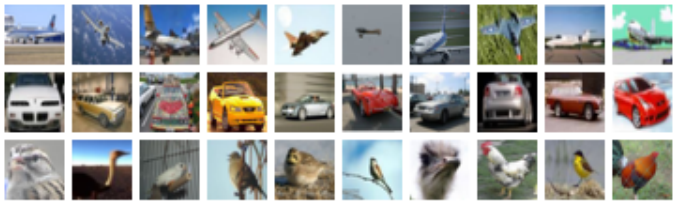
other

bird

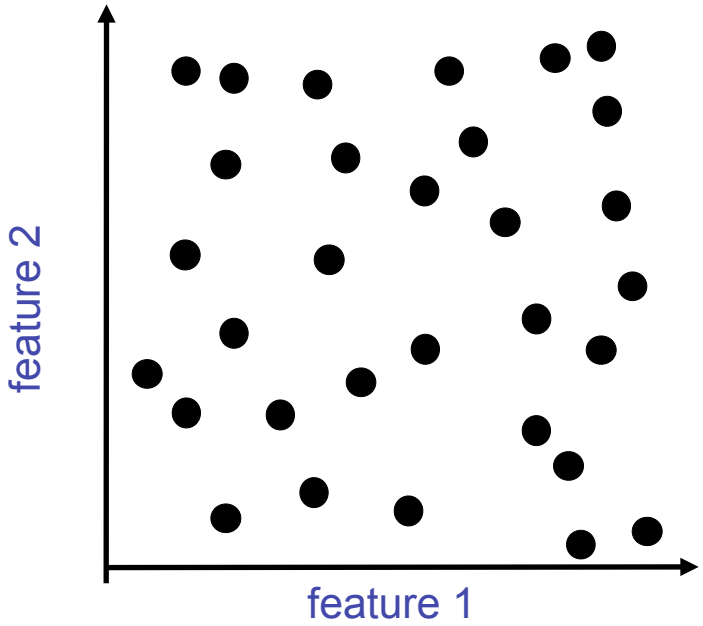
```
birds = []
planes = []
other = []
for image in images:
    if bird in image:
        birds.append(image)
    elseif plane in image:
        planes.append(image)
    else:
        other.append(tweet)
return birds, planes, other
```

Machine learning algorithms

Write a program that sorts image
into those containing "**birds**", "**airplanes**", or ***other.***



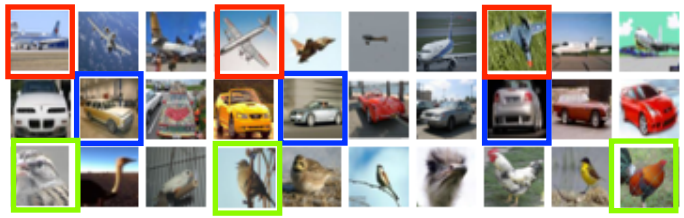
airplane
other
bird



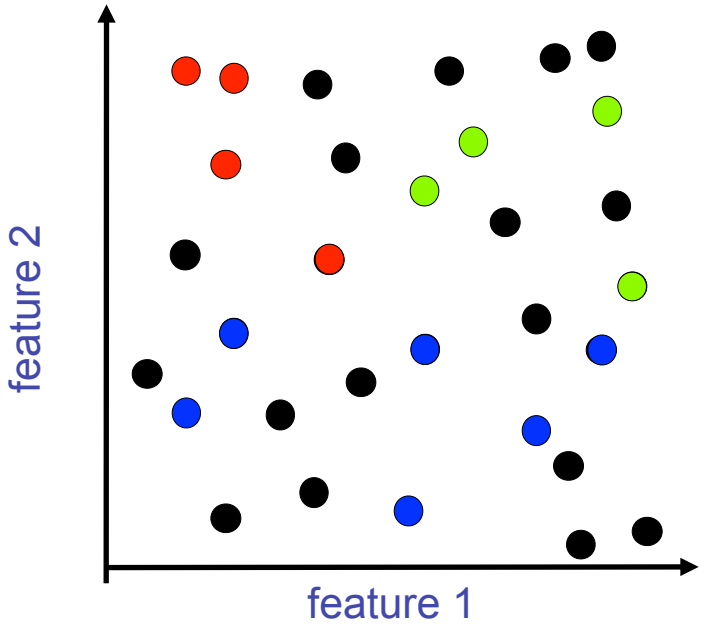

```
birds = []
planes = []
other = []
for image in images:
    if bird in image:
        birds.append(image)
    elseif plane in image:
        planes.append(image)
    else:
        other.append(tweet)
return birds, planes, other
```

Machine learning algorithms

Write a program that sorts image
into those containing "**birds**",
"**airplanes**", or ***other.***



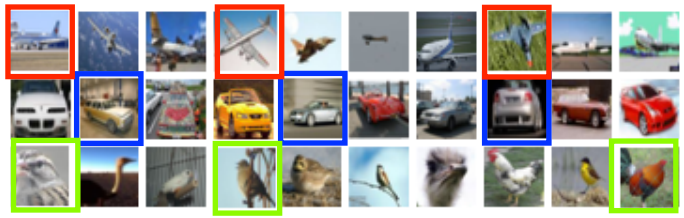
airplane
other
bird



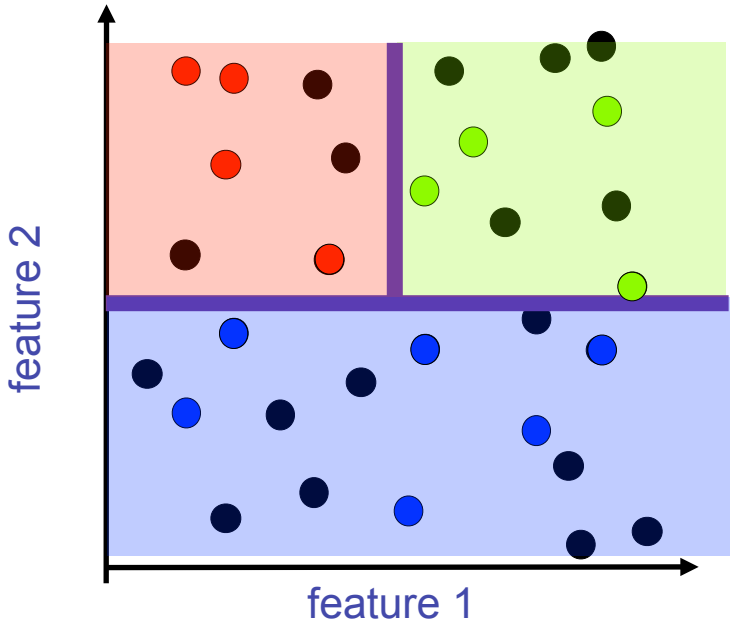

```
birds = []
planes = []
other = []
for image in images:
    if bird in image:
        birds.append(image)
    elseif plane in image:
        planes.append(image)
    else:
        other.append(tweet)
return birds, planes, other
```

Machine learning algorithms

Write a program that sorts image
into those containing "**birds**",
"**airplanes**", or ***other.***



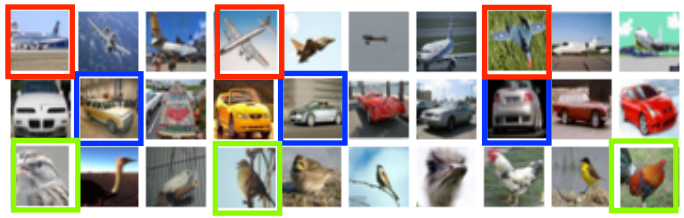
airplane
other
bird



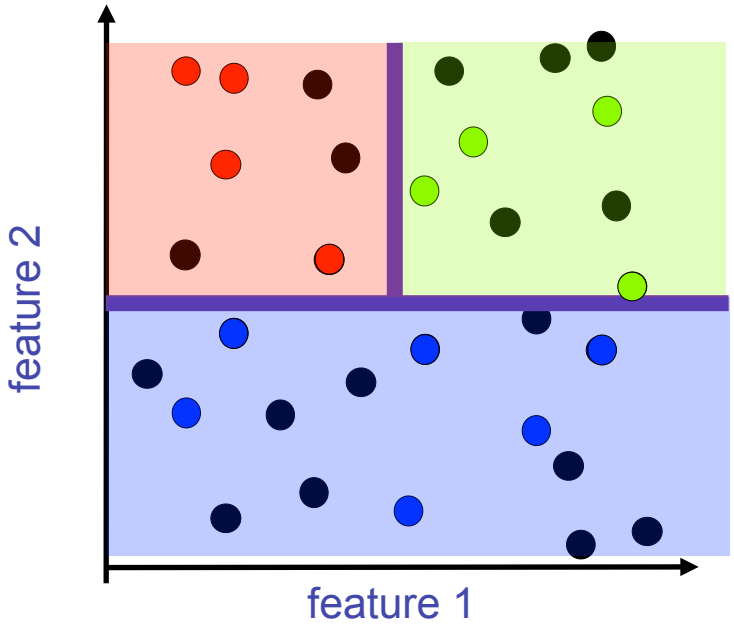

```
birds = []
planes = []
other = []
for image in images:
    if bird in image:
        birds.append(image)
    elseif plane in image:
        planes.append(image)
    else:
        other.append(tweet)
return birds, planes, other
```

Machine learning algorithms

Write a program that sorts image
into those containing "**birds**",
"**airplanes**", or ***other.***



airplane
other
bird




```
birds = []
planes = []
other = []
for image in images:
    if bird in image:
        birds.append(image)
    elseif plane in image:
        planes.append(image)
    else:
        other.append(tweet)
return birds, planes, other
```

The decision rule of
`if "cat" in tweet:`
is **hard coded by expert.**

The decision rule of
`if bird in image:`
is **LEARNED using DATA**

Machine Learning Ingredients

- **Data**: past observations
- **Hypotheses/Models**: devised to capture the patterns in data
- **Prediction**: apply model to forecast future observations

Machine Learning Ingredients

- **Data**: past observations
- **Hypotheses/Models**: devised to capture the patterns in data
- **Prediction**: apply model to forecast future observations

ML uses past data to make personalized predictions

Machine Learning Ingredients

- **Data**: past observations
- **Hypotheses/Models**: devised to capture the patterns in data
- **Prediction**: apply model to forecast future observations

ML uses past data to make personalized predictions



Machine Learning Ingredients

- **Data**: past observations
- **Hypotheses/Models**: devised to capture the patterns in data
- **Prediction**: apply model to forecast future observations

ML uses past data to make personalized predictions





You may also like…

Machine Learning Ingredients

- **Data**: past observations
- **Hypotheses/Models**: devised to capture the patterns in data
- **Prediction**: apply model to forecast future observations

ML uses past data to make personalized predictions





You may also like…



Machine Learning Ingredients

- **Data**: past observations
- **Hypotheses/Models**: devised to capture the patterns in data
- **Prediction**: apply model to forecast future observations

ML uses past data to make personalized predictions





You may also like…





Machine Learning Ingredients

- **Data**: past observations
- **Hypotheses/Models**: devised to capture the patterns in data
- **Prediction**: apply model to forecast future observations

ML uses past data to make personalized predictions





You may also like…







Machine Learning Ingredients

- **Data**: past observations
- **Hypotheses/Models**: devised to capture the patterns in data
- **Prediction**: apply model to forecast future observations

ML uses past data to make personalized predictions





You may also like…











Machine Learning Ingredients

- **Data**: past observations
- **Hypotheses/Models**: devised to capture the patterns in data
- **Prediction**: apply model to forecast future observations

ML uses past data to make personalized predictions





You may also like…









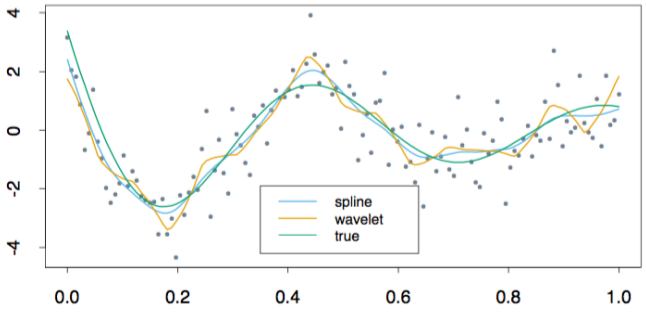


Machine learning is incredibly powerful and can have significant (unintended) negative consequences on society through targeting, excluding, and misusing.

Learning objectives of this course:

- introduction to the fundamental concepts of machine learning
- analysis and implementation of machine learning algorithms
- knowing how to use machine learning responsibly and robustly

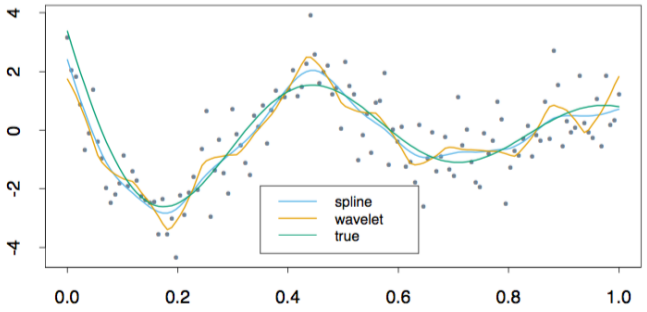
Flavors of ML



Regression

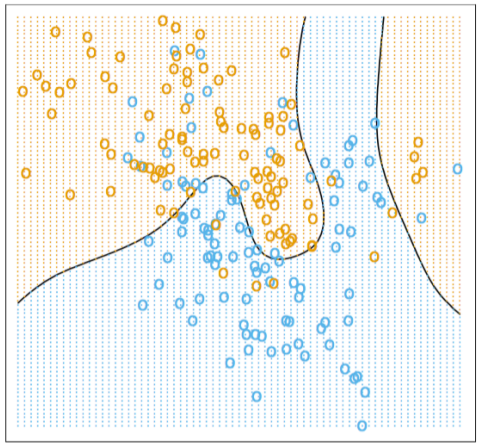
Predict continuous value:
ex: stock market, credit score, temperature, Netflix rating

Flavors of ML



Regression

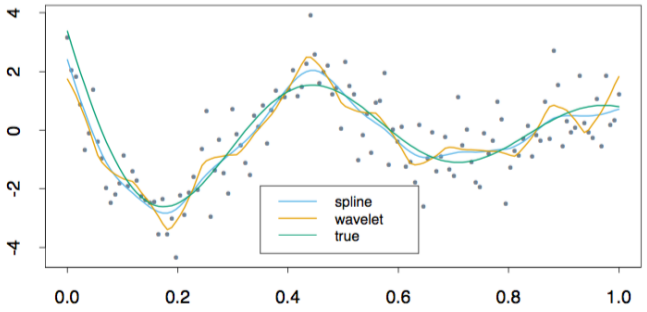
Predict continuous value:
ex: stock market, credit score, temperature, Netflix rating



Classification

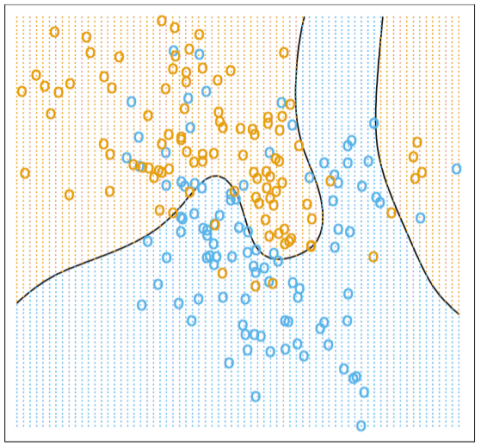
Predict categorical value:
loan or not? spam or not? what disease is this?

Flavors of ML



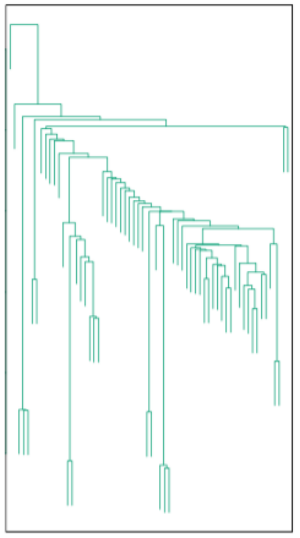
Regression

Predict continuous value:
ex: stock market, credit score, temperature, Netflix rating



Classification

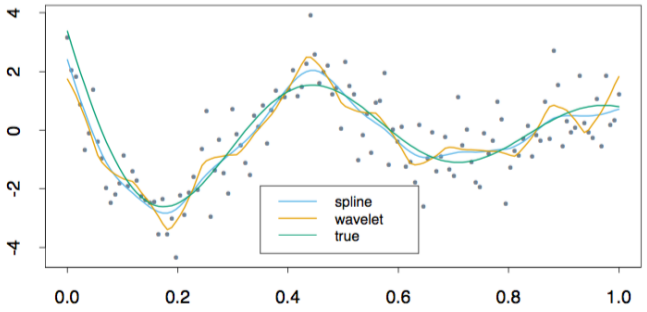
Predict categorical value:
loan or not? spam or not? what disease is this?



Unsupervised Learning

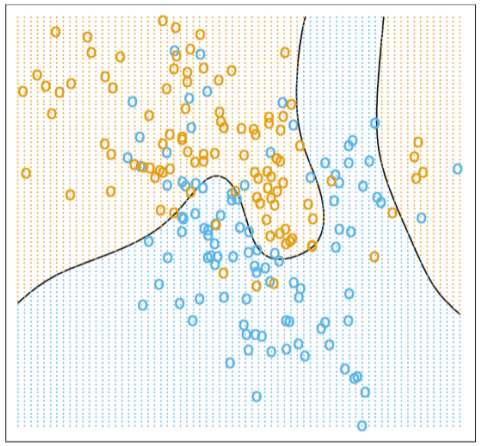
Predict structure:
tree of life from DNA, find similar images, community detection

Flavors of ML



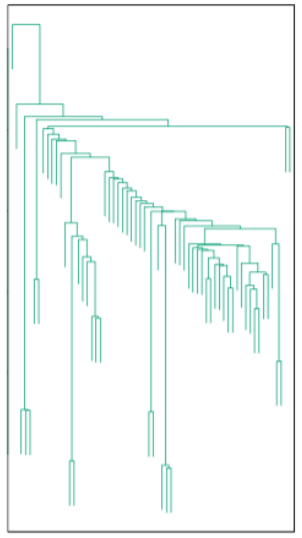
Regression

Predict continuous value:
ex: stock market, credit score, temperature, Netflix rating



Classification

Predict categorical value:
loan or not? spam or not? what disease is this?

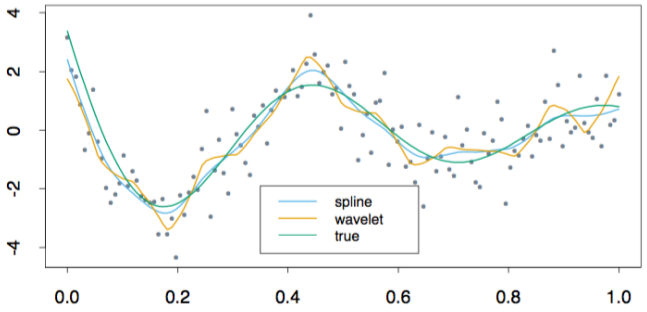


Unsupervised Learning

Predict structure:
tree of life from DNA, find similar images, community detection

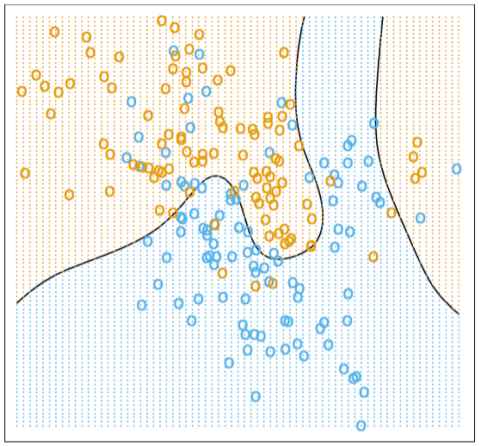
Self-supervised:
Model distribution of large-scale data, like text or images (large vision and language models)

Flavors of ML



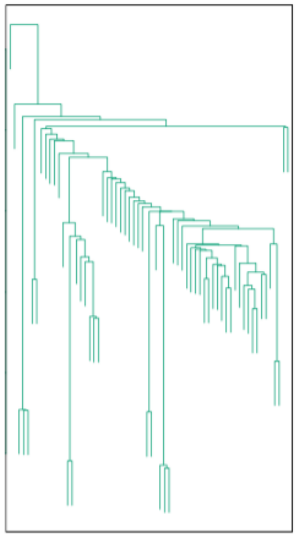
Regression

Predict continuous value:
ex: stock market, credit score, temperature, Netflix rating



Classification

Predict categorical value:
loan or not? spam or not? what disease is this?



Unsupervised Learning

Predict structure:
tree of life from DNA, find similar images, community detection

Self-supervised:
Model distribution of large-scale data, like text or images (large vision and language models)

Mix of statistics (conceptual) and algorithms (programming)

Probability review



Definitions

- **Random Variable**: A variable that takes on different values determined randomly.
  - Example: The height of a person from the US

- **Distribution**: The different values a random variable can take on along with the probability of that value.

- We talk about **sampling** from a distribution:
  - “Consider a sample of 100 different heights of people from the US drawn randomly from the distribution of all heights.”

Definitions

- **Random Variable**: A variable that takes on different values determined randomly.
  - Example: The height of a person from the US $P(\ )$
- **Distribution**: The different values a random variable can take on along with the probability of that value.
- We talk about **sampling** from a distribution:
  - "Consider a sample of 100 different heights of people from the US drawn randomly from the distribution of all heights."

Definitions

- **Random Variable**: A variable that takes on different values determined randomly.
  - Example: The height of a person from the US $P(\ )$
- **Distribution**: The different values a random variable can take on along with the probability of that value.

$$P(X=H) = 1/2$$
$$P(X=T) = 1/2$$

- We talk about **sampling** from a distribution:
  - "Consider a sample of 100 different heights of people from the US drawn randomly from the distribution of all heights."

Independence

Let X and Y be **random variables**

Ex. X is the outcome of the first roll of a 6-sided dice, Y is the outcome of the second roll of the dice

(X and Y take values in {1,2,3,4,5,6} each with equal probability)

An **event** is statement about the world that holds or not:

Define events $A = \{X \in \{3,4\}\}$,

$B = \{X = 1\}$,

$C = \{Y \in \{3,4\}\}$

Every event is assigned a **probability**:

$P(A) = P(X \in \{3,4\}) =$

Independence

Let X and Y be **random variables**

Ex. X is the outcome of the first roll of a 6-sided dice, Y is the outcome of the second roll of the dice

(X and Y take values in {1,2,3,4,5,6} each with equal probability)

An **event** is statement about the world that holds or not:

Define events $A = \{X \in \{3,4\}\}$,

$B = \{X = 1\}$,

$C = \{Y \in \{3,4\}\}$

Every event is assigned a **probability**:

$P(A) = P(X \in \{3,4\}) =$ *2/6 = 1/3*

Independence

Let X and Y be **random variables**

Ex. X is the outcome of the first roll of a 6-sided dice, Y is the outcome of the second roll of the dice

(X and Y take values in {1,2,3,4,5,6} each with equal probability)

An **event** is statement about the world that holds or not:

Define events $A = \{X \in \{3,4\}\}$,

$B = \{X = 1\}$,

$C = \{Y \in \{3,4\}\}$

Every event is assigned a **probability**:

$P(A) = P(X \in \{3,4\}) = 1/3$

For any events U, V we have $P(U \cup V) = P(U) + P(V) - P(U \cap V)$

Independence

Let X and Y be **random variables**

Ex. X is the outcome of the first roll of a 6-sided dice, Y is the outcome of the second roll of the dice

(X and Y take values in {1,2,3,4,5,6} each with equal probability)

An **event** is statement about the world that holds or not:

Define events $A = \{X \in \{3,4\}\}$,

$B = \{X = 1\}$,

$C = \{Y \in \{3,4\}\}$

Any events U, V are **independent** if $P(U \cap V) = P(U)P(V)$

Are A, B independent?

B, C?

A, C?

Independence

Let X and Y be **random variables**

Ex. X is the outcome of the first roll of a 6-sided dice, Y is the outcome of the second roll of the dice

(X and Y take values in {1,2,3,4,5,6} each with equal probability)

An **event** is statement about the world that holds or not:

Define events $A = \{X \in \{3,4\}\}$,

$B = \{X = 1\}$,

$C = \{Y \in \{3,4\}\}$

Any events U, V are **independent** if $P(U \cap V) = P(U)P(V)$

Are A, B independent? $P(A)=$

B, C?

A, C?

Independence

Let X and Y be **random variables**

Ex. X is the outcome of the first roll of a 6-sided dice, Y is the outcome of the second roll of the dice

(X and Y take values in {1,2,3,4,5,6} each with equal probability)

An **event** is statement about the world that holds or not:

Define events $A = \{X \in \{3,4\}\}$,

$B = \{X = 1\}$,

$C = \{Y \in \{3,4\}\}$

Any events U, V are **independent** if $P(U \cap V) = P(U)P(V)$

Are A, B independent? $P(A) = 1/3$

B, C?

A, C?

Independence

Let X and Y be **random variables**

Ex. X is the outcome of the first roll of a 6-sided dice, Y is the outcome of the second roll of the dice

(X and Y take values in {1,2,3,4,5,6} each with equal probability)

An **event** is statement about the world that holds or not:

Define events $A = \{X \in \{3,4\}\}$,

$B = \{X = 1\}$,

$C = \{Y \in \{3,4\}\}$

Any events U, V are **independent** if $P(U \cap V) = P(U)P(V)$

Are A, B independent? $P(A) = 1/3$ $P(B) =$

B, C?

A, C?

Independence

Let X and Y be **random variables**

Ex. X is the outcome of the first roll of a 6-sided dice, Y is the outcome of the second roll of the dice

(X and Y take values in {1,2,3,4,5,6} each with equal probability)

An **event** is statement about the world that holds or not:

Define events $A = \{X \in \{3,4\}\}$,

$B = \{X = 1\}$,

$C = \{Y \in \{3,4\}\}$

Any events U, V are **independent** if $P(U \cap V) = P(U)P(V)$

Are A, B independent? $P(A) = 1/3$ $P(B) = 1/6$

B, C?

A, C?

Independence

Let X and Y be **random variables**

Ex. X is the outcome of the first roll of a 6-sided dice, Y is the outcome of the second roll of the dice

(X and Y take values in {1,2,3,4,5,6} each with equal probability)

An **event** is statement about the world that holds or not:

Define events $A = \{X \in \{3,4\}\}$,

$B = \{X = 1\}$,

$C = \{Y \in \{3,4\}\}$

Any events U, V are **independent** if $P(U \cap V) = P(U)P(V)$

Are A, B independent? $P(A) = 1/3$ $P(B) = 1/6$

$P(A \cap B) =$

B, C?

A, C?

Independence

Let X and Y be **random variables**

Ex. X is the outcome of the first roll of a 6-sided dice, Y is the outcome of the second roll of the dice

(X and Y take values in {1,2,3,4,5,6} each with equal probability)

An **event** is statement about the world that holds or not:

Define events $A = \{X \in \{3,4\}\}$,

$B = \{X = 1\}$,

$C = \{Y \in \{3,4\}\}$

Any events U, V are **independent** if $P(U \cap V) = P(U)P(V)$

Are A, B independent? $P(A) = 1/3$ $P(B) = 1/6$

$P(A \cap B) = 0$

B, C?

A, C?

Independence

Let X and Y be **random variables**

Ex. X is the outcome of the first roll of a 6-sided dice, Y is the outcome of the second roll of the dice

(X and Y take values in {1,2,3,4,5,6} each with equal probability)

An **event** is statement about the world that holds or not:

Define events $A = \{X \in \{3,4\}\}$,

$B = \{X = 1\}$,

$C = \{Y \in \{3,4\}\}$

Any events U, V are **independent** if $P(U \cap V) = P(U)P(V)$

Are A, B independent? $P(A) = 1/3$ $P(B) = 1/6$ $P(A \cap B) = 0$ *No*

B, C?

A, C?

Independence

Let X and Y be **random variables**

Ex. X is the outcome of the first roll of a 6-sided dice, Y is the outcome of the second roll of the dice

(X and Y take values in {1,2,3,4,5,6} each with equal probability)

An **event** is statement about the world that holds or not:

Define events $A = \{X \in \{3,4\}\}$,

$B = \{X = 1\}$,

$C = \{Y \in \{3,4\}\}$

Any events U, V are **independent** if $P(U \cap V) = P(U)P(V)$

Are A, B independent? $P(A) = 1/3$ $P(B) = 1/6$ $P(A \cap B) = 0$ *No*

B, C? $P(C) =$

A, C?

Independence

Let X and Y be **random variables**

Ex. X is the outcome of the first roll of a 6-sided dice, Y is the outcome of the second roll of the dice

(X and Y take values in {1,2,3,4,5,6} each with equal probability)

An **event** is statement about the world that holds or not:

Define events $A = \{X \in \{3,4\}\}$,

$B = \{X = 1\}$,

$C = \{Y \in \{3,4\}\}$

Any events U, V are **independent** if $P(U \cap V) = P(U)P(V)$

Are A, B independent? $P(A) = 1/3$ $P(B) = 1/6$ $P(A \cap B) = 0$ *No*

B, C? $P(C) = 1/3$

A, C?

Independence

Let X and Y be **random variables**

Ex. X is the outcome of the first roll of a 6-sided dice, Y is the outcome of the second roll of the dice

(X and Y take values in {1,2,3,4,5,6} each with equal probability)

An **event** is statement about the world that holds or not:

Define events $A = \{X \in \{3,4\}\}$,

$B = \{X = 1\}$,

$C = \{Y \in \{3,4\}\}$

Any events U, V are **independent** if $P(U \cap V) = P(U)P(V)$

Are A, B independent? $P(A) = 1/3$ $P(B) = 1/6$ $P(A \cap B) = 0$ *No*

B, C? $P(C) = 1/3$ $P(B) =$

A, C?

Independence

Let X and Y be **random variables**

Ex. X is the outcome of the first roll of a 6-sided dice, Y is the outcome of the second roll of the dice

(X and Y take values in {1,2,3,4,5,6} each with equal probability)

An **event** is statement about the world that holds or not:

Define events $A = \{X \in \{3,4\}\}$,

$B = \{X = 1\}$,

$C = \{Y \in \{3,4\}\}$

Any events U, V are **independent** if $P(U \cap V) = P(U)P(V)$

Are A, B independent? $P(A) = 1/3$ $P(B) = 1/6$ $P(A \cap B) = 0$ *No*

B, C? $P(C) = 1/3$ $P(B) = 1/6$

A, C?

Independence

Let X and Y be **random variables**

Ex. X is the outcome of the first roll of a 6-sided dice, Y is the outcome of the second roll of the dice

(X and Y take values in {1,2,3,4,5,6} each with equal probability)

An **event** is statement about the world that holds or not:

Define events $A = \{X \in \{3,4\}\}$,

$B = \{X = 1\}$,

$C = \{Y \in \{3,4\}\}$

Any events U, V are **independent** if $P(U \cap V) = P(U)P(V)$

Are A, B independent? $P(A) = 1/3$ $P(B) = 1/6$

$P(A \cap B) = 0$

No

B, C? $P(C) = 1/3$ $P(B) = 1/6$

$P(A \cap B) =$

A, C?

Independence

Let X and Y be **random variables**

Ex. X is the outcome of the first roll of a 6-sided dice, Y is the outcome of the second roll of the dice

(X and Y take values in {1,2,3,4,5,6} each with equal probability)

An **event** is statement about the world that holds or not:

Define events $A = \{X \in \{3,4\}\}$,

$B = \{X = 1\}$,

$C = \{Y \in \{3,4\}\}$

Any events U, V are **independent** if $P(U \cap V) = P(U)P(V)$

Are A, B independent? $P(A) = 1/3 \quad P(B) = 1/6$

$P(A \cap B) = 0$

No

B, C? $P(C) = 1/3 \quad P(B) = 1/6$

$P(A \cap B) = 1/3 \times 1/6$

A, C?

Independence

Let X and Y be **random variables**

Ex. X is the outcome of the first roll of a 6-sided dice, Y is the outcome of the second roll of the dice

(X and Y take values in {1,2,3,4,5,6} each with equal probability)

An **event** is statement about the world that holds or not:

Define events $A = \{X \in \{3,4\}\}$,

$B = \{X = 1\}$,

$C = \{Y \in \{3,4\}\}$

Any events U, V are **independent** if $P(U \cap V) = P(U)P(V)$

Are A, B independent? $P(A) = 1/3 \quad P(B) = 1/6$

$P(A \cap B) = 0$

No

B, C? $P(C) = 1/3 \quad P(B) = 1/6$

$P(A \cap B) = 1/3 \times 1/6$

Yes

A, C?

Independence

Let X and Y be **random variables**

Ex. X is the outcome of the first roll of a 6-sided dice, Y is the outcome of the second roll of the dice

(X and Y take values in {1,2,3,4,5,6} each with equal probability)

An **event** is statement about the world that holds or not:

Define events $A = \{X \in \{3,4\}\}$,

$B = \{X = 1\}$,

$C = \{Y \in \{3,4\}\}$

Any events U, V are **independent** if $P(U \cap V) = P(U)P(V)$

Are A, B independent? $P(A) = 1/3 \quad P(B) = 1/6$

$P(A \cap B) = 0$

No

B, C? $P(C) = 1/3 \quad P(B) = 1/6$

$P(A \cap B) = 1/3 \times 1/6$

Yes

A, C?

Yes

Independence

Let X and Y be **random variables**

Ex. X is the outcome of the first roll of a 6-sided dice, Y is the outcome of the second roll of the dice

(X and Y take values in {1,2,3,4,5,6} each with equal probability)

An **event** is statement about the world that holds or not:

Define events $A = \{X \in \{3,4\}\}$,

$B = \{X = 1\}$,

$C = \{Y \in \{3,4\}\}$

Any events U, V are **independent** if $P(U \cap V) = P(U)P(V)$

We define the **conditional probability** of event U given V as

$$P(U \mid V) = \frac{P(U \cap V)}{P(V)}$$

What is $P(X \leq 4 \mid X \geq 3)$?

Independence

Let X and Y be **random variables**

Ex. X is the outcome of the first roll of a 6-sided dice, Y is the outcome of the second roll of the dice

(X and Y take values in {1,2,3,4,5,6} each with equal probability)

An **event** is statement about the world that holds or not:

Define events $A = \{X \in \{3,4\}\}$,

$B = \{X = 1\}$,

$C = \{Y \in \{3,4\}\}$

Any events U, V are **independent** if $P(U \cap V) = P(U)P(V)$

We define the **conditional probability** of event U given V as

$$P(U \mid V) = \frac{P(U \cap V)}{P(V)}$$

What is $P(X \leq 4 \mid X \geq 3)$

Independence

Let X and Y be **random variables**

Ex. X is the outcome of the first roll of a 6-sided dice, Y is the outcome of the second roll of the dice

(X and Y take values in {1,2,3,4,5,6} each with equal probability)

An **event** is statement about the world that holds or not:

Define events $A = \{X \in \{3,4\}\}$,

$B = \{X = 1\}$,

$C = \{Y \in \{3,4\}\}$

Any events U, V are **independent** if $P(U \cap V) = P(U)P(V)$

We define the **conditional probability** of event U given V as

$$P(U \mid V) = \frac{P(U \cap V)}{P(V)}$$

What is $P(X \leq 4 \mid X \geq 3) = \frac{P(3 \leq X \leq 4)}{P(X \geq 3)}$

Independence

Let X and Y be **random variables**

Ex. X is the outcome of the first roll of a 6-sided dice, Y is the outcome of the second roll of the dice

(X and Y take values in {1,2,3,4,5,6} each with equal probability)

An **event** is statement about the world that holds or not:

Define events $A = \{X \in \{3,4\}\}$,

$B = \{X = 1\}$,

$C = \{Y \in \{3,4\}\}$

Any events U, V are **independent** if $P(U \cap V) = P(U)P(V)$

We define the **conditional probability** of event U given V as

$$P(U \mid V) = \frac{P(U \cap V)}{P(V)}$$

What is $P(X \leq 4 \mid X \geq 3) = \frac{P(3 \leq X \leq 4)}{P(X \geq 3)} = \frac{1/3}{2/3}$

Independence

Let X and Y be **random variables**

Ex. X is the outcome of the first roll of a 6-sided dice, Y is the outcome of the second roll of the dice

(X and Y take values in {1,2,3,4,5,6} each with equal probability)

An **event** is statement about the world that holds or not:

Define events $A = \{X \in \{3,4\}\}$,

$B = \{X = 1\}$,

$C = \{Y \in \{3,4\}\}$

Any events U, V are **independent** if $P(U \cap V) = P(U)P(V)$

We define the **conditional probability** of event U given V as

$$P(U \mid V) = \frac{P(U \cap V)}{P(V)}$$

What is $P(X \leq 4 \mid X \geq 3) = \frac{P(3 \leq X \leq 4)}{P(X \geq 3)} = \frac{1/3}{2/3} = 1/2$

Independence

Let X and Y be **random variables**

Ex. X is the outcome of the first roll of a 6-sided dice, Y is the outcome of the second roll of the dice

(X and Y take values in {1,2,3,4,5,6} each with equal probability)

An **event** is statement about the world that holds or not:

Define events $A = \{X \in \{3,4\}\}$,
$B = \{X = 1\}$,
$C = \{Y \in \{3,4\}\}$

Any events U, V are **independent** if $P(U \cap V) = P(U)P(V)$

We define the **conditional probability** of event U given V as

$$P(U \mid V) = \frac{P(U \cap V)}{P(V)}$$

Observe: if U, V are independent then $P(U \mid V) = P(U)$.
In words: if independent, V tells you nothing about U (and vice versa)

Independence

Let X and Y be **random variables**

Ex. X is the outcome of the first roll of a 6-sided dice, Y is the outcome of the second roll of the dice

(X and Y take values in {1,2,3,4,5,6} each with equal probability)

An **event** is statement about the world that holds or not:

Define events $A = \{X \in \{3,4\}\}$,
$B = \{X = 1\}$,
$C = \{Y \in \{3,4\}\}$

Any events U, V are **independent** if $P(U \cap V) = P(U)P(V)$

We define the **conditional probability** of event U given V as

$$P(U \mid V) = \frac{P(U \cap V)}{P(V)}$$

If independent then:

Observe: if U, V are independent then $P(U \mid V) = P(U)$.

In words: if independent, V tells you nothing about U (and vice versa)

Independence

Let X and Y be **random variables**

Ex. X is the outcome of the first roll of a 6-sided dice, Y is the outcome of the second roll of the dice

(X and Y take values in {1,2,3,4,5,6} each with equal probability)

An **event** is statement about the world that holds or not:

Define events $A = \{X \in \{3,4\}\}$,

$B = \{X = 1\}$,

$C = \{Y \in \{3,4\}\}$

Any events U, V are **independent** if $P(U \cap V) = P(U)P(V)$

We define the **conditional probability** of event U given V as

$$P(U \mid V) = \frac{P(U \cap V)}{P(V)}$$

If independent then: $= \frac{P(\mathrm{U})P(\mathrm{V})}{P(\mathrm{V})}$

Observe: if U, V are independent then $P(U \mid V) = P(U)$.

In words: if independent, V tells you nothing about U (and vice versa)

Independence

Let X and Y be **random variables**

Ex. X is the outcome of the first roll of a 6-sided dice, Y is the outcome of the second roll of the dice

(X and Y take values in {1,2,3,4,5,6} each with equal probability)

An **event** is statement about the world that holds or not:

Define events $A = \{X \in \{3,4\}\}$,

$B = \{X = 1\}$,

$C = \{Y \in \{3,4\}\}$

Any events U, V are **independent** if $P(U \cap V) = P(U)P(V)$

We define the **conditional probability** of event U given V as

$$P(U \mid V) = \frac{P(U \cap V)}{P(V)}$$

If independent then: $= \frac{P(U)P(V)}{P(V)}$

Observe: if U, V are independent then $P(U \mid V) = P(U)$.

In words: if independent, V tells you nothing about U (and vice versa)

Independence

Let X and Y be **random variables**

Ex. X is the outcome of the first roll of a 6-sided dice, Y is the outcome of the second roll of the dice

(X and Y take values in {1,2,3,4,5,6} each with equal probability)

An **event** is statement about the world that holds or not:

Define events $A = \{X \in \{3,4\}\}$,

$B = \{X = 1\}$,

$C = \{Y \in \{3,4\}\}$

Any events U, V are **independent** if $P(U \cap V) = P(U)P(V)$

We define the **conditional probability** of event U given V as

$$P(U \mid V) = \frac{P(U \cap V)}{P(V)}$$

If independent then: $= \frac{P(U)P(V)}{P(V)}$

Observe: if U, V are independent then $P(U \mid V) = P(U)$.

In words: if independent, V tells you nothing about U (and vice versa)

Independence

Let X and Y be **random variables**

Ex. X is the outcome of the first roll of a 6-sided dice, Y is the outcome of the second roll of the dice

(X and Y take values in {1,2,3,4,5,6} each with equal probability)

An **event** is statement about the world that holds or not:

Define events $A = \{X \in \{3,4\}\}$,

$B = \{X = 1\}$,

$C = \{Y \in \{3,4\}\}$

Any events U, V are **independent** if $P(U \cap V) = P(U)P(V)$

We define the **conditional probability** of event U given V as

$$P(U \mid V) = \frac{P(U \cap V)}{P(V)}$$

If independent then: $= \frac{P(U)P(V)}{P(V)} = P(U)$

Observe: if U, V are independent then $P(U \mid V) = P(U)$.

In words: if independent, V tells you nothing about U (and vice versa)

Mean, variance

Mean $\mathbb{E}[X], \mu$

The expected value of X, each value is weighted by the probability of seeing it.

$$\mathbb{E}[X] = \sum_x P(X = x)x$$

Variance $\mathrm{Var}(X), \sigma^2$

The expected squared deviation of X from its mean.

$$\mathbb{E}[(X - \mathbb{E}[X])^2]$$

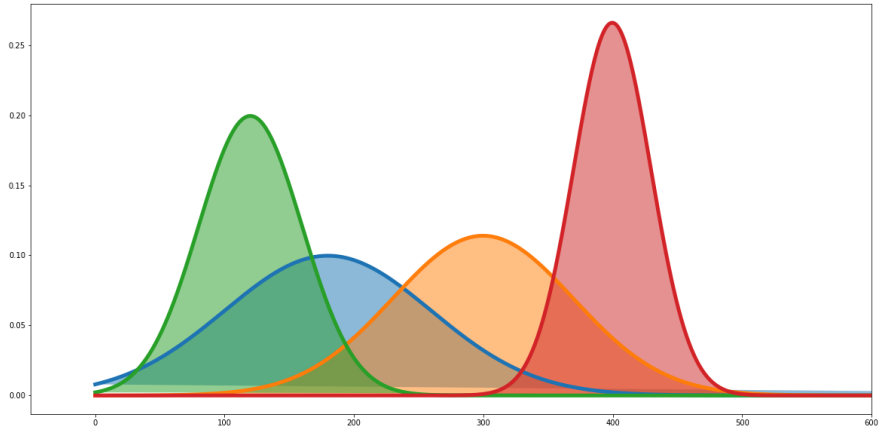
Median M

The value of X that is separating the higher half of it's range from the lower half.

$$P(X \leq M) = .5$$

Mean, variance

The mean is a prediction of the value of the random variable. Answers the question "What do I expect the height of a random person to be?"

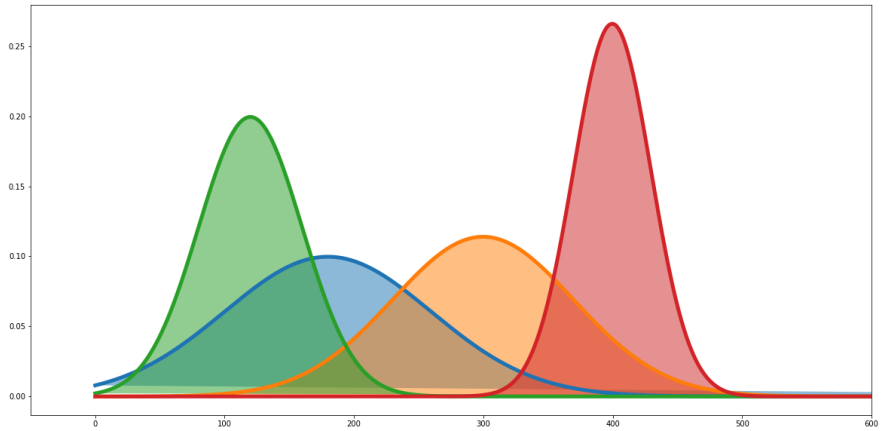


The variance captures the spread in your data. Also captures the error in the prediction using the mean. "How much do people's heights deviate?"

$$\mathbb{E}[(X - \mathbb{E}[X])^2]$$

Mean, variance

The mean is a prediction of the value of the random variable. Answers the question "What do I expect the height of a random person to be?"



Which distribution has a:

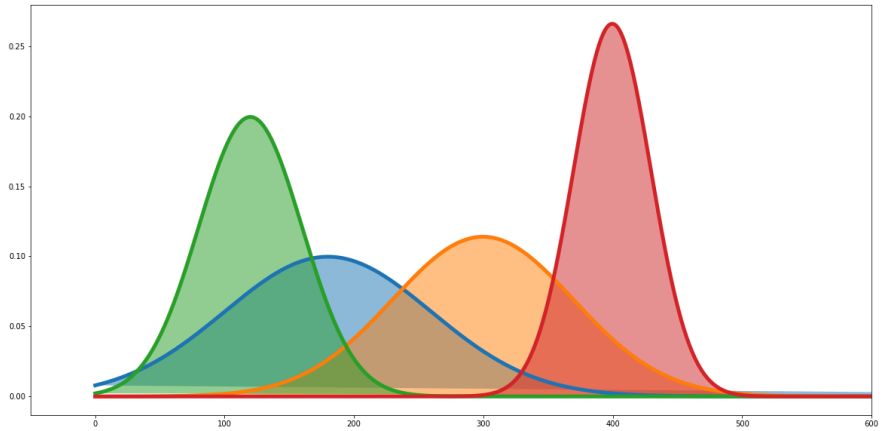
- *Large mean, small variance?*
- *Small mean, large variance?*

The variance captures the spread in your data. Also captures the error in the prediction using the mean. "How much do people's heights deviate?"

$$\mathbb{E}[(X - \mathbb{E}[X])^2]$$

Mean, variance

The mean is a prediction of the value of the random variable. Answers the question "What do I expect the height of a random person to be?"



Which distribution has a:

- *Large mean, small variance?*

  Red

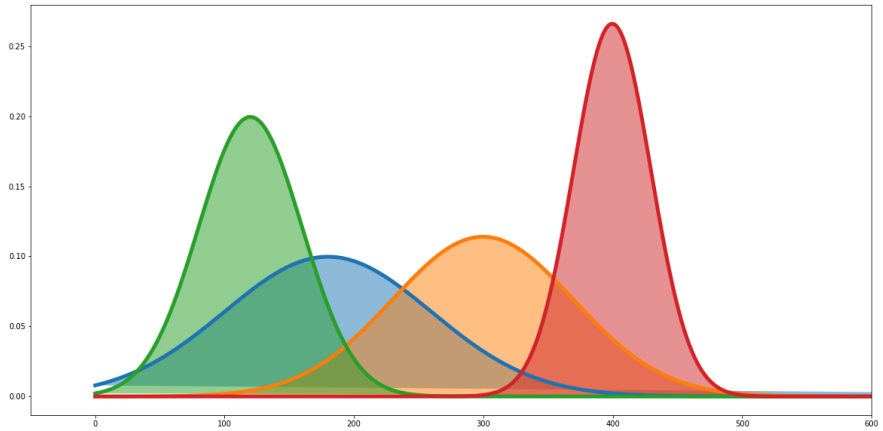
- *Small mean, large variance?*

The variance captures the spread in your data. Also captures the error in the prediction using the mean. "How much do people's heights deviate?"

$$\mathbb{E}[(X - \mathbb{E}[X])^2]$$

Mean, variance

The mean is a prediction of the value of the random variable. Answers the question "What do I expect the height of a random person to be?"



Which distribution has a:

- *Large mean, small variance?*

  Red

- *Small mean, large variance?*

  Blue

The variance captures the spread in your data. Also captures the error in the prediction using the mean. "How much do people's heights deviate?"

$$\mathbb{E}[(X - \mathbb{E}[X])^2]$$

Maximum Likelihood Estimation



Your first consulting job

- *Client*: I have special coin, if I flip it, what's the probability it will be heads?
- *You (a machine learner)*: I need to collect **data.**

- *You (a frequentist)*: The probability is:

Your first consulting job

- *Client*: I have special coin, if I flip it, what's the probability it will be heads?
- *You (a machine learner)*: I need to collect **data.**

H H

- *You (a frequentist)*: The probability is:

Your first consulting job

- *Client*: I have special coin, if I flip it, what's the probability it will be heads?
- *You (a machine learner)*: I need to collect **data.**

H H

- *You (a frequentist)*: The probability is:

100% heads!

Your first consulting job

- *Client*: Uhhhh…. You sure about that? I just got a tails.
- *You (a machine learner)*: I need to collect **more data.**
  - *flips coin 5 times, get HHTHT

- *You*: The probability is: 60% Heads, 40% Tails!

Your first consulting job

- *Client*: Uhhhh…. You sure about that? I just got a tails.
- *You (a machine learner)*: I need to collect **more data.**
  - *flips coin 10000 times, it comes up Heads 60% of the time

- *You*: The probability is: 60% Heads, 40% Tails!

- *Client*: **Why should I believe you?**

- *You (a machine learner)*: Let's do some math!

Coin – Bernoulli Distribution

- **Data**: sequence *D= (HHTHT…),* **k heads** out of **n flips**
- **Hypothesis:** P(Heads) = θ, P(Tails) = 1-θ
  - Flips are i.i.d.:
    - Independent events
    - Identically distributed according to Bernoulli distribution

- $P(\mathcal{D}|\theta) =$

Coin – Bernoulli Distribution

- **Data**: sequence *D= (HHTHT…),* **k heads** out of **n flips**
- **Hypothesis:** P(Heads) = θ, P(Tails) = 1-θ
  - Flips are i.i.d.:
    - Independent events
    - Identically distributed according to Bernoulli distribution

- $P(\mathcal{D}|\theta) =$

= P(HHTHT | θ)

Coin – Bernoulli Distribution

- **Data**: sequence *D= (HHTHT…),* **k heads** out of **n flips**
- **Hypothesis:** P(Heads) = θ, P(Tails) = 1-θ
  - Flips are i.i.d.:
    - Independent events
    - Identically distributed according to Bernoulli distribution

- $P(\mathcal{D}|\theta) =$

$= P(HHTHT \mid \theta)$

$= P(H)P(H)P(T)P(H)P(T)$ # by independence

Coin – Bernoulli Distribution

- **Data**: sequence *D= (HHTHT…),* **k heads** out of **n flips**
- **Hypothesis:** P(Heads) = θ, P(Tails) = 1-θ
  - Flips are i.i.d.:
    - Independent events
    - Identically distributed according to Bernoulli distribution

- $P(\mathcal{D}|\theta) =$

$= P(HHTHT \mid \theta)$

$= P(H)P(H)P(T)P(H)P(T)$ # by independence

$= \theta^k(1-\theta)^{n-k}$

Maximum Likelihood Estimation

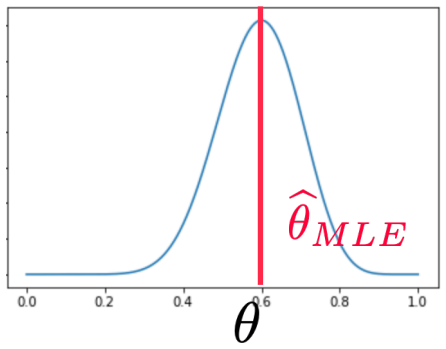
- **Data**: sequence *D= (HHTHT…),* **k heads** out of **n flips**
- **Hypothesis:** P(Heads) = θ, P(Tails) = 1-θ
- **Likelihood:**

$$P(\mathcal{D}|\theta) = \theta^k(1-\theta)^{n-k}$$

- **Maximum likelihood estimation (MLE):** Choose θ that maximizes the probability of observed data:

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(\mathcal{D}|\theta)$$

$$= \arg\max_{\theta} \log P(\mathcal{D}|\theta)$$



Maximum Likelihood Estimation

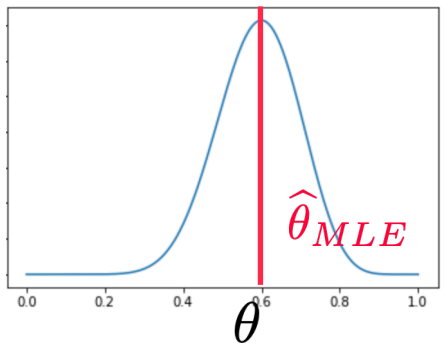
- **Data**: sequence *D= (HHTHT…),* **k heads** out of **n flips**
- **Hypothesis:** P(Heads) = θ, P(Tails) = 1-θ
- **Likelihood:**

$$P(\mathcal{D}|\theta) = \theta^k(1-\theta)^{n-k}$$ # likelihood

- **Maximum likelihood estimation (MLE):** Choose θ that maximizes the probability of observed data:

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(\mathcal{D}|\theta)$$

$$= \arg\max_{\theta} \log P(\mathcal{D}|\theta)$$



Maximum Likelihood Estimation

- **Data**: sequence *D= (HHTHT…),* **k heads** out of **n flips**
- **Hypothesis:** P(Heads) = θ, P(Tails) = 1-θ
- **Likelihood:**

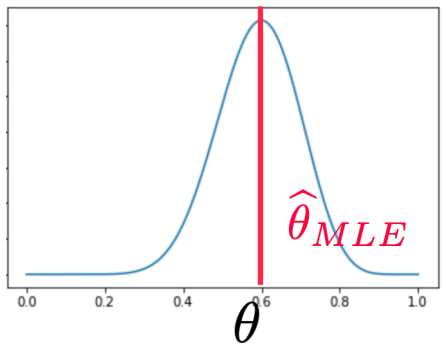
$$P(\mathcal{D}|\theta) = \theta^k(1-\theta)^{n-k}$$ # likelihood

- **Maximum likelihood estimation (MLE):** Choose θ that maximizes the probability of observed data:

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(\mathcal{D}|\theta)$$

Why take the log?

$$= \arg\max_{\theta} \log P(\mathcal{D}|\theta)$$



Maximum Likelihood Estimation

- **Data**: sequence *D= (HHTHT…),* **k heads** out of **n flips**
- **Hypothesis:** P(Heads) = θ, P(Tails) = 1-θ
- **Likelihood:**

$$P(\mathcal{D}|\theta) = \theta^k(1-\theta)^{n-k}$$ # likelihood

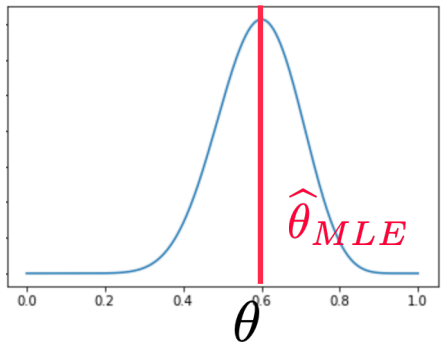
- **Maximum likelihood estimation (MLE):** Choose θ that maximizes the probability of observed data:

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(\mathcal{D}|\theta)$$

$$= \arg\max_{\theta} \log P(\mathcal{D}|\theta)$$

Why take the log?

- Easier to work with



Maximum Likelihood Estimation

- **Data**: sequence *D= (HHTHT…)*, **k heads** out of **n flips**
- **Hypothesis:** P(Heads) = θ, P(Tails) = 1-θ
- **Likelihood:**

$$P(\mathcal{D}|\theta) = \theta^k(1-\theta)^{n-k}$$ # likelihood

- **Maximum likelihood estimation (MLE):** Choose θ that maximizes the probability of observed data:

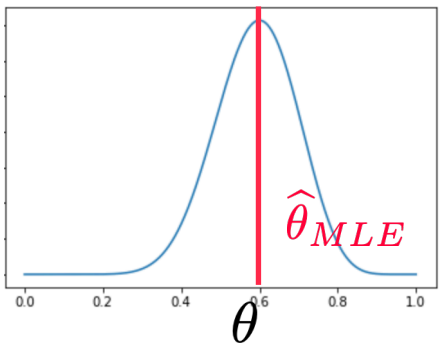
$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(\mathcal{D}|\theta)$$

$$= \arg\max_{\theta} \log P(\mathcal{D}|\theta)$$

$$= \arg\max_{\theta} \log\left[\theta^k(1-\theta)^{n-k}\right]$$

Why take the log?

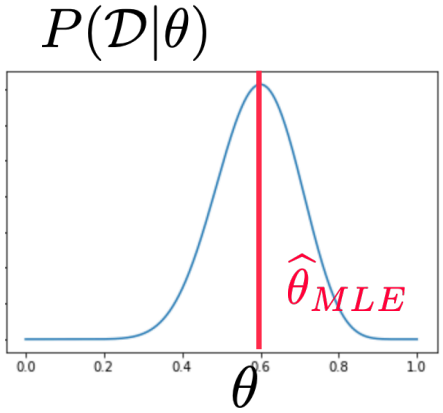
- Easier to work with



MLE: Your first learning algorithm

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \log P(\mathcal{D}|\theta)$$

$$= \arg\max_{\theta} \log \theta^{k}(1-\theta)^{n-k}$$



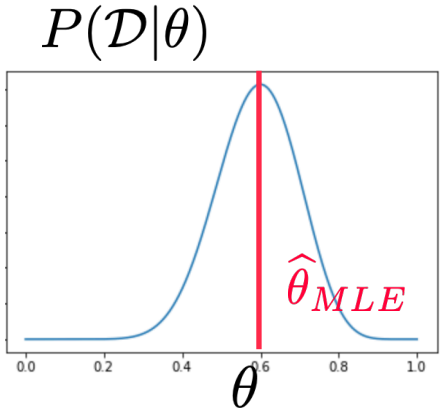
- How do we find θ that maximizes likelihood?
- Use the fact that derivative is zero at maxima (also at minima)
- Set derivative to zero, and find θ satisfying:

$$\frac{d}{d\theta} \log P(\mathcal{D}|\theta) = 0$$

MLE: Your first learning algorithm

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \log P(\mathcal{D}|\theta)$$

$$= \arg\max_{\theta} \log \theta^{k}(1-\theta)^{n-k}$$



- How do we find θ that maximizes likelihood?
- Use the fact that derivative is zero at maxima (also at minima)
- Set derivative to zero, and find θ satisfying:

$$\frac{d}{d\theta} \log P(\mathcal{D}|\theta) = 0$$

MLE

MLE: Your first learning algorithm

- **First manipulate the log likelihood to make it easy to work with:**

$$\begin{aligned}\widehat{\theta}_{MLE} &= \arg\max_{\theta} \log P(\mathcal{D}|\theta) \\ &= \arg\max_{\theta} \log \theta^k (1-\theta)^{n-k}\end{aligned}$$

- **Then set derivative to 0, and find θ satisfying:** $\frac{d}{d\theta} \log P(\mathcal{D}|\theta) = 0$

MLE: Your first learning algorithm

- **First manipulate the log likelihood to make it easy to work with:**

$$\begin{aligned}\widehat{\theta}_{MLE} &= \arg\max_{\theta} \log P(\mathcal{D}|\theta) \\ &= \arg\max_{\theta} \log \theta^k (1-\theta)^{n-k} \\ &= \arg\max_{\theta} k\log\theta + (n-k)\log(1-\theta)\end{aligned}$$

- **Then set derivative to 0, and find θ satisfying:** $\frac{d}{d\theta} \log P(\mathcal{D}|\theta) = 0$

MLE: Your first learning algorithm

- **First manipulate the log likelihood to make it easy to work with:**

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \log P(\mathcal{D}|\theta)$$

$$= \arg\max_{\theta} \log \theta^k (1-\theta)^{n-k}$$

$$= \arg\max_{\theta} k\log\theta + (n-k)\log(1-\theta)$$

- **Then set derivative to 0, and find θ satisfying:** $\frac{d}{d\theta} \log P(\mathcal{D}|\theta) = 0$

$$\frac{d}{d\theta}\log P(\mathscr{D}\,|\,\theta) = 0$$

MLE: Your first learning algorithm

- **First manipulate the log likelihood to make it easy to work with:**

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \log P(\mathcal{D}|\theta)$$

$$= \arg\max_{\theta} \log \theta^k (1-\theta)^{n-k}$$

$$= \arg\max_{\theta} k\log\theta + (n-k)\log(1-\theta)$$

- **Then set derivative to 0, and find θ satisfying:** $\frac{d}{d\theta} \log P(\mathcal{D}|\theta) = 0$

$$\frac{d}{d\theta}\log P(\mathcal{D}\,|\,\theta) = 0$$

for your formula sheet

$$\frac{d}{dx}\log x = \frac{1}{x}$$

MLE: Your first learning algorithm

- **First manipulate the log likelihood to make it easy to work with:**

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \log P(\mathcal{D}|\theta)$$

$$= \arg\max_{\theta} \log \theta^{k}(1-\theta)^{n-k}$$

$$= \arg\max_{\theta} k\log\theta + (n-k)\log(1-\theta)$$

- **Then set derivative to 0, and find θ satisfying:** $\frac{d}{d\theta}\log P(\mathcal{D}|\theta) = 0$

$$\frac{d}{d\theta}\log P(\mathscr{D}\,|\,\theta) = 0 \longrightarrow \frac{k}{\theta} - \frac{(n-k)}{(1-\theta)} = 0$$

for your formula sheet

$$\frac{d}{dx}\log x = \frac{1}{x}$$

MLE: Your first learning algorithm

- **First manipulate the log likelihood to make it easy to work with:**

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \log P(\mathcal{D}|\theta)$$

$$= \arg\max_{\theta} \log \theta^{k}(1-\theta)^{n-k}$$

$$= \arg\max_{\theta} k\log\theta + (n-k)\log(1-\theta)$$

- **Then set derivative to 0, and find θ satisfying:** $\frac{d}{d\theta}\log P(\mathcal{D}|\theta) = 0$

$$\frac{d}{d\theta}\log P(\mathcal{D}|\theta) = 0 \longrightarrow \frac{k}{\theta} - \frac{(n-k)}{(1-\theta)} = 0$$

$$k - k\theta = n\theta - k\theta$$

for your formula sheet

$$\frac{d}{dx}\log x = \frac{1}{x}$$

MLE: Your first learning algorithm

- **First manipulate the log likelihood to make it easy to work with:**

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \log P(\mathcal{D}|\theta)$$

$$= \arg\max_{\theta} \log \theta^{k}(1-\theta)^{n-k}$$

$$= \arg\max_{\theta} k\log\theta + (n-k)\log(1-\theta)$$

- **Then set derivative to 0, and find θ satisfying:** $\frac{d}{d\theta}\log P(\mathcal{D}|\theta) = 0$

$$\frac{d}{d\theta}\log P(\mathcal{D}|\theta) = 0 \longrightarrow \frac{k}{\theta} - \frac{(n-k)}{(1-\theta)} = 0$$

$$k - k\theta = n\theta - k\theta$$

$$\hat{\theta}_{\text{MLE}} = \frac{k}{n}$$

for your formula sheet

$$\frac{d}{dx}\log x = \frac{1}{x}$$

MLE: Your first learning algorithm

- **First manipulate the log likelihood to make it easy to work with:**

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \log P(\mathcal{D}|\theta)$$

$$= \arg\max_{\theta} \log \theta^k (1-\theta)^{n-k}$$

$$= \arg\max_{\theta} k\log\theta + (n-k)\log(1-\theta)$$

- **Then set derivative to 0, and find θ satisfying:** $\frac{d}{d\theta} \log P(\mathcal{D}|\theta) = 0$

$$\frac{d}{d\theta}\log P(\mathscr{D}\,|\,\theta) = 0 \longrightarrow \frac{k}{\theta} - \frac{(n-k)}{(1-\theta)} = 0$$

for your formula sheet

$$\frac{d}{dx}\log x = \frac{1}{x}$$

$$k - k\theta = n\theta - k\theta$$

our example:

$$\hat{\theta}_{\text{MLE}} = \frac{k}{n}$$

MLE: Your first learning algorithm

- **First manipulate the log likelihood to make it easy to work with:**

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \log P(\mathcal{D}|\theta)$$

$$= \arg\max_{\theta} \log \theta^{k}(1-\theta)^{n-k}$$

$$= \arg\max_{\theta} k\log\theta + (n-k)\log(1-\theta)$$

- **Then set derivative to 0, and find θ satisfying:** $\frac{d}{d\theta}\log P(\mathcal{D}|\theta) = 0$

$$\frac{d}{d\theta}\log P(\mathcal{D}\,|\,\theta) = 0 \longrightarrow \frac{k}{\theta} - \frac{(n-k)}{(1-\theta)} = 0$$

for your formula sheet

$$\frac{d}{dx}\log x = \frac{1}{x}$$

$$k - k\theta = n\theta - k\theta$$

$$\hat{\theta}_{\text{MLE}} = \frac{k}{n}$$

our example:

$$\hat{\theta}_{\text{MLE}} = \frac{3}{5} = 60\%$$

Your first consulting job

- *Client*: Uhhhh…. You sure about that? I just got a tails.
- *You (a machine learner)*: I need to collect **more data.**
  - *flips coin 10000 times, it comes up Heads 60% of the time

- *You*: The probability is: 60% Heads, 40% Tails!

- *Client*: **Why should I believe you?**

- *You (a machine learner)*: Let's do some math!

Your first consulting job

- *Client*: Uhhhh…. You sure about that? I just got a tails.
- *You (a machine learner)*: I need to collect **more data.**
  - *flips coin 10000 times, it comes up Heads 60% of the time

- *You*: The probability is: 60% Heads, 40% Tails!

- *Client*: **Why should I believe you?**

- *You (a machine learner)*: Let's do some math!

$$\hat{\theta}_{\text{MLE}} = \frac{3}{5} = 60\%$$

How good is MLE? Well, it's unbiased

- We treat MLE $\widehat{\theta}_{\text{MLE}}$ as a random variable, where there is a ground truth parameter θ^* that generates the data $\mathscr{D} = (HHTTH\ldots)$ of a fixed size n

- What can we say about this random variable $\widehat{\theta}_{\text{MLE}}$?
- First good property of MLE for Binomial: **unbiased**
  - Definition: **bias** of our MLE is

$$\text{Bias}(\widehat{\theta}_{\text{MLE}}) := \mathbb{E}_{\mathscr{D}\sim P_{\theta^*}}\left[\hat{\theta}_{\text{MLE}}\right] - \theta^* =$$

- **Expectation** describes how the estimator behaves *on average*

How good is MLE? Well, it's unbiased

- We treat MLE $\widehat{\theta}_{\text{MLE}}$ as a random variable, where there is a ground truth parameter θ^* that generates the data $\mathscr{D} = (HHTTH\ldots)$ of a fixed size n

$$\hat{\theta}_{\text{MLE}} = \frac{k}{n}$$

- What can we say about this random variable $\widehat{\theta}_{\text{MLE}}$?
- First good property of MLE for Binomial: **unbiased**
  - Definition: **bias** of our MLE is

$$\text{Bias}(\widehat{\theta}_{\text{MLE}}) := \mathbb{E}_{\mathscr{D}\sim P_{\theta^*}}\left[\hat{\theta}_{\text{MLE}}\right] - \theta^* =$$

- **Expectation** describes how the estimator behaves *on average*

How good is MLE? Well, it's unbiased

- We treat MLE $\widehat{\theta}_{\text{MLE}}$ as a random variable, where there is a ground truth parameter θ^* that generates the data $\mathscr{D} = (HHTTH\ldots)$ of a fixed size n

$$\hat{\theta}_{\text{MLE}} = \frac{k}{n}$$ # random variable

- What can we say about this random variable $\widehat{\theta}_{\text{MLE}}$?
- First good property of MLE for Binomial: **unbiased**
  - Definition: **bias** of our MLE is

$$\text{Bias}(\widehat{\theta}_{\text{MLE}}) := \mathbb{E}_{\mathscr{D}\sim P_{\theta^*}}\left[\hat{\theta}_{\text{MLE}}\right] - \theta^* =$$

- **Expectation** describes how the estimator behaves *on average*

How good is MLE? Well, it's unbiased

- We treat MLE $\widehat{\theta}_{\text{MLE}}$ as a random variable, where there is a ground truth parameter θ^* that generates the data $\mathscr{D} = (HHTTH\ldots)$ of a fixed size n

$$\hat{\theta}_{\text{MLE}} = \frac{k}{n}$$ # random variable

- What can we say about this random variable $\widehat{\theta}_{\text{MLE}}$?
- First good property of MLE for Binomial: **unbiased**
  - Definition: **bias** of our MLE is # "true predictor"

$$\text{Bias}(\widehat{\theta}_{\text{MLE}}) := \mathbb{E}_{\mathscr{D}\sim P_{\theta^*}}\left[\hat{\theta}_{\text{MLE}}\right] - \theta^* =$$

- **Expectation** describes how the estimator behaves *on average*

How good is MLE? Well, it's unbiased

- We treat MLE $\hat{\theta}_{\text{MLE}}$ as a random variable, where there is a ground truth parameter θ^* that generates the data $\mathcal{D} = (HHTTH\ldots)$ of a fixed size n

$$\hat{\theta}_{\text{MLE}} = \frac{k}{n}$$ # random variable

- What can we say about this random variable $\hat{\theta}_{\text{MLE}}$?
- First good property of MLE for Binomial: **unbiased**
  - Definition: **bias** of our MLE is # "true predictor"

$$\text{Bias}(\hat{\theta}_{\text{MLE}}) := \mathbb{E}_{\mathcal{D}\sim P_{\theta^*}}\left[\hat{\theta}_{\text{MLE}}\right] - \theta^* =$$

Unbiased means bias = 0

- **Expectation** describes how the estimator behaves *on average*

How good is MLE? Well, it's unbiased

- We treat MLE $\widehat{\theta}_{\text{MLE}}$ as a random variable, where there is a ground truth parameter θ^* that generates the data $\mathscr{D} = (HHTTH\ldots)$ of a fixed size n

$$\hat{\theta}_{\text{MLE}} = \frac{k}{n}$$ # random variable

- What can we say about this random variable $\widehat{\theta}_{\text{MLE}}$?
- First good property of MLE for Binomial: **unbiased**
  - Definition: **bias** of our MLE is

"true predictor"

$$\text{Bias}(\widehat{\theta}_{\text{MLE}}) := \mathbb{E}_{\mathscr{D}\sim P_{\theta^*}}\left[\hat{\theta}_{\text{MLE}}\right] - \theta^* = E\left[\frac{k}{n}\right] - \theta^*$$

Unbiased means bias = 0

- **Expectation** describes how the estimator behaves *on average*

How good is MLE? Well, it's unbiased

- We treat MLE $\widehat{\theta}_{\text{MLE}}$ as a random variable, where there is a ground truth parameter θ^* that generates the data $\mathscr{D} = (HHTTH\ldots)$ of a fixed size n

$$\hat{\theta}_{\text{MLE}} = \frac{k}{n}$$ # random variable

- What can we say about this random variable $\widehat{\theta}_{\text{MLE}}$?
- First good property of MLE for Binomial: **unbiased**
  - Definition: **bias** of our MLE is

"true predictor"

$$\text{Bias}(\widehat{\theta}_{\text{MLE}}) := \mathbb{E}_{\mathscr{D}\sim P_{\theta^*}}\left[\hat{\theta}_{\text{MLE}}\right] - \theta^* = E\left[\frac{k}{n}\right] - \theta^*$$

$$= \frac{\theta^* n}{n} - \theta^*$$

Unbiased means bias = 0

- **Expectation** describes how the estimator behaves *on average*

How good is MLE? Well, it's unbiased

- We treat MLE $\hat{\theta}_{\text{MLE}}$ as a random variable, where there is a ground truth parameter θ^* that generates the data $\mathcal{D} = (HHTTH\ldots)$ of a fixed size n

$$\hat{\theta}_{\text{MLE}} = \frac{k}{n}$$ # random variable

- What can we say about this random variable $\hat{\theta}_{\text{MLE}}$?
- First good property of MLE for Binomial: **unbiased**
  - Definition: **bias** of our MLE is

"true predictor"

$$\text{Bias}(\hat{\theta}_{\text{MLE}}) := \mathbb{E}_{\mathcal{D}\sim P_{\theta^*}}[\hat{\theta}_{\text{MLE}}] - \theta^* = E\left[\frac{k}{n}\right] - \theta^*$$

$$= \frac{\theta^* \cancel{n}}{\cancel{n}} - \theta^*$$

Unbiased means bias = 0

- **Expectation** describes how the estimator behaves *on average*

How good is MLE? Well, it's unbiased

- We treat MLE $\widehat{\theta}_{\text{MLE}}$ as a random variable, where there is a ground truth parameter θ^* that generates the data $\mathscr{D} = (HHTTH\ldots)$ of a fixed size n

$$\hat{\theta}_{\text{MLE}} = \frac{k}{n}$$ # random variable

- What can we say about this random variable $\widehat{\theta}_{\text{MLE}}$?
- First good property of MLE for Binomial: **unbiased**
  - Definition: **bias** of our MLE is

"true predictor"

$$\text{Bias}(\widehat{\theta}_{\text{MLE}}) := \mathbb{E}_{\mathscr{D}\sim P_{\theta^*}}[\hat{\theta}_{\text{MLE}}] - \theta^* = E\left[\frac{k}{n}\right] - \theta^*$$

$$= \frac{\theta^* \cancel{n}}{\cancel{n}} - \theta^* = 0$$

Unbiased means bias = 0

- **Expectation** describes how the estimator behaves *on average*

How good is MLE? Well, it's unbiased

- We treat MLE $\hat{\theta}_{\text{MLE}}$ as a random variable, where there is a ground truth parameter θ^* that generates the data $\mathcal{D} = (HHTTH\ldots)$ of a fixed size n

$$\hat{\theta}_{\text{MLE}} = \frac{k}{n}$$ # random variable

- What can we say about this random variable $\hat{\theta}_{\text{MLE}}$?
- First good property of MLE for Binomial: **unbiased**
  - Definition: **bias** of our MLE is

"true predictor"

$$\text{Bias}(\hat{\theta}_{\text{MLE}}) := \mathbb{E}_{\mathcal{D}\sim P_{\theta^*}}[\hat{\theta}_{\text{MLE}}] - \theta^* = E\left[\frac{k}{n}\right] - \theta^*$$

$$= \frac{\theta^* \cancel{n}}{\cancel{n}} - \theta^* = 0$$



Unbiased means bias = 0

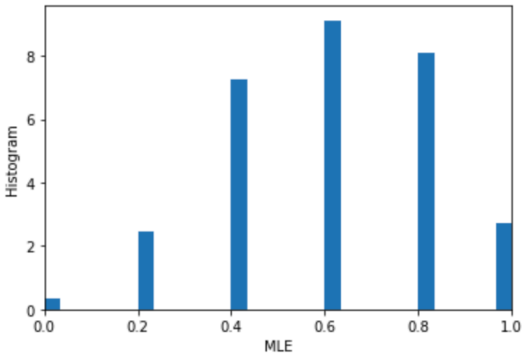
- **Expectation** describes how the estimator behaves *on average*

How many flips do I need?

$$\widehat{\theta}_{MLE} = \frac{k}{n}$$

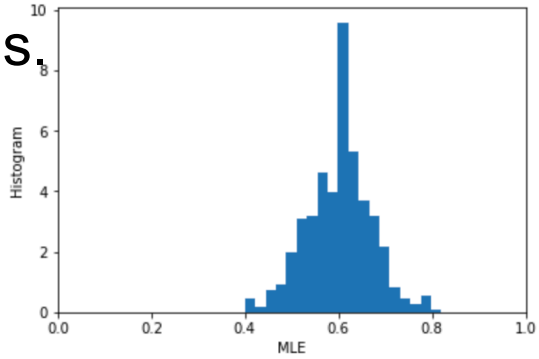
- *Client*: I flipped the coin 5 times and got 2 heads.

$$\widehat{\theta}_{MLE} =$$



- *Client*: I flipped the coin 50 times and got 30 heads.

$$\widehat{\theta}_{MLE} =$$



- *Client:* they are both unbiased, which one is right? Why?

Quantifying Uncertainty

- The **Variance** is the expected squared deviation from the mean:

$$\text{Variance}(\widehat{\theta}_{MLE}) := \mathbb{E}\left[\left(\widehat{\theta}_{MLE} - \mathbb{E}[\widehat{\theta}_{MLE}]\right)^2\right]$$

- As a rule of thumb

$$\widehat{\theta}_{\text{MLE}} \simeq \mathbb{E}\left[\widehat{\theta}_{\text{MLE}}\right] \pm \sqrt{\text{Variance}(\widehat{\theta}_{\text{MLE}})}$$

- Second good property of MLE: **minimum (asymptotic) variance**
- **Exercise**: compute the $\text{Variance}(\widehat{\theta}_{\text{MLE}})$

Expectation versus High Probability

- Tail bound of a random variable
- For any ε>0 can we bound $\mathbb{P}(|\widehat{\theta}_{MLE} - \mathbb{E}[\widehat{\theta}_{MLE}]| \geq \epsilon)$?

Markov's inequality

For any $t > 0$ and non-negative random variable X

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$$

- **Exercise**: Apply Markov's inequality to obtain bound. (Hint: set $X = |\widehat{\theta}_{MLE} - \theta^*|^2$)

Maximum Likelihood Estimation

Observe $X_1, X_2, \ldots, X_n$ drawn IID from $f(x;\theta)$ for some "true" $\theta = \theta_*$

Likelihood function $L_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta)$

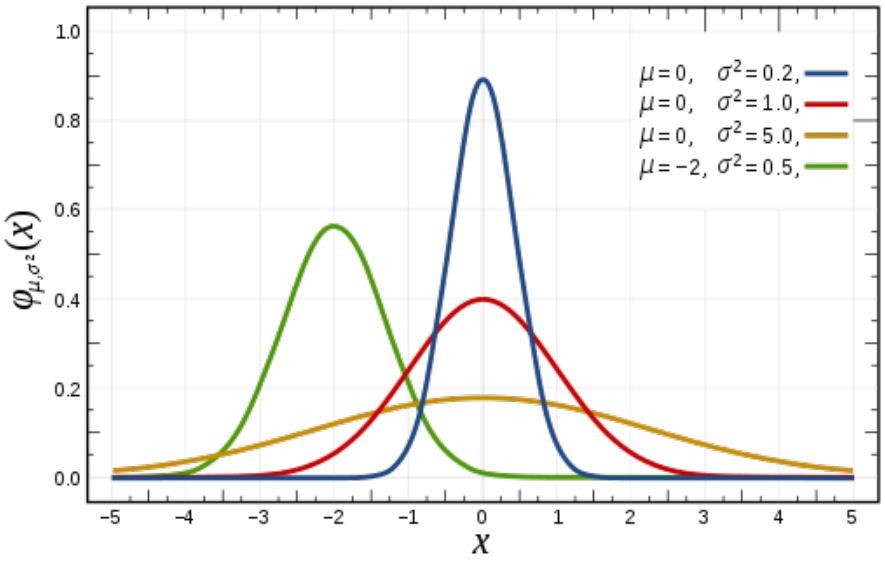
Log-Likelihood function $l_n(\theta) = \log(L_n(\theta)) = \sum_{i=1}^{n} \log(f(X_i;\theta))$

Maximum Likelihood Estimator (MLE) $\widehat{\theta}_{MLE} = \arg\max_{\theta} L_n(\theta)$

What about continuous variables?

- *Client*: What if I am measuring a **continuous variable**?
- ***You*: Let me tell you about Gaussians…**

$$P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

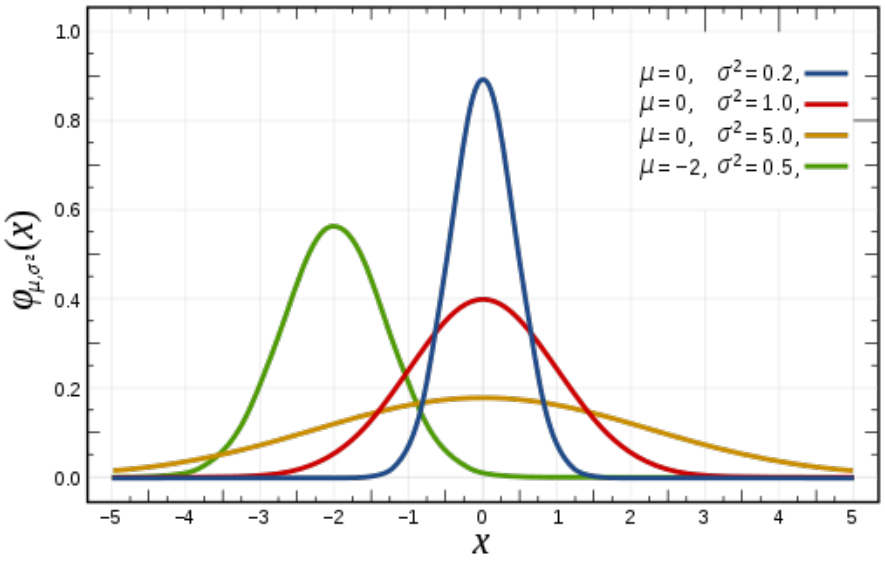


What about continuous variables?

- *Client*: What if I am measuring a **continuous variable**?
- ***You*: Let me tell you about Gaussians…**

$$P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

PDF of Gaussian



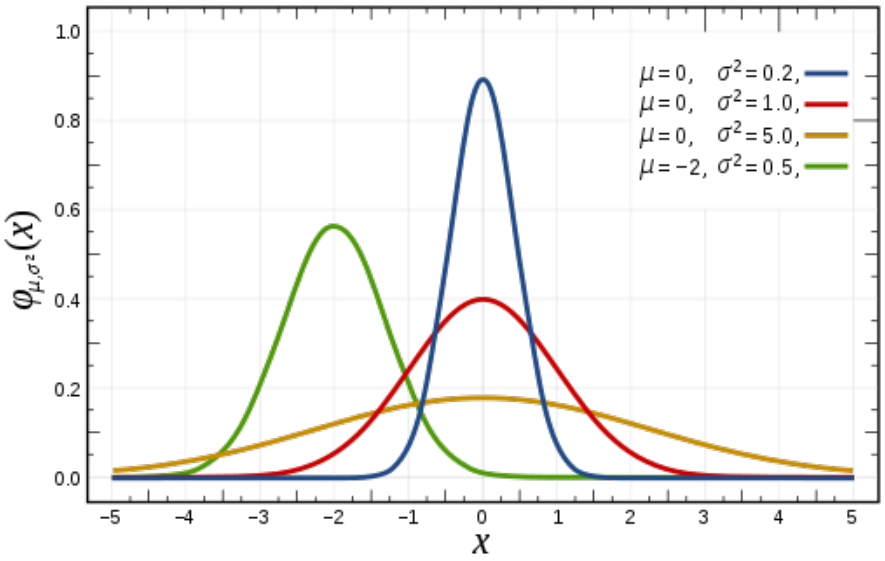
What about continuous variables?

- *Client*: What if I am measuring a **continuous variable**?
- ***You*: Let me tell you about Gaussians…**

$$P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

PDF of Gaussian

(good candidate for formula sheet)



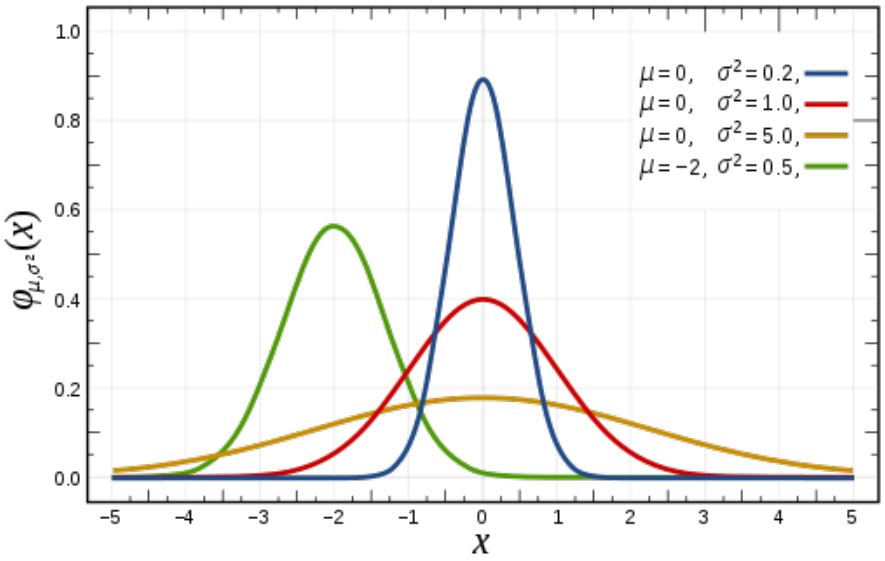
What about continuous variables?

- *Client*: What if I am measuring a **continuous variable**?
- ***You*: Let me tell you about Gaussians…**

$$P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

PDF of Gaussian

(good candidate for formula sheet)



Given a set of i.i.d. samples from a Gaussian, fit what parameters?

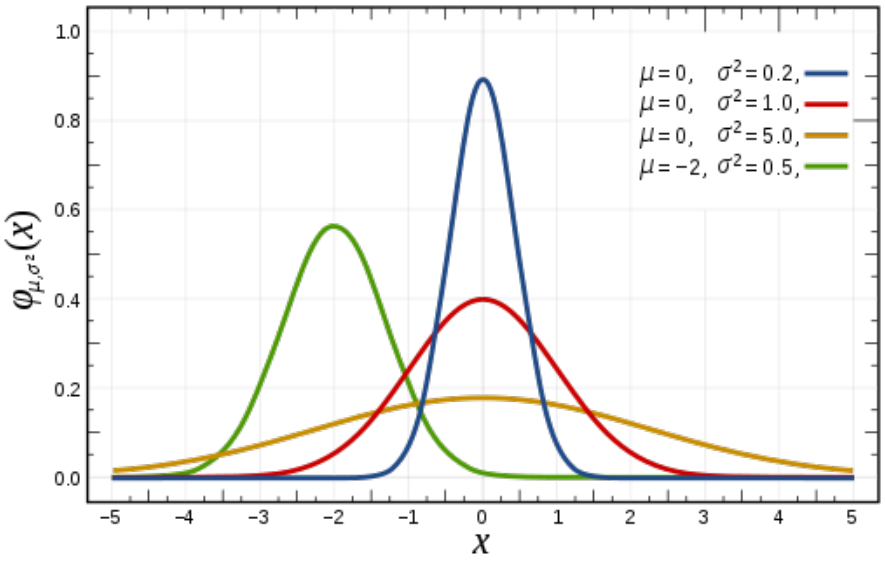
What about continuous variables?

- *Client*: What if I am measuring a **continuous variable**?
- ***You*: Let me tell you about Gaussians…**

$$P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

PDF of Gaussian

(good candidate for formula sheet)



Given a set of i.i.d. samples from a Gaussian, fit what parameters?

$$\theta = [\mu, \sigma]$$

Some properties of Gaussians

- affine transformation (multiplying by scalar and adding a constant)
  - $X \sim N(\mu, \sigma^2)$
  - $Y = aX + b \quad \rightarrow \quad Y \sim N(a\mu+b, a^2\sigma^2)$

- Sum of Gaussians
  - $X \sim N(\mu_X, \sigma^2_X)$
  - $Y \sim N(\mu_Y, \sigma^2_Y)$
  - $Z = X+Y \quad \rightarrow \quad Z \sim N(\mu_X+\mu_Y, \sigma^2_X+\sigma^2_Y)$

MLE for Gaussian

- Prob. of i.i.d. samples $D=\{x_1,\ldots,x_n\}$ (e.g., temperature):

$$P(\mathcal{D}|\mu,\sigma) = P(x_1,\ldots,x_n|\mu,\sigma)$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

- Log-likelihood of data:

$$\log P(\mathcal{D}|\mu,\sigma) = -n\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}$$

- What is $\widehat{\theta}_{MLE}$ for $\theta = (\mu,\sigma^2)$? Draw a picture!

MLE for Gaussian

- Prob. of i.i.d. samples $D=\{x_1,\ldots,x_n\}$ (e.g., temperature):

$$P(\mathcal{D}|\mu,\sigma) = P(x_1,\ldots,x_n|\mu,\sigma)$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

Wait, why can I just multiply all samples together?

- Log-likelihood of data:

$$\log P(\mathcal{D}|\mu,\sigma) = -n\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}$$

- What is $\widehat{\theta}_{MLE}$ for $\theta = (\mu,\sigma^2)$? Draw a picture!

MLE for Gaussian

- Prob. of i.i.d. samples $D=\{x_1,\ldots,x_n\}$ (e.g., temperature):

$$P(\mathcal{D}|\mu,\sigma) = P(x_1,\ldots,x_n|\mu,\sigma)$$
$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

Wait, why can I just multiply all samples together?

→ By i.i.d. assumption

- Log-likelihood of data:

$$\log P(\mathcal{D}|\mu,\sigma) = -n\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}$$

- What is $\widehat{\theta}_{MLE}$ for $\theta=(\mu,\sigma^2)$? Draw a picture!

MLE for Gaussian

- Prob. of i.i.d. samples $D=\{x_1,\ldots,x_n\}$ (e.g., temperature):

$$P(\mathcal{D}|\mu,\sigma) = P(x_1,\ldots,x_n|\mu,\sigma)$$

Likelihood:

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

Wait, why can I just multiply all samples together?

→ By i.i.d. assumption

- Log-likelihood of data:

$$\log P(\mathcal{D}|\mu,\sigma) = -n\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}$$

- What is $\widehat{\theta}_{MLE}$ for $\theta = (\mu,\sigma^2)$? Draw a picture!

MLE for Gaussian

- Prob. of i.i.d. samples $D=\{x_1,\ldots,x_n\}$ (e.g., temperature):

$$P(\mathcal{D}|\mu,\sigma) = P(x_1,\ldots,x_n|\mu,\sigma)$$

Likelihood:

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

Wait, why can I just multiply all samples together?

→ By i.i.d. assumption

- Log-likelihood of data:

LL:

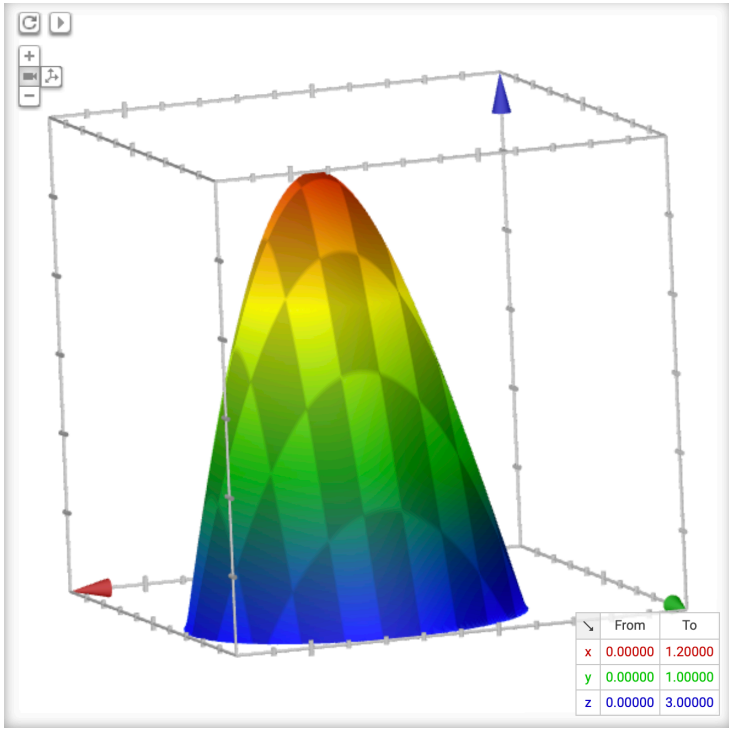
$$\log P(\mathcal{D}|\mu,\sigma) = -n\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}$$

- What is $\widehat{\theta}_{MLE}$ for $\theta = (\mu,\sigma^2)$? Draw a picture!

MLE for variance

- Again, set derivative to zero:

$$\frac{d}{d\sigma}\log P(\mathscr{D};\mu,\sigma) = \frac{d}{d\sigma}\left[-n\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$



MLE for Gaussian

Generate $\mathcal{D} = \{x_1, ..., x_n\}$, where

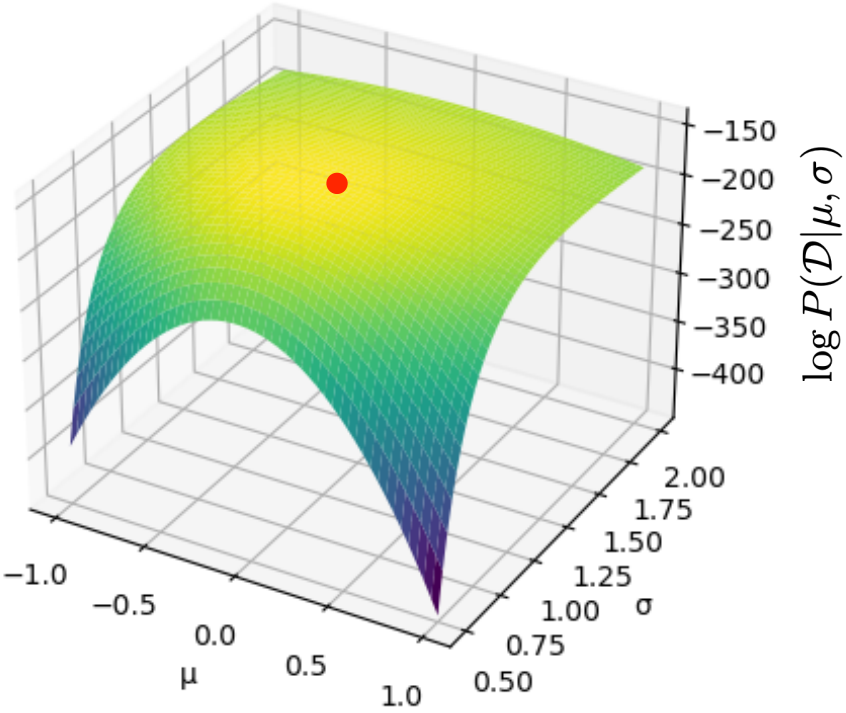
$$n = 100$$

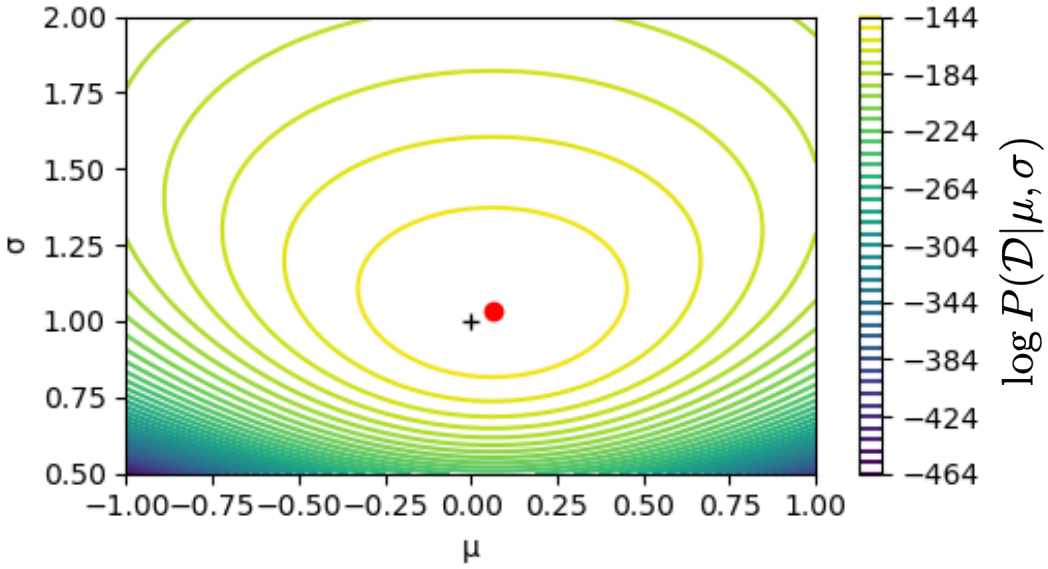
$$x_i \sim \mathcal{N}(\mu, \sigma^2)$$

$$\mu = 0$$

$$\sigma^2 = 1$$

$$\log P(\mathcal{D}|\mu, \sigma) = -n \log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}$$





+ $(\mu_{True}, \sigma_{True})$

● $(\hat{\mu}_{MLE}, \hat{\sigma}_{MLE})$

Your second learning algorithm: MLE for mean of a Gaussian

- What's MLE for mean? Set **partial derivative** to zero.

$$\frac{\partial}{\partial \mu} \log P(\mathcal{D} \mid \mu, \sigma) = \frac{\partial}{\partial \mu} \left[-n \log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} \right]$$

Your second learning algorithm: MLE for mean of a Gaussian

- What's MLE for mean? Set **partial derivative** to zero.

$$\frac{\partial}{\partial \mu} \log P(\mathcal{D} \mid \mu, \sigma) = \frac{\partial}{\partial \mu} \left[-n \log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} \right]$$

Your second learning algorithm: MLE for mean of a Gaussian

- What's MLE for mean? Set **partial derivative** to zero.

$$\frac{\partial}{\partial \mu} \log P(\mathcal{D} \mid \mu, \sigma) = \frac{\partial}{\partial \mu} \left[-n \log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} \right]$$

$$= -\sum_{i=1}^{n} \frac{-2(x_i - \mu)}{2\sigma^2}$$

Your second learning algorithm: MLE for mean of a Gaussian

- What's MLE for mean? Set **partial derivative** to zero.

$$\frac{\partial}{\partial \mu} \log P(\mathcal{D} \mid \mu, \sigma) = \frac{\partial}{\partial \mu} \left[-n \log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} \right]$$

$$= -\sum_{i=1}^{n} \frac{-2(x_i - \mu)}{2\sigma^2}$$

Your second learning algorithm: MLE for mean of a Gaussian

- What's MLE for mean? Set **partial derivative** to zero.

$$\frac{\partial}{\partial \mu} \log P(\mathcal{D} \mid \mu, \sigma) = \frac{\partial}{\partial \mu} \left[-n \log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} \right]$$

$$= \sum_{i=1}^{n} \frac{2(x_i - \mu)}{2\sigma^2}$$

Your second learning algorithm: MLE for mean of a Gaussian

- What's MLE for mean? Set **partial derivative** to zero.

$$\frac{\partial}{\partial \mu} \log P(\mathcal{D} \mid \mu, \sigma) = \frac{\partial}{\partial \mu} \left[-n \log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} \right]$$

$$= \sum_{i=1}^{n} \frac{2(x_i - \mu)}{2\sigma^2}$$

$$= \frac{-n\mu + \sum_{i=1}^{n} x_i}{\sigma^2} = 0$$

Your second learning algorithm: MLE for mean of a Gaussian

- What's MLE for mean? Set **partial derivative** to zero.

$$\frac{\partial}{\partial \mu} \log P(\mathcal{D} \mid \mu, \sigma) = \frac{\partial}{\partial \mu} \left[-n \log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} \right]$$

$$= \sum_{i=1}^{n} \frac{2(x_i - \mu)}{2\sigma^2}$$

$$= \frac{-n\mu + \sum_{i=1}^{n} x_i}{\sigma^2} = 0$$

$$\hat{\mu}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Your second learning algorithm: MLE for mean of a Gaussian

- What's MLE for mean? Set **partial derivative** to zero.

$$\frac{\partial}{\partial \mu} \log P(\mathcal{D} \mid \mu, \sigma) = \frac{\partial}{\partial \mu} \left[-n \log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} \right]$$

$$= \sum_{i=1}^{n} \frac{2(x_i - \mu)}{2\sigma^2}$$

$$= \frac{-n\mu + \sum_{i=1}^{n} x_i}{\sigma^2} = 0$$

$$\hat{\mu}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Your second learning algorithm: MLE for mean of a Gaussian

- What's MLE for mean? Set **partial derivative** to zero.

$$\frac{\partial}{\partial \mu} \log P(\mathcal{D} \mid \mu, \sigma) = \frac{\partial}{\partial \mu} \left[-n \log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} \right]$$

$$= \sum_{i=1}^{n} \frac{2(x_i - \mu)}{2\sigma^2}$$

$$= \frac{-n\mu + \sum_{i=1}^{n} x_i}{\sigma^2} = 0$$

$$\hat{\mu}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

empirical mean!

MLE for variance

- Again, set partial derivative to zero:

$$\frac{\partial}{\partial \sigma} \log P(\mathcal{D} \mid \mu, \sigma) = \frac{\partial}{\partial \sigma} \left[-n \log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} \right]$$

MLE for variance

- Again, set partial derivative to zero:

$$\frac{\partial}{\partial \sigma} \log P(\mathcal{D} \mid \mu, \sigma) = \frac{\partial}{\partial \sigma} \left[-n \log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} \right]$$

MLE for variance

reminders for formulas:

$$\frac{d}{dx}\log x = \frac{1}{x}$$

- Again, set partial derivative to zero:

$$\frac{\partial}{\partial \sigma} \log P(\mathcal{D} \mid \mu, \sigma) = \frac{\partial}{\partial \sigma} \left[-n \log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} \right]$$

MLE for variance

reminders for formulas:

$$\frac{d}{dx}\log x = \frac{1}{x} \qquad \frac{d}{dx}\frac{1}{x^2} = \frac{d}{dx}x^{-2} = -2x^{-3}$$

- Again, set partial derivative to zero:

$$\frac{\partial}{\partial \sigma} \log P(\mathcal{D} \mid \mu, \sigma) = \frac{\partial}{\partial \sigma}\left[-n \log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}\right]$$

MLE for variance

reminders for formulas:

$$\frac{d}{dx}\log x = \frac{1}{x} \qquad \frac{d}{dx}\frac{1}{x^2} = \frac{d}{dx}x^{-2} = -2x^{-3}$$

- Again, set partial derivative to zero:

$$\frac{\partial}{\partial \sigma}\log P(\mathcal{D} \mid \mu, \sigma) = \frac{\partial}{\partial \sigma}\left[-n\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= \frac{-n}{\sigma} + \sum_{i=1}^{n}\frac{2(x_i-\mu)^2}{2\sigma^3}$$

MLE for variance

reminders for formulas:

$$\frac{d}{dx}\log x = \frac{1}{x} \qquad \frac{d}{dx}\frac{1}{x^2} = \frac{d}{dx}x^{-2} = -2x^{-3}$$

- Again, set partial derivative to zero:

$$\frac{\partial}{\partial \sigma}\log P(\mathcal{D} \mid \mu, \sigma) = \frac{\partial}{\partial \sigma}\left[-n\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= \frac{-n}{\sigma} + \sum_{i=1}^{n}\frac{2(x_i-\mu)^2}{2\sigma^3}$$

MLE for variance

reminders for formulas:

$$\frac{d}{dx}\log x = \frac{1}{x} \qquad \frac{d}{dx}\frac{1}{x^2} = \frac{d}{dx}x^{-2} = -2x^{-3}$$

- Again, set partial derivative to zero:

$$\frac{\partial}{\partial\sigma}\log P(\mathcal{D} \mid \mu, \sigma) = \frac{\partial}{\partial\sigma}\left[-n\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= \frac{-n}{\sigma} + \sum_{i=1}^{n}\frac{2(x_i-\mu)^2}{2\sigma^3}$$

$$= \frac{-n\sigma^2 + \sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^3}$$

MLE for variance

reminders for formulas:

$$\frac{d}{dx}\log x = \frac{1}{x} \qquad \frac{d}{dx}\frac{1}{x^2} = \frac{d}{dx}x^{-2} = -2x^{-3}$$

- Again, set partial derivative to zero:

$$\frac{\partial}{\partial \sigma}\log P(\mathcal{D} \mid \mu, \sigma) = \frac{\partial}{\partial \sigma}\left[-n\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= \frac{-n}{\sigma} + \sum_{i=1}^{n}\frac{2(x_i-\mu)^2}{2\sigma^3}$$

$$= \frac{-n\sigma^2 + \sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^3} = 0$$

MLE for variance

reminders for formulas:

$$\frac{d}{dx}\log x = \frac{1}{x} \qquad \frac{d}{dx}\frac{1}{x^2} = \frac{d}{dx}x^{-2} = -2x^{-3}$$

- Again, set partial derivative to zero:

$$\frac{\partial}{\partial\sigma}\log P(\mathcal{D} \mid \mu, \sigma) = \frac{\partial}{\partial\sigma}\left[-n\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= \frac{-n}{\sigma} + \sum_{i=1}^{n}\frac{2(x_i-\mu)^2}{2\sigma^3}$$

$$= \frac{-n\sigma^2 + \sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^3} = 0$$

MLE for variance

reminders for formulas:

$$\frac{d}{dx}\log x = \frac{1}{x} \qquad \frac{d}{dx}\frac{1}{x^2} = \frac{d}{dx}x^{-2} = -2x^{-3}$$

- Again, set partial derivative to zero:

$$\frac{\partial}{\partial\sigma}\log P(\mathcal{D} \mid \mu, \sigma) = \frac{\partial}{\partial\sigma}\left[-n\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= \frac{-n}{\sigma} + \sum_{i=1}^{n}\frac{2(x_i-\mu)^2}{2\sigma^3}$$

$$= \frac{-n\sigma^2 + \sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^3} = 0$$

$$= \sigma^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_{\text{MLE}})^2$$

MLE for variance

reminders for formulas:

$$\frac{d}{dx}\log x = \frac{1}{x} \qquad \frac{d}{dx}\frac{1}{x^2} = \frac{d}{dx}x^{-2} = -2x^{-3}$$

- Again, set partial derivative to zero:

$$\frac{\partial}{\partial\sigma}\log P(\mathcal{D} \mid \mu, \sigma) = \frac{\partial}{\partial\sigma}\left[-n\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= \frac{-n}{\sigma} + \sum_{i=1}^{n}\frac{2(x_i-\mu)^2}{2\sigma^3}$$

$$= \frac{-n\sigma^2 + \sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^3} = 0$$

$$= \sigma^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_{\text{MLE}})^2$$

sub in $\hat{\mu}_{\text{MLE}}$

MLE for variance

reminders for formulas:

$$\frac{d}{dx}\log x = \frac{1}{x} \qquad \frac{d}{dx}\frac{1}{x^2} = \frac{d}{dx}x^{-2} = -2x^{-3}$$

- Again, set partial derivative to zero:

$$\frac{\partial}{\partial\sigma}\log P(\mathcal{D} \mid \mu, \sigma) = \frac{\partial}{\partial\sigma}\left[-n\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= \frac{-n}{\sigma} + \sum_{i=1}^{n}\frac{2(x_i-\mu)^2}{2\sigma^3}$$

$$= \frac{-n\sigma^2 + \sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^3} = 0$$

$$= \sigma^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_{\text{MLE}})^2$$

sub in $\hat{\mu}_{\text{MLE}}$

Learning Gaussian parameters

- MLE:

$$\widehat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\widehat{\sigma^2}_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{\mu}_{MLE})^2$$

- MLE for the variance of a Gaussian is **biased**

$$\mathbb{E}[\widehat{\sigma^2}_{MLE}] \neq \sigma^2$$

- Unbiased variance estimator:

$$\widehat{\sigma^2}_{unbiased} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \widehat{\mu}_{MLE})^2$$

Maximum Likelihood Estimation

Observe $X_1, X_2, \ldots, X_n$ drawn IID from $f(x;\theta)$ for some "true" $\theta = \theta_*$

Likelihood function $L_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta)$

Log-Likelihood function $l_n(\theta) = \log(L_n(\theta)) = \sum_{i=1}^{n} \log(f(X_i;\theta))$

Maximum Likelihood Estimator (MLE) $\widehat{\theta}_{MLE} = \arg\max_{\theta} L_n(\theta)$

Properties (under benign regularity conditions—smoothness, identifiability, etc.):

- Asymptotically consistent and normal: $\frac{\widehat{\theta}_{MLE} - \theta_*}{\widehat{se}} \sim \mathcal{N}(0,1)$
- Asymptotic Optimality, minimum variance (see Cramer-Rao lower bound)

The MLE is a "recipe" that begins with a *model* for data $f(x;\theta)$

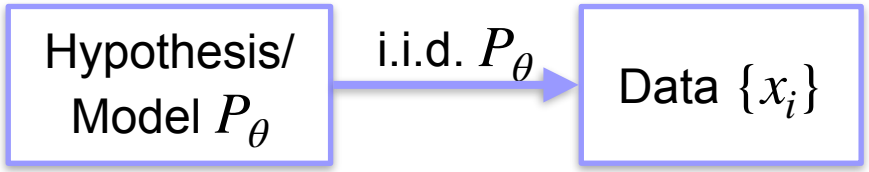
Recap

- Learning is…
  - Collect some data
    - E.g., coin flips

Data $\{x_i\}$

Recap

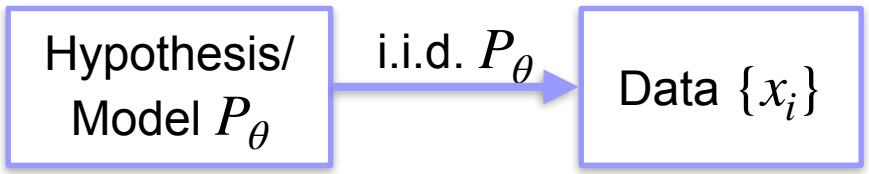
- Learning is…
  - Collect some data
    - E.g., coin flips
  - Choose a hypothesis class or model
    - E.g., binomial



Recap

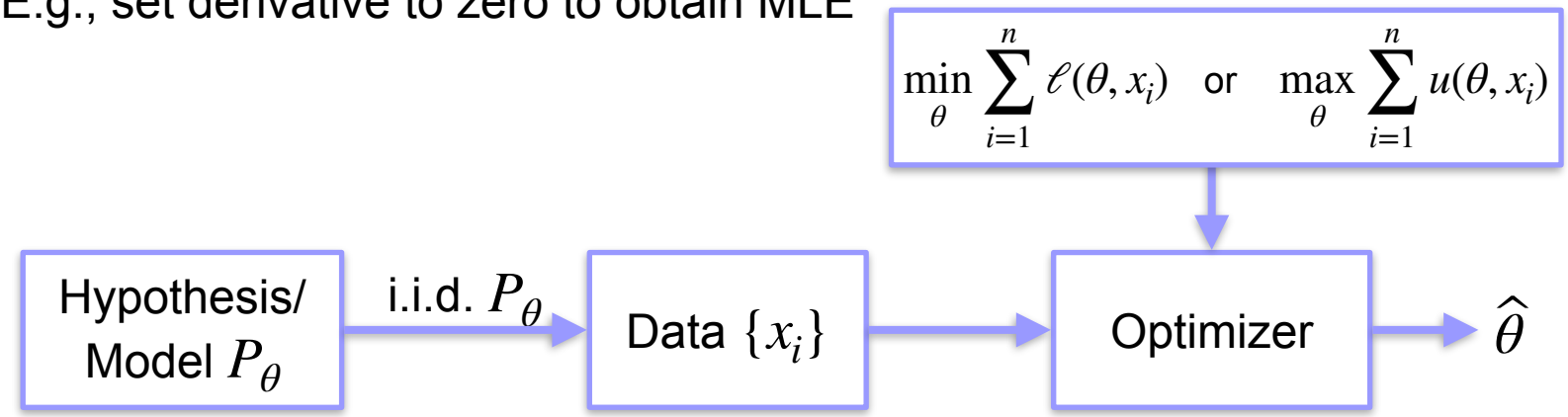
- Learning is…
  - Collect some data
    - E.g., coin flips
  - Choose a hypothesis class or model
    - E.g., binomial
  - Choose a loss function
    - E.g., data likelihood

$$\min_{\theta} \sum_{i=1}^{n} \ell(\theta, x_i) \quad \text{or} \quad \max_{\theta} \sum_{i=1}^{n} u(\theta, x_i)$$



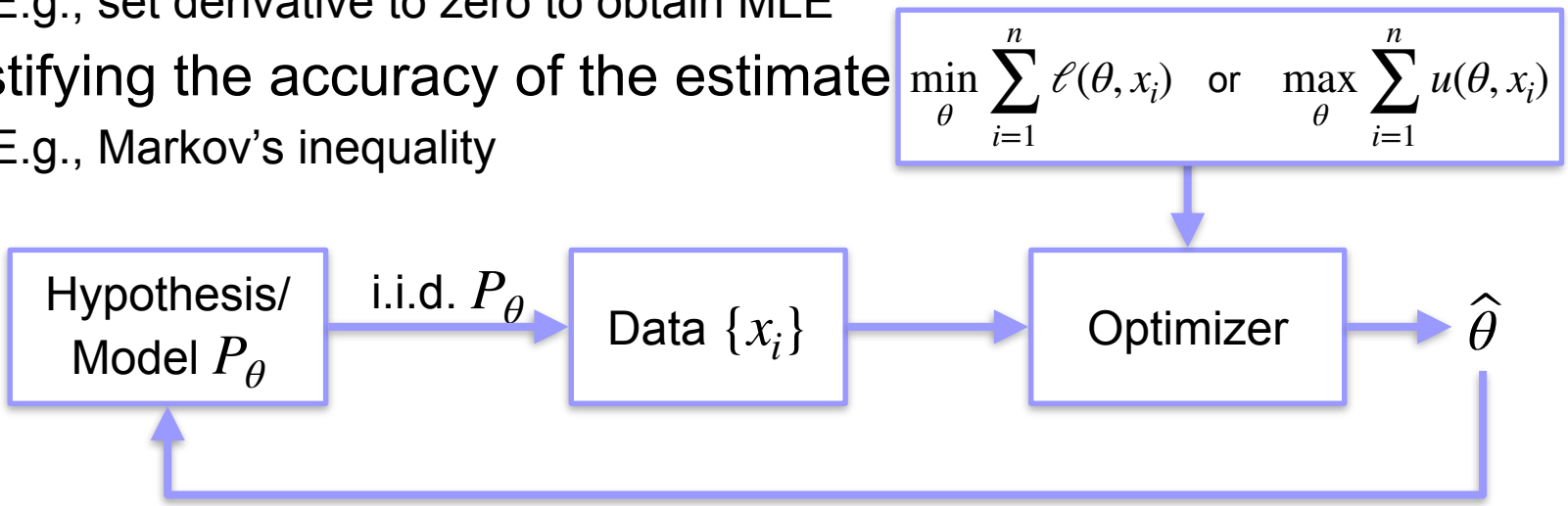
Recap

- Learning is…
  - Collect some data
    - E.g., coin flips
  - Choose a hypothesis class or model
    - E.g., binomial
  - Choose a loss function
    - E.g., data likelihood
  - Choose an optimization procedure
    - E.g., set derivative to zero to obtain MLE



Recap

- Learning is…
  - Collect some data
    - E.g., coin flips
  - Choose a hypothesis class or model
    - E.g., binomial
  - Choose a loss function
    - E.g., data likelihood
  - Choose an optimization procedure
    - E.g., set derivative to zero to obtain MLE
  - Justifying the accuracy of the estimate
    - E.g., Markov's inequality



Applications preview

Maximum Likelihood Estimation

Maximum Likelihood Estimation

Why is it useful to recover the "true" parameters θ_* of a probabilistic model?

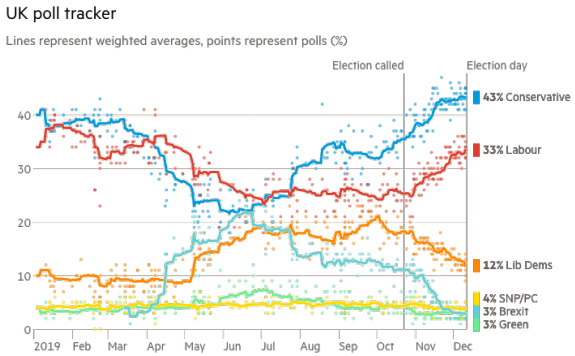
- **Estimation** of the parameters θ_* is the goal
- Help **interpret** or summarize large datasets
- Make **predictions** about future data
- **Generate** new data $X \sim f(\cdot\,; \hat{\theta}_{\text{MLE}})$

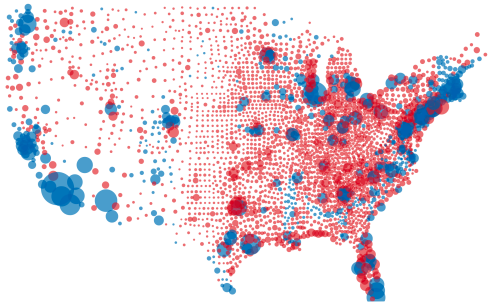
Estimation

Observe $X_1, X_2, \ldots, X_n$ drawn IID from $f(x;\theta)$ for some "true" $\theta = \theta_*$

Opinion polls
How does the greater population feel about an issue?
Correct for over-sampling?

- θ_* is "true" average opinion
- $X_1, X_2, \ldots$ are sample calls



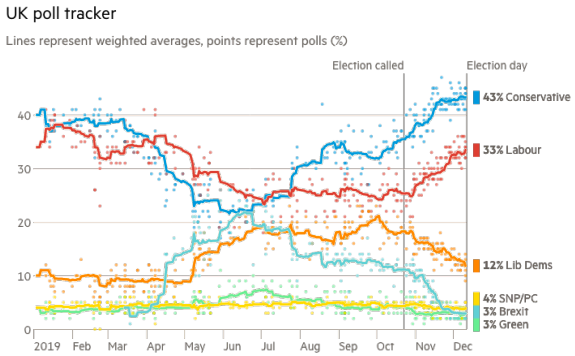


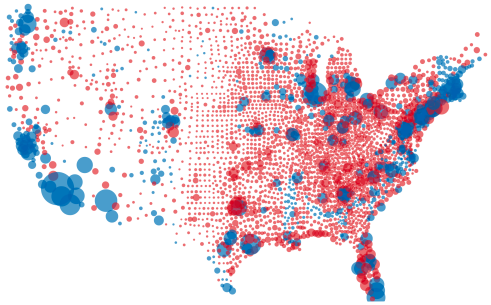
Estimation

Observe $X_1, X_2, \ldots, X_n$ drawn IID from $f(x;\theta)$ for some "true" $\theta = \theta_*$

Opinion polls
How does the greater population feel about an issue? Correct for over-sampling?

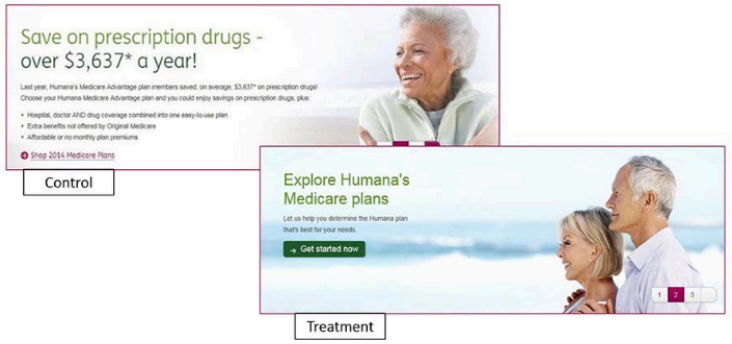
- θ_* is "true" average opinion
- $X_1, X_2, \ldots$ are sample calls





A/B testing
How do we figure out which ad results in more click-through?

- θ_* are the "true" average rates
- $X_1, X_2, \ldots$ are binary "clicks"



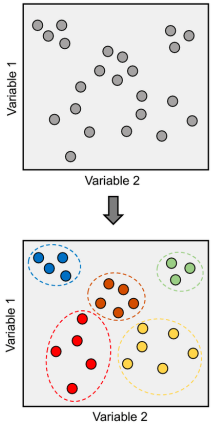
Interpret

Observe $X_1, X_2, \ldots, X_n$ drawn IID from $f(x;\theta)$ for some "true" $\theta = \theta_*$

Customer segmentation / clustering
Can we identify distinct groups of customers by their behavior?

- θ_* describes "center" of distinct groups
- $X_1, X_2, \ldots$ are individual customers





Interpret

Observe $X_1, X_2, \ldots, X_n$ drawn IID from $f(x;\theta)$ for some "true" $\theta = \theta_*$

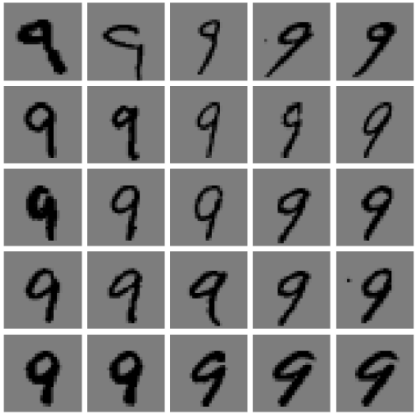
Customer segmentation / clustering
Can we identify distinct groups of customers by their behavior?

- θ_* describes "center" of distinct groups
- $X_1, X_2, \ldots$ are individual customers



Data exploration
What are the degrees of freedom of the dataset?

- θ_* describes the principle directions of variation
- $X_1, X_2, \ldots$ are the individual images



Predict

Observe $X_1, X_2, \ldots, X_n$ drawn IID from $f(x;\theta)$ for some "true" $\theta = \theta_*$

Content recommendation
Can we predict how much someone will like a movie based on past ratings?

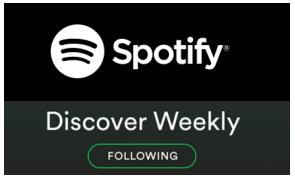
- θ_* describes user's preferences
- $X_1, X_2, \ldots$ are (movie, rating) pairs

Predict

Observe $X_1, X_2, \ldots, X_n$ drawn IID from $f(x;\theta)$ for some "true" $\theta = \theta_*$

Content recommendation
Can we predict how much someone will like a movie based on past ratings?

- θ_* describes user's preferences
- $X_1, X_2, \ldots$ are (movie, rating) pairs



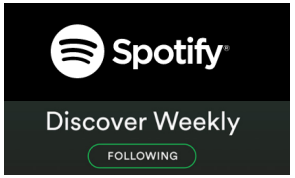
Predict

Observe $X_1, X_2, \ldots, X_n$ drawn IID from $f(x;\theta)$ for some "true" $\theta = \theta_*$

Content recommendation
Can we predict how much someone will like a movie based on past ratings?

- θ_* describes user's preferences
- $X_1, X_2, \ldots$ are (movie, rating) pairs





Predict

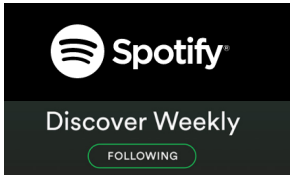
Observe $X_1, X_2, \ldots, X_n$ drawn IID from $f(x;\theta)$ for some "true" $\theta = \theta_*$

Content recommendation
Can we predict how much someone will like a movie based on past ratings?

- θ_* describes user's preferences
- $X_1, X_2, \ldots$ are (movie, rating) pairs







Predict

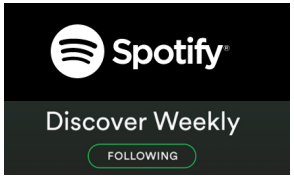
Observe $X_1, X_2, \ldots, X_n$ drawn IID from $f(x;\theta)$ for some "true" $\theta = \theta_*$

Content recommendation
Can we predict how much someone will like a movie based on past ratings?

- θ_* describes user's preferences
- $X_1, X_2, \ldots$ are (movie, rating) pairs







Object recognition / classification
Identify a flower given just its picture?

- θ_* describes the characteristics of each kind of flower
- $X_1, X_2, \ldots$ are the (image, label) pairs



(a) (b) (c)

Figure 1.1: Three types of Iris flowers: Setosa, Versicolor and Virginica. Used with kind permission of Dennis Kramb and SIGNA.

index	sl	sw	pl	pw	label
0	5.1	3.5	1.4	0.2	Setosa
1	4.9	3.0	1.4	0.2	Setosa
	...				
50	7.0	3.2	4.7	1.4	Versicolor
	...				
149	5.9	3.0	5.1	1.8	Virginica

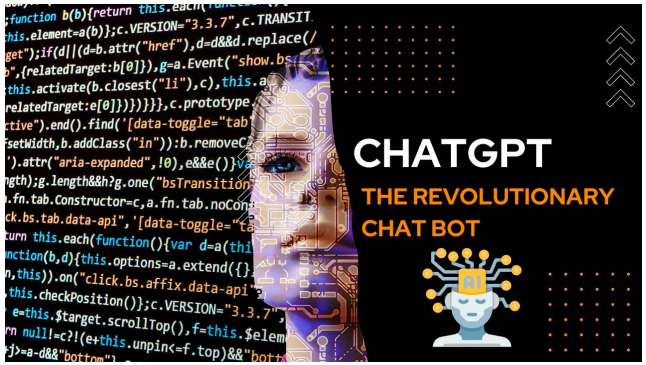
Generate

Observe $X_1, X_2, \ldots, X_n$ drawn IID from $f(x;\theta)$ for some "true" $\theta = \theta_*$

Text generation
Can AI generate text that could have been written like a human?

- θ_* describes language structure
- $X_1, X_2, \ldots$ are text snippets found online

"Kaia the dog wasn't a natural pick to go to mars. No one could have predicted she would..."



https://chat.openai.com/chat

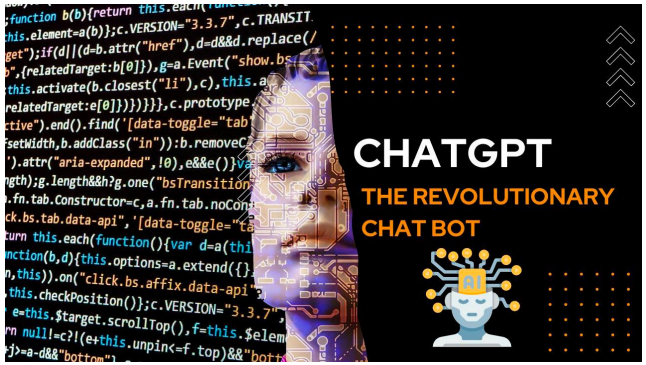
Generate

Observe $X_1, X_2, \ldots, X_n$ drawn IID from $f(x;\theta)$ for some "true" $\theta = \theta_*$

Text generation
Can AI generate text that could have been written like a human?

- θ_* describes language structure
- $X_1, X_2, \ldots$ are text snippets found online

"Kaia the dog wasn't a natural pick to go to mars. No one could have predicted she would..."



https://chat.openai.com/chat

Image to text generation
Can AI generate an image from a prompt?

- θ_* describes the coupled structure of images and text
- $X_1, X_2, \ldots$ are the (image, caption) pairs found online

"dog talking on cell phone under water, oil painting"



https://labs.openai.com/

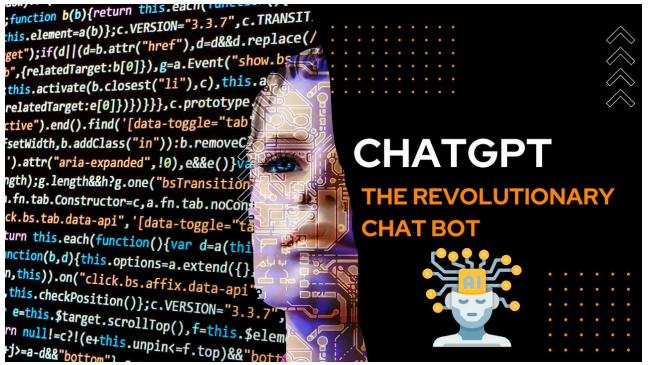
Generate

Observe $X_1, X_2, \ldots, X_n$ drawn IID from $f(x;\theta)$ for some "true" $\theta = \theta_*$

Text generation
Can AI generate text that could have been written like a human?

- θ_* describes language structure
- $X_1, X_2, \ldots$ are text snippets found online

"Kaia the dog wasn't a natural pick to go to mars. No one could have predicted she would..."



https://chat.openai.com/chat

MLE!

Image to text generation
Can AI generate an image from a prompt?

- θ_* describes the coupled structure of images and text
- $X_1, X_2, \ldots$ are the (image, caption) pairs found online

"dog talking on cell phone under water, oil painting"



https://labs.openai.com/

Linear Regression

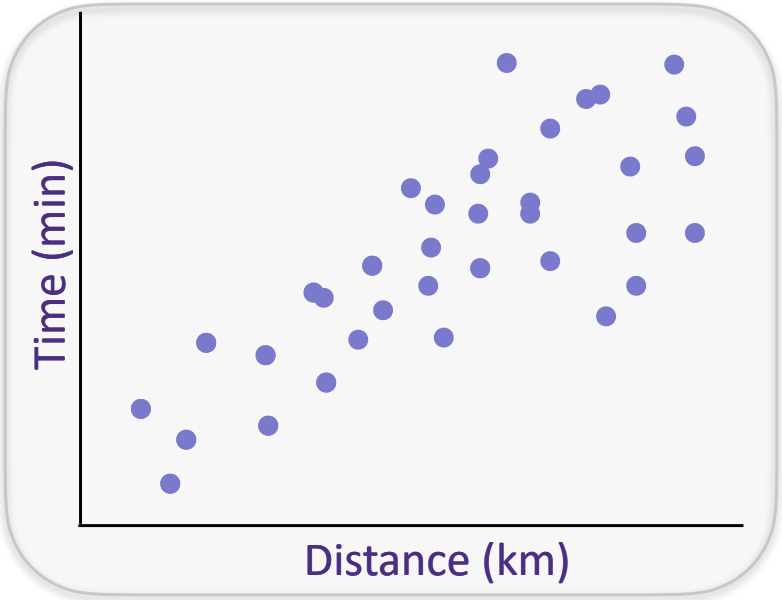


The regression problem, 1-dimensional

Potato delivery logs, predict:

y = delivery time (minutes)

x = distance (km)



Training Data:

$$\{(x_i, y_i)\}_{i=1}^{n} \qquad x_i \in \mathbb{R}, \; y_i \in \mathbb{R}$$

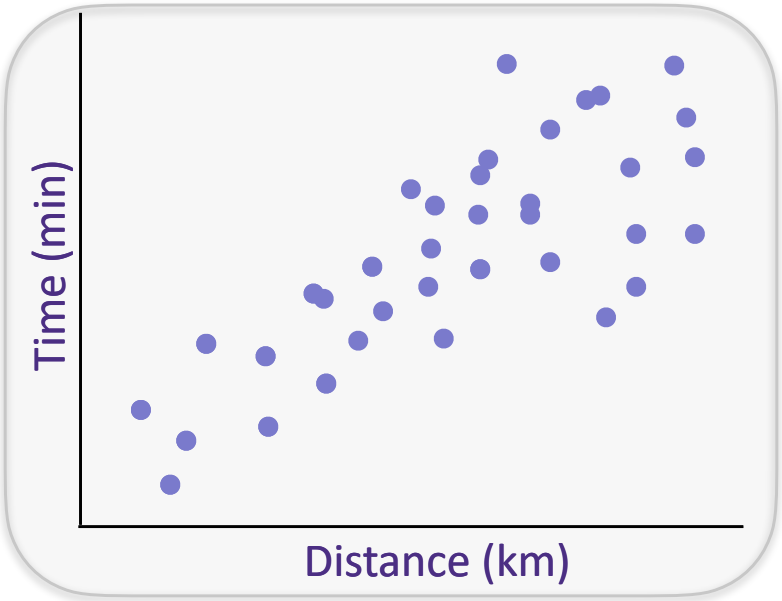
The regression problem, 1-dimensional

Potato delivery logs, predict:

y = delivery time (minutes)

x = distance (km)

previously we just had x



Training Data:

$$\{(x_i, y_i)\}_{i=1}^{n} \qquad x_i \in \mathbb{R} \quad y_i \in \mathbb{R}$$

The regression problem, 1-dimensional

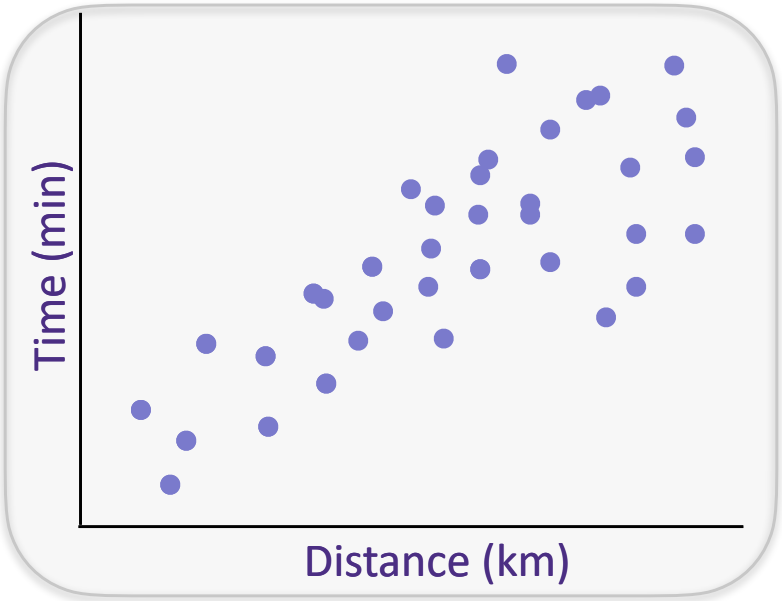
Potato delivery logs, predict:

y = delivery time (minutes)

x = distance (km)

now we have label y

previously we just had x



Training Data:

$$\{(x_i, y_i)\}_{i=1}^{n} \qquad x_i \in \mathbb{R} \quad y_i \in \mathbb{R}$$

The regression problem, 1-dimensional

Potato delivery logs, predict:

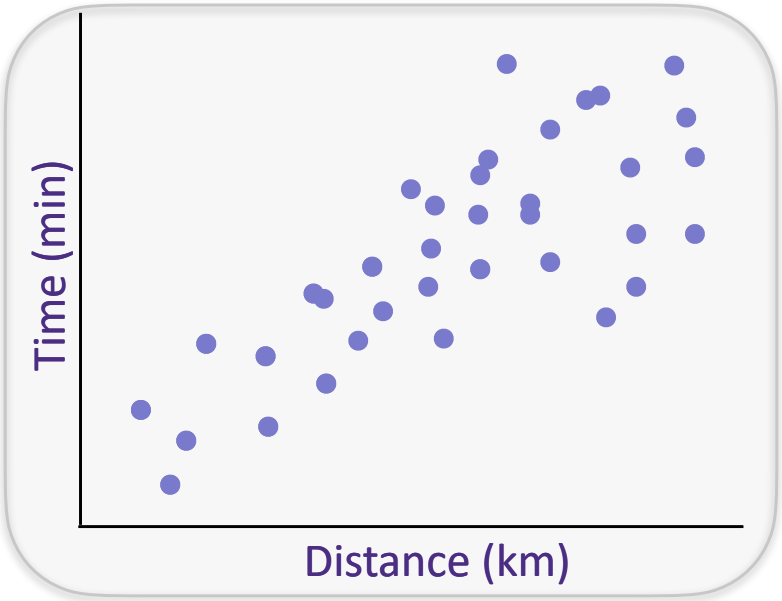
y = delivery time (minutes)

x = distance (km)

Goal: learn p(y|x)

now we have label y

previously we just had x



Training Data:

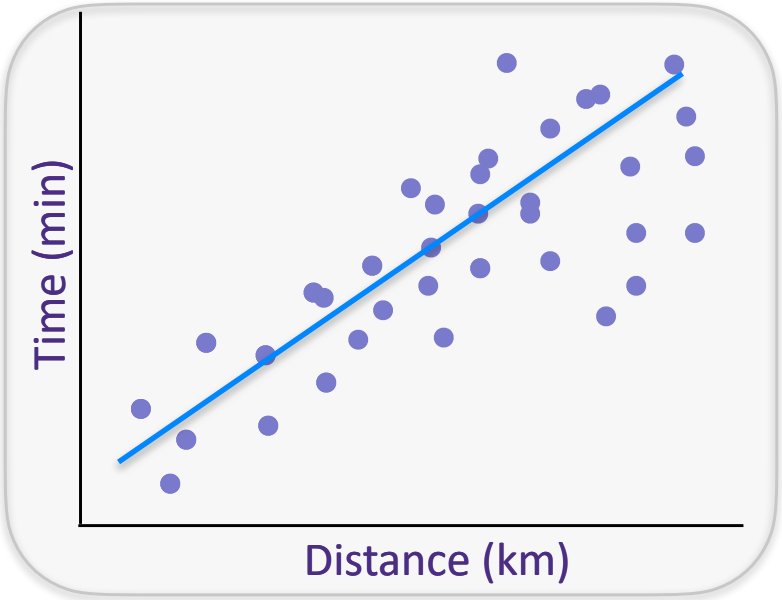
$$\{(x_i, y_i)\}_{i=1}^{n} \qquad x_i \in \mathbb{R} \quad y_i \in \mathbb{R}$$

The regression problem, 1-dimensional

Potato delivery logs, predict:

y = delivery time (minutes)

x = distance (km)



Goal: learn p(y|x)

now we have label y

previously we just had x

Training Data:

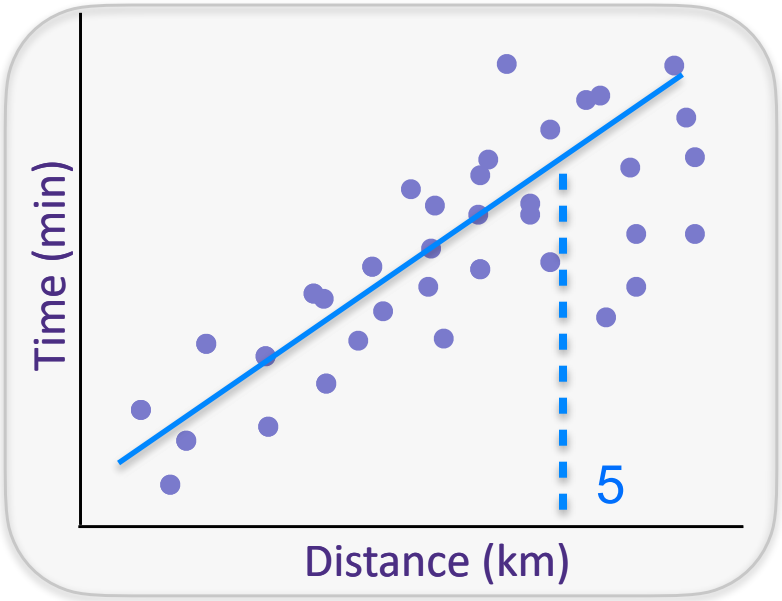
$$\{(x_i, y_i)\}_{i=1}^{n} \qquad x_i \in \mathbb{R} \quad y_i \in \mathbb{R}$$

The regression problem, 1-dimensional

Potato delivery logs, predict:

y = delivery time (minutes)

x = distance (km)



Goal: learn p(y|x)

now we have label y

previously we just had x

Training Data:

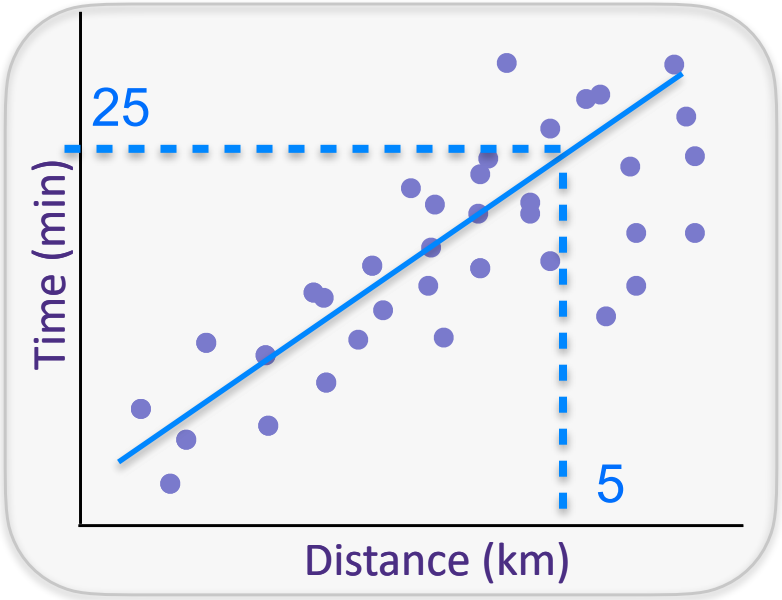
$$\{(x_i, y_i)\}_{i=1}^{n} \qquad x_i \in \mathbb{R} \quad y_i \in \mathbb{R}$$

The regression problem, 1-dimensional

Potato delivery logs, predict:

y = delivery time (minutes)

x = distance (km)



Goal: learn p(y|x)

now we have label y

previously we just had x

Training Data:

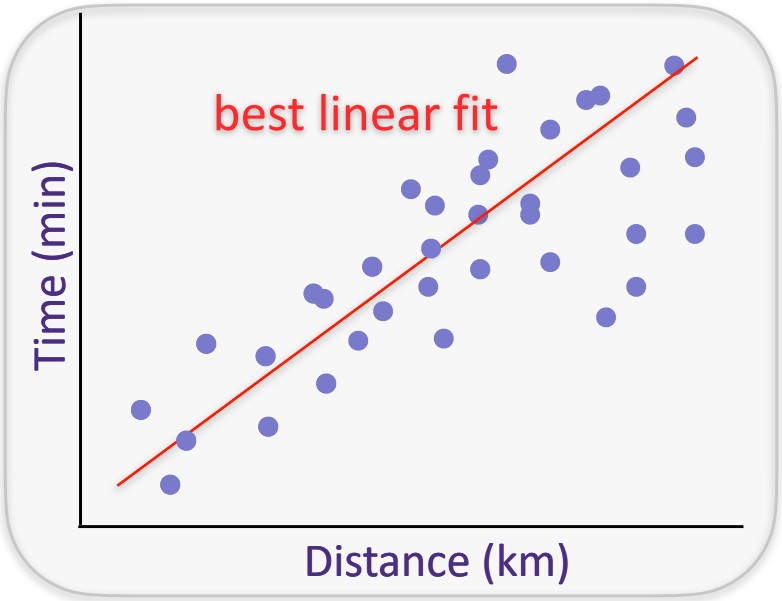
$$\{(x_i, y_i)\}_{i=1}^{n} \qquad x_i \in \mathbb{R} \quad y_i \in \mathbb{R}$$

Fit a function to our data, 1-d

Potato delivery logs, predict:

y = delivery time (minutes)

x = distance (km)



Training Data: $\{(x_i, y_i)\}_{i=1}^{n}$ $x_i \in \mathbb{R}$ $y_i \in \mathbb{R}$

Hypothesis/Model: linear

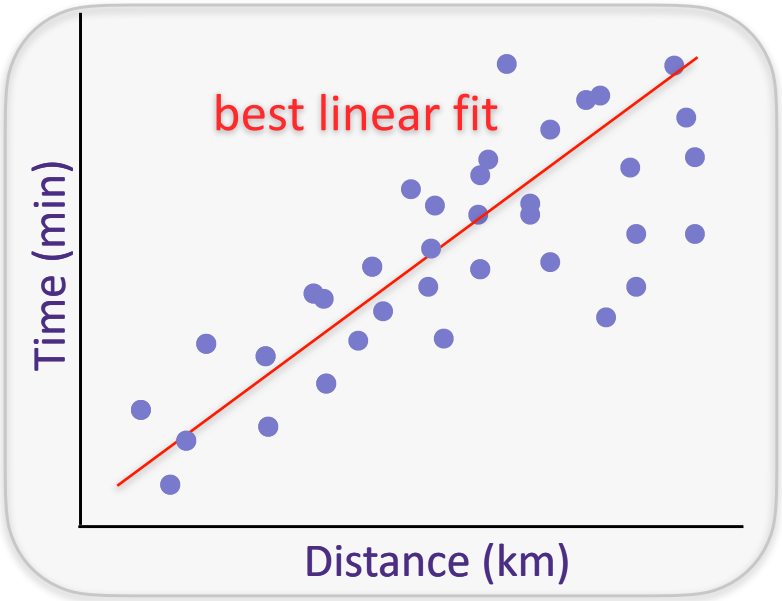
$$y_i = x_i w + \epsilon_i \qquad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

Fit a function to our data, 1-d

Potato delivery logs, predict:

y = delivery time (minutes)

x = distance (km)



Training Data: $\{(x_i, y_i)\}_{i=1}^{n}$ $x_i \in \mathbb{R}$ $y_i \in \mathbb{R}$

Hypothesis/Model: linear

$$y_i = x_i w + \epsilon_i \qquad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

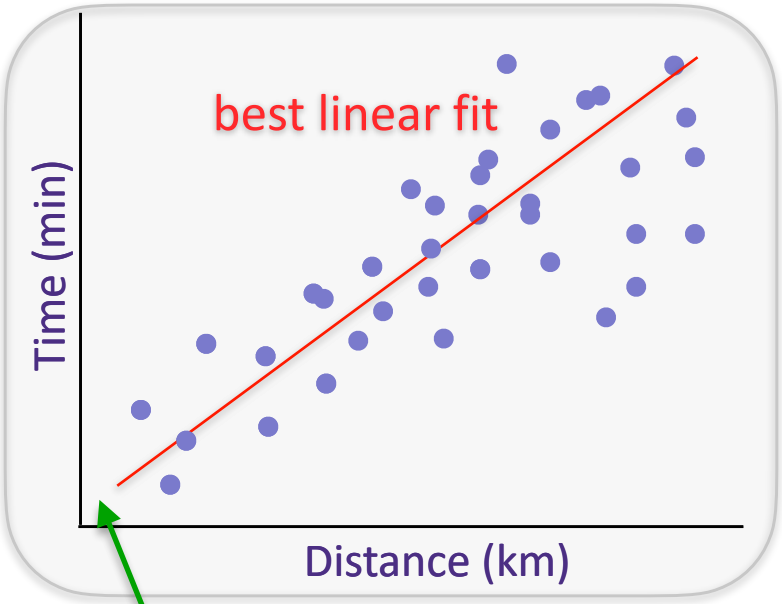
$w \in \mathbb{R}$ *slope of line*

Fit a function to our data, 1-d

Potato delivery logs, predict:

y = delivery time (minutes)

x = distance (km)



Training Data: $\{(x_i, y_i)\}_{i=1}^{n}$ $x_i \in \mathbb{R}$ $y_i \in \mathbb{R}$

Hypothesis/Model: linear

$$y_i = x_i w + \epsilon_i \qquad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

$w \in \mathbb{R}$ *slope of line*

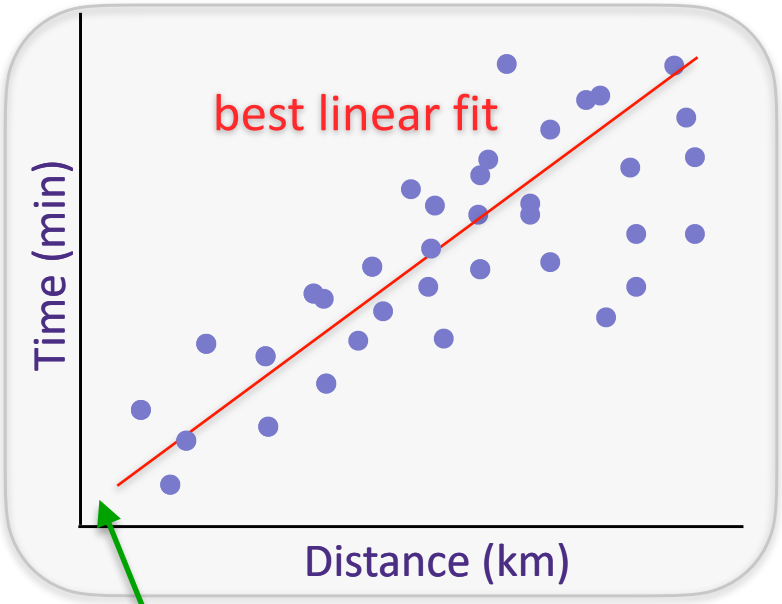
(no intercept for now to make the math easier)

Fit a function to our data, 1-d

Potato delivery logs, predict:

y = delivery time (minutes)

x = distance (km)



Training Data: $\{(x_i, y_i)\}_{i=1}^{n}$ $x_i \in \mathbb{R}$ $y_i \in \mathbb{R}$

Hypothesis/Model: linear

$$y_i = x_i w + \epsilon_i \qquad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

$w \in \mathbb{R}$ *slope of line*

- why the noise?

(no intercept for now to make the math easier)

Process

Decide on a **model**

assume delivery time is a linear function of distance.

Find the function which fits the data best

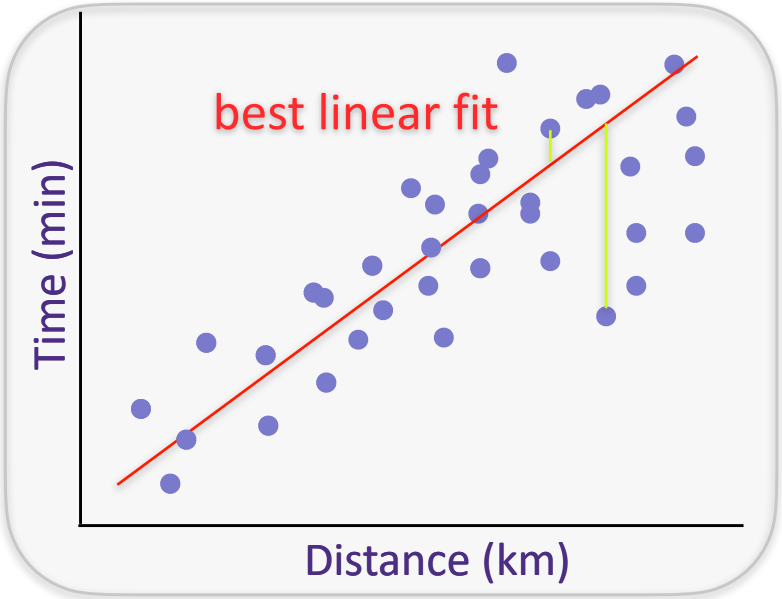
Use function to make prediction on new examples

Fit a function to our data, 1-dimension

Potato delivery logs, predict:

y = delivery time (minutes)

x = distance (km)



Error

$$y_i = x_i w + \epsilon_i$$

Training Data:

$$\{(x_i, y_i)\}_{i=1}^{n}$$

$$x_i \in \mathbb{R}$$
$$y_i \in \mathbb{R}$$

Hypothesis/Model: linear

$$y_i \approx x_i w$$

Loss: least squares solution

$$\min_w \sum_{i=1}^{n} (y_i - x_i w)^2$$

Predicting potato delivery time

- *Target: delivery time in minutes*
- *Candidate features:*
- *Distance to the destination*
- *Potato weight / payload mass*
- *Nearby delivery-drone traffic*
- *Humidity, wind and daylight*
- *Measure features available at dispatch*
- *Test which ones improve predictions*

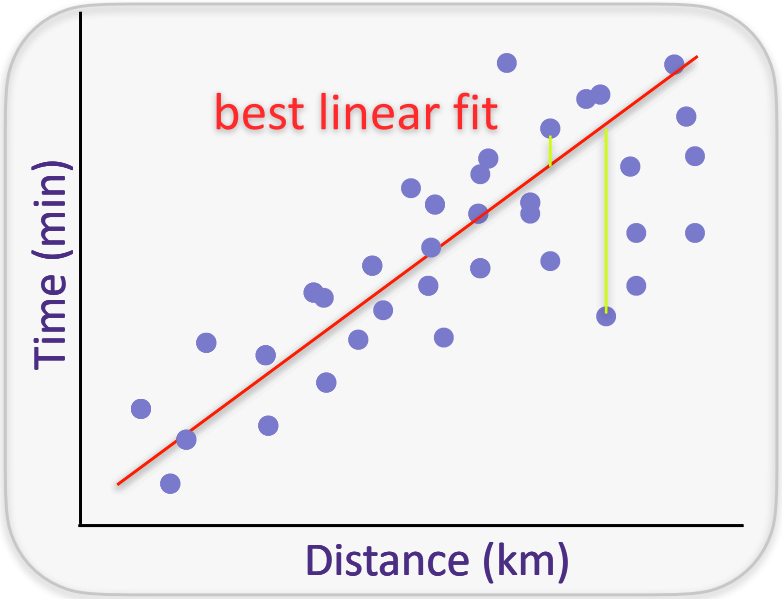


The regression problem, d-dimensions

Potato delivery logs, predict:

y = delivery time (minutes)

x = distance, traffic, prep time



Error:

$$y_i = x_i w + \epsilon_i$$

Training Data:

$$\{(x_i, y_i)\}_{i=1}^{n} \qquad x_i \in \mathbb{R}^d \quad y_i \in \mathbb{R}$$

Hypothesis/Model: linear

$$y_i \approx x_i^T w$$

Loss: least squares solution

$$\min_w \sum_{i=1}^{n} \left(y_i - x_i^T w\right)^2$$

The regression problem in matrix notation

Data:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$$

d : # of features

n : # of examples/datapoints

The regression problem in matrix notation

Data:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$$

d : # of features

n : # of examples/datapoints

Model:

$$\begin{aligned} y_1 &= x_1^T w + \epsilon_1 \\ y_2 &= x_2^T w + \epsilon_2 \\ &\vdots \\ y_n &= x_n^T w + \epsilon_n \end{aligned}$$

$$\mathbf{y} = \mathbf{X}w + \epsilon$$

The regression problem in matrix notation

Data:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$$

d : # of features

n : # of examples/datapoints

Model:

$$\begin{aligned} y_1 &= x_1^T w + \epsilon_1 \\ y_2 &= x_2^T w + \epsilon_2 \\ &\vdots \\ y_n &= x_n^T w + \epsilon_n \end{aligned}$$

$$\mathbf{y} = \mathbf{X}w + \epsilon$$

Loss: $\widehat{w}_{LS} = \arg\min_w \sum_{i=1}^{n} \left(y_i - x_i^T w\right)^2$

The regression problem in matrix notation

Data:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$$

d : # of features

n : # of examples/datapoints

Model:

$$\begin{aligned} y_1 &= x_1^T w + \epsilon_1 \\ y_2 &= x_2^T w + \epsilon_2 \\ &\vdots \\ y_n &= x_n^T w + \epsilon_n \end{aligned}$$

$$\mathbf{y} = \mathbf{X}w + \epsilon$$

Loss: $\widehat{w}_{LS} = \arg\min_w \sum_{i=1}^{n} \left(y_i - x_i^T w\right)^2 = \arg\min_w ||\mathbf{y} - \mathbf{X}w||_2^2$

$$= \arg\min_w (\mathbf{y} - \mathbf{X}w)^T (\mathbf{y} - \mathbf{X}w)$$

The regression problem in matrix notation

$$\begin{aligned}\widehat{w}_{LS} &= \arg\min_{w} ||\mathbf{y} - \mathbf{X}w||_2^2 \\ &= \arg\min_{w} (\mathbf{y} - \mathbf{X}w)^T (\mathbf{y} - \mathbf{X}w)\end{aligned}$$

Set gradient w.r.t. w to zero to find the minima:

The regression problem in matrix notation

$$\begin{aligned}\widehat{w}_{LS} &= \arg\min_{w} ||\mathbf{y} - \mathbf{X}w||_2^2 \\ &= \arg\min_{w} (\mathbf{y} - \mathbf{X}w)^T(\mathbf{y} - \mathbf{X}w) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\end{aligned}$$

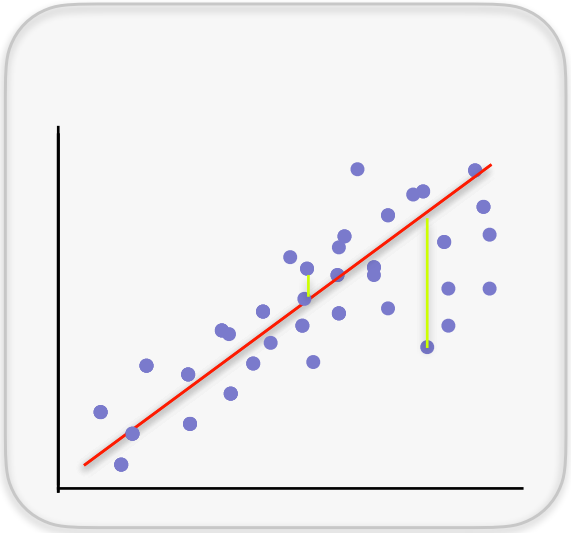
"Closed form" solution!

The regression problem in matrix notation

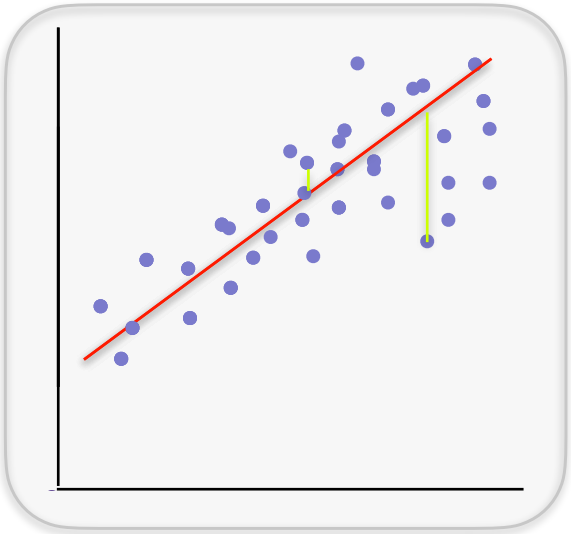
Linear model: $y_i = x_i^T w + \epsilon_i$

Least squares solution:

$$\widehat{w}_{LS} = \arg\min_w ||\mathbf{y} - \mathbf{X}w||_2^2$$
$$= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$



What about an offset (a.k.a intercept)?

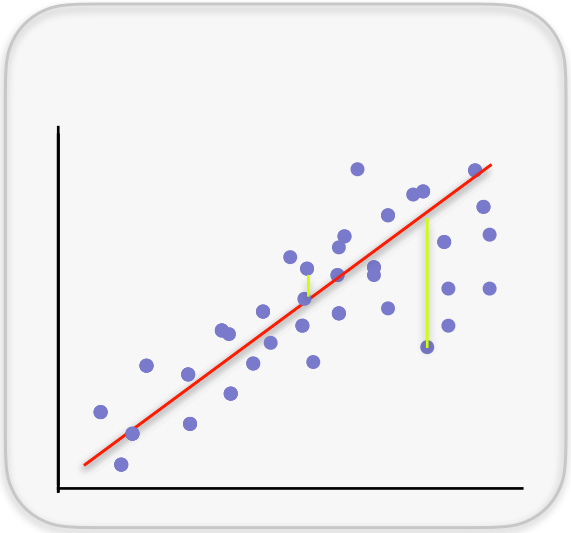


The regression problem in matrix notation

Linear model: $y_i = x_i^T w + \epsilon_i$

Least squares solution:

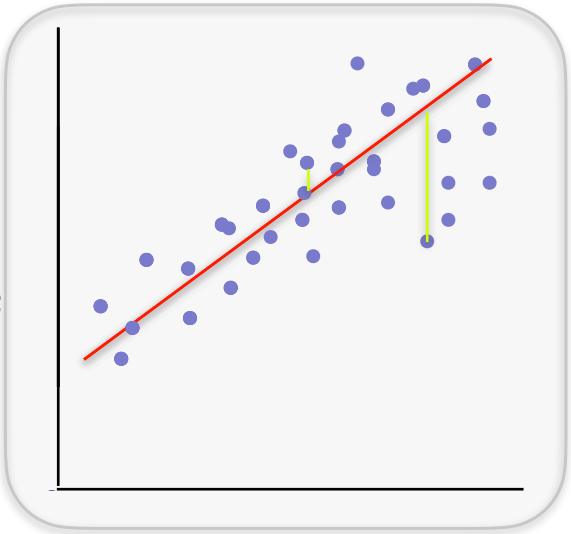
$$\widehat{w}_{LS} = \arg\min_w ||\mathbf{y} - \mathbf{X}w||_2^2$$
$$= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$



Affine model: $y_i = x_i^T w + b + \epsilon_i$

Least squares solution:

$$\widehat{w}_{LS}, \widehat{b}_{LS} = \arg\min_{w,b} \sum_{i=1}^{n} \left(y_i - (x_i^T w + b)\right)^2$$
$$= \arg\min_{w,b} ||\mathbf{y} - (\mathbf{X}w + \mathbf{1}b)||_2^2$$



Dealing with an offset

$$\widehat{w}_{LS}, \widehat{b}_{LS} = \arg\min_{w,b} ||\mathbf{y} - (\mathbf{X}w + \mathbf{1}b)||_2^2$$

Set gradient w.r.t. w and b to zero to find the minima:

Dealing with an offset

$$\widehat{w}_{LS}, \widehat{b}_{LS} = \arg\min_{w,b} ||\mathbf{y} - (\mathbf{X}w + \mathbf{1}b)||_2^2$$

$$\mathbf{X}^T\mathbf{X}\widehat{w}_{LS} + \widehat{b}_{LS}\mathbf{X}^T\mathbf{1} = \mathbf{X}^T\mathbf{y}$$

$$\mathbf{1}^T\mathbf{X}\widehat{w}_{LS} + \widehat{b}_{LS}\mathbf{1}^T\mathbf{1} = \mathbf{1}^T\mathbf{y}$$

If $\mathbf{X}^T\mathbf{1} = 0$, if the features have zero mean,

$$\widehat{w}_{LS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

$$\widehat{b}_{LS} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Dealing with an offset

$$\widehat{w}_{LS}, \widehat{b}_{LS} = \arg\min_{w,b} ||\mathbf{y} - (\mathbf{X}w + \mathbf{1}b)||_2^2$$

$$\mathbf{X}^T\mathbf{X}\widehat{w}_{LS} + \widehat{b}_{LS}\mathbf{X}^T\mathbf{1} = \mathbf{X}^T\mathbf{y}$$

$$\mathbf{1}^T\mathbf{X}\widehat{w}_{LS} + \widehat{b}_{LS}\mathbf{1}^T\mathbf{1} = \mathbf{1}^T\mathbf{y}$$

If $\mathbf{X}^T\mathbf{1} = 0$,

$$\widehat{w}_{LS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

$$\widehat{b}_{LS} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

In general, when $\mathbf{X}^T\mathbf{1} \neq 0$,

Dealing with an offset

$$\widehat{w}_{LS}, \widehat{b}_{LS} = \arg\min_{w,b} ||\mathbf{y} - (\mathbf{X}w + \mathbf{1}b)||_2^2$$

$$\mathbf{X}^T\mathbf{X}\widehat{w}_{LS} + \widehat{b}_{LS}\mathbf{X}^T\mathbf{1} = \mathbf{X}^T\mathbf{y}$$
$$\mathbf{1}^T\mathbf{X}\widehat{w}_{LS} + \widehat{b}_{LS}\mathbf{1}^T\mathbf{1} = \mathbf{1}^T\mathbf{y}$$

If $\mathbf{X}^T\mathbf{1} = 0$,

$$\widehat{w}_{LS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

$$\widehat{b}_{LS} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

In general, when $\mathbf{X}^T\mathbf{1} \neq 0$,

$$\mu = \frac{1}{n}\mathbf{X}^T\mathbf{1}$$

$$\widetilde{\mathbf{X}} = \mathbf{X} - \mathbf{1}\mu^T$$

$$\widehat{w}_{\mathrm{LS}} = (\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T\mathbf{y}$$

$$\widehat{b}_{\mathrm{LS}} = \frac{1}{n}\mathbf{1}^T\mathbf{y} - \mu^T\widehat{w}_{\mathrm{LS}}$$

Process

Decide on a **model:** $y_i = x_i^T w + b + \epsilon_i$

Choose a loss function - least squares

Pick the function which minimizes loss on data

$$\widehat{w}_{LS}, \widehat{b}_{LS} = \arg\min_{w,b} \sum_{i=1}^{n} \left(y_i - (x_i^T w + b)\right)^2$$

Use function to make prediction on new examples

$$\hat{y}_{\text{new}} = x_{\text{new}}^T \hat{w}_{LS} + \hat{b}_{LS}$$

Dealing with an offset - Revisted

$$\widehat{w}_{LS}, \widehat{b}_{LS} = \arg\min_{w,b} ||\mathbf{y} - (\mathbf{X}w + \mathbf{1}b)||_2^2$$

$$\widetilde{\mathbf{X}} = (\mathbf{X}, \mathbf{1})$$

$$\widetilde{\mathbf{X}}\widetilde{w} = \mathbf{X}w + b\mathbf{1} \quad \text{with} \quad \widetilde{w} = (w, b)$$

Why is least squares a good loss function?

$$\begin{aligned}\widehat{w}_{LS} &= \arg\min_{w} ||\mathbf{y} - \mathbf{X}w||_2^2 \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\end{aligned}$$

$$P(y_i; x_i, w, \sigma) \; =$$

Why is least squares a good loss function?

$$\widehat{w}_{LS} = \arg\min_{w} ||\mathbf{y} - \mathbf{X}w||_2^2$$
$$= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Consider $y_i = x_i^T w + \epsilon_i$ where $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

$P(y_i; x_i, w, \sigma) =$

Why is least squares a good loss function?

Maximum Likelihood Estimator:

$$\begin{aligned}\widehat{w}_{\mathrm{MLE}} &= \arg\max_{w} \log P(\{y_i\}_{i=1}^n; \{x_i\}_{i=1}^n, w, \sigma) \\ &= \arg\max_{w} -n\log(\sigma\sqrt{2\pi}) + \sum_{i=1}^{n} -\frac{(y_i - x_i^T w)^2}{2\sigma^2}\end{aligned}$$

Why is least squares a good loss function?

Maximum Likelihood Estimator:

$$\begin{aligned}
\widehat{w}_{\text{MLE}} &= \arg\max_{w} \log P(\{y_i\}_{i=1}^n; \{x_i\}_{i=1}^n, w, \sigma) \\
&= \arg\max_{w} -n\log(\sigma\sqrt{2\pi}) + \sum_{i=1}^{n} -\frac{(y_i - x_i^T w)^2}{2\sigma^2} \\
&= \arg\min_{w} \sum_{i=1}^{n} (y_i - x_i^T w)^2
\end{aligned}$$

Recall: $$\widehat{w}_{LS} = \arg\min_{w} \sum_{i=1}^{n} \left(y_i - x_i^T w\right)^2$$

$$\boxed{\widehat{w}_{LS} = \widehat{w}_{MLE} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}}$$

Analysis of Error

$$\text{if} \quad y_i = x_i^T w + \epsilon_i \quad \text{and} \quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2) \qquad \mathbf{Y} = \mathbf{X}w + \epsilon$$

$$\begin{aligned} \widehat{w}_{MLE} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}w + \epsilon) \\ &= w + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon \end{aligned}$$

Maximum Likelihood Estimator is unbiased:

Analysis of Error

$$\text{if} \quad y_i = x_i^T w + \epsilon_i \quad \text{and} \quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2) \qquad \mathbf{Y} = \mathbf{X}w + \epsilon$$

$$\begin{aligned} \widehat{w}_{MLE} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}w + \epsilon) \\ &= w + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon \end{aligned}$$

Covariance is:

Analysis of Error

$$\text{if} \quad y_i = x_i^T w + \epsilon_i \quad \text{and} \quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2) \qquad \mathbf{Y} = \mathbf{X}w + \epsilon$$

$$\begin{aligned} \widehat{w}_{MLE} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}w + \epsilon) \\ &= w + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon \end{aligned}$$

$$\mathbb{E}[\widehat{w}_{MLE}] = w$$

$$\text{Cov}(\widehat{w}_{MLE}) = \mathbb{E}[(\widehat{w} - \mathbb{E}[\widehat{w}])(\widehat{w} - \mathbb{E}[\widehat{w}])^T] = (\mathbf{X}^T\mathbf{X})^{-1}$$

$$\widehat{w}_{MLE} \sim \mathcal{N}(w, (\mathbf{X}^T\mathbf{X})^{-1})$$