CSE 446/546: Machine Learning

Matt Golub
Hunter Schafer

UNIVERSITY of WASHINGTON

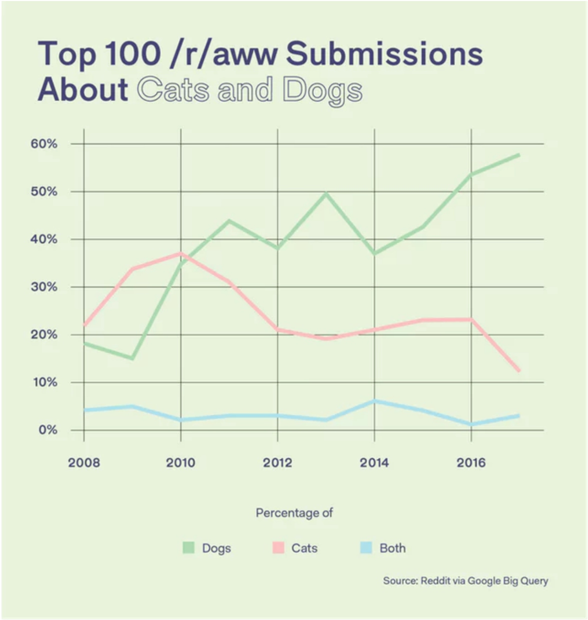
Traditional algorithms

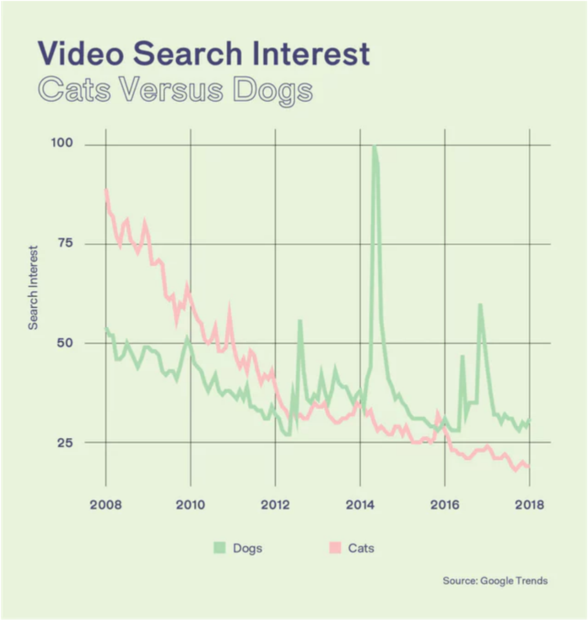
Social media mentions of Cats vs. Dogs

Reddit

Google

Twitter?



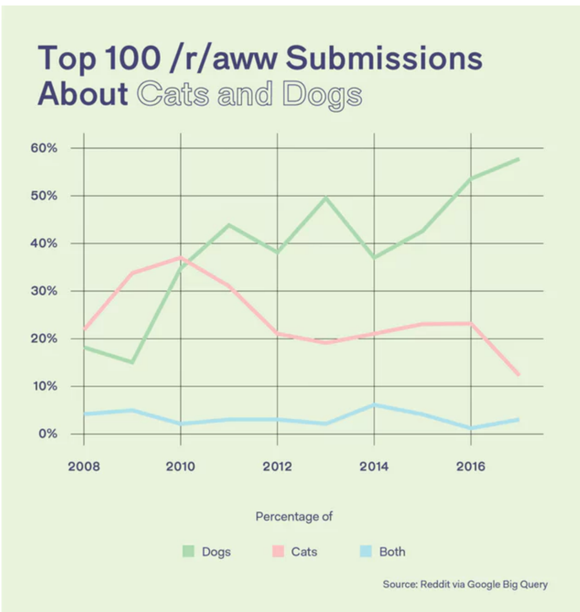


Graphics courtesy of https://theoutline.com/post/3128/dogs-cats-internet-popularity?zd=1

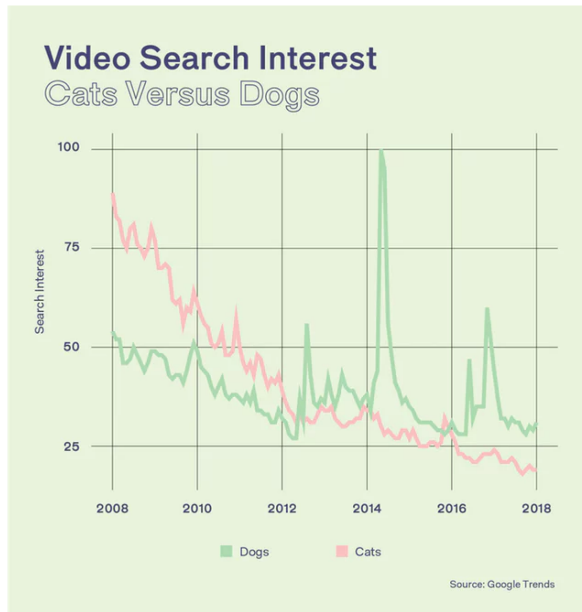
Traditional algorithms

Social media mentions of Cats vs. Dogs

Reddit



Google



Twitter?

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

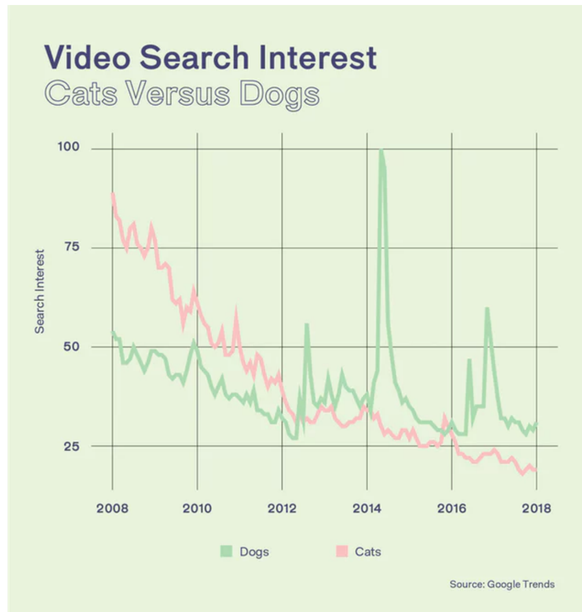
Graphics courtesy of https://theoutline.com/post/3128/dogs-cats-internet-popularity?zd=1

Traditional algorithms

Social media mentions of Cats vs. Dogs

Reddit

Google



Twitter?

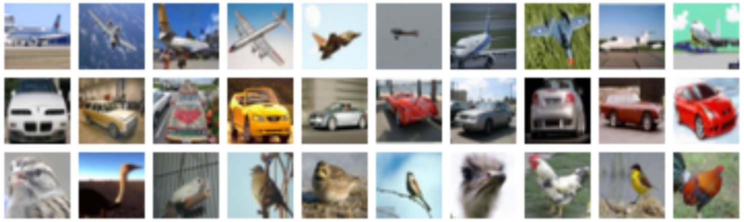
```
cats = []
dogs = []
other = []
for tweet in tweets:
    if "cat" in tweet:
        cats.append(tweet)
    elif "dog" in tweet:
        dogs.append(tweet)
    else:
        other.append(tweet)
return cats, dogs, other
```

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

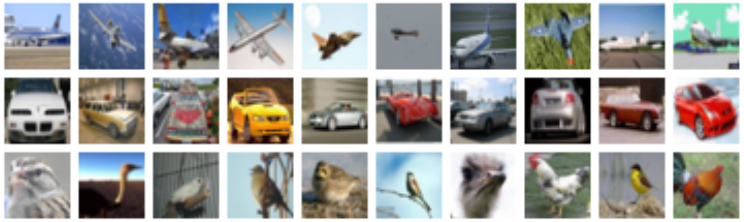
Write a program that sorts images into those containing "**birds**", "**airplanes**", or ***other.***



airplane
other
bird

Machine learning algorithms

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

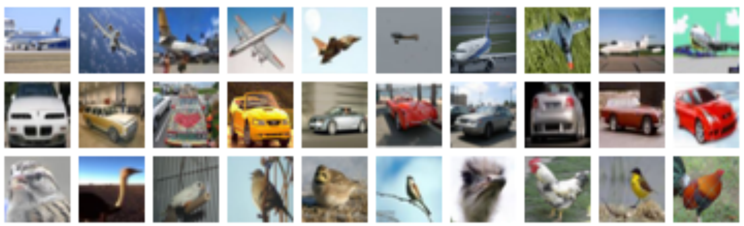


airplane
other
bird

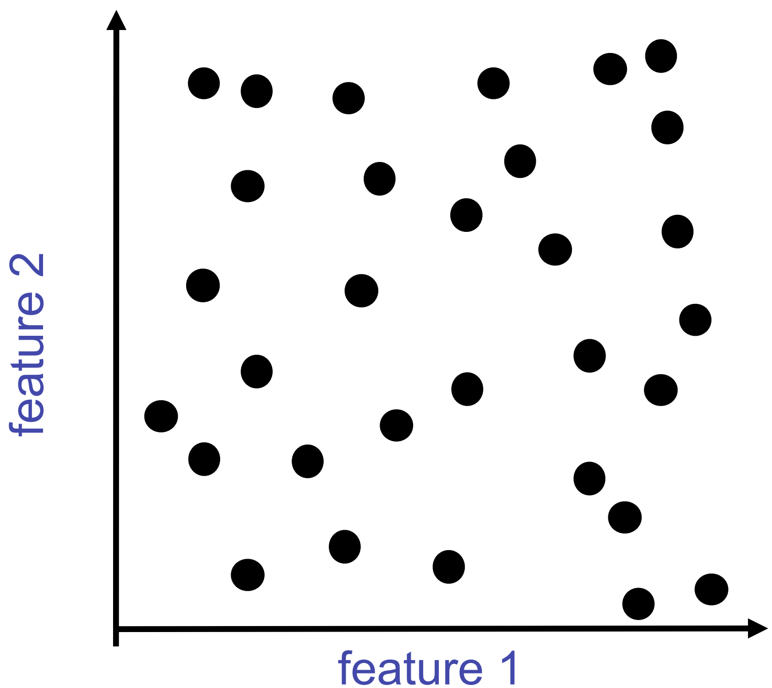
```
birds = []
planes = []
other = []
for image in images:
    if bird in image:
        birds.append(image)
    elif plane in image:
        planes.append(image)
    else:
        other.append(image)
return birds, planes, other
```

Machine learning algorithms

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



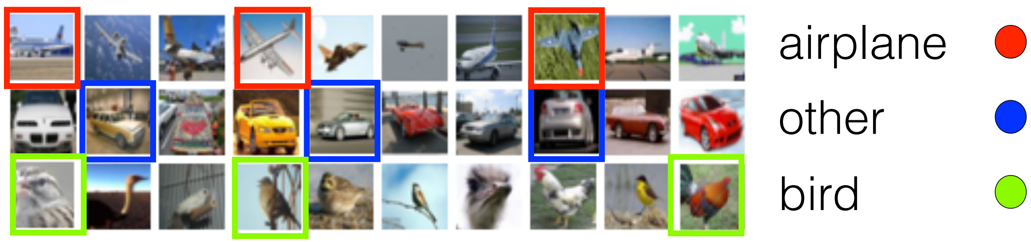
airplane
other
bird

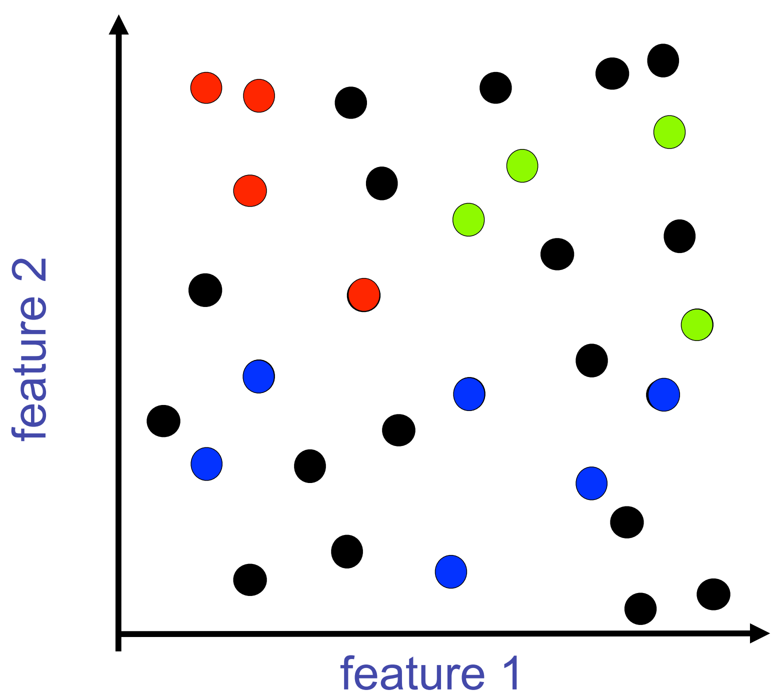



```
birds = []
planes = []
other = []
for image in images:
    if bird in image:
        birds.append(image)
    elif plane in image:
        planes.append(image)
    else:
        other.append(image)
return birds, planes, other
```

Machine learning algorithms

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

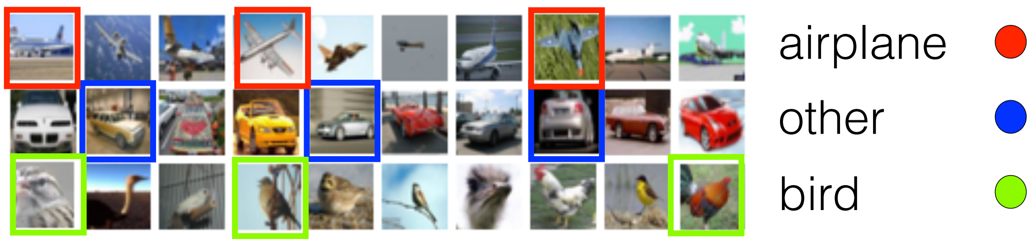


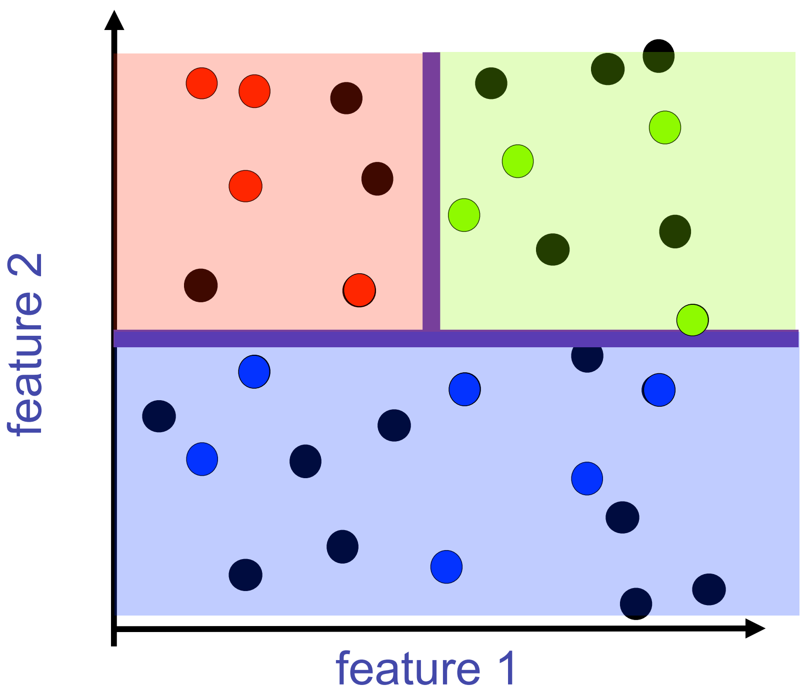



```
birds = []
planes = []
other = []
for image in images:
    if bird in image:
        birds.append(image)
    elif plane in image:
        planes.append(image)
    else:
        other.append(image)
return birds, planes, other
```

Machine learning algorithms

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

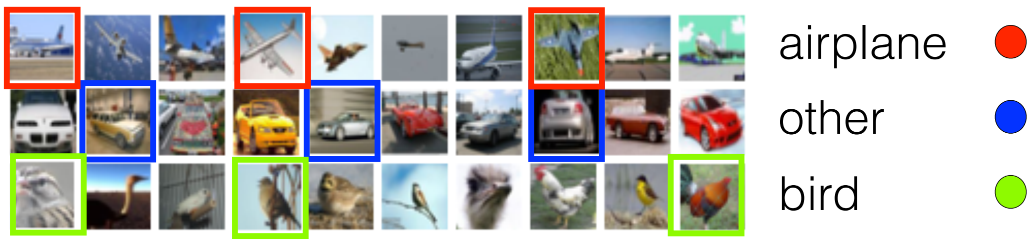


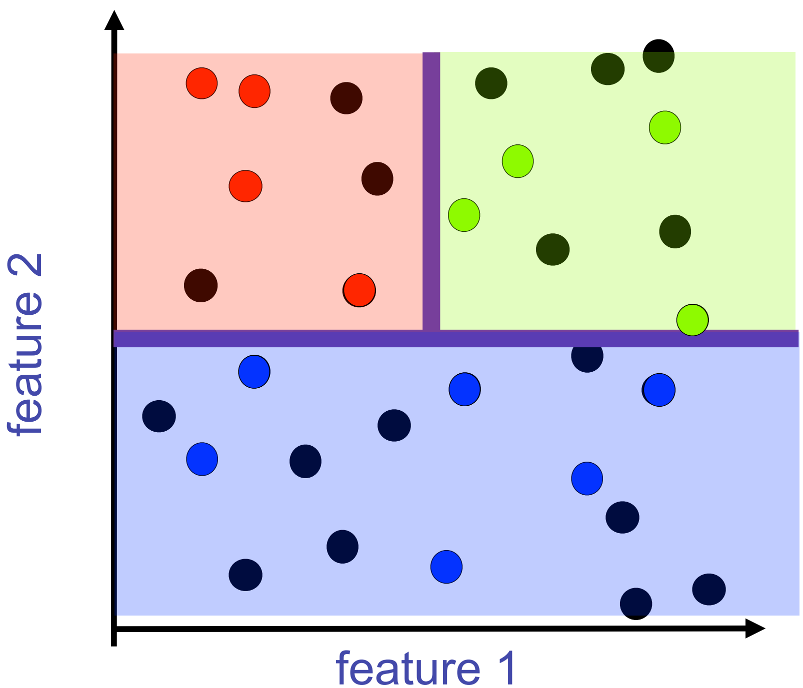



```
birds = []
planes = []
other = []
for image in images:
    if bird in image:
        birds.append(image)
    elif plane in image:
        planes.append(image)
    else:
        other.append(image)
return birds, planes, other
```

Machine learning algorithms

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






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

NETFLIX

98% Match 2017 TV

Spotify

Discover Weekly

FOLLOWING

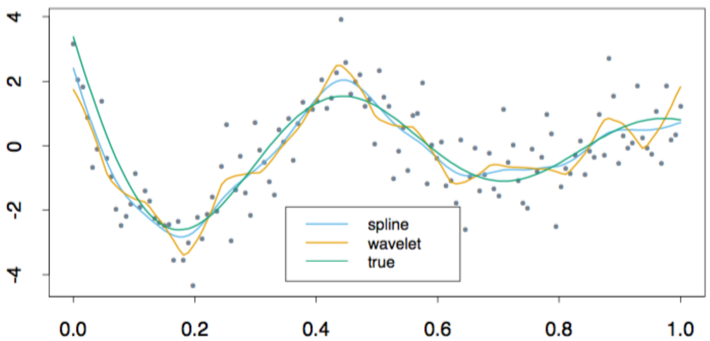
amazon prime

You may also like…

ML uses past data to make personalized predictions

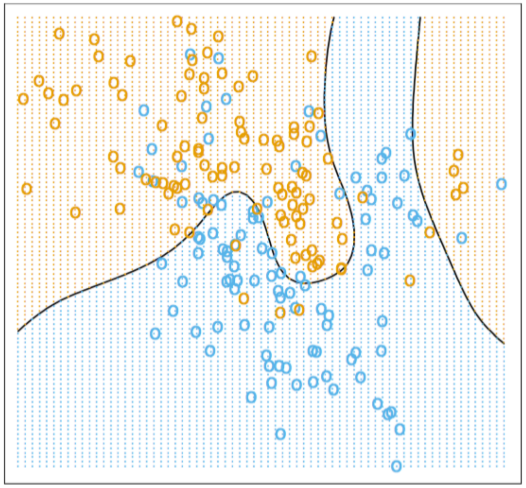
Mix of statistics (conceptual) and algorithms (programming)

Flavors of ML



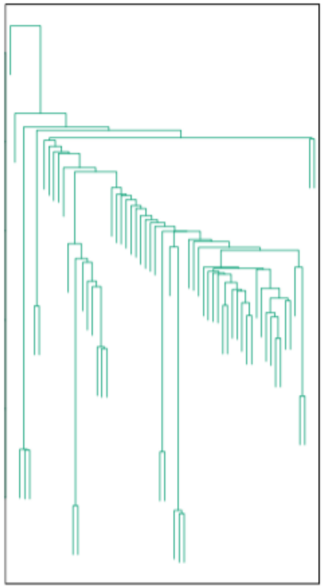
Regression

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

Mix of statistics (conceptual) and algorithms (programming)

CSE 446/546: Machine Learning

Instructors: Matt Golub and Hunter Schafer
Contact: cse446-staff@cs.washington.edu
Website: https://courses.cs.washington.edu/courses/cse446/24wi/

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 learn the foundations of ML to understand the latest results

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

Course Staff - Instructors

Matt Golub (he / him)
Assistant Professor in CSE

Hunter Schafer (he / him)
Associate Teaching Professor in CSE

Course Staff - Teaching Assistants

- Varich Boonsanong, varicb at cs.washington.edu
- Shawn Collinge, shawnc6 at cs.washington.edu
- Anderson Lee, lee0618 at cs.washington.edu
- Hannah Lee, hannahyk at cs.washington.edu
- Mingyu Lu, mingyulu at cs.washington.edu
- Esteban Safranchik, estebans at cs.washington.edu
- Abosh Upadhyaya, mihiru at cs.washington.edu
- Andrew Wagenmaker, ajwagen at cs.washington.edu
- Guang Yang, gyang1 at cs.washington.edu
- Wuwei Zhang, wz86 at cs.washington.edu

Course Registration

- As of today, all enrollment restrictions have been dropped.
- If there are open spaces, any UW student may register for those spaces.
- All CSE course registration processes are managed centrally by CSE. Do **not** email instructors, we cannot help.
- Resources:

- https://www.cs.washington.edu/academics/ugrad/advising/
- https://www.cs.washington.edu/academics/ugrad/courses/petition

Lectures

- Will be broadcast on Zoom.
- Will be recorded and posted shortly after class.
- In-person attendance is encouraged.
- **For exams, in-person attendance is mandatory.**

Prerequisites

- Familiarity with:
  - Linear algebra
    - linear dependence, rank, linear equations
  - Multivariate calculus
  - Probability and statistics
    - Distributions, densities, marginalization, moments
  - Algorithms
    - Basic data structures, complexity
- Use HW0 to judge skills
- **See assigned reading and website for additional review materials!**

CSE 446 vs 546

Course	Lecture	Section	Homework	Grading
446	CSE2 G20 (Amazon Auditorium) MW 9:00 -- 10:20am	Attend the section you are registered.	**A** problems only. No credit will be rewarded for completing **B** problems.	You will be graded (e.g., curved) against your peers in 446 only (on a 4.0 scale). For example, if you you recieved a (curved) score of 0.9 on the **A** problems, then your full grade on your transcript will be (4.0)*(0.9) = 3.6. Any attempt of the **B** problems will not influence your grade in any way.
546	CSE2 G20 (Amazon Auditorium) MW 9:00 -- 10:20am	None	**A** and **B** problems.	You will be graded (e.g., curved) against your peers in 546 only. Your grade on the **A** and **B** problems will be curved seperately, and then summed. For example, if you you recieved a (curved) score of 0.9 on the **A** problems, and a (curved) score of 0.8 on the **B** problems, then your full grade on your transcript will be (3.8)*(0.9) + (0.2)*(0.8) = 3.58, rounded to 3.6. If only the **A** problems on the homework are attempted, the highest score attainable is a 3.8. If only the **B** problems are attempted, the highest score attainable is a 0.2.

Grading

- 5 homeworks (60%)
  - Each contains both theoretical questions and will have programming.
  - Collaboration okay. You must write, submit, and understand your answers and code (which we may run).
  - Do not Google for answers or ask chatGPT to do it.
  - **READ COLLABORATION POLICY ON WEBSITE**
- Midterm (20%) and Final (20%)
- Separate grading curves for 446 and 546.

Homeworks

- **HW 0** is out (**Due next Wednesday 1/10 @ 11:59pm**)
  - Should be review (but being rusty is expected)
  - Work individually, treat as guide for what to brush up on
- **HW 1,2,3,4**. They are not easy or short. Start early.
- **Submit** to Gradescope.
- **Regrade requests** on Gradescope.
- **Late days:** 5 days total over the quarter; no more than 2 per assignment.
- **Assignments due at 11:59pm**, submit early and often (do not email us at 12:05).

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

- **Announcements, questions about class, homework help**
  - EdStem (https://edstem.org/)
  - "I think there is a typo in the homework?"
  - "What does this notation mean?"
  - "Is this an accurate description of how this works?
- **Personal concerns (**cse446-staff@cs.washington.edu)
  - "Was in hospital…", "Laptop was stolen…"
- **Office hours**
  - "How do I get started on problem 2?"
  - "Am I on the right track?"
  - "I have this problem at work—can you point me in the right direction?"
- **Regrade requests**
  - Directly submit on Gradescope
- **Anonymous feedback** (https://feedback.cs.washington.edu/)
  - "Your real-world example X lacked nuance. I would like you to…"

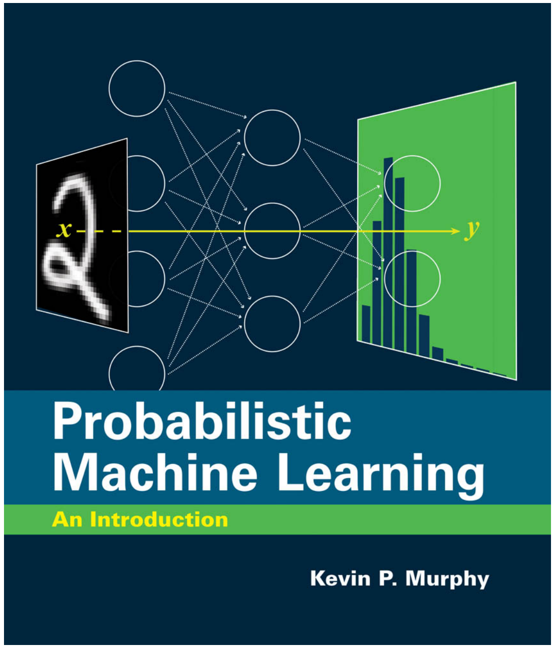
Textbooks

- **Free PDF** Textbook I will assign most reading from:

Probabilistic Machine Learning: An Introduction

Kevin Murphy

  - PDF linked to on website, also in print



- So many more resources on the website!
- I may occasionally point you to other (free) readings

Enjoy!

- ML is becoming ubiquitous in science, engineering and beyond
- This class should give you a basic foundation for understanding and applying ML

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."

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**:

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

$= P(3) + P(4)$ ~~$- P(X=3 \wedge X=4)$~~

$= \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$

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

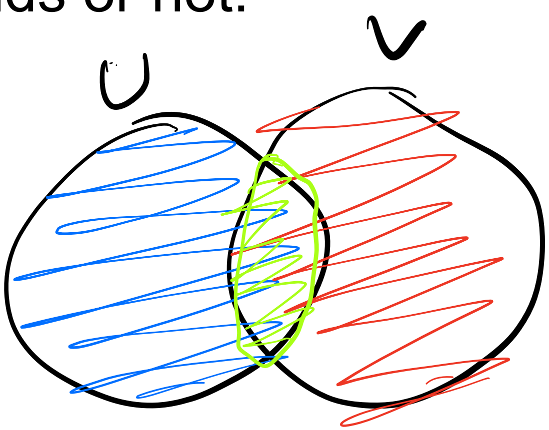
(X and Y take values in {1,2,3,4,5,6} each with equal probability)

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?

$$P(A \cap B) = P(X \in \{3,4\} \cap X = 1) = 0$$

$$\neq P(A)P(B)$$

$$= \frac{1}{3} \cdot \frac{1}{6} = \frac{1}{18}$$

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 (no)? B,C (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 $P(X \leq 4 \mid X \geq 3)$? $= \dfrac{P(X \in \{3,4\})}{P(X \in \{3,4,5,6\})} = \dfrac{1/3}{2/3} = \dfrac{1}{2}$

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

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

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

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

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

The expected squared deviation of X from its mean.

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

$$= E(X^2) - E(X)^2$$

Median M

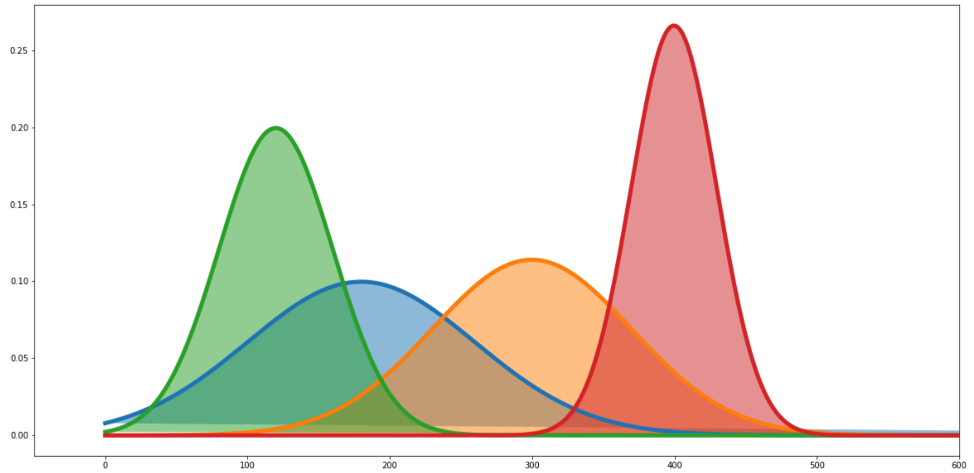
The value of X that is separating the higher half of it's range from the lower half.

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

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

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?"

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

Maximum Likelihood Estimation

Your first consulting job

- *Billionaire*: I have special coin, if I flip it, what's the probability it will be heads?
- *You*: Please flip it a few times:

H H T H T

- *You*: The probability is: 3/5

- *Billionaire:* Why?

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

- $P(\mathcal{D}|\theta) = P(H, H, T, H, T \mid \theta)$

$$= P(H|\theta)\,P(H, T, H, T \mid \theta)$$

$$= P(H|\theta)^k\, P(T|\theta)^{n-k}$$

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

Maximum Likelihood Estimation

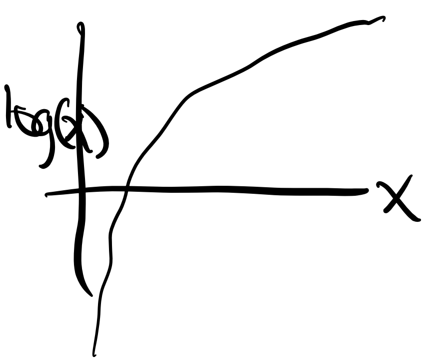
- **Data**: sequence *D= (HHTHT…),* **k heads** out of **n flips**
- **Hypothesis:** P(Heads) = θ, P(Tails) = 1-θ

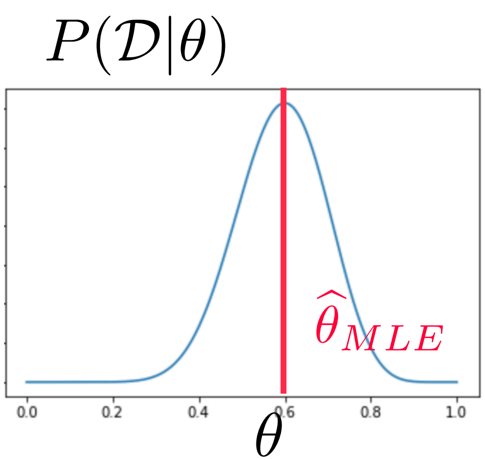
$$P(\mathcal{D}|\theta) = \theta^k(1-\theta)^{n-k}$$

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

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

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





Your first learning algorithm

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

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

- Set derivative to zero:

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

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

$$\frac{d}{d\theta}(\cdots) = \frac{k}{\theta} + \frac{(n-k)}{1-\theta}(-1) = 0$$

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

$$k - \cancel{k\theta} = n\theta - \cancel{k\theta}$$

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

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

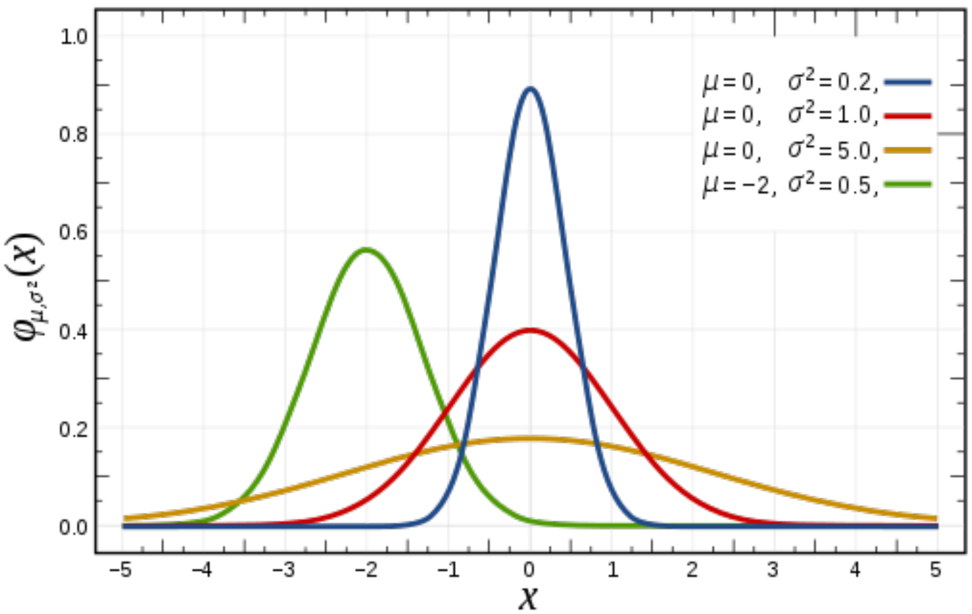
Recap

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

What about continuous variables?

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

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



Some properties of Gaussians

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

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

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

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

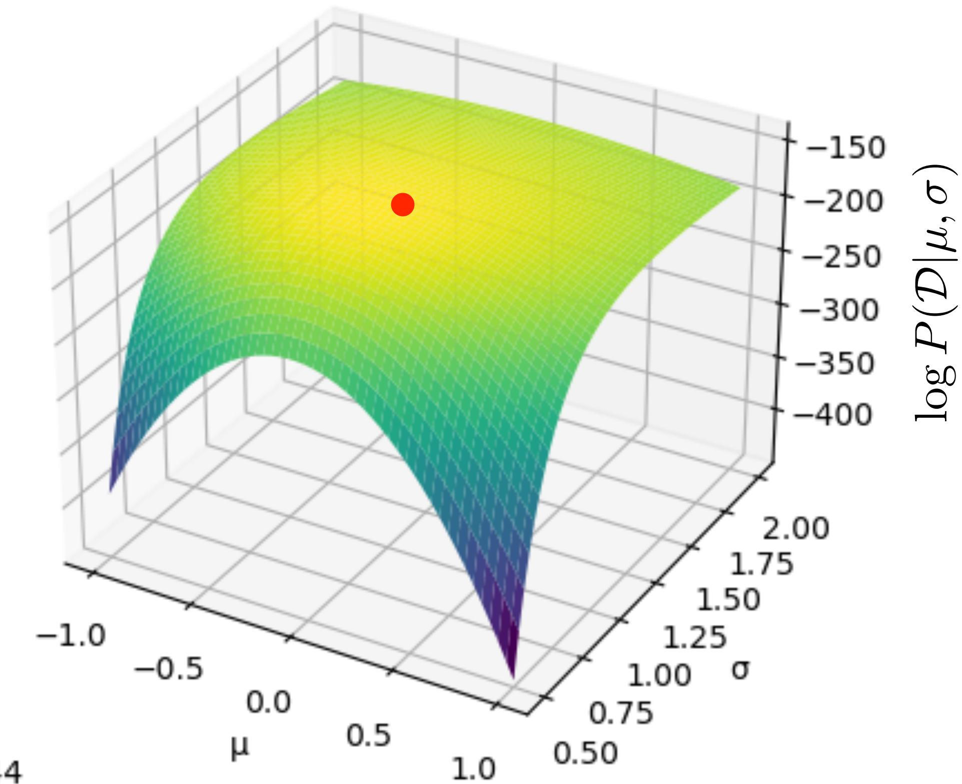
$$n = 100$$

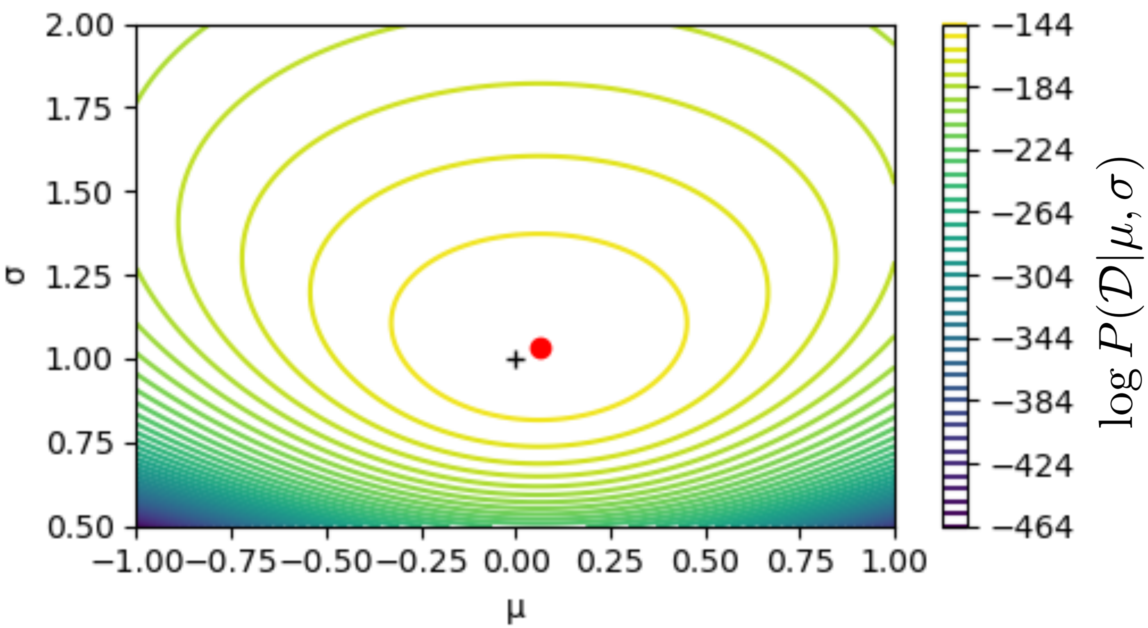
$$x_i \sim \mathcal{N}(\mu, \sigma^2)$$

$$\mu = 0$$

$$\sigma^2 = 1$$

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





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

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

Your second learning algorithm: MLE for mean of a Gaussian

- What's MLE for mean?

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

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

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

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

MLE for variance

$$\frac{d}{d\sigma}\sigma^{-2} = -2\sigma^{-3}$$

- Again, set derivative to zero:

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

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

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

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

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

Learning Gaussian parameters

- MLE:

$E(\hat{\mu}_{MLE}) = \mu$

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

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

- 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) $\widehat{\theta}_{MLE} = \arg\max_{\theta} L_n(\theta)$

Under benign assumptions, as the number of observations $n \to \infty$ we have $\widehat{\theta}_{MLE} \to \theta_*$

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