

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:

- *You*: The probability is:

- *Billionaire*: Why?

Coin – Binomial Distribution

- **Data:** sequence $D = (HHTHT\dots)$, **k heads** out of **n flips**
- **Hypothesis:** $P(\text{Heads}) = \theta$, $P(\text{Tails}) = 1 - \theta$
 - Flips are i.i.d.:
 - Independent events
 - Identically distributed according to Binomial distribution

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

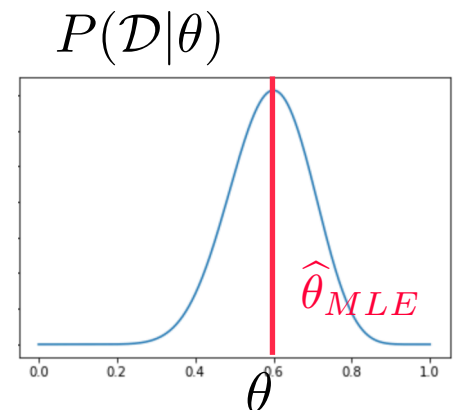
Maximum Likelihood Estimation

- **Data:** sequence $D = (HHTHT\dots)$, **k heads** out of **n flips**
- **Hypothesis:** $P(\text{Heads}) = \theta$, $P(\text{Tails}) = 1 - \theta$

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

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

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



Your first learning algorithm

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

- Set derivative to zero:

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

Maximum Likelihood Estimation

Observe X_1, X_2, \dots, 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) $\hat{\theta}_{MLE} = \arg \max_{\theta} L_n(\theta)$

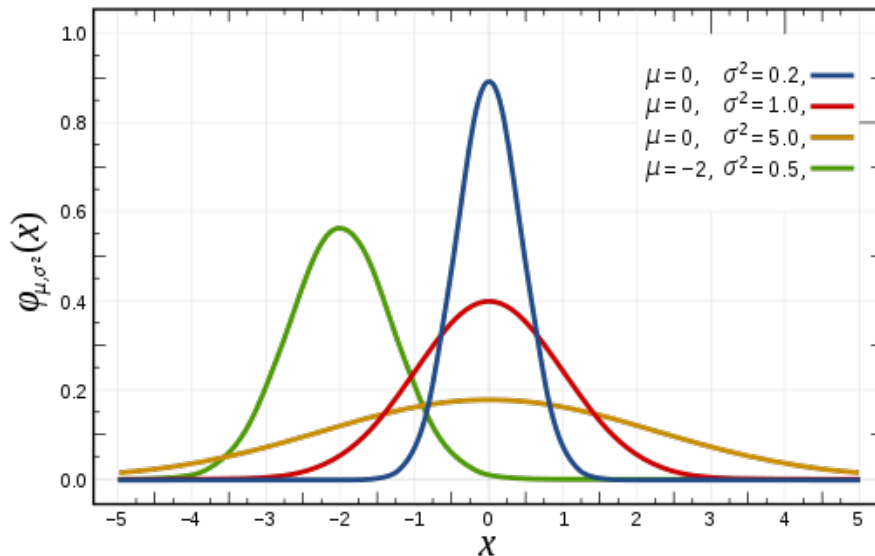
Recap

- Learning is...
 - Collect some data
 - E.g., coin flips
 - Choose a hypothesis class or model
 - E.g., binomial
 - 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)
 - $X \sim N(\mu, \sigma^2)$
 - $Y = aX + b \rightarrow Y \sim N(a\mu + b, a^2\sigma^2)$
- Sum of Gaussians
 - $X \sim N(\mu_X, \sigma_X^2)$
 - $Y \sim N(\mu_Y, \sigma_Y^2)$
 - $Z = X + Y \rightarrow Z \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$

MLE for Gaussian

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

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

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

Your second learning algorithm: MLE for mean of a Gaussian

- What's MLE for mean?

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

MLE for variance

- Again, set derivative to zero:

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

Learning Gaussian parameters

- MLE:

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

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

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

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

- Unbiased variance estimator:

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

Maximum Likelihood Estimation

Observe X_1, X_2, \dots, 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) $\hat{\theta}_{MLE} = \arg \max_{\theta} L_n(\theta)$

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

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

Applications preview

Maximum Likelihood Estimation

Observe X_1, X_2, \dots, 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) $\hat{\theta}_{MLE} = \arg \max_{\theta} L_n(\theta)$

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

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 $X \sim f(\cdot; \hat{\theta}_{MLE})$

Estimation

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

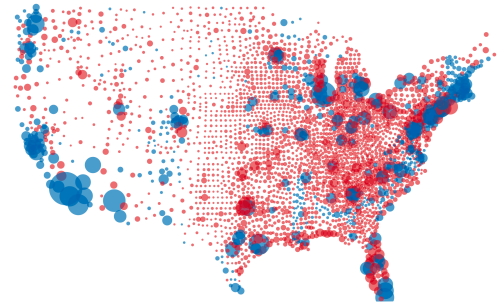
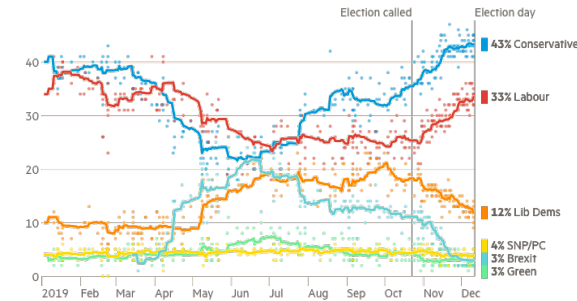
Opinion polls

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

- θ_* is “true” average opinion
- X_1, X_2, \dots are sample calls

UK poll tracker

Lines represent weighted averages, points represent polls (%)



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, \dots are binary “clicks”

Save on prescription drugs - over \$3,637* a year!

Last year, Humana's Medicare Advantage plan members saved, on average, \$3,637* on prescription drugs! Choose your Humana Medicare Advantage plan and you could enjoy savings on prescription drugs, plus:

- Hospital, doctor AND drug coverage combined into one easy-to-use plan
- Extra benefits not offered by Original Medicare
- Affordable or no monthly plan premiums

Shop 2014 Medicare Plans

Control

Explore Humana's Medicare plans

Let us help you determine the Humana plan that's best for your needs.

Get started now

Treatment

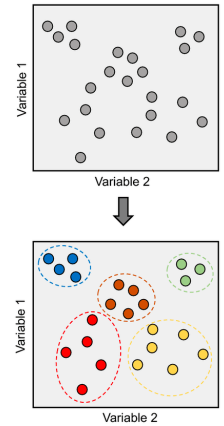
Interpret

Observe X_1, X_2, \dots, 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?

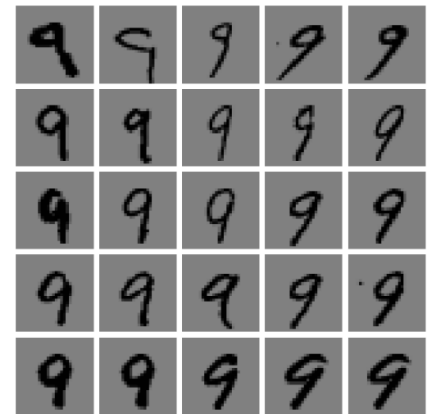
- θ_* describes “center” of distinct groups
- X_1, X_2, \dots are individual customers



Data exploration

What are the degrees of freedom of the dataset?

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



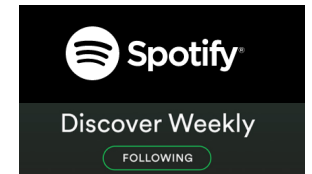
Predict

Observe X_1, X_2, \dots, 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, \dots 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, \dots 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 Krumb 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

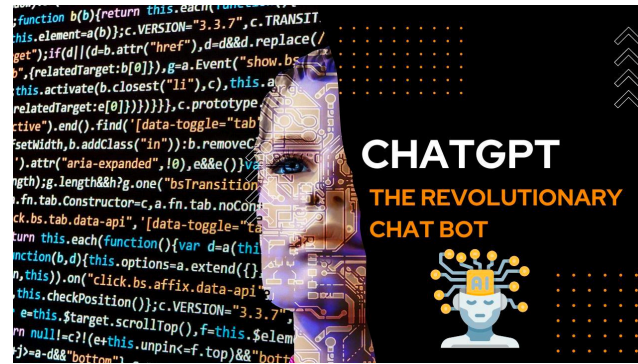
Observe X_1, X_2, \dots, X_n drawn IID from $f(x; \theta)$ for some “true” $\theta = \theta_*$

Text generation

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

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

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



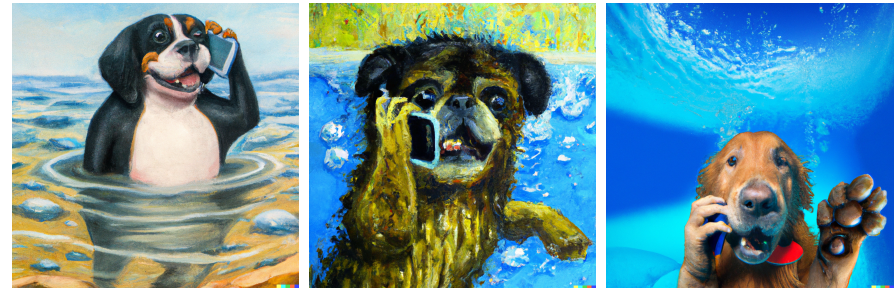
<https://chat.openai.com/chat>

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, \dots are the (image, caption) pairs found online

“dog talking on cell phone under water, oil painting”



<https://labs.openai.com/>

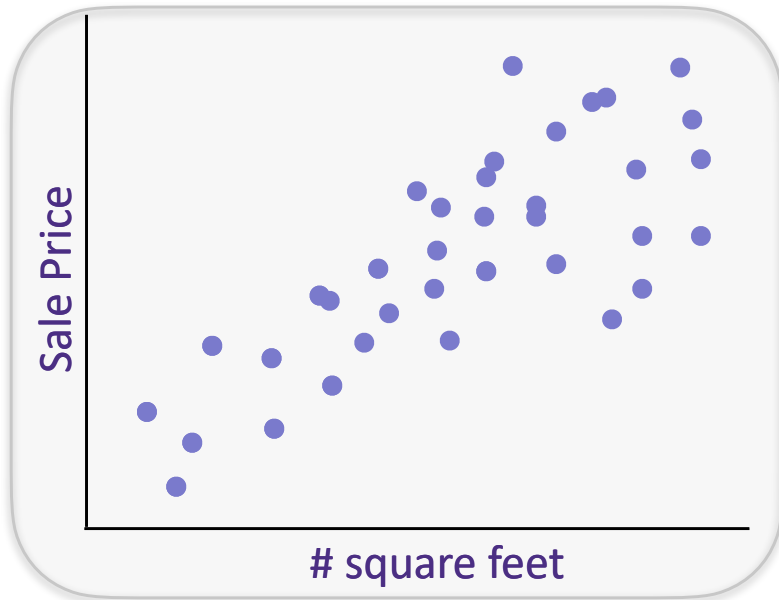
Linear Regression

The regression problem, 1-dimensional

Given past sales data on [zillow.com](https://www.zillow.com), predict:

y = House sale price *from*

x = {# sq. ft.}



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

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

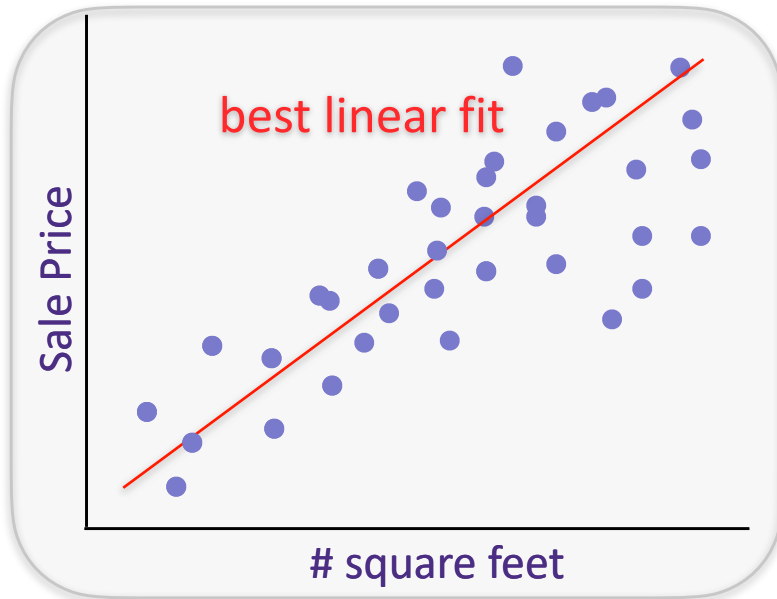
$$y_i \in \mathbb{R}$$

Fit a function to our data, 1-d

Given past sales data on [zillow.com](https://www.zillow.com), predict:

$y =$ House sale price *from*

$x =$ {# sq. ft.}



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

Hypothesis/Model: linear

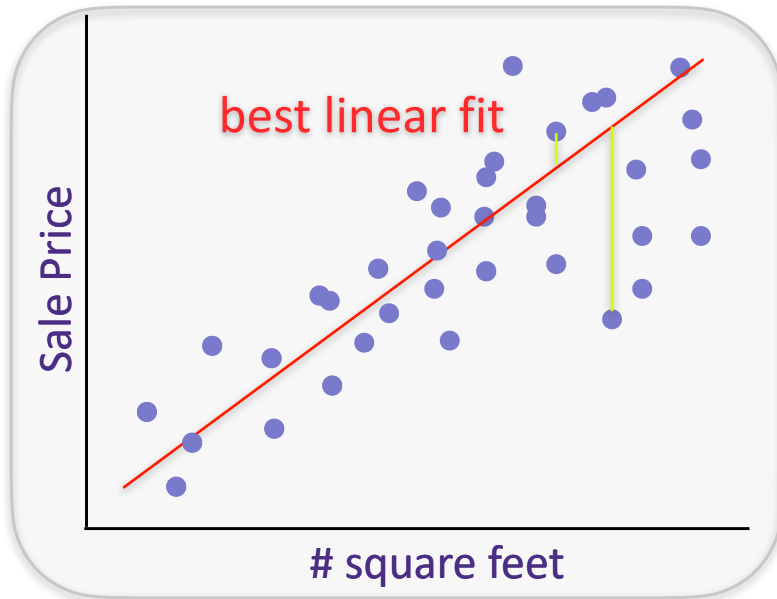
$$y_i = x_i w + \epsilon_i \quad \epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

Fit a function to our data, 1-d

Given past sales data on [zillow.com](https://www.zillow.com), predict:

$y =$ House sale price *from*

$x =$ {# sq. ft.}



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

Hypothesis/Model: linear

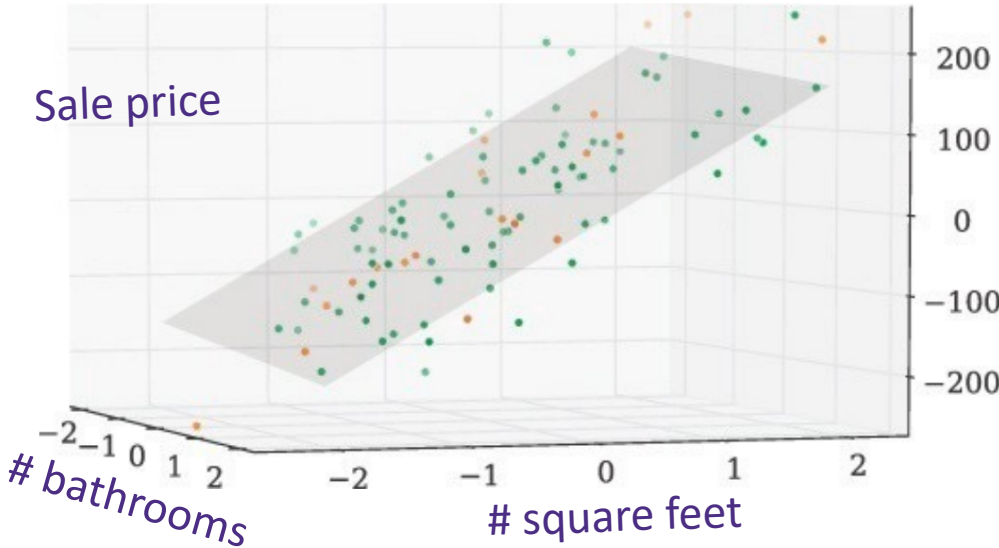
$$y_i = x_i w + \epsilon_i \quad \epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

The regression problem, d-dim

Given past sales data on [zillow.com](https://www.zillow.com), predict:

$y =$ House sale price *from*

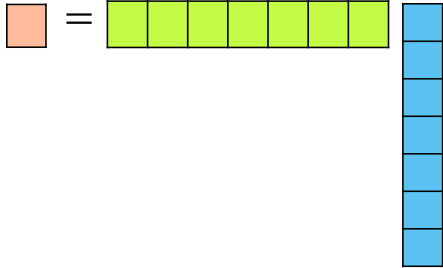
$x =$ {# sq. ft., zip code, date of sale, etc.}



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

Hypothesis/Model: linear

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

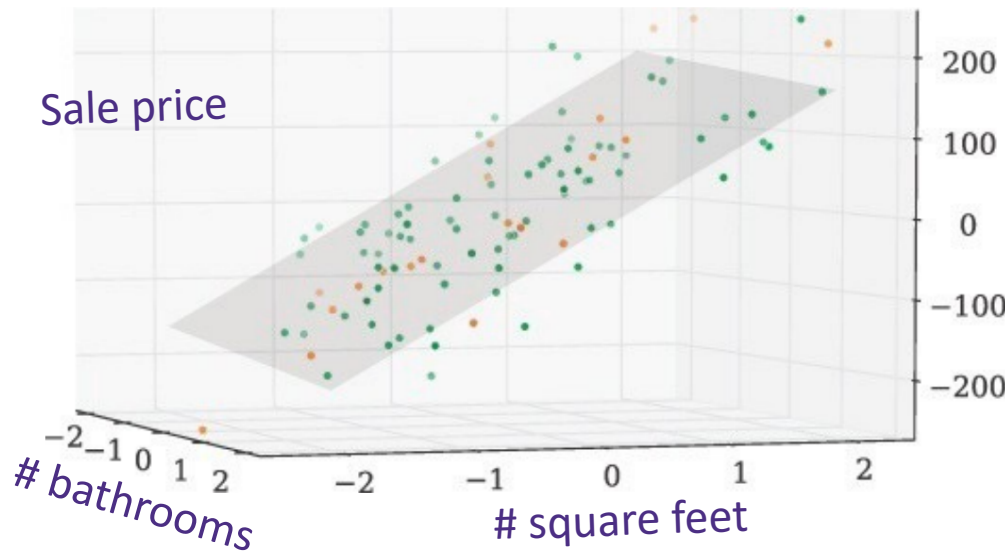


The regression problem, d-dim

Given past sales data on [zillow.com](https://www.zillow.com), predict:

$y =$ House sale price *from*

$x =$ {# sq. ft., zip code, date of sale, etc.}



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

Hypothesis/Model: linear

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

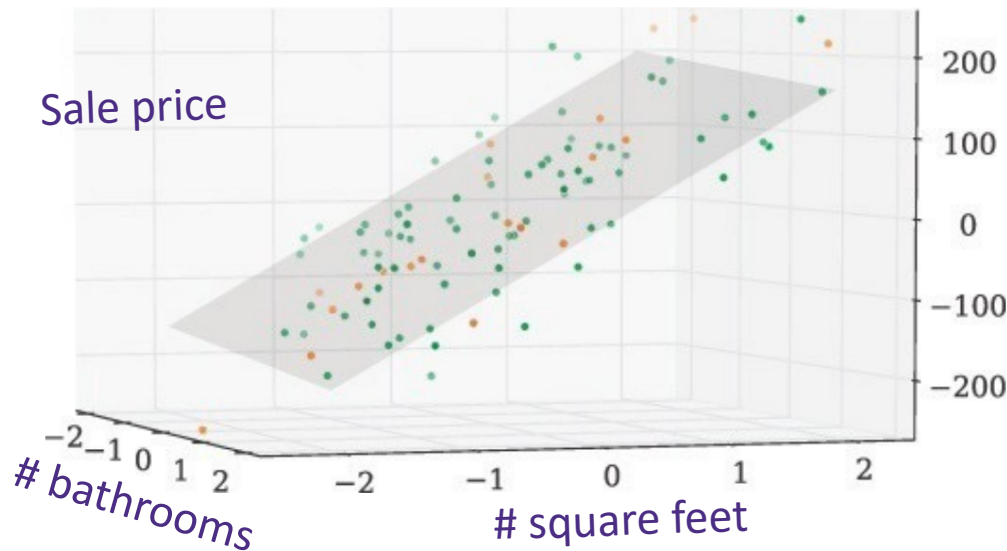
$$p(y|x, w, \sigma) =$$

The regression problem, d-dim

Given past sales data on [zillow.com](https://www.zillow.com), predict:

$y =$ House sale price *from*

$x =$ {# sq. ft., zip code, date of sale, etc.}



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

Hypothesis/Model: linear

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

$$p(y|x, w, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-x^T w)^2/2\sigma^2}$$

Maximizing log-likelihood

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

$$p(y|x, w, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-x^\top w)^2/2\sigma^2}$$

Likelihood: $P(\mathcal{D}|w, \sigma) = \prod_{i=1}^n p(y_i|x_i, w, \sigma) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i-x_i^\top w)^2/2\sigma^2}$

Maximum Likelihood Estimation

Observe X_1, X_2, \dots, 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) $\hat{\theta}_{MLE} = \arg \max_{\theta} L_n(\theta)$

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

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 $X \sim f(\cdot; \hat{\theta}_{MLE})$

Maximizing log-likelihood

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

$$p(y|x, w, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-x^\top w)^2/2\sigma^2}$$

Likelihood: $P(\mathcal{D}|w, \sigma) = \prod_{i=1}^n p(y_i|x_i, w, \sigma) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i-x_i^\top w)^2/2\sigma^2}$

Maximize (wrt w): $\log P(\mathcal{D}|w, \sigma) = \log \left(\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i-x_i^\top w)^2/2\sigma^2} \right)$

Maximizing log-likelihood

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

$$p(y|x, w, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-x^\top w)^2/2\sigma^2}$$

Likelihood: $P(\mathcal{D}|w, \sigma) = \prod_{i=1}^n p(y_i|x_i, w, \sigma) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i-x_i^\top w)^2/2\sigma^2}$

Maximize (wrt w): $\log P(\mathcal{D}|w, \sigma) = \log \left(\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i-x_i^\top w)^2/2\sigma^2} \right)$

$$\hat{w}_{MLE} = \arg \min_w \sum_{i=1}^n (y_i - x_i^\top w)^2$$

Maximizing log-likelihood

$$\hat{w}_{MLE} = \arg \min_w \sum_{i=1}^n (y_i - x_i^\top w)^2$$

Set derivate=0, solve for w

Maximizing log-likelihood

$$\hat{w}_{MLE} = \arg \min_w \sum_{i=1}^n (y_i - x_i^\top w)^2$$

Set derivate=0, solve for w

$$\hat{w}_{MLE} = \left(\sum_{i=1}^n x_i x_i^\top \right)^{-1} \sum_{i=1}^n x_i y_i$$

The regression problem in matrix notation

$$\hat{w}_{MLE} = \arg \min_w \sum_{i=1}^n (y_i - x_i^\top w)^2$$

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

d : # of features

n : # of examples/datapoints

The regression problem in matrix notation

$$\hat{w}_{MLE} = \arg \min_w \sum_{i=1}^n (y_i - x_i^\top w)^2$$

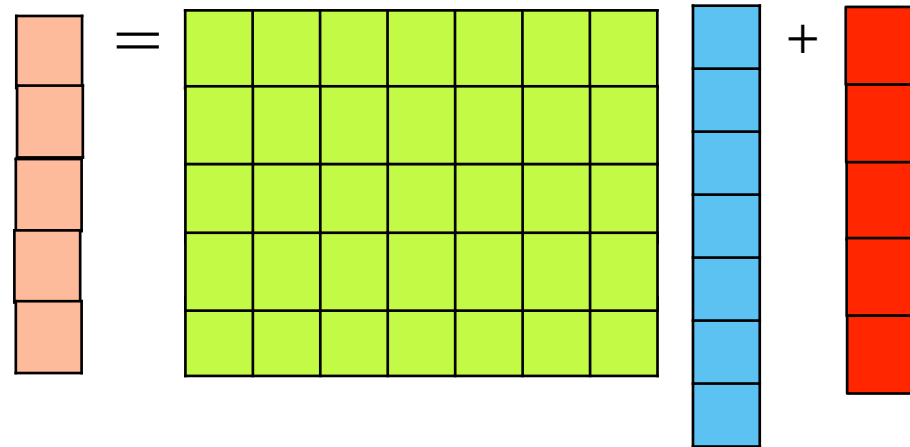
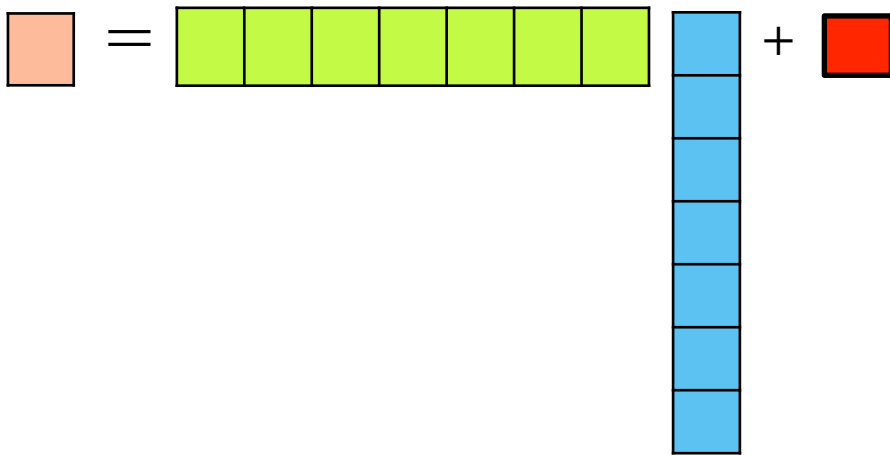
$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{bmatrix}$$

d : # of features

n : # of examples/datapoints

$$y_i = x_i^\top w + \epsilon_i$$

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



The regression problem in matrix notation

$$\hat{w}_{MLE} = \arg \min_w \sum_{i=1}^n (y_i - x_i^\top w)^2$$

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

d : # of features

n : # of examples/datapoints

$$y_i = x_i^\top w + \epsilon_i$$

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

$$\begin{aligned} \hat{w}_{LS} &= \arg \min_w \|\mathbf{y} - \mathbf{X}w\|_2^2 \\ &= \arg \min_w (\mathbf{y} - \mathbf{X}w)^\top (\mathbf{y} - \mathbf{X}w) \end{aligned}$$

$$\ell_2 \text{ norm: } \|z\|_2 = \sqrt{\sum_{i=1}^n z_i^2} = \sqrt{z^\top z}$$

The regression problem in matrix notation

$$\hat{w}_{MLE} = \arg \min_w \sum_{i=1}^n (y_i - x_i^\top w)^2$$

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

d : # of features

n : # of examples/datapoints

$$y_i = x_i^T w + \epsilon_i$$

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

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

The regression problem in matrix notation

$$\hat{w}_{MLE} = \arg \min_w \sum_{i=1}^n (y_i - x_i^\top w)^2$$

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

d : # of features

n : # of examples/datapoints

$$y_i = x_i^T w + \epsilon_i$$

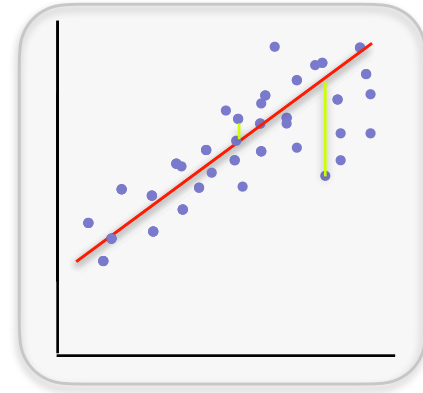
$$\mathbf{y} = \mathbf{X}w + \epsilon$$

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

$$\hat{w}_{LS} = \hat{w}_{MLE} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

The regression problem in matrix notation

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



What about an offset?

$$\begin{aligned}\hat{w}_{LS}, \hat{b}_{LS} &= \arg \min_{w,b} \sum_{i=1}^n (y_i - (x_i^T w + b))^2 \\ &= \arg \min_{w,b} \|\mathbf{y} - (\mathbf{X}w + \mathbf{1}b)\|_2^2\end{aligned}$$

Dealing with an offset

$$\hat{w}_{LS}, \hat{b}_{LS} = \arg \min_{w, b} \|\mathbf{y} - (\mathbf{X}w + \mathbf{1}b)\|_2^2$$

Dealing with an offset

$$\hat{w}_{LS}, \hat{b}_{LS} = \arg \min_{w, b} \|\mathbf{y} - (\mathbf{X}w + \mathbf{1}b)\|_2^2$$

$$\mathbf{X}^T \mathbf{X} \hat{w}_{LS} + \hat{b}_{LS} \mathbf{X}^T \mathbf{1} = \mathbf{X}^T \mathbf{y}$$

$$\mathbf{1}^T \mathbf{X} \hat{w}_{LS} + \hat{b}_{LS} \mathbf{1}^T \mathbf{1} = \mathbf{1}^T \mathbf{y}$$

If $\mathbf{X}^T \mathbf{1} = 0$ (i.e., if each feature is mean-zero) then

$$\hat{w}_{LS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

$$\hat{b}_{LS} = \frac{1}{n} \sum_{i=1}^n y_i$$

Make Predictions

$$\hat{\mathbf{w}}_{LS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

$$\hat{b}_{LS} = \frac{1}{n} \sum_{i=1}^n y_i$$

A new house is about to be listed. What should it sell for?

$$\hat{y}_{\text{new}} = x_{\text{new}}^T \hat{\mathbf{w}}_{LS} + \hat{b}_{LS}$$

Process

Decide on a **model** for the likelihood function $f(x; \theta)$

Find the function which fits the data best

Choose a loss function- least squares

Pick the function which minimizes loss on data

Use function to make prediction on new examples