Maximum Likelihood Estimation

H H T T H

θ if H
$1-\theta$ if T

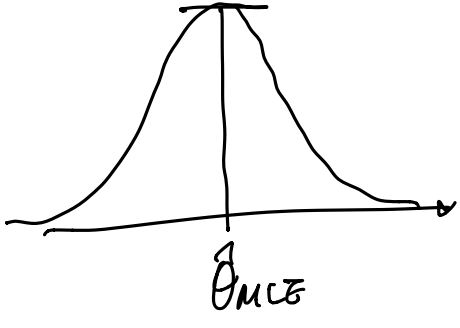
Observe $X_1, X_2, \ldots, X_n$ drawn IID from $f(x;\theta)$ for some "true" $\theta = \theta_*$

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

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

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

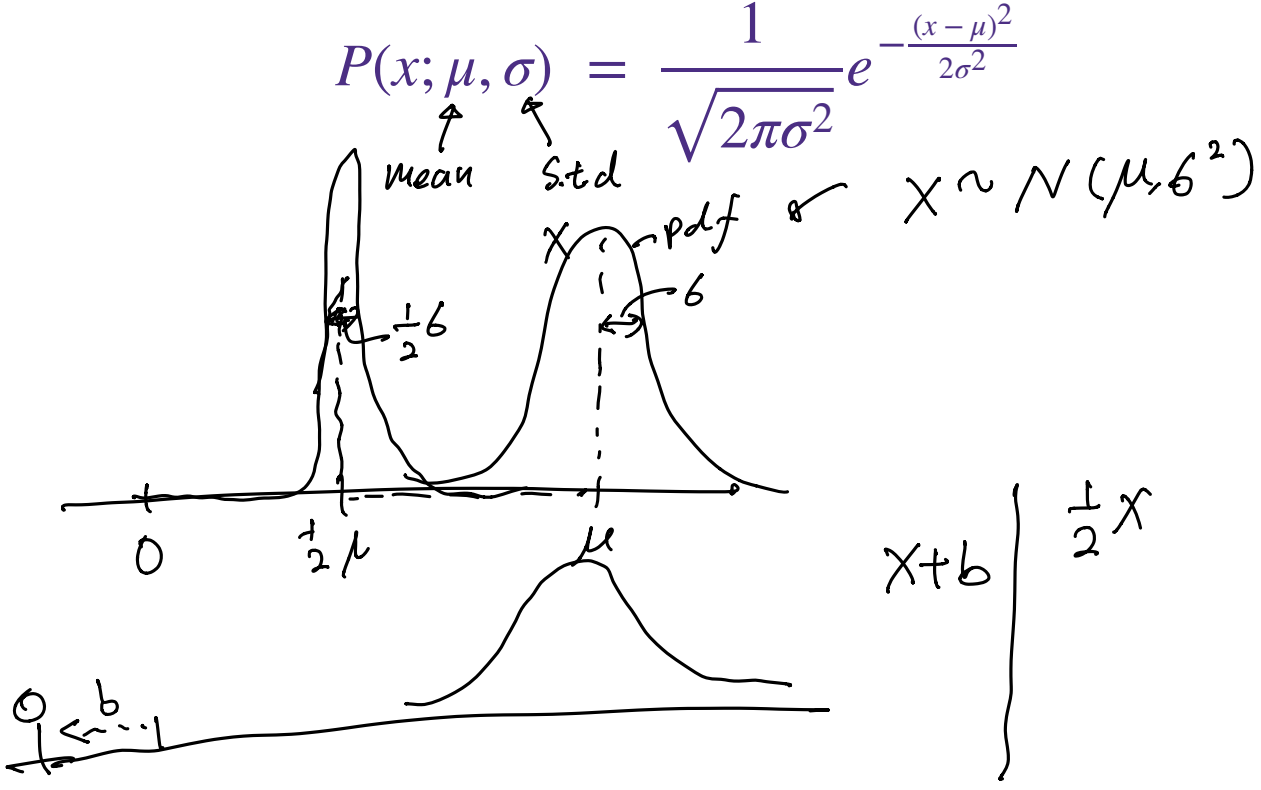
$= \frac{K}{n}$



What about continuous variables?

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

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



Some properties of Gaussians

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

- 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):

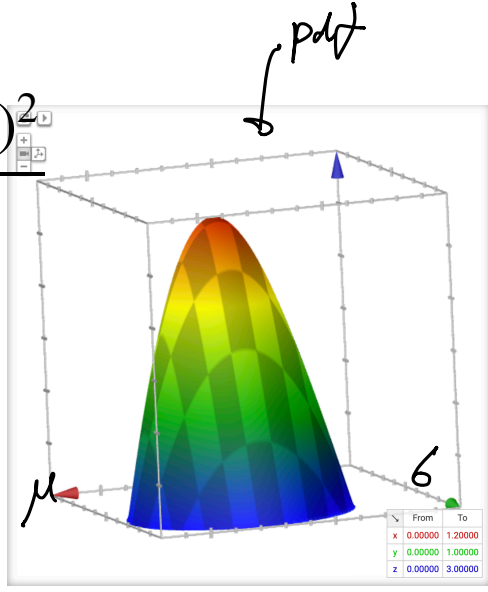
$$P(\mathcal{D};\mu,\sigma) = P(x_1,\ldots,x_n;\mu,\sigma) = P(x_1|\mu,\sigma)\times\cdots\times P(x_n|\mu,\sigma)$$

$$= \prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = f(x_1|\mu,\sigma)$$

- 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 $\theta=(\mu,\sigma^2)$?



Your second learning algorithm: MLE for mean of a Gaussian

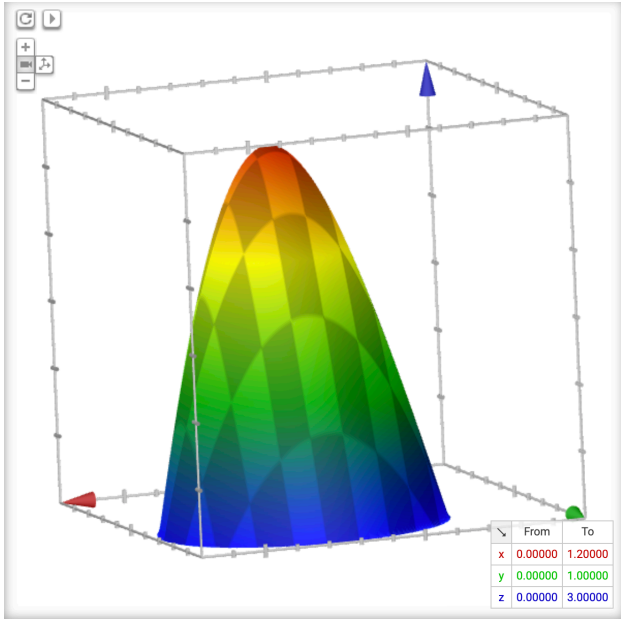
- What's MLE for mean?

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

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

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}x_i - n\cdot\mu\right)\Big|_{\mu=\hat{\mu}} = 0$$

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



MLE for variance

- Again, set derivative to zero:

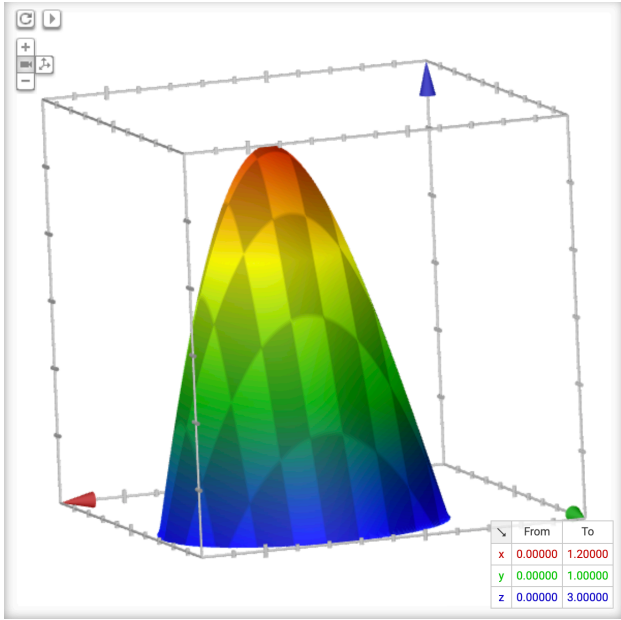
$$\frac{d}{d\sigma}\log P(\mathcal{D};\mu,\sigma) = \frac{d}{d\sigma}\left[-n\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

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

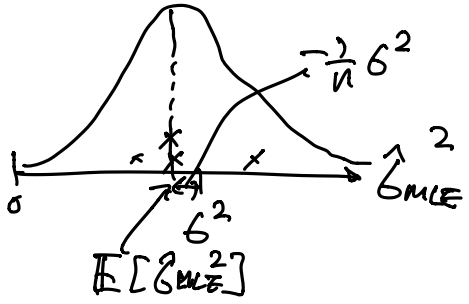
$$= \frac{1}{\sigma^3}\left(-n\sigma^2 + \sum_{i=1}^{n}(x_i-\mu)^2\right)\Bigg|_{\substack{\mu=\hat{\mu}\\ \sigma=\hat{\sigma}}} = 0$$

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

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



What can we say about the MLE?



- MLE:

	mean μ	variance σ^2
$\mathbb{E}$	✓	✓
VAR	?	?

$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad , \quad \mathbb{E}_{x_i \sim N(\mu,\sigma)}[\hat{\mu}_{MLE}] = \mu$$

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

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

$$\frac{n-1}{n}\cdot\sigma^2 = \mathbb{E}[\widehat{\sigma^2}_{MLE}] \neq \sigma^2$$

$$\left(1-\frac{1}{n}\right)\sigma^2 = \mathbb{E}[\hat{\sigma}^2_{MLE}]$$

$$\text{bias} = -\frac{1}{n}\sigma^2$$

- Unbiased variance estimator:

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

Maximum Likelihood Estimation

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

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)

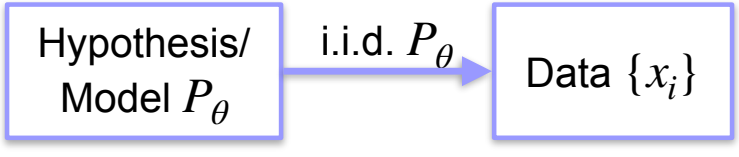
Recap

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

Data $\{x_i\}$

Recap

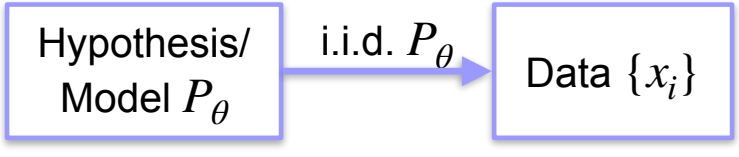
- Learning is…
  - Collect some data
    - E.g., coin flips
  - Choose a hypothesis class or model
    - E.g., binomial



Recap

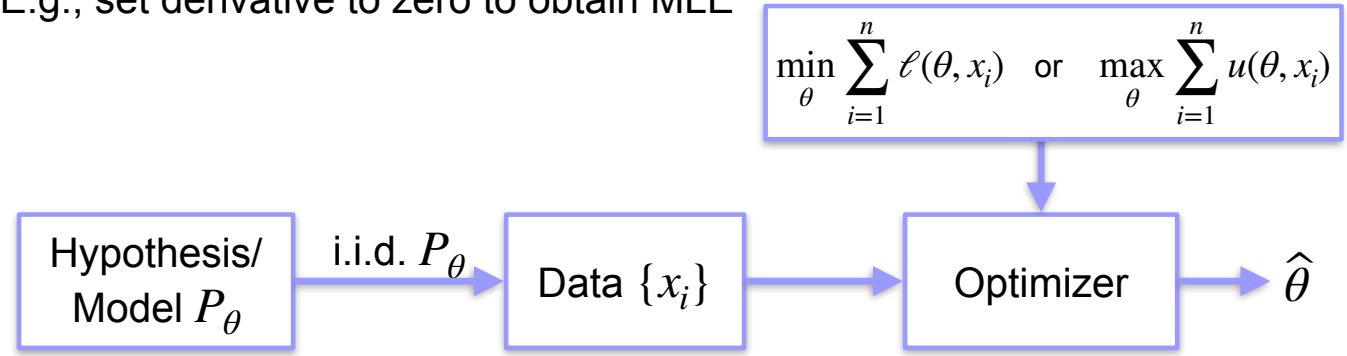
- Learning is…
  - Collect some data
    - E.g., coin flips
  - Choose a hypothesis class or model
    - E.g., binomial
  - Choose a loss function
    - E.g., data likelihood

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



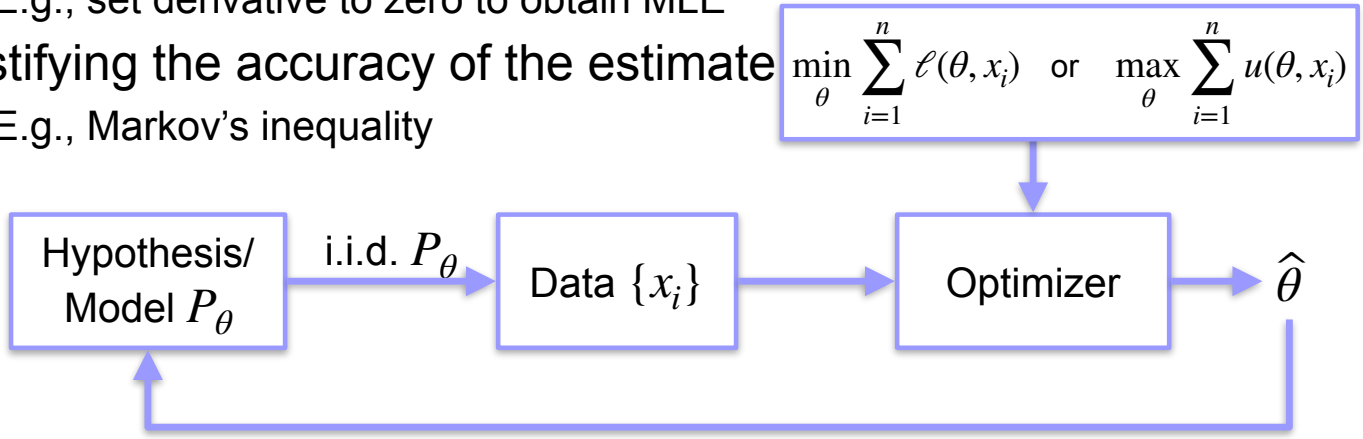
Recap

- 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



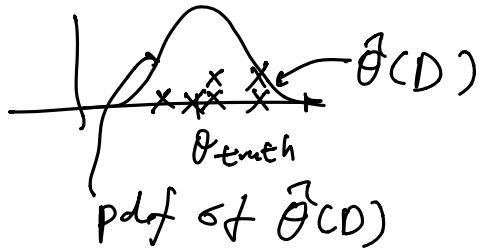
Recap

- 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



usually, $D = \{x_i\}_{i=1}^n \rightarrow \hat{\theta}(D)$

Random variable



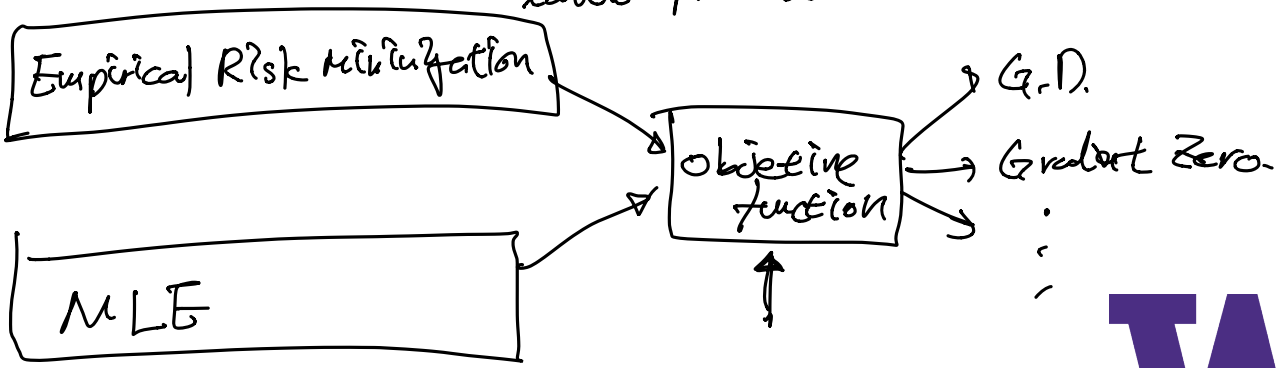
Imagine, again,

Hypothesis class is linear function

$$\sum_{i=1}^{d} w_i \cdot x_{1,i} = y_1$$

Linear Regression

labels predicted real-valued

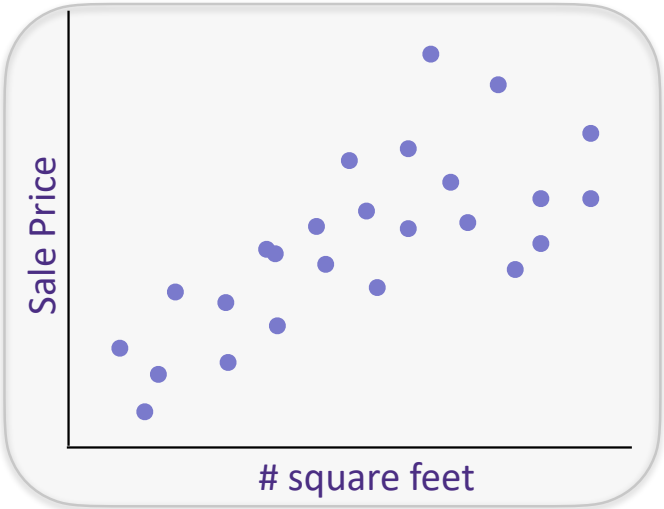


The regression problem, 1-dimensional

Given past sales data on zillow.com, predict:

y = **House sale price** *from*

x = **{# sq. ft.}**



Training Data:

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

input — Labels

$x_i \in \mathbb{R}$

$y_i \in \mathbb{R}$

Process

Decide on a **model**

Y_i

assume house sale price is a linear function of square feet.

X_i

Find the function which fits the data best

Use function to make prediction on new examples

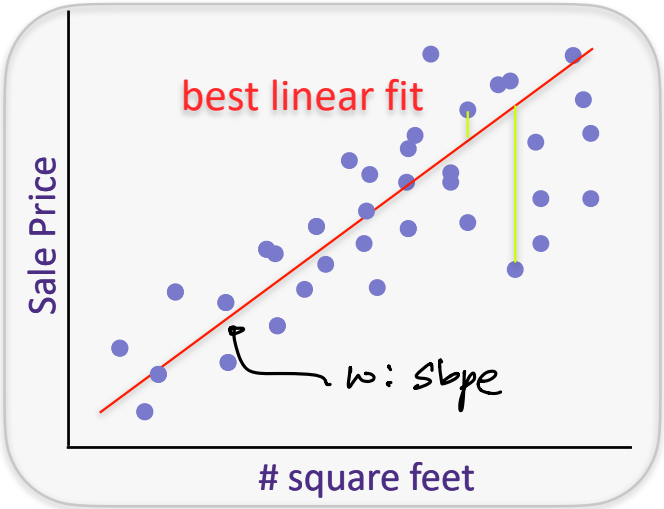
$Y_{new} = f(X_{new})$

Fit a function to our data, 1-dimension

Given past sales data on zillow.com, predict:

y = **House sale price** *from*

x = **{# sq. ft.}**



Error

$$y_i = x_i w + \epsilon_i$$

Training Data:

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

Hypothesis/Model: linear

$$y_i \approx x_i w$$

Loss: least squares solution

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

ϵ_i : error

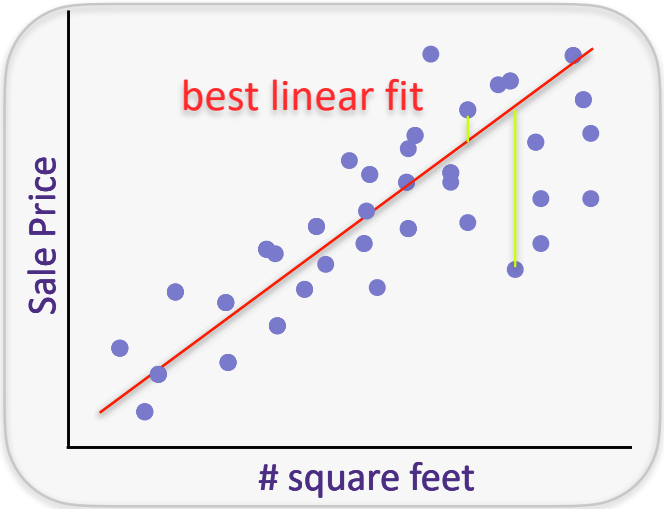
The regression problem, d-dimensions

Given past sales data on zillow.com, predict:

y = House sale price *from*

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

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



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

Hypothesis/Model: linear

$y_i \approx x_i^T w = x_{i,1} w_1 + x_{i,2} w_2 + \cdots + x_{i,d} w_d$

Loss: least squares solution

$$\min_{w} \sum_{i=1}^{n} \underbrace{\left(y_i - x_i^T w\right)}_{\text{error}}^2$$

$w \in \mathbb{R}^d$

The regression problem in matrix notation

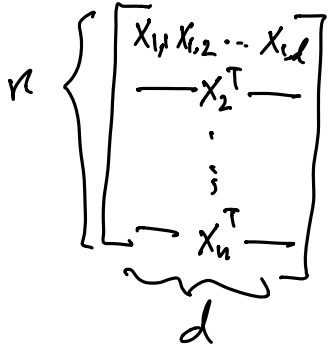
Data:

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

d : # of features

n : # of examples/datapoints

$n > d$



The regression problem in matrix notation

Data:

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

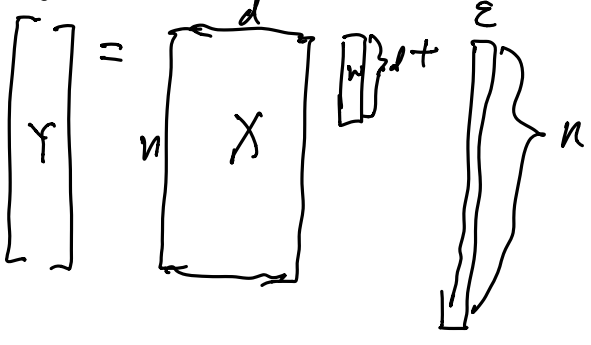
d : # of features

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

$\leftarrow \sum_i v_i^2 = \|v\|_2^2 = v^T \cdot v$

$\|v\|_2 = \sqrt{v_1^2 + v_2^2 + \cdots}$

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:

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

error.

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

$$\begin{bmatrix} y_1 - x_1^T w \\ \vdots \\ y_n - x_n^T w \end{bmatrix} = y - Xw$$

$$\|v\|_2 \triangleq \sqrt{v_1^2 + \cdots + v_d^2}$$

$$\|v\|_2^2 = (\|v\|_2)^2$$

$$v_1^2 + \cdots + v_d^2$$

$$v^T v$$

Loss:

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

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

The regression problem in matrix notation

parameter $\theta = w$

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

$\mathcal{L}(w)$

$f(w) = w^T w = w_1^2 + w_2^2 + \dots$

$\nabla_w f(w) = 2 \cdot w$

$f(w) = (Aw)^T Aw$ (*)

$\nabla_w f(w) = 2 A^T A w$

$f(w) = (Aw+b)^T (Aw+b)$

$\nabla_w f(w) = 2 A^T (Aw+b)$

$\|v\|_p = (|v_1|^p + \dots)^{1/p}$

$p = 0, 1, 2, \infty$

NNZ

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

$$\frac{1}{2}\nabla_w \mathcal{L}(w) = -X^T(y - Xw) \quad \leftarrow ?$$
$$= -X^T y + X^T X w \Big|_{w=\hat{w}_{LS}} = 0$$

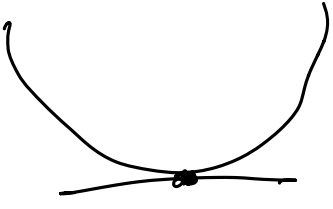
$$X^T X \cdot \hat{w} = X^T \cdot y \longrightarrow \hat{w} = (X^T X)^{-1} \cdot X^T y$$

$d \square \; \square d = d \square n$

d invertible.

$\square \cdot \square$

Closed-form



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!