2nd MLE example

Lecture 2: MLE for Gaussian and linear regression

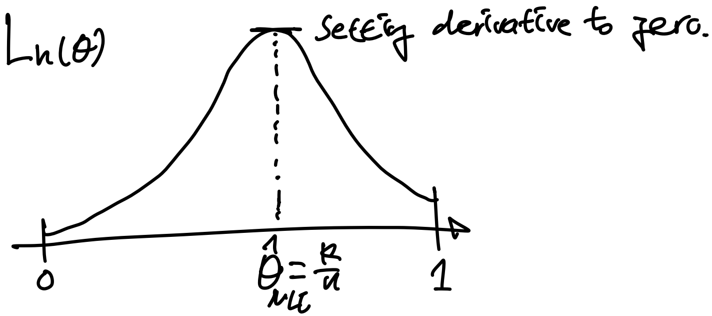
1st ML algorithm.

Recap: Maximum Likelihood Estimation

$\text{Bern}(\theta) \rightarrow \mathbb{P}(H;\theta) = \theta$

$\mathbb{P}(T;\theta) = 1-\theta$

- **Observe** $\overset{H,}{X_1}, \overset{T,}{X_2}, \ldots, \overset{H}{X_n}$ drawn i.i.d. from $P(X_i; \theta)$ for some true $\theta = \theta^*$
- **Likelihood function**: $L_n(\theta) = \prod_{i=1}^{n} P(X_i; \theta) = \theta^K (1-\theta)^{n-K}$ if K heads.
- **Log-likelihood function**: $\ell_n(\theta) = \log L_n(\theta) = \sum_{i=1}^{n} \log P(X_i; \theta) = K \cdot \log\theta + (n-K)\log(1-\theta)$
- **Maximum Likelihood Estimator (MLE)**: $\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \ell_n(\theta) = \frac{K}{n}$



What about continuous variables?

- *Client*: What if I am measuring a **continuous variable**?
- ***You*: Let me tell you about Gaussians…** $X \sim N(\mu, \sigma^2)$

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

← p.d.f

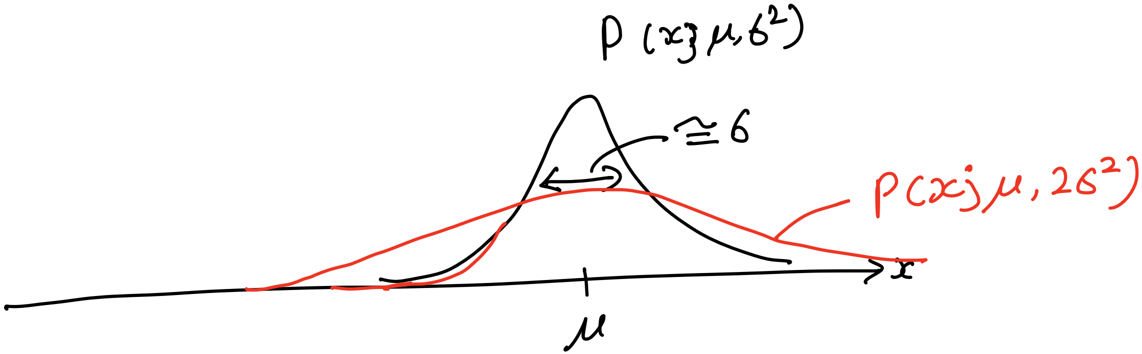
Probability density function

mean

Variance

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

Defined as



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):

$$P(\mathscr{D};\mu,\sigma) = P(x_1,\ldots,x_n;\mu,\sigma)$$

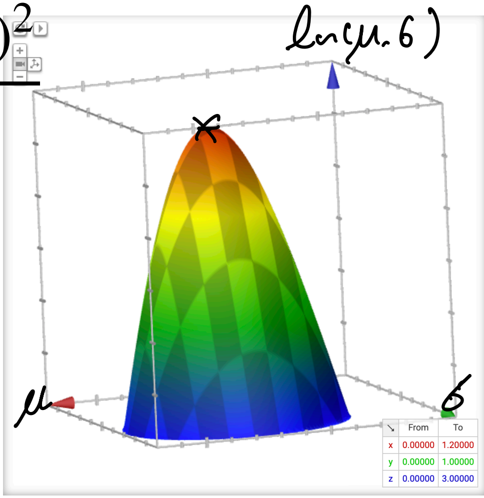
Independence $\longrightarrow$ $= P(x_1;\mu,\sigma^2) \times \cdots P(x_n;\mu,\sigma^2)$

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

- Log-likelihood of data:

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

$= \ell_n(\mu,\sigma)$



- 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?

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

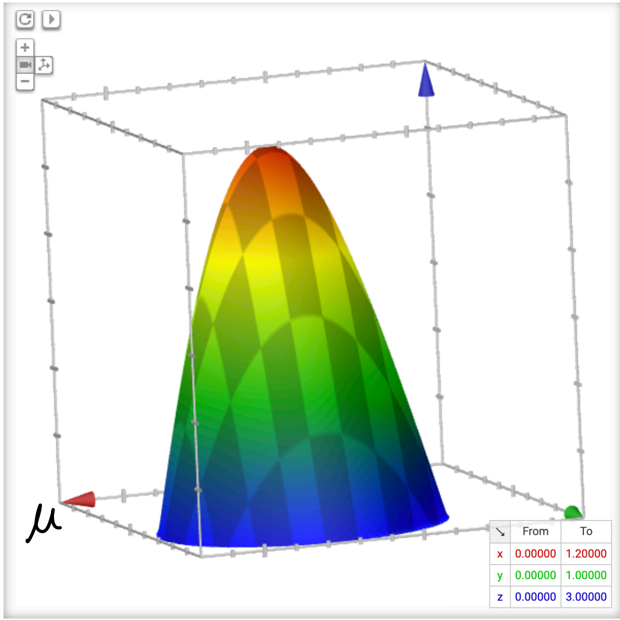
$$= -\sum_{i=1}^{n}\frac{2(x_i-\mu)}{2\sigma^2}$$

for $\hat{\mu}_{MLE}$

$$= -\frac{1}{\sigma^2}\left\{\sum_{i=1}^{n}x_i - n\cdot\mu\right\} \stackrel{?}{=} 0$$

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

* Note that $\hat{\mu}_{MLE}$ does not depend on $\hat{\sigma}^2_{MLE}$ for Gaussian.



MLE for variance

- Again, set derivative to zero:

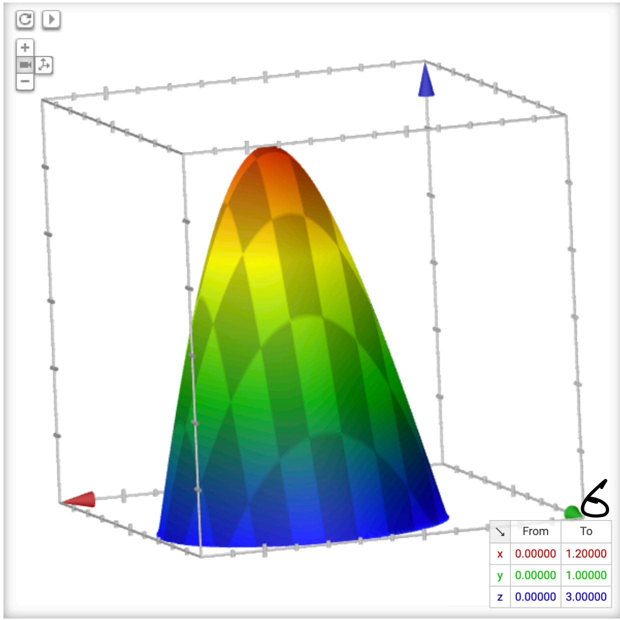
$$\frac{d}{d\sigma}\log P(\mathcal{D};\mu,\sigma) = \frac{d}{d\sigma}\Big[-n\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\Big]$$

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

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

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

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



What can we say about the MLE?

*Treat $(\hat{\mu}_{MLE}, \hat{\sigma}^2_{MLE})$ as Random Variables.
Imagine we run many experiments.

	$\hat{\mu}$	$\hat{\sigma}^2$
$\mathbb{E}$	√	√
Var	Not in lecture	Not in lecture

- MLE:

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

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

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \mathbb{E}\left[\frac{1}{n}\cdot n\cdot \mu\right] = \mu$$

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

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

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

$$\text{bias} = \mathbb{E}[\hat{\sigma}^2_{MLE}] - \sigma^2 = -\frac{1}{n}\sigma^2.$$

- 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)

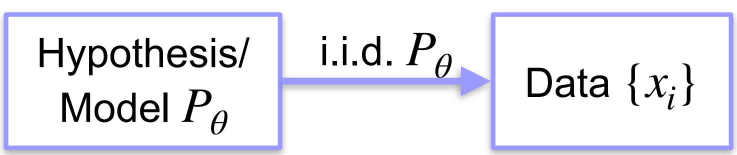
Recap

- 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

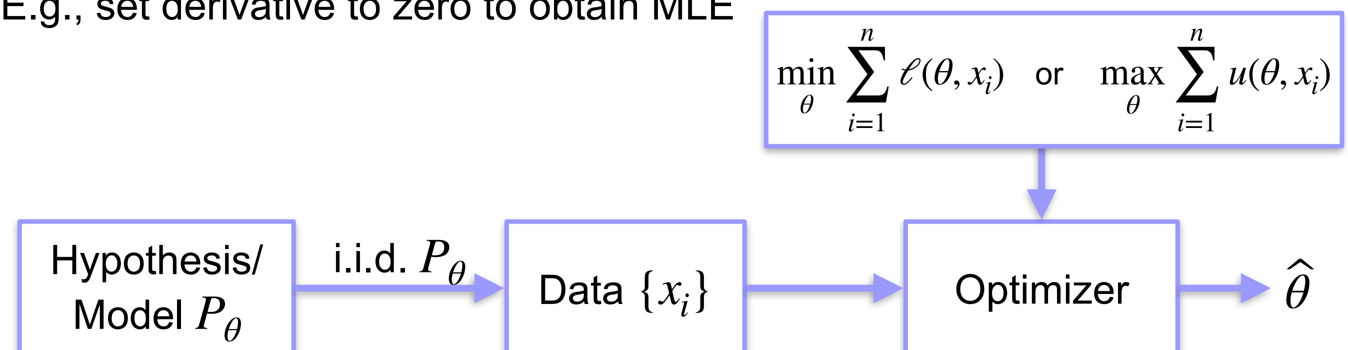
- Learning is…
  - Collect some data
    - E.g., coin flips
  - Choose a hypothesis class or model
    - E.g., binomial
  - Choose a loss function
    - E.g., data likelihood

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

$$\boxed{\text{Hypothesis/ Model } P_\theta} \xrightarrow{\text{i.i.d. } P_\theta} \boxed{\text{Data } \{x_i\}}$$

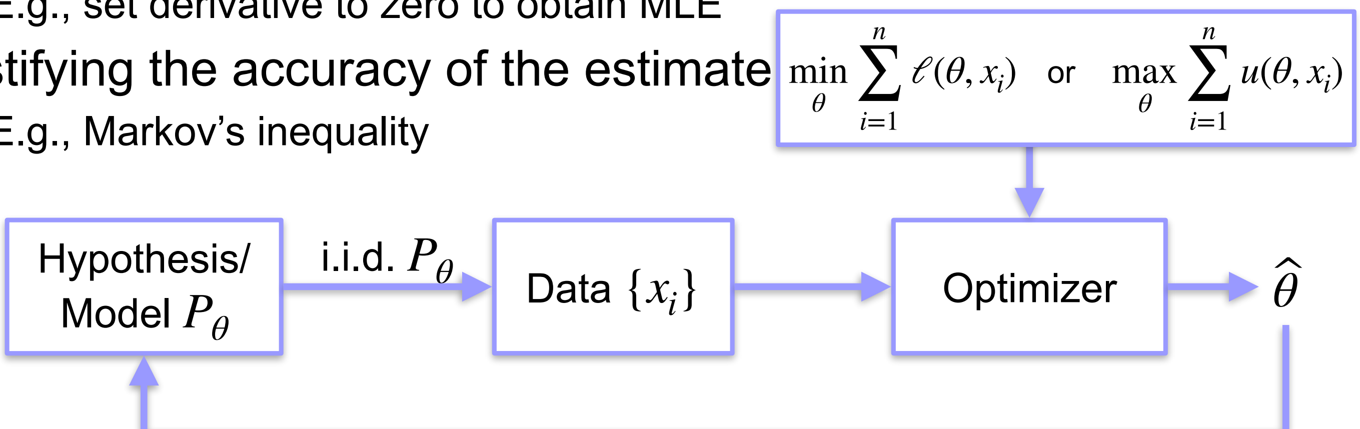
Recap

- 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



hypothesis class is linear functions

supervised learning with continuous labels.

Linear Regression

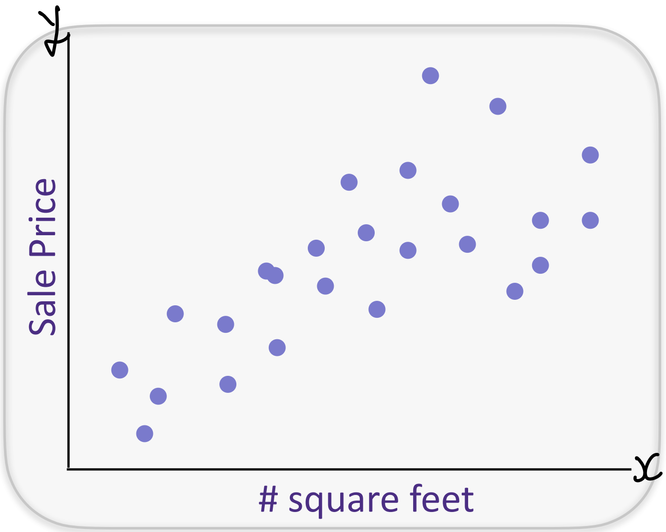
The regression problem, 1-dimensional

You are trying to sell your house, what is the right price?

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 label

$x_i \in \mathbb{R}$

$y_i \in \mathbb{R}$

Process

Decide on a **model** / Hypothesis explains Data, i.e. how price is related to sqft.

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

x_i y_i

Find the function which fits the data best

or model

Use function to make prediction on new examples

Fit a function to our data, 1-dimension

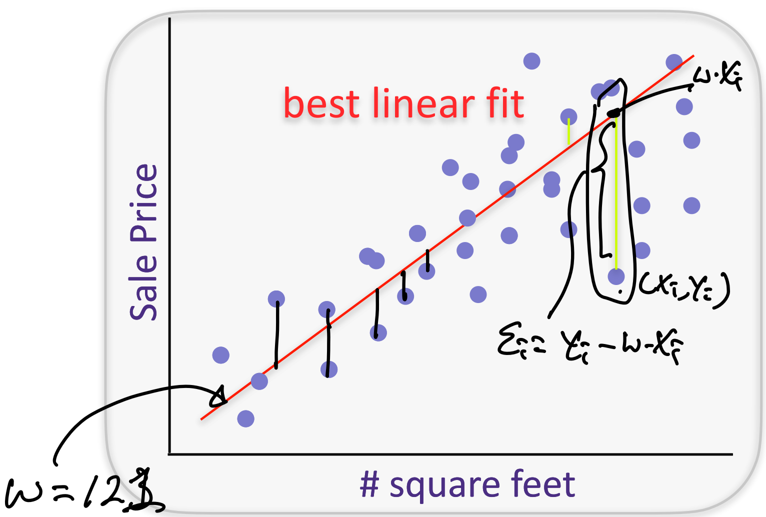
Given past sales data on zillow.com, predict:

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

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

Error

$$y_i = x_i w + \epsilon_i$$



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

Hypothesis/Model: linear

$$\underset{\text{label}}{y_i} \approx \underset{\text{model parameter}}{\cancel{x_i w}}\ w \cdot x_i \ (\text{input}) + \epsilon_i \ (\text{error/noise})$$

Loss: least squares solution

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

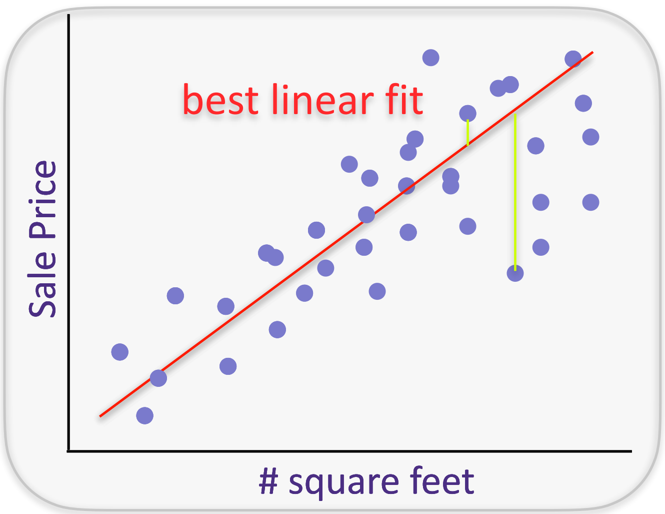
The regression problem, d-dimensions

Given past sales data on zillow.com, predict:

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

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

Error:
$y_i = x_i w + \epsilon_i$



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

$x_i \in \mathbb{R}^d$ $\quad x_i = \underbrace{\begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{bmatrix}}_{d}$

$y_i \in \mathbb{R}$

Hypothesis/Model: linear

$$y_i \approx x_i^T w = [x_{i1} \; \cdots \; x_{id}] \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix} = x_{i1} \cdot w_1 + x_{i2} \cdot w_2 + \cdots$$

Loss: least squares solution

$$\min_{\substack{w \\ \in \\ \mathbb{R}^d}} \sum_{i=1}^n \left(y_i - x_i^T w\right)^2$$

The regression problem in matrix notation

Data:

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

$$= n\left\{ \underbrace{\begin{bmatrix} x_{11} & \cdots & x_{1d} \\ — & x_2^T & — \\ & \vdots & \\ — & x_n^T & — \end{bmatrix}}_{d} \right.$$

d : # of features

n : # of examples/datapoints

* it is really important to keep track of dimensions.

The regression problem in matrix notation

d : # of features

n : # of examples/datapoints

Data:

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

Model:

$$\begin{aligned} y_1 &= x_1^T w + \epsilon_1 \\ y_2 &= x_2^T w + \epsilon_2 \\ &\vdots \\ y_n &= x_n^T w + \epsilon_n \end{aligned} \quad \xLeftrightarrow{\text{equivalent}} \quad \mathbf{y} = \mathbf{X} w + \epsilon$$

label input · model noise

$$n\left\{\begin{bmatrix} y \end{bmatrix}\right. = n \underbrace{\begin{bmatrix} X & & \\ & & \\ & & \end{bmatrix}}_{d} \begin{bmatrix} w \end{bmatrix}_d + \begin{bmatrix} \varepsilon \end{bmatrix}$$

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:

$$y_1 = x_1^T w + \epsilon_1$$

$$y_2 = x_2^T w + \epsilon_2$$

$$\vdots$$

$$y_n = x_n^T w + \epsilon_n$$

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

ℓ_2-norm

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

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:

$$y_1 = x_1^T w + \epsilon_1$$
$$y_2 = x_2^T w + \epsilon_2$$
$$\vdots$$
$$y_n = x_n^T w + \epsilon_n$$

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

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

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

The regression problem in matrix notation

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

model parameter θ

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

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

$$\Rightarrow X^T X \cdot w = X^T y \in \mathbb{R}^d \Rightarrow w = (X^T X)^{-1} \cdot X^T y$$

* We will assume that $X^T X$ is full rank and invertible

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!

Questions?