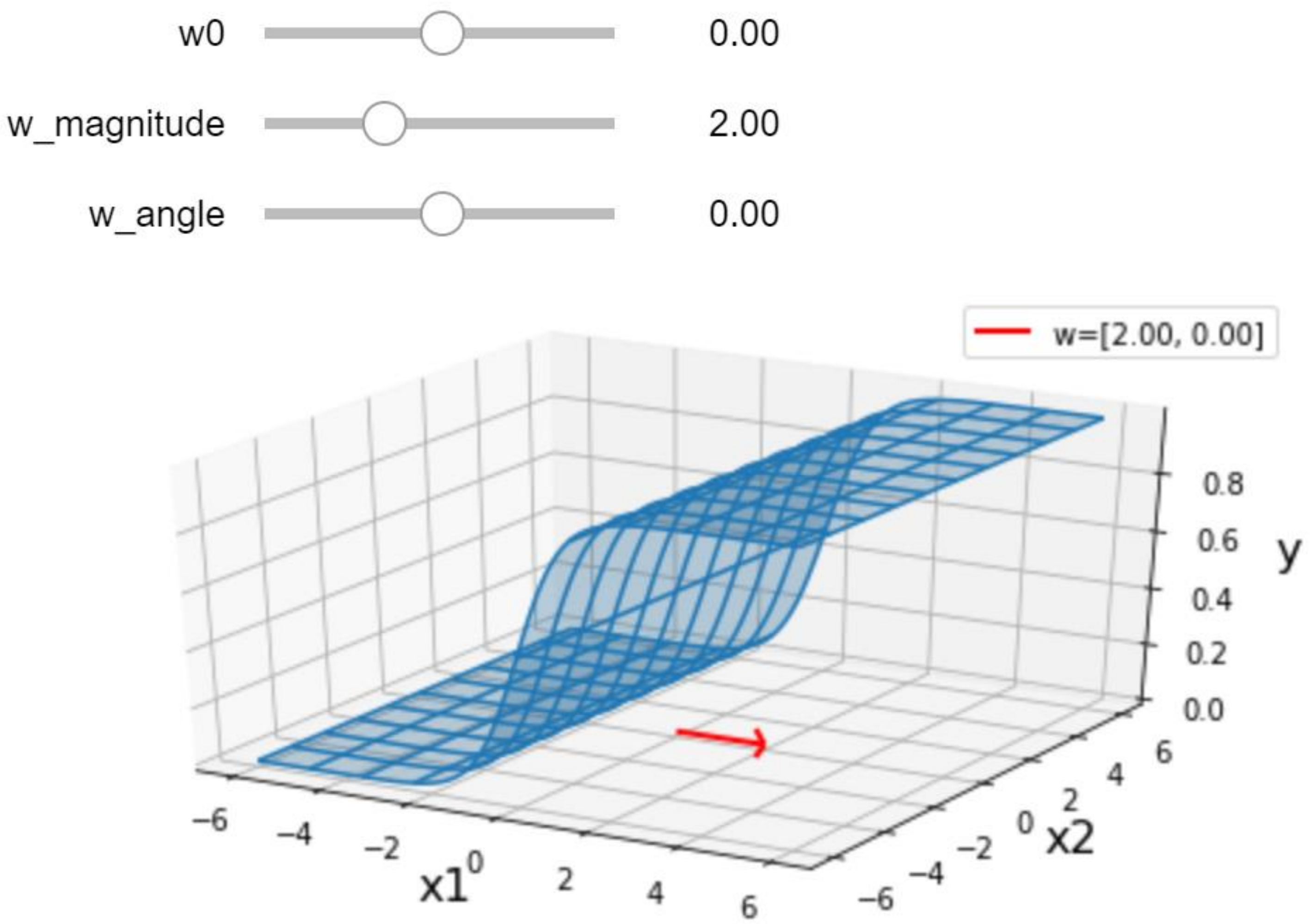
Warm-up as You Log In

Interact with the lec8.ipynb posted on the course website schedule



Announcements

Assignments

- HW3
  - Solution Session: Fri, 10/2, 8 pm

Schedule change this week

- Recitation slots this Friday will all be lecture (all three)

Midterm 1

- Practice exam
  - Timed (~~90~~ 80 min) exam in Gradescope
  - Open for a 24 hour window only, Tue 7 pm to Wed 7 pm
  - Need to take the practice exam to have access to the questions

Plan

Last time

- Logistic Regression
- Likelihood ← $p(y \mid \theta)$

Today

- Likelihood ← $p(y \mid \underline{x}, \theta)$
- MLE
- Conditional Likelihood and M(C)LE
- Solving Linear Regression

Friday

- Multiclass Logistic Regression

Introduction to Machine Learning

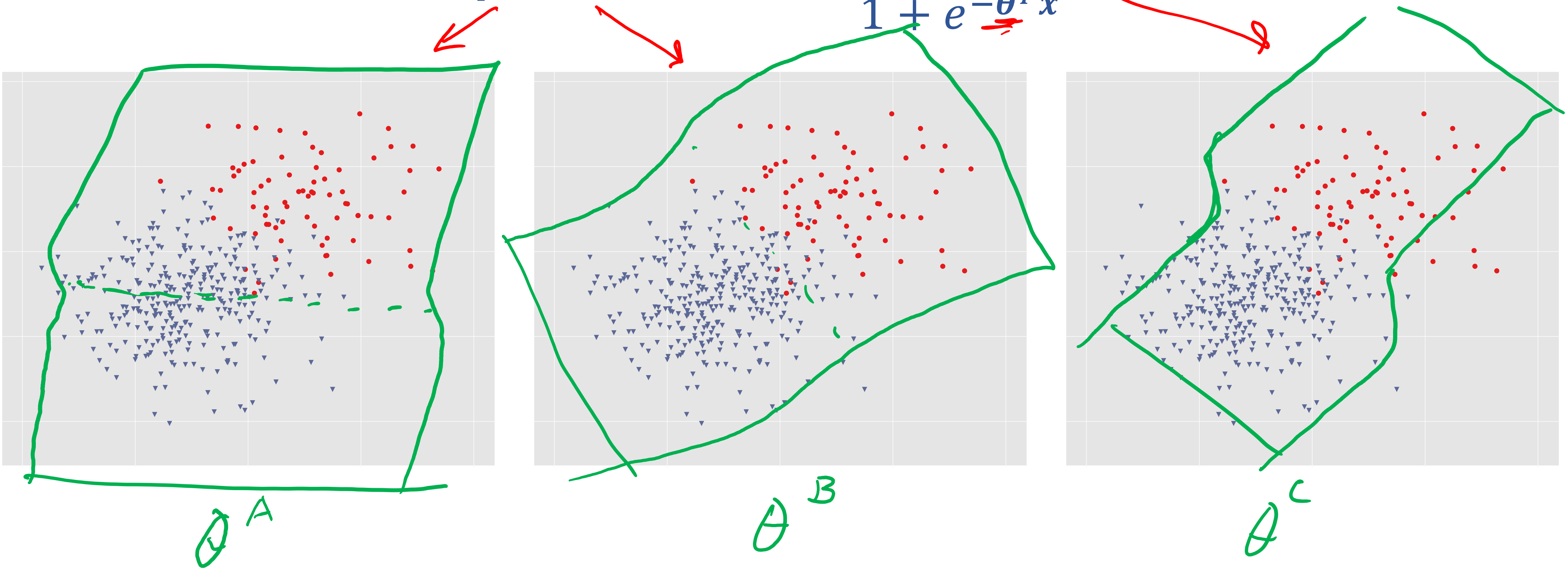
Logistic Regression

Instructor: Pat Virtue

Prediction for Cancer Diagnosis

Learn to predict if a patient has cancer ($Y = 1$) or not ($Y = 0$) given the input of just one test result, X_A.

$$p(Y = 1 \mid \boldsymbol{x}, \boldsymbol{\theta}) = \frac{1}{1 + e^{-\boldsymbol{\theta}^T \boldsymbol{x}}}$$



Likelihood

Likelihood: The probability (or density) of random variable Y taking on value y given the distribution parameters, $\boldsymbol{\theta}$.

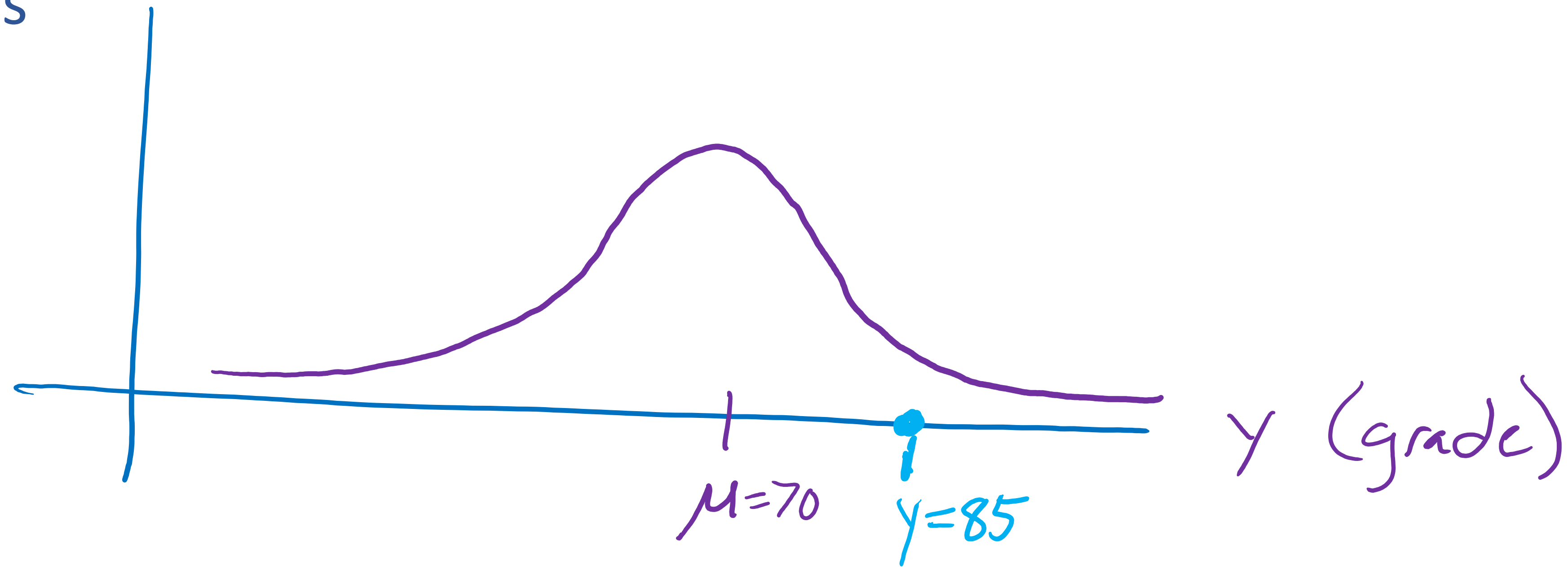
$$p(Y = y \mid \underline{\theta})$$

Likelihood

Likelihood: The probability (or density) of random variable Y taking on value y given the distribution parameters, $\boldsymbol{\theta}$.

$$p(Y=y \mid \mu=70, \sigma=10)$$

Grades



Warm-up as You Log In

Assume that exam scores are drawn independently from the same Gaussian (Normal) distribution.

Given three exam scores 75, 80, 90, which pair of parameters is a better fit?

A) Mean 80, standard deviation 3

B) Mean 85, standard deviation 7

Use a calculator/computer.

Gaussian PDF: $p\left(y \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$

Likelihood

Likelihood: The probability (or density) of random variable Y taking on value y given the distribution parameters, $\boldsymbol{\theta}$.

i.i.d.: Independent and identically distributed

$$P(Y^{(1)}=y^{(1)}, Y^{(2)}=y^{(2)}, Y^{(3)}=y^{(3)})$$ ← joint distribution not iid

identical

$$P(Y=y^{(1)}, Y=y^{(2)}, Y=y^{(3)})$$

independent

$$= P(Y=y^{(1)})\, P(Y=y^{(2)})\, P(Y=y^{(3)})$$

Piazza Poll 1

Assume that exam scores are drawn independently from the same Gaussian (Normal) distribution.

Given three exam scores 75, 80, 90, which pair of parameters is a better fit?

A) Mean 80, standard deviation 3 $\prod p(y^{(i)} \mid \mu = 80, \sigma^2 = 3)$ $\theta^{(A)}$

B) Mean 85, standard deviation 7 74% $\prod p(y \mid \quad)$ $\theta^{(B)}$

Use a calculator/computer.

Gaussian PDF: $p(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$ $\rightarrow \prod_{i=1}^{3} p(y^{(i)} \mid \mu, \sigma^2)$

Likelihood

Trick coin

$Y=1$ Heads

$[1, 0, 1, 1]$

All T	1/3 H	Fair	2/3 H	All H
$\phi^{(A)} = 0$	$\phi^{(B)} = 1/3$	$\phi^{(C)} = 1/2$	$\phi^{(D)} = 2/3$	$\phi^{(E)} = 1$

$$p(y^{(1)} \dots y^{(4)} \mid \phi^{A}) = \underline{\prod p(y^{(i)} \mid \phi^{(A)})}$$

ϕ heads

$(1-\phi)$ fail

$$= \phi^{A} \cdot (1-\phi^{A}) \cdot \phi^{A} \cdot \phi^{A}$$

$$= 0 \cdot 1 \cdot 0 \cdot 0 = 0$$

Piazza Poll 2

We model the outcome of a single mysterious weighted-coin flip as a Bernoulli random variable:

$$Y \sim Bern(\phi)$$

$$p(y \mid \phi) = \begin{cases} \phi, & y = 1 \;\; (Heads) \\ 1 - \phi, & y = 0 \;\; (Tails) \end{cases}$$

Given the ordered sequence of coin flip outcomes:

$$[1, 0, 1, 1]$$

What is the estimate of parameter $\hat{\phi}$?

Piazza Poll 2

We model the outcome of a single mysterious weighted-coin flip as a Bernoulli random variable:

$$Y \sim Bern(\phi)$$

$$p(y \mid \phi) = \begin{cases} \phi, & y = 1 \ \ (Heads) \\ 1 - \phi, & y = 0 \ \ (Tails) \end{cases}$$

Given the ordered sequence of coin flip outcomes:

$$[1, 0, 1, 1]$$

What is the estimate of parameter $\hat{\phi}$?

A. 0.0 B. 1/8 C. 1/4 D. 1/2 E. 3/4 F. 3/8 G. 1.0

Why?

Piazza Poll 2

We model the outcome of a single mysterious weighted-coin flip as a Bernoulli random variable:

$$Y \sim Bern(\phi)$$

$$p(y \mid \phi) = \begin{cases} \phi, & y = 1 \ \ (Heads) \\ 1 - \phi, & y = 0 \ \ (Tails) \end{cases}$$

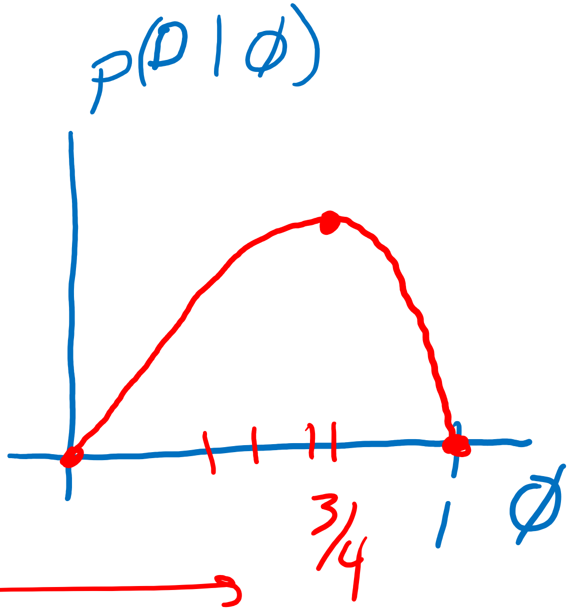
Given the ordered sequence of coin flip outcomes:

$$\mathcal{D} = [1, 0, 1, 1]$$

What is the estimate of parameter $\hat{\phi}$ for any possible ϕ?

A. 0.0 B. 1/8 C. 1/4 D. 1/2 E. 3/4 F. 3/8 G. 1.0

80%



$$\frac{N_{y=1}}{N}$$

Why?

$P_\theta(y) = p(y;\theta) = p(y|\theta)$ $\log xz = \log x + \log z$

Likelihood and Maximum Likelihood Estimation

Likelihood: The probability (or density) of random variable Y taking on value y given the distribution parameters, $\boldsymbol{\theta}$.

$$p(D|\theta) = \prod p(y^{(i)}|\theta)$$

Likelihood function: The value of likelihood as we change theta

(same as likelihood, but conceptually we are considering many different values of the parameters)

NOT $p(\theta|D)$

likelihood $\mathcal{L}(\theta;D) = p(D|\theta) = \prod p(y^{(i)}|\theta)$

log likelihood $\ell(\theta;D) = \log p(D|\theta) = \sum \log p(y^{(i)}|\theta)$

Likelihood and Log Likelihood

Bernouli distribution:

$$Y \sim Bern(z)$$

$$p(y \mid z) = \begin{cases} z, & y = 1 \\ 1 - z, & y = 0 \end{cases}$$

What is the log likelihood for three i.i.d. samples, given parameter z:

$$\mathcal{D} = \{y^{(1)} = 1, y^{(2)} = 0, y^{(3)} = 1, y^{(4)} = 1\}$$

$$L(z) =$$

$$\ell(z) =$$

Likelihood and Log Likelihood

Bernoulli distribution:

$$Y \sim Bern(z)$$

$$p(y) = \begin{cases} z, & y = 1 \\ 1 - z, & y = 0 \end{cases}$$

What is the log likelihood for three i.i.d. samples, given parameter z?

$$\mathcal{D} = \{y^{(1)} = 1, y^{(2)} = 1, y^{(3)} = 0\}$$

0 1

$$L(z) = z \cdot z \cdot (1 - z) \qquad = \prod_n z^{y^{(n)}} (1 - z)^{(1 - y^{(n)})}$$

$$\ell(z) = \log z + \log z + \log(1 - z) \qquad = \sum_n \underbrace{y^{(n)} \log z + \left(1 - y^{(n)}\right) \log(1 - z)}$$

$\partial \ell / \partial z = 0 \qquad z = \frac{N_{y=1}}{N}$

Previous Piazza Poll

We model the outcome of a single mysterious weighted-coin flip as a Bernoulli random variable:

$$Y \sim Bern(\phi)$$

Given the ordered sequence of coin flip outcomes:

$$[1, 0, 1, 1]$$

What is the estimate of parameter $\hat{\phi}$?

A. 0.0 B. 1/8 C. 1/4 D. 1/2 E. 3/4 F. 3/8 G. 1.0

Why?

Warm-up as You Log In

Assume that exam scores are drawn independently from the same Gaussian (Normal) distribution.

Given three exam scores 75, 80, 90, which pair of parameters is a better fit?

A) Mean 80, standard deviation 3

B) Mean 85, standard deviation 7

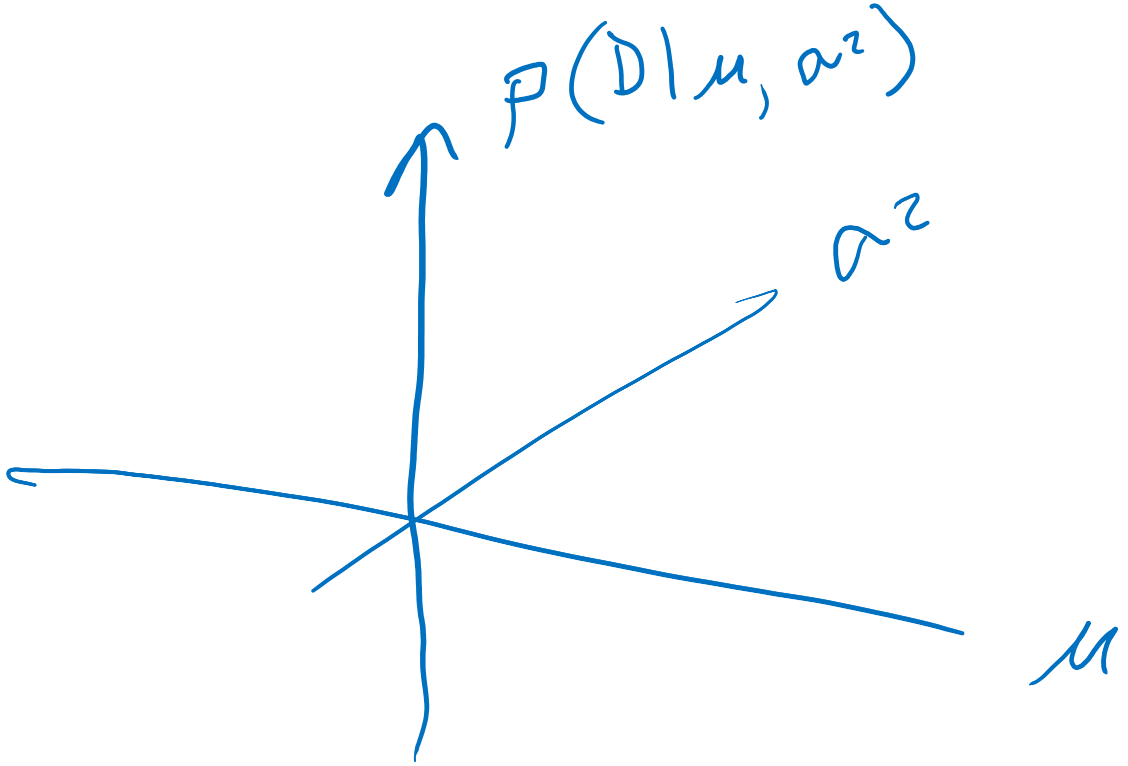
Use a calculator/computer.

Gaussian PDF: $p(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$

Warm-up as You Log In

Assume that exam scores are drawn independently from the same Gaussian (Normal) distribution.

Given three exam scores 75, 80, 90, which pair of parameters is a better fit?



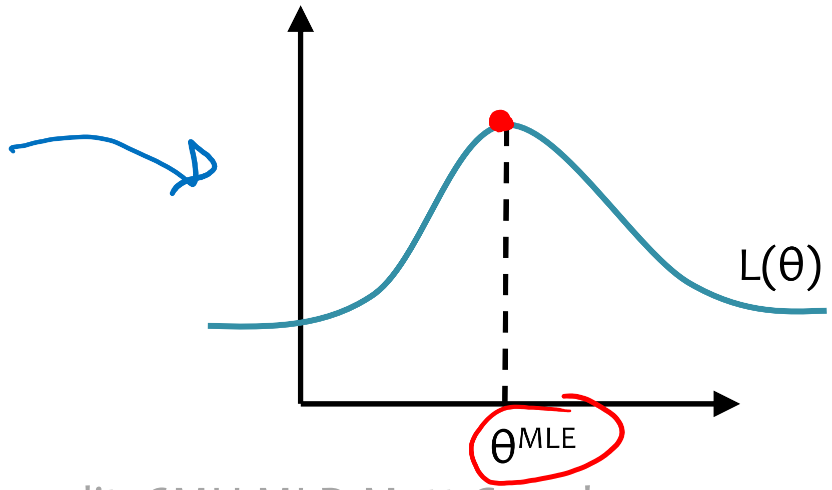
MLE

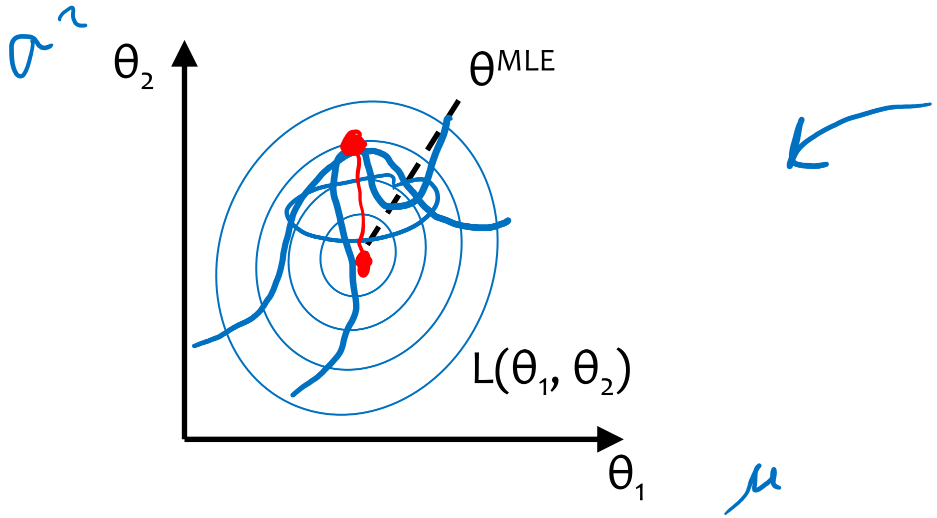
Suppose we have data $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$

Principle of Maximum Likelihood Estimation:
Choose the parameters that maximize the likelihood of the data.

$$\boldsymbol{\theta}^{\text{MLE}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \prod_{i=1}^{N} p(\mathbf{x}^{(i)} | \boldsymbol{\theta})$$

Maximum Likelihood Estimate (MLE)





Slide credit: CMU MLD Matt Gormley

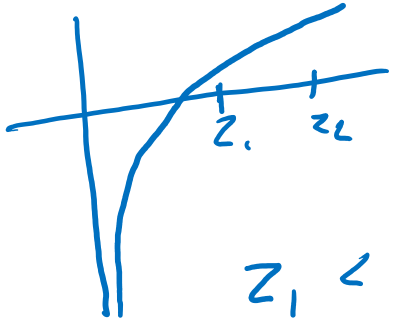
Maximum Likelihood Estimation

MLE of parameter θ for i.i.d. dataset $\mathcal{D} = \{y^{(i)}\}_{i=1}^{N}$

$$
\begin{aligned}
\hat{\theta}_{MLE} &= \operatorname*{argmax}_{\theta} p(\mathcal{D} \mid \theta) \\
&= \operatorname*{argmax}_{\theta} \log p(\mathcal{D} \mid \theta) \\
&= \operatorname*{argmax}_{\theta} \log \prod p(y^{(i)} \mid \theta) \quad \leftarrow \text{i.i.d.} \\
&= \operatorname*{argmax}_{\theta} \sum \log p(y^{(i)} \mid \theta) \\
&= \operatorname*{argmin}_{\theta} -\sum \log p(y^{(i)} \mid \theta)
\end{aligned}
$$

SGD negative log likelihood

log monotonic



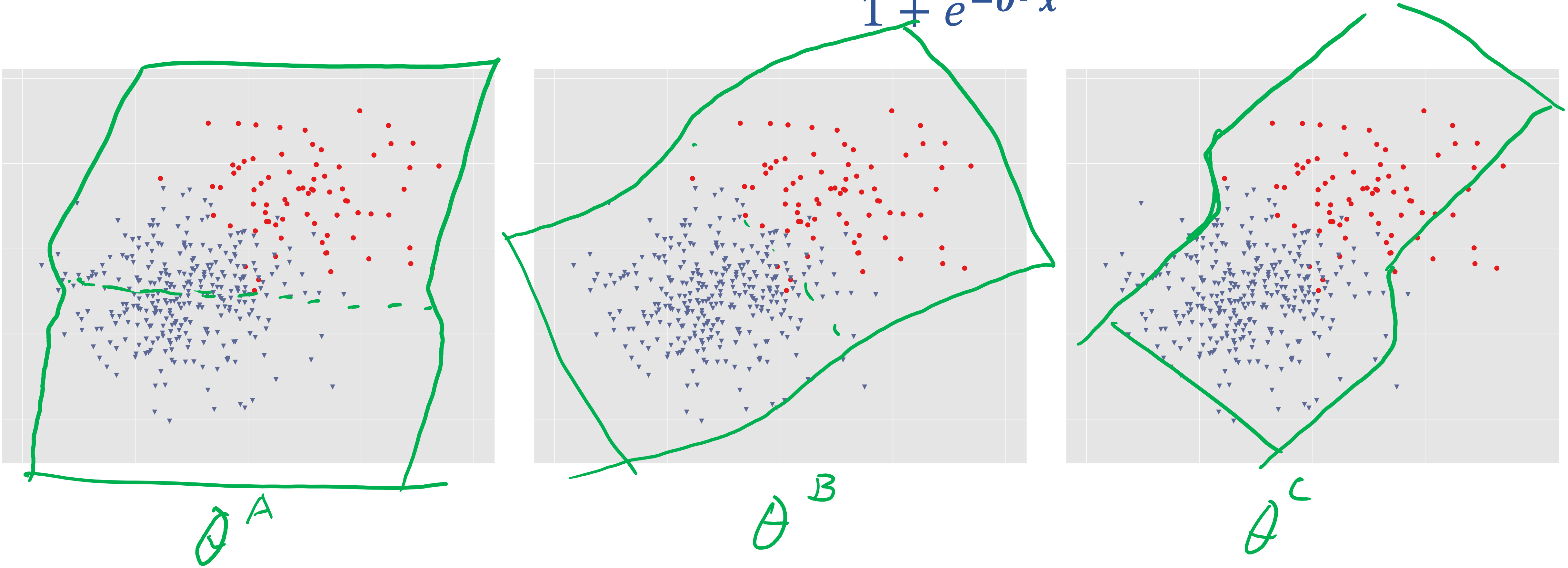
$z_1 < z_2$

$\log z_1 < \log z_2$

Prediction for Cancer Diagnosis

Learn to predict if a patient has cancer ($Y = 1$) or not ($Y = 0$) given the input of just one test result, X_A.

$$p(Y = 1 \mid \boldsymbol{x}, \boldsymbol{\theta}) = \frac{1}{1 + e^{-\boldsymbol{\theta}^T \boldsymbol{x}}}$$



OVERLY-SIMPLE PROBABILISTIC CLASSIFIER

Overly-simple Probabilistic Classifier

Y	X_1	X_2
1	0.3	9
0	3	4
1	2	1
1	1	-3

1) Model: $Y \sim Bern(\phi)$ ignore x

$$p(y \mid \cancel{\boldsymbol{x}}, \phi) = \begin{cases} \phi, & y = 1 \\ 1 - \phi, & y = 0 \end{cases}$$

2) Objective $J(\phi)$ = negative log likelihood

$$J(\phi) = -\sum_i \log p(y^i \mid x^i, \phi)$$

Bernoulli

$$= -\sum_i y^{(i)} \log \phi + (1-y^{(i)}) \log(1-\phi)$$

$$\log\left(\phi^{y^{(i)}} (1-\phi)^{(1-y^{(i)})}\right)$$

3) Solve

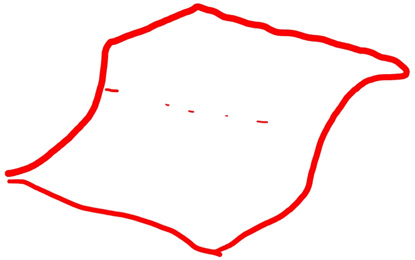
Closed-form solution $\frac{dJ}{d\phi} = 0$

and solve for $\hat{\phi}$

BINARY LOGISTIC REGRESSION

Binary Logistic Regression

logistic



1) Model: $Y \sim Bern(\mu)$ $\quad \mu = \sigma(\boldsymbol{\theta}^T \boldsymbol{x})$ $\quad \sigma(z) = \frac{1}{1+e^{-z}}$

$$p(Y=y \mid \vec{x}, \vec{\theta}) = \begin{cases} \mu & \text{if } y=1 \\ 1-\mu & \text{if } y=0 \end{cases}$$

2) Objective function: negative log likelihood

$$\ell(\vec{\theta}) = \sum_{i=1}^{N} \log p(Y=y^{(i)} \mid \vec{x}, \vec{\theta}) \leftarrow \text{log likelihood}$$

$$J(\vec{\theta}) = -\frac{1}{N} \ell(\vec{\theta})$$

3) Solve for $\hat{\theta}$ SGD

Binary Logistic Regression

Gradient

$$J(\vec{\theta})$$

$$J^{(i)}(\theta)$$

$$\nabla J^{(i)}(\theta) =$$

Solve Logistic Regression

$\mu = \sigma(\boldsymbol{\theta}^T \boldsymbol{x})$ $\sigma(z) = \frac{1}{1+e^{-z}}$

$$J(\boldsymbol{\theta}) = -\frac{1}{N}\sum_n \left(y^{(n)} \log \mu^{(n)} + \left(1 - y^{(n)}\right) \log\left(1 - \mu^{(n)}\right)\right)$$

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = -\frac{1}{N}\sum_n \left(y^{(n)} - \mu^{(n)}\right) \boldsymbol{x}^{(n)}$$

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{w}) = 0?$$

No closed form solution ☹

Back to iterative methods. Solve with (stochastic) gradient descent, Newton's method, or Iteratively Reweighted Least Squares (IRLS)