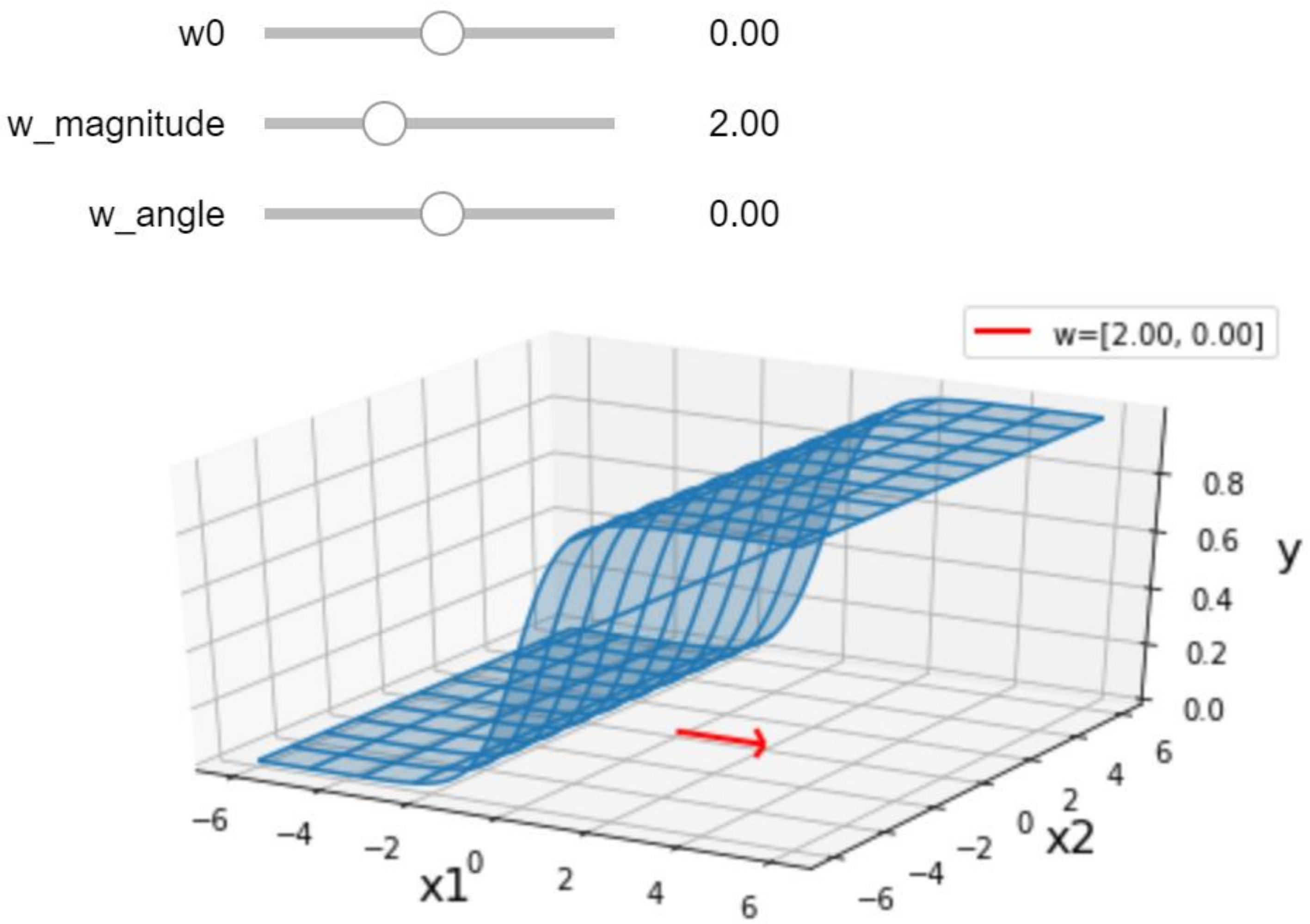
Warm-up as You Log In

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 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

Today

- Likelihood
- MLE
- Conditional Likelihood and M(C)LE
- Solving Linear Regression

Friday

- Multiclass Logistic Regression

Introduction to Machine Learning

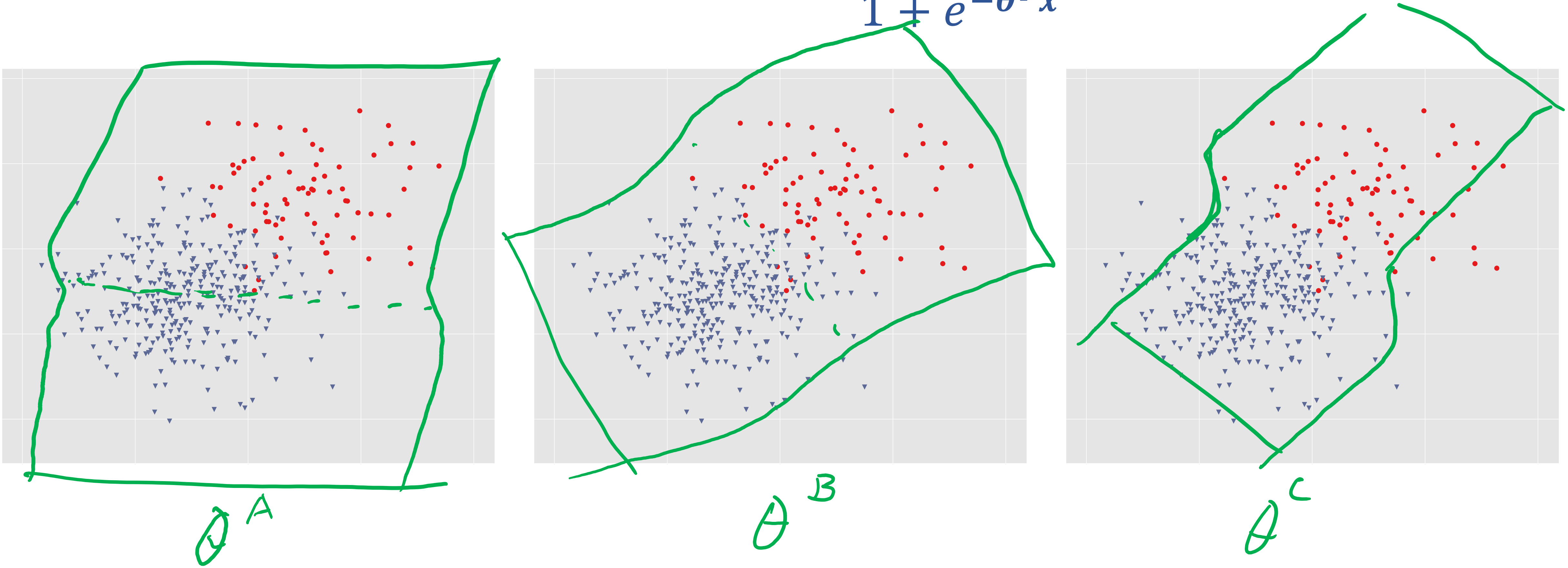
Logistic Regression

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, $\boldsymbol{\theta}$.

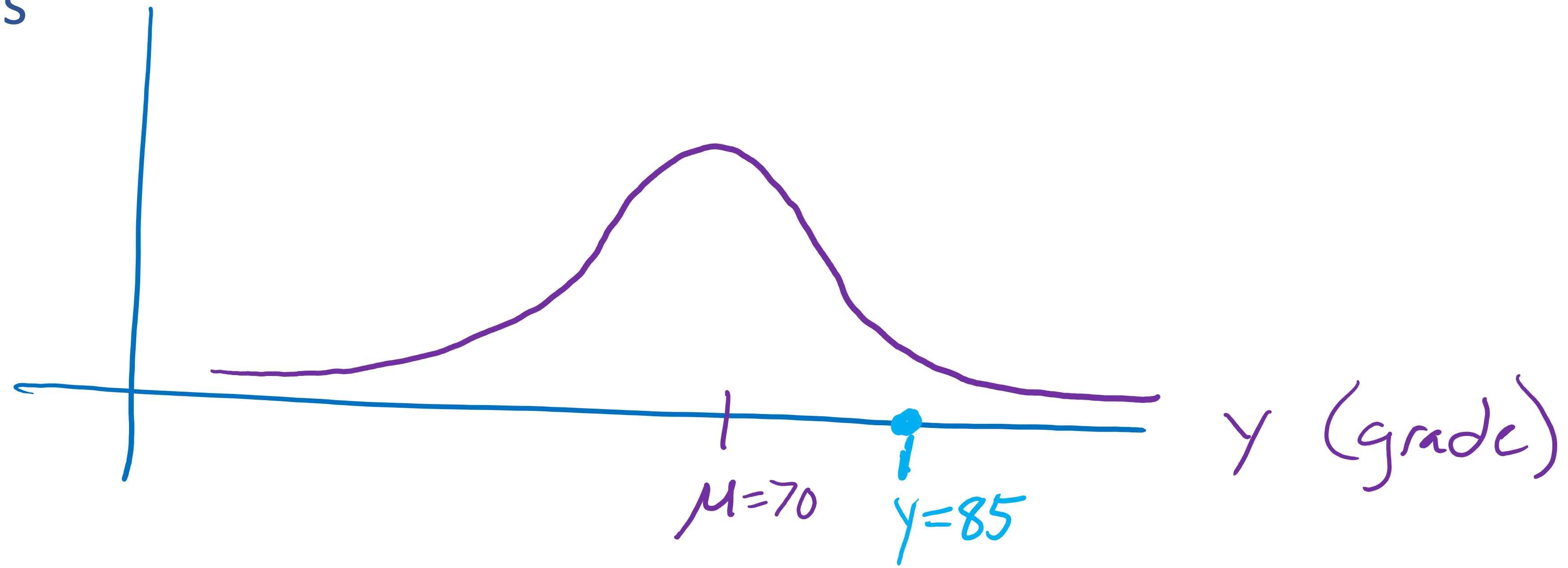
$$p(Y = y \mid \theta)$$

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

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

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

B) Mean 85, standard deviation 7

Use a calculator/computer.

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

Likelihood

Trick coin

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:

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

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

Why?

Likelihood and Maximum Likelihood Estimation

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

Likelihood function: The value of likelihood as we change theta

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

Likelihood and Log Likelihood

Bernouli distribution:

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

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

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?

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

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

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

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?

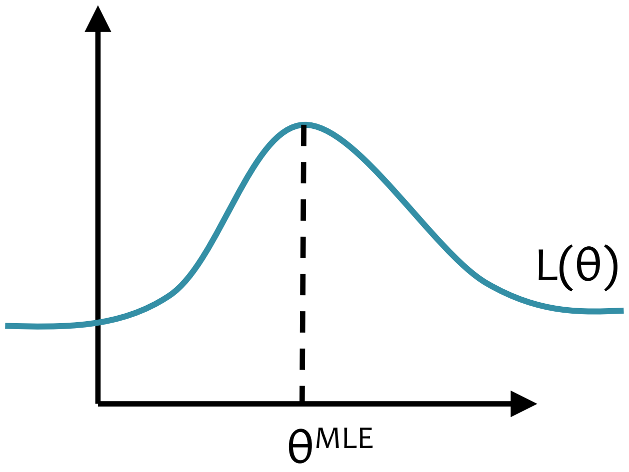
MLE

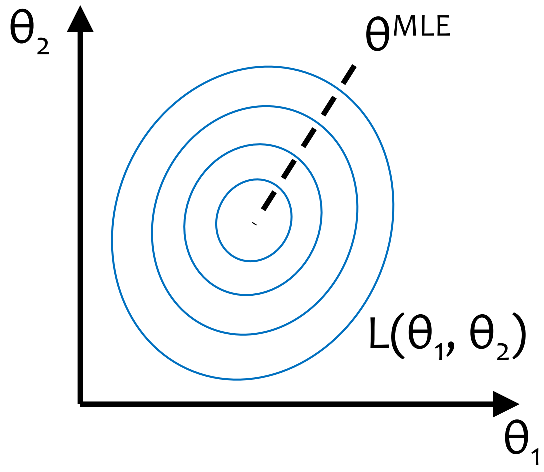
Suppose we have data $\mathcal{D} = \{x^{(i)}\}_{i=1}^N$

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

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

Maximum Likelihood Estimate (MLE)





Slide credit: CMU MLD Matt Gormley

Maximum Likelihood Estimation

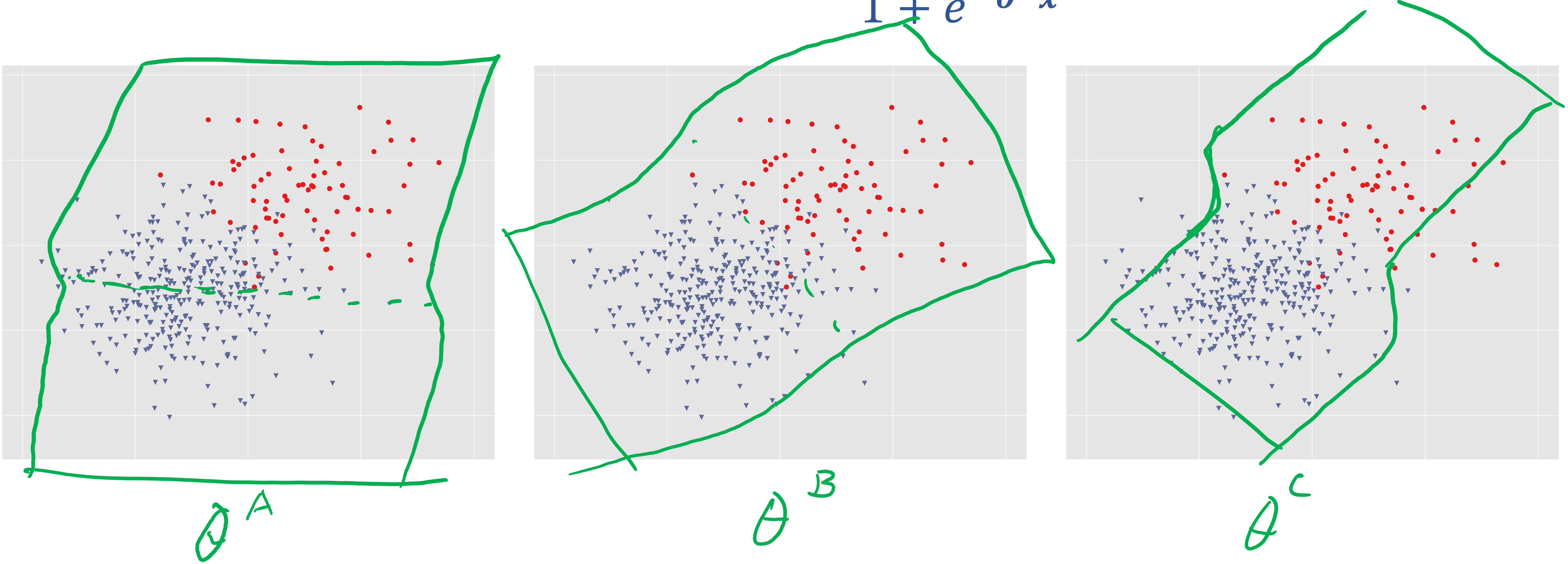
MLE of parameter θ for i.i.d. dataset $\mathcal{D} = \left\{y^{(i)}\right\}_{i=1}^{N}$

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}}\, p(\mathcal{D} \mid \theta)$$

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

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

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

2)

BINARY LOGISTIC REGRESSION

Binary Logistic Regression

1) Model: $Y \sim Bern(\mu) \quad \mu = \sigma(\boldsymbol{\theta}^T \boldsymbol{x}) \qquad \sigma(z) = \frac{1}{1+e^{-z}}$
2)

Binary Logistic Regression

Gradient

Solve Logistic Regression

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

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

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

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

No closed form solution ☹

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