Announcements

Assignments

- HW6
  - Due Today, 11:59 pm

Midterm 2

- Mon, 11/9, during lecture
- See Piazza for details on practice exam tomorrow

Schedule change

- Lecture on Friday instead of recitation

Plan

Last Time

- M(C)LE $\underset{\theta}{\operatorname{argmax}}\, p(\underline{y \mid \boldsymbol{x}}, \boldsymbol{\theta})$
- MAP $\underset{\theta}{\operatorname{argmax}}\, p(\underline{\,y \mid \boldsymbol{x}}, \boldsymbol{\theta}\,)\, p(\boldsymbol{\theta})$

Ridge Regression

Linear Reg.

Gaussian prior

Today

- Generative models $\underset{\theta}{\operatorname{argmax}}\, p(\,\underline{\boldsymbol{x} \mid y}, \theta\,)\, p(\underline{y} \mid \theta)$
- Naïve Bayes $\underset{\theta}{\operatorname{argmax}}\, \prod_{m=1}^{M} p(\,x_m \mid y, \theta\,)\, p(y \mid \theta)$

$p(x, y \mid \theta)$

Introduction to Machine Learning

Generative Models & Naïve Bayes

Instructor: Pat Virtue

Recall: Fisher Iris Dataset

https://en.wikipedia.org/wiki/Iris_flower_data_set

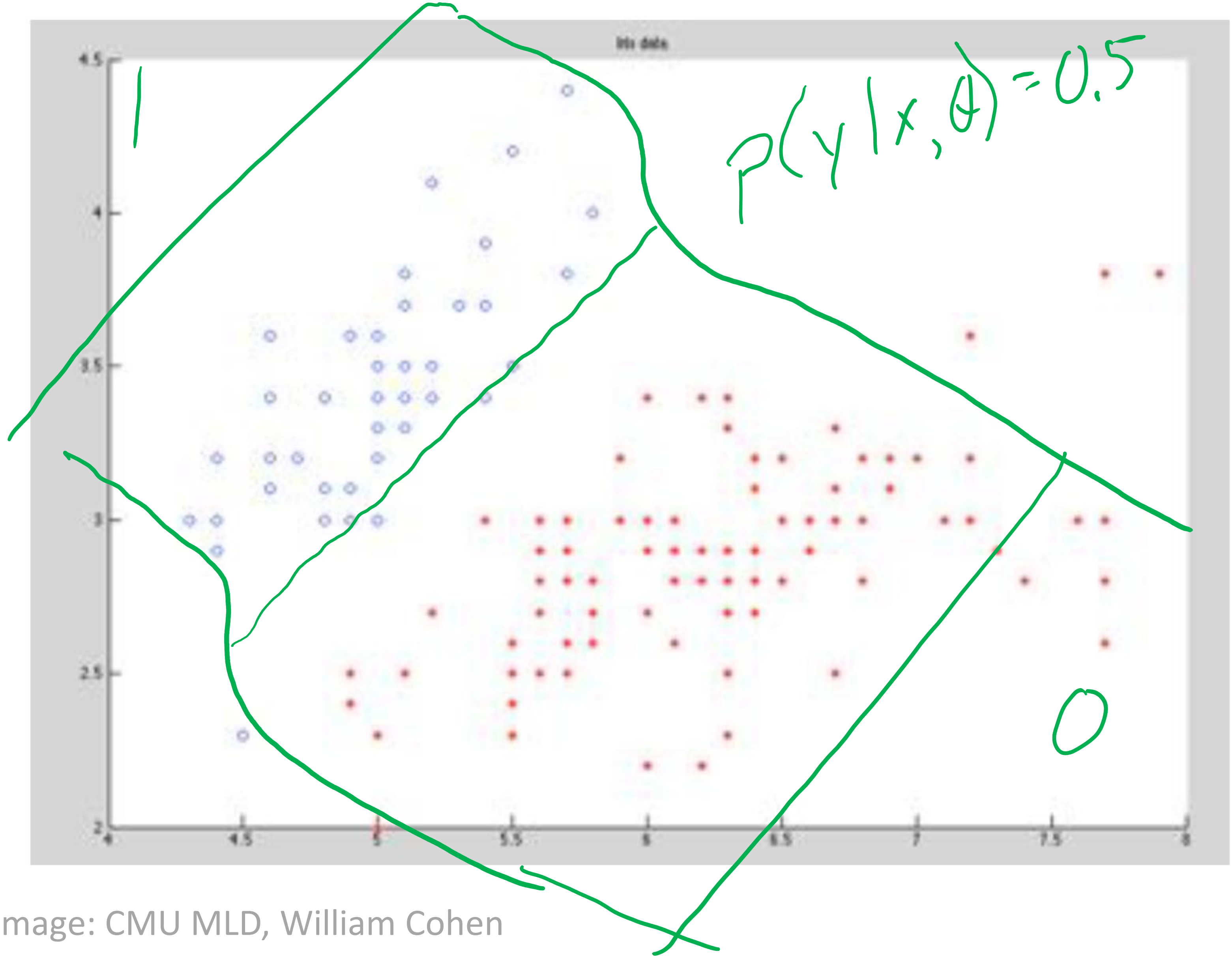
Recall: Fisher Iris Dataset

https://en.wikipedia.org/wiki/Iris_flower_data_set

Fisher (1936) used 150 measurements of flowers from 3 different species: Iris setosa (0), Iris virginica (1), Iris versicolor (2) collected by Anderson (1936)

Species	Sepal Length	Sepal Width	Petal Length	Petal Width
0	4.3	3.0	1.1	0.1
0	4.9	3.6	1.4	0.1
0	5.3	3.7	1.5	0.2
1	4.9	2.4	3.3	1.0
1	5.7	2.8	4.1	1.3
1	6.3	3.3	4.7	1.6
1	6.7	3.0	5.0	1.7

Modeling the Fisher Iris Dataset

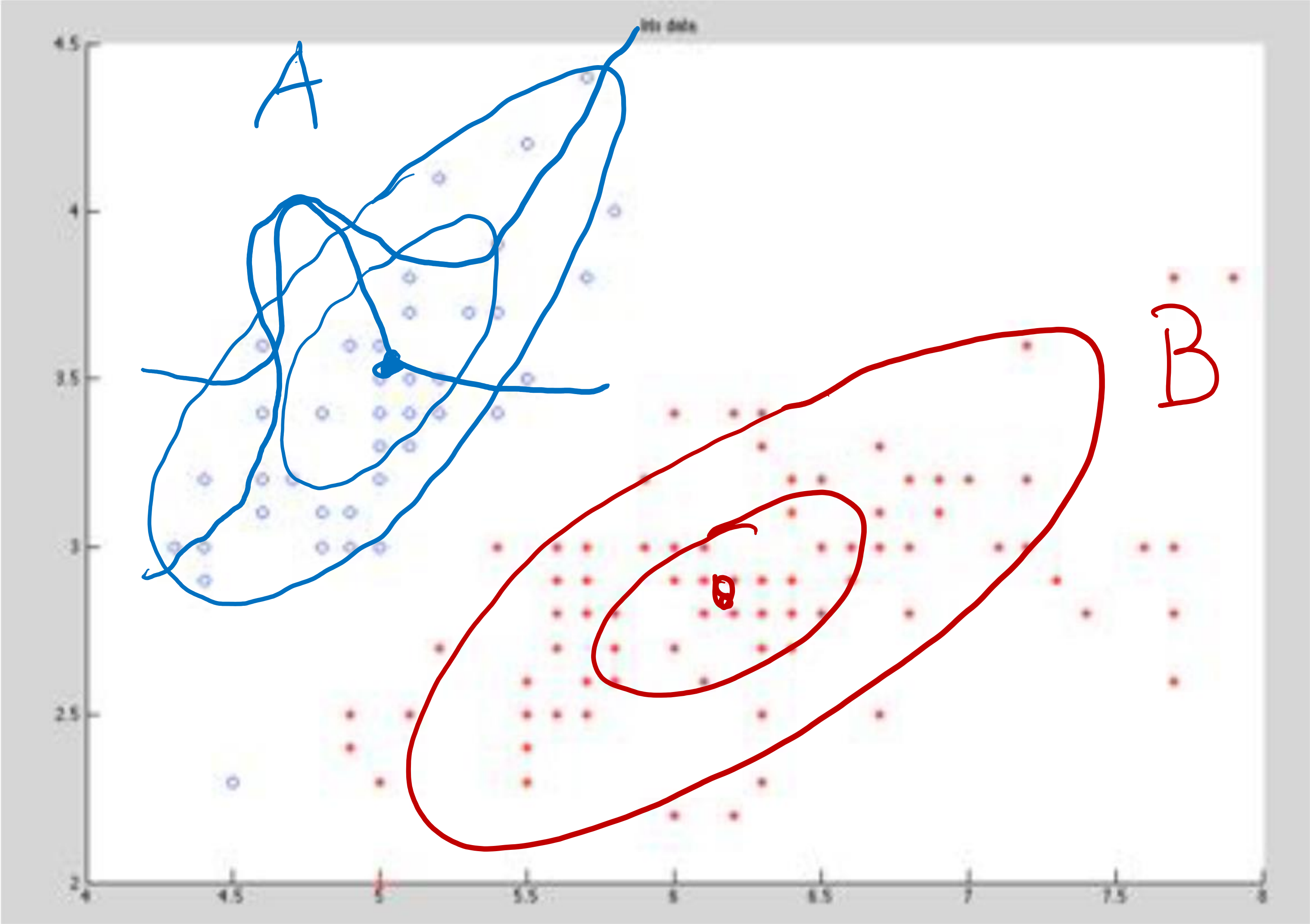


$$P(y = \text{blue} \mid \vec{x}, \vec{\theta})$$

$$= \frac{1}{1 + e^{-\theta^T x}}$$

Image: CMU MLD, William Cohen

Modeling the Fisher Iris Dataset



$$\vec{\mu}_A = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

$$\Sigma_B = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$$

$$\vec{\mu}_B = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

$$\Sigma_B = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$$

Image: CMU MLD, William Cohen

Modeling the Fisher Iris Dataset

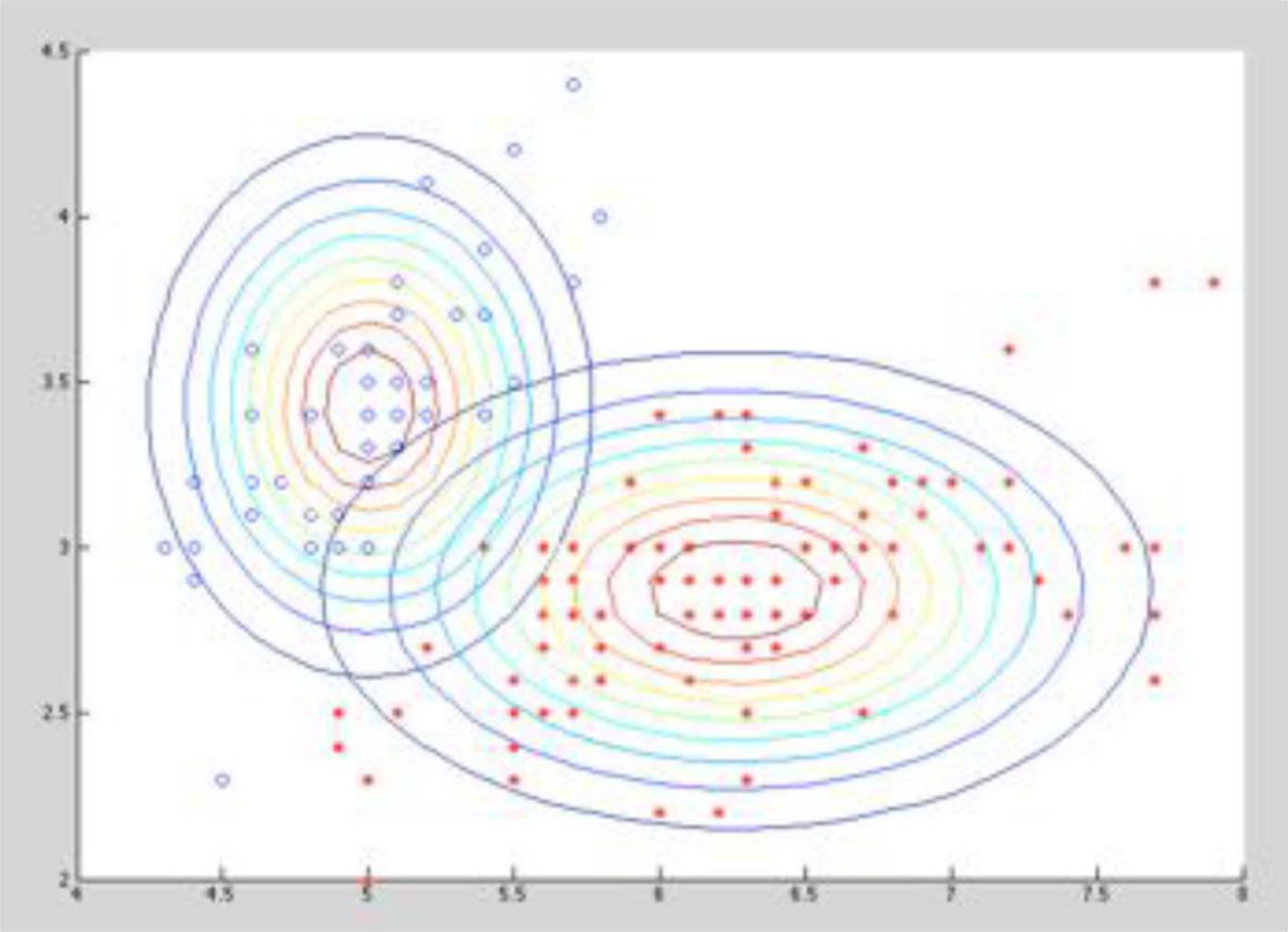
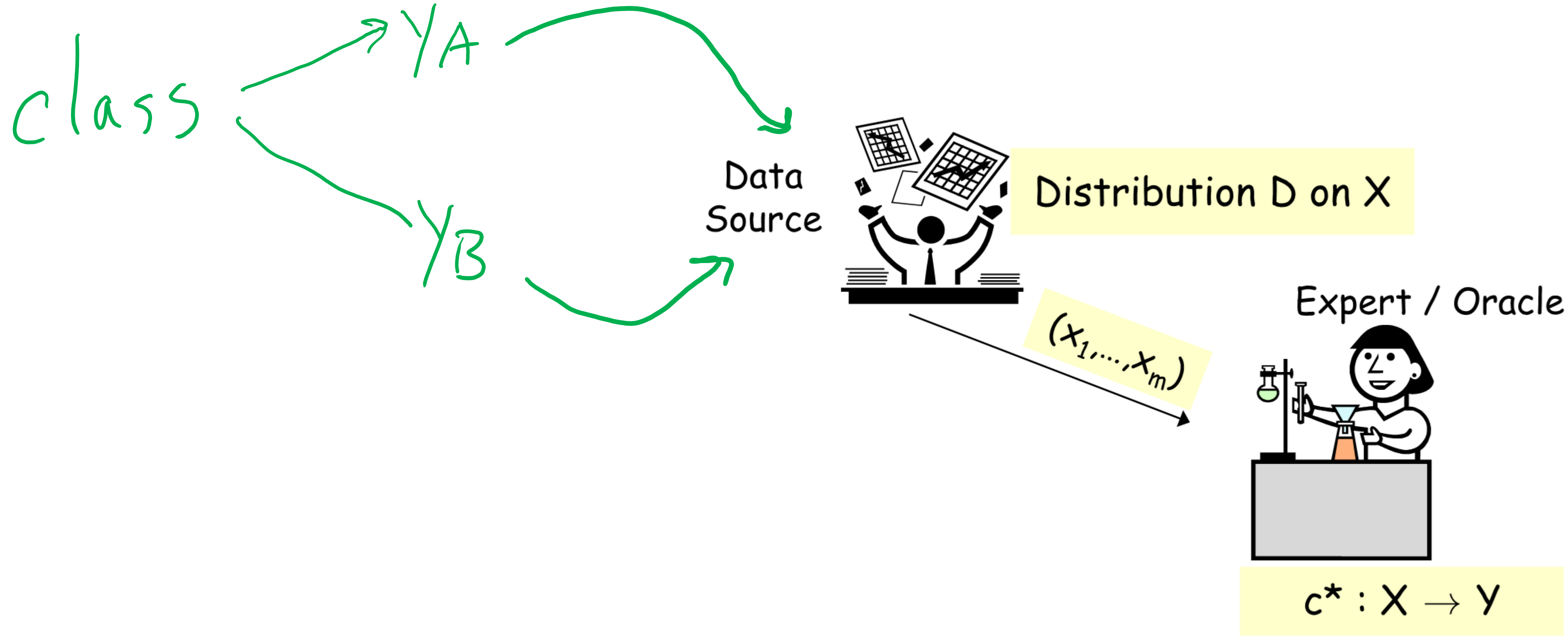


Image: CMU MLD, William Cohen

Generative vs Discriminative Modeling

Discriminative: modeling X → Y directly

Generative: Stronger modeling assumptions about where data came from



Generative vs Discriminative Modeling

$$\alpha = \frac{1}{p(x)}$$

$$\alpha = \frac{1}{\sum_y p(x,y)}$$

Discriminative: model $p(y \mid x, \theta)$ directly

- Learn parameters θ from data

Generative: model $p(y \mid \theta_{class})$ and $p(x \mid y, \theta_{class\ conditional})$

- Learn parameters θ_{class} and $\theta_{class\ conditional}$ from data
- Use Bayes rule to compute $p(y \mid x, \theta_{class}, \theta_{class\ conditional})$

$$p(y \mid x, \theta_{class}, \theta_{class\ conditional}) \propto p(x \mid y, \theta_{class\ conditional})\, p(y \mid \theta_{class})$$

$$p(y|x) = \alpha\, p(x|y)\, p(y) = \alpha\, p(x,y)$$

Generative Story

News article topic classification

Y
- Document class: Business, Entertainment, Politics

X
- Words in the document

SPAM classification

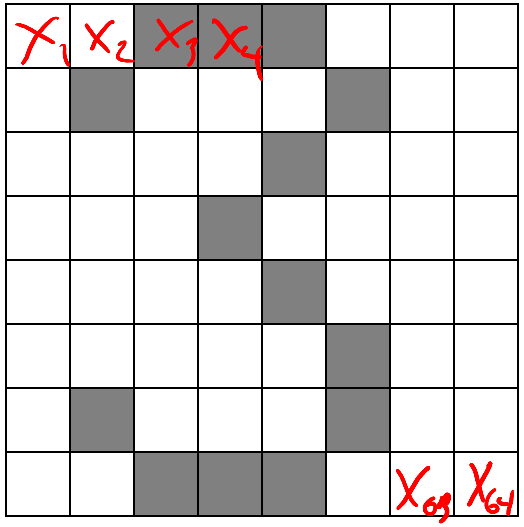
Y
- Document class: SPAM or not

X
- Words in the document

Generative Story

Hand-written digits

- Y Digit class: 0-9
- X Pixels in images



Generative vs Discriminative Modeling

Discriminative: $p(y \mid x)$

Generative: $p(y \mid x) = \alpha\, p(x, y) = \alpha\, \underline{p(x \mid y)}\, \underline{p(y)}$

Modeling assumptions vs. Data assumptions

~~Assumptions vs Data~~

- Discriminative: fewer modeling assumptions, more data
- Generative: more model assumptions, less data

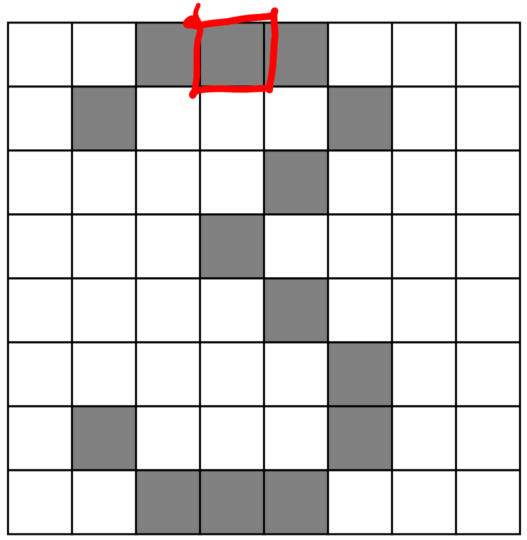
Quick Check

How many parameters?

- $P(Y)$, Y represents outcome of a 6-sided die roll

$$\theta = \begin{bmatrix} 0.1 \\ 0.1 \\ 0.4 \\ 0.2 \\ 0.1 \end{bmatrix}$$

Multivariate Generative Models



Hand-written digits: How many parameters?

- $P(Y)$ 10 $P(Y=3) = \phi_3 = 0.1$
- $P(\boldsymbol{X} \mid Y = 3)$
  $= \underline{p(X_1, X_2, \ldots X_{64} \mid Y} = 3)$

$P(X_1=0, X_2=0, \ldots X_{64}=0 \mid Y=3) = \phi_{3,00000\ldots0}$

$P(X_1=1, X_2=0, \ldots X_{64}=0 \mid Y=3) = \phi_{3,1,000\ldots0}$

$2^{64} \times 10 + 10$

Naïve Bayes assumption (bag of pixels)

All X_m are independent given Y

- $P(Y)$
- $P(\boldsymbol{X} \mid Y = 3)$
  $= \underbrace{p(X_1 \mid Y = 3)}_{\phi} \underbrace{p(X_2 \mid Y = 3)}_{\phi} \ldots \underline{p(X_{64} \mid Y = 3)}$ $64 \times 10 + 10$

Conditional Independence and Naïve Bayes

Independence

$$A \perp\!\!\!\perp B$$

$$P(A|B) = P(A)$$

$$P(B|A) = P(B)$$

$$P(A,B) = P(A)P(B)$$

Conditional independence

$$A \perp\!\!\!\perp B \mid C$$

$$P(A|B,C) = P(A|C)$$

$$P(B|A,C) = P(B|C)$$

$$P(A,B|C) = P(A|C)P(B|C)$$

Conditional Independence

Example: Fire, Smoke, Alarm

Fire and Alarm are independent given Smoke

$F \perp\!\!\!\perp A \mid S$

$$P(\underline{A \mid S}, F) = P(\underline{A \mid S})$$

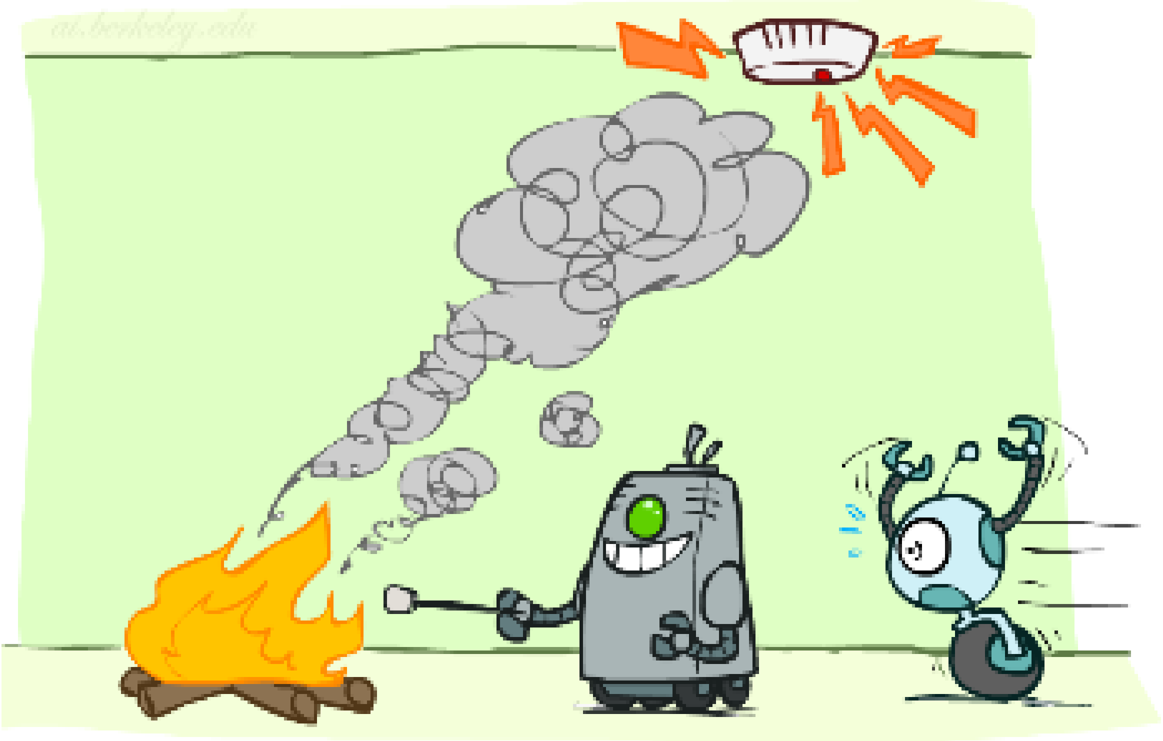


Image: http://ai.berkeley.edu/

Conditional Independence and Naïve Bayes

Independence

$$P(A \mid B) = P(A) \qquad P(A, B) = P(A)P(B)$$
$$P(B \mid A) = P(B)$$

Conditional independence

$$P(A \mid B, C) = P(A \mid C) \qquad P(A, B \mid C) = P(A \mid C)P(B \mid C)$$
$$P(B \mid A, C) = P(B \mid C)$$

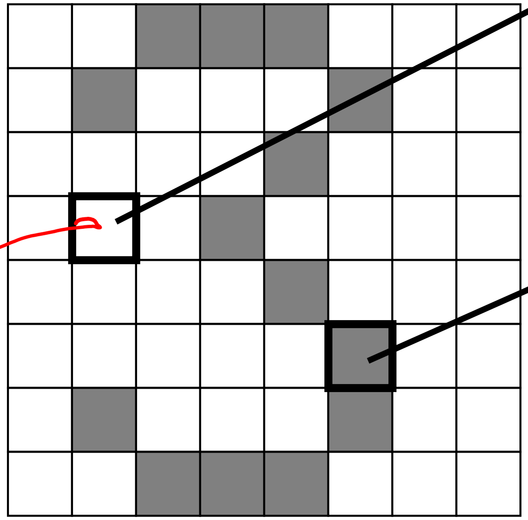
Naïve Bayes assumption

All X_m are cond. indep. give Y

$$P(X_1, X_2, \ldots, X_M \mid Y) = \prod_{m=1}^{M} P(X_m \mid Y)$$

Naïve Bayes for Digits

y	$P(Y)$
1	0.1
2	0.1
3	0.1
4	0.1
5	0.1
6	0.1
7	0.1
8	0.1
9	0.1
0	0.1



y	$P(X_{3,1} = 1 \mid y)$
1	0.01
2	0.05
3	0.05
4	0.30
5	0.80
6	0.90
7	0.05
8	0.60
9	0.50
0	0.80

y	$P(X_{5,5} = 1 \mid y)$
1	0.05
2	0.01
3	0.90
4	0.80
5	0.90
6	0.90
7	0.25
8	0.85
9	0.60
0	0.80

SPAM Classification

Breakout room exercise

https://tinyurl.com/301601spam

SPAM classification

- Y: Binary random variable

  Document is SPAM ($Y = 1$) or not ($Y = 0$)
- X_m: Binary random variable

  Word m appears in document ($X_m = 1$) or not ($X_m = 0$)

Page 1: Estimate parameters from data

Page 2: Calculate probabilities of new sentence given page 1 parameters

SPAM Classification

Breakout room exercise

$$P(Y|X) = \frac{\prod P(X_m|Y)P(Y)}{Z} = \frac{P(X,Y)}{\sum_Y P(X,Y)}$$

$$\rightarrow P(Y=0|X) = \frac{P(X, Y=0)}{(P(X, Y=0) + P(X, Y=1))}$$

$$P(Y=0, X_{301}=0, X_{free}=0 \ldots X_{teach}=1 \ldots)$$

$$= P(X_1 \ldots X_m | Y=0) P(Y=0)$$

$$= \prod_m P(X_m|Y) P(Y=0)$$

m	x	$P(X_m = x \mid Y = 0)$	$P(X_m = x \mid Y = 1)$
301	0	1/4	3/3
free	0	3/4	1/3
is	0	4/4	2/3
money	0	4/4	0/3 ←
now	1	1/4	2/3
Pat	1	3/4	1/3
Sir	0	4/4	2/3
teach	1	4/4	1/3
to	0	2/4	2/3
tomorrow	0	4/4	3/3

$P(Y=0)$ $P(Y=1)$

$P(Y = 0, X_1, \ldots, X_M)$	$P(Y = 1, X_1, \ldots, X_M)$
0,01 ←	0

$P(Y = 0 \mid X_1, \ldots, X_M)$	$P(Y = 1 \mid X_1, \ldots, X_M)$
$\frac{0.01}{(0.01+0)} = 1$ ←	$\frac{0}{(0.01+0)} = 0$

Reminder MLE for Bernoulli

Bernoulli distribution:

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

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

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

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

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

$$L(\phi) = \phi^2 \cdot (1 - \phi)^1 \qquad = \phi^{N_{y=1}} (1 - \phi)^{N_{y=0}}$$

Naïve Bayes MLE

$p(y \mid x, \theta)$

$\mathcal{D} = \{y^{(n)}, \boldsymbol{x}^{(n)}\}_{n=1}^{N}$

$y^{(n)} \in \{0,1\}$

$\boldsymbol{x}^{(n)} \in \{0,1\}^{M}$

$\phi \in [0,1]$

$\boldsymbol{\Theta} \in [0,1]^{Mx2}$

$$L(\phi, \boldsymbol{\Theta}) = p(\mathcal{D} \mid \phi, \boldsymbol{\Theta})$$

$$= \Pi_{n=1}^{N} p\left(\mathcal{D}^{(n)} \mid \phi, \boldsymbol{\Theta}\right)$$ i.i.d assumption

$$= \Pi_{n=1}^{N} p\left(y^{(n)}, \boldsymbol{x}^{(n)} \mid \phi, \boldsymbol{\Theta}\right)$$

$$= \Pi_{n=1}^{N} p\left(y^{(n)} \mid \phi\right) p\left(\boldsymbol{x}^{(n)} \mid y^{(n)}, \boldsymbol{\Theta}\right)$$ Generative model

$$= \Pi_{n=1}^{N} p\left(y^{(n)} \mid \phi\right) p\left(x_1^{(n)}, x_2^{(n)}, \dots, x_M^{(n)} \mid y^{(n)}, \boldsymbol{\Theta}\right)$$

$$= \Pi_{n=1}^{N} p\left(y^{(n)} \mid \phi\right) \Pi_{m=1}^{M} p\left(x_m^{(n)} \mid y^{(n)}, \theta_{m,y}\right)$$ Naïve Bayes

Naïve Bayes MLE

$$\mathcal{D} = \{y^{(n)}, \boldsymbol{x}^{(n)}\}_{n=1}^{N}$$
$$y^{(n)} \in \{0,1\}$$
$$\boldsymbol{x}^{(n)} \in \{0,1\}^M$$
$$\phi \in [0,1]$$
$$\boldsymbol{\Theta} \in [0,1]^{Mx2}$$

$$L(\phi, \boldsymbol{\Theta}) = p(\mathcal{D} \mid \phi, \boldsymbol{\Theta})$$

$$= \prod_{n=1}^{N} p\left(\mathcal{D}^{(n)} \mid \phi, \boldsymbol{\Theta}\right)$$ i.i.d assumption

$$= \prod_{n=1}^{N} p\left(y^{(n)}, \boldsymbol{x}^{(n)} \mid \phi, \boldsymbol{\Theta}\right)$$

$$= \prod_{n=1}^{N} p\left(y^{(n)} \mid \phi\right) p\left(\boldsymbol{x}^{(n)} \mid y^{(n)}, \boldsymbol{\Theta}\right)$$ Generative model

$$= \prod_{n=1}^{N} p\left(y^{(n)} \mid \phi\right) p\left(x_1^{(n)}, x_2^{(n)}, \dots, x_M^{(n)} \mid y^{(n)}, \boldsymbol{\Theta}\right)$$

$$= \prod_{n=1}^{N} p\left(y^{(n)} \mid \phi\right) \prod_{m=1}^{M} p\left(x_m^{(n)} \mid y^{(n)}, \theta_{m,y}\right)$$ Naïve Bayes

$$= \prod_{n=1}^{N} \phi^{y^{(n)}} (1-\phi)^{1-y^{(n)}} \prod_{m=1}^{M} \theta_{m,1}^{\mathbb{I}\left(y^{(n)}=1 \,\wedge\, x_m^{(n)}=1\right)} \left(1-\theta_{m,1}\right)^{\mathbb{I}\left(y^{(n)}=1 \,\wedge\, x_m^{(n)}=0\right)} \theta_{m,0}^{\mathbb{I}\left(y^{(n)}=0 \,\wedge\, x_m^{(n)}=1\right)} \left(1-\theta_{m,0}\right)^{\mathbb{I}\left(y^{(n)}=0 \,\wedge\, x_m^{(n)}=0\right)}$$

$$= \phi^{N_{y=1}} (1-\phi)^{N_{y=0}} \prod_{m=1}^{M} \theta_{m,1}^{N_{y=1,x_m=1}} \left(1-\theta_{m,1}\right)^{N_{y=1,x_m=0}} \theta_{m,0}^{N_{y=0,x_m=1}} \left(1-\theta_{m,0}\right)^{N_{y=0,x_m=0}}$$

Generative Models

SPAM:

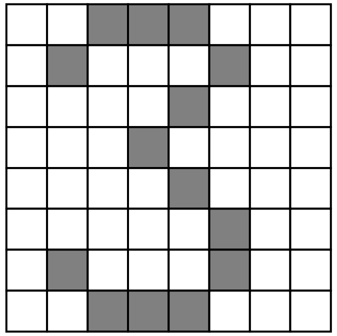
- Class distribution: $Y \sim \ Bern(\phi)$
- Class conditional distribution: $X_m \sim \ Bern(\theta_{m,y})$
- Naïve Bayes X_i conditionally independent X_j given Y for all $i \neq j$

$$p(X_i, X_j \mid Y) \ = \ p(X_i \mid Y)\, p(X_j \mid Y)$$

Digits:

- Class distribution: $Y \sim Categorical(\boldsymbol{\phi})$
- Class conditional distribution: $X_m \sim \ Bern(\theta_{m,y})$
- Naïve Bayes X_i conditionally independent X_j given Y for all $i \neq j$

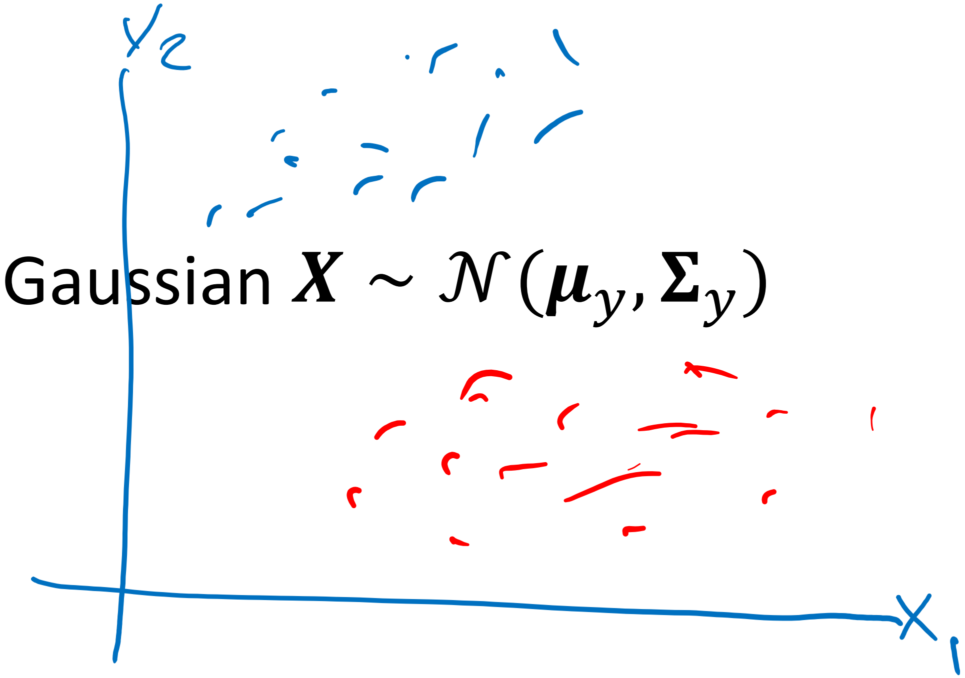
$$p(X_i, X_j \mid Y) \ = \ p(X_i \mid Y)\, p(X_j \mid Y)$$



Generative Models with Continuous Features

Iris dataset:

- Class distribution: $Y \sim \ Bern(\phi)$
- Class conditional distribution: Multivariate Gaussian $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$
- Naïve Bayes assumption?



Piazza Poll 1

Iris dataset:

- Class distribution: $Y \sim \; Bern(\phi)$
- Class conditional distribution: $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$
- Naïve Bayes assumption? $x_1 \perp\!\!\!\perp x_2 \mid y$

Which of the following pairs of Gaussian class conditional distributions satisfy the Naïve Bayes assumptions? Select ALL that apply.

$A.\;\; \boldsymbol{\mu}_{y=0} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma}_{y=0} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \boldsymbol{\mu}_{y=1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma}_{y=1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

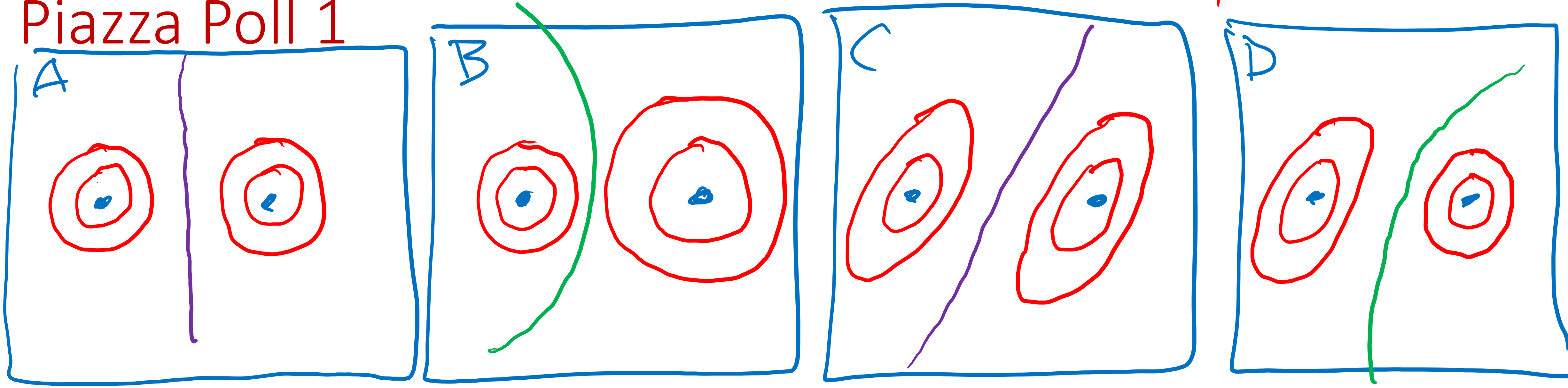
$B.\;\; \boldsymbol{\mu}_{y=0} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma}_{y=0} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \boldsymbol{\mu}_{y=1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma}_{y=1} = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$

$C.\;\; \boldsymbol{\mu}_{y=0} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma}_{y=0} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}, \qquad \boldsymbol{\mu}_{y=1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma}_{y=1} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$

$D.\;\; \boldsymbol{\mu}_{y=0} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma}_{y=0} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}, \qquad \boldsymbol{\mu}_{y=1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma}_{y=1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

Piazza Poll 1

$x_1 \perp\!\!\!\perp x_2 \mid y$



A. $\boldsymbol{\mu}_{y=0} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma}_{y=0} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{\mu}_{y=1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma}_{y=1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

B. $\boldsymbol{\mu}_{y=0} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma}_{y=0} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{\mu}_{y=1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma}_{y=1} = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$

C. $\boldsymbol{\mu}_{y=0} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma}_{y=0} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}, \quad \boldsymbol{\mu}_{y=1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma}_{y=1} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$

D. $\boldsymbol{\mu}_{y=0} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma}_{y=0} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}, \quad \boldsymbol{\mu}_{y=1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma}_{y=1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

Class-conditional Gaussian Distributions

Iris dataset:

- Class distribution: $Y \sim \ Bern(\phi)$ (or Categorical)
- Class conditional distribution: $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$

- Naïve Bayes assumption: diagonal $\Sigma_{y=0}$ $\Sigma_{y=1}$
- Linear Decision Boundary: $\Sigma_{y=0} = \Sigma_{y=1}$
- Quadradic Decision Boundary: $\Sigma_{y=0} \neq \Sigma_{y=1}$

MLE vs MAP vs Generative vs Discriminative

	MLE	MAP
Discriminative	$p(D\|\theta)$ $p(y\|x,\theta)$	$p(D\|\theta)p(\theta)$ $p(y\|x,\theta)p(\theta)$
Generative	$p(D\|\theta)$ $p(x\|y,\theta)p(y\|\theta)$	$p(D\|\theta)p(\theta)$ $p(x\|y,\theta)p(y\|\theta)p(\theta)$