

# CS 360: Machine Learning

Prof. Sara Mathieson

Fall 2019



# Admin

- Lab 4 due **Sunday night** (short Naïve Bayes exercise)
  - In lab: finish Naïve Bayes exercise with partner
  - Begin lab (make sure you can read in the data)
- Midterm next Thursday (short in-lab part + take home)
- Start labs early!
- It is okay not to finish all parts of every lab
- **TODAY**: make sure you have a study guide, Handout 7&8, and reading quiz

# Midterm Notes

- For the in-lab portion, you may bring a one page (front and back) “cheat-sheet”
- Handwritten, created by you
- You may also bring a calculator, but shouldn’t need it
- Take-home portion: open notes/internet, but no discussion with another human!
- Budget around 2-3 hours

# Technology and Justice Keynote: TODAY!

**Keynote:** Vulnerabilities: How Social Media and Data Infrastructure are Exploited for Fun, Profit, and Politics

**Thursday, September 26, 2019**

4:30 p.m.

Sharpless Auditorium

danah boyd is a Principal Researcher at Microsoft Research, the founder and president of Data & Society, and a Visiting Professor at New York University. Her research is focused on addressing social and cultural inequities by understanding the relationship between technology and society.





# Outline for September 26

- Reading Quiz
- Introduction to conditional probability
  - Clinical trials example
- Naïve Bayes
- Confusion matrices

# Outline for September 26

- Reading Quiz
- Introduction to conditional probability
  - Clinical trials example
- Naïve Bayes
- Confusion matrices

# Reading Quiz #4

1. How would you say  $P(A, B)$  in words?
2. Based on class on Tuesday, what is Bayes rule?

$$P(A, B) =$$

3. What is the Naive Bayes assumption? (circle one)
  - (a) The label and the features are independent given the model
  - (b) Features are independent given the label
  - (c) Training examples are independent given their labels
4. If I want to predict the label ( $y$ ) of an example based on its features ( $\vec{x}$ ), which of the following expressions would I want to compute? (circle the best one)
  - (a)  $p(\vec{x}, y)$
  - (b)  $p(\vec{x} \mid y)$
  - (c)  $p(y \mid \vec{x})$

# Reading Quiz #4

1. How would you say  $P(A, B)$  in words?

Probability of A and B

2. Based on class on Tuesday, what is Bayes rule?

$$P(A, B) =$$

3. What is the Naive Bayes assumption? (circle one)

- (a) The label and the features are independent given the model
- (b) Features are independent given the label
- (c) Training examples are independent given their labels

4. If I want to predict the label ( $y$ ) of an example based on its features ( $\vec{x}$ ), which of the following expressions would I want to compute? (circle the best one)

- (a)  $p(\vec{x}, y)$
- (b)  $p(\vec{x} \mid y)$
- (c)  $p(y \mid \vec{x})$

# Reading Quiz #4

1. How would you say  $P(A, B)$  in words?

Probability of A and B

2. Based on class on Tuesday, what is Bayes rule?

$$P(A, B) = P(A) P(B|A) \quad \text{or} \quad P(B) P(A|B)$$

3. What is the Naive Bayes assumption? (circle one)

- (a) The label and the features are independent given the model
- (b) Features are independent given the label
- (c) Training examples are independent given their labels

4. If I want to predict the label ( $y$ ) of an example based on its features ( $\vec{x}$ ), which of the following expressions would I want to compute? (circle the best one)

- (a)  $p(\vec{x}, y)$
- (b)  $p(\vec{x} \mid y)$
- (c)  $p(y \mid \vec{x})$

# Reading Quiz #4

1. How would you say  $P(A, B)$  in words?

Probability of A and B

2. Based on class on Tuesday, what is Bayes rule?

$$P(A, B) = P(A) P(B|A) \quad \text{or} \quad P(B) P(A|B)$$

3. What is the Naive Bayes assumption? (circle one)

(a) The label and the features are independent given the model

(b) Features are independent given the label

(c) Training examples are independent given their labels

4. If I want to predict the label ( $y$ ) of an example based on its features ( $\vec{x}$ ), which of the following expressions would I want to compute? (circle the best one)

(a)  $p(\vec{x}, y)$

(b)  $p(\vec{x} \mid y)$

(c)  $p(y \mid \vec{x})$

# Reading Quiz #4

1. How would you say  $P(A, B)$  in words?

Probability of A and B

2. Based on class on Tuesday, what is Bayes rule?

$$P(A, B) = P(A) P(B|A) \quad \text{or} \quad P(B) P(A|B)$$

3. What is the Naive Bayes assumption? (circle one)

(a) The label and the features are independent given the model

(b) Features are independent given the label

(c) Training examples are independent given their labels

4. If I want to predict the label ( $y$ ) of an example based on its features ( $\vec{x}$ ), which of the following expressions would I want to compute? (circle the best one)

(a)  $p(\vec{x}, y)$

(b)  $p(\vec{x} \mid y)$

(c)  $p(y \mid \vec{x})$

# Outline for September 26

- Reading Quiz
- Introduction to conditional probability
  - Clinical trials example
- Naïve Bayes
- Confusion matrices



# Probability

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

## Bayes Rule

(statement about  $P(A, B)$ )

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

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

Independence:  $P(A, B) = P(A)P(B)$   
independent

Apply Bayes Rule many times:  
(“chain rule” of probability)

$$P(A, B, C) = P(C) P(\underline{A, B} | \underline{C})$$

$$y = P(\underline{C}) P(\underline{B|C}) P(\underline{A|B, C})$$

## Conditional Independence

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

"A is independent of B given C"

↑                      ↑                      ↑

thunder                  rain                  lightning

A " " " C " B

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



→ if conditional independence  
⇒ get rid of  $\bar{B}$

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

## Marginal Probability

$$P(A) = \sum_{b \in \text{vals}(B)} P(A, B=b)$$

$$P(U) = P(U, \text{rain}) + P(U, \text{sun})$$

↑  
umbrella

Only 2 weathers.

$$P(\text{spam} | \text{words}) = \frac{P(\text{spam}) P(\text{words} | \text{spam})}{P(\text{words})}$$

"posterior"      "prior"      "likelihood"      "evidence"      "normalizer"

$$P(\text{words}) = P(\text{spam}, \text{words}) + P(\bar{\text{spam}}, \text{words})$$

↓

$$P(\text{spam}) P(\text{words} | \text{spam}) + P(\bar{\text{spam}}) P(\cdot | \cdot)$$

↑  
"not spam"

$$P(S|w) = \frac{A}{A+B}$$

$$P(\bar{S}|w) = \frac{B}{A+B}$$

} sum = 1



# Handout 7

$$P(D|pos) = \frac{P(D)P(pos|D)}{P(H, pos) + P(D, pos)}$$

$$P(H)P(pos|H) + P(D)P(pos|D)$$

$$= \frac{1}{100} \cdot \frac{9}{10}$$

$$= \frac{\frac{99}{100} \cdot \frac{1}{10} + \frac{1}{100} \cdot \frac{9}{10}}{\frac{9}{108} = \frac{1}{12} \approx 0.0833 \approx 8\%}$$

prob  
of jth  
feature having  
value v,  
& label k

$$\ominus_{j,v,k} =$$

$$\frac{N_{j,v,k} + 1}{N_k + |f_j|}$$

$$= \frac{P(x_j=v, y=k) + 1}{P(y=k) + |f_j|}$$

# of  
values of  
feature j

# Outline for September 26

- Reading Quiz
- Introduction to conditional probability
  - Clinical trials example
- **Naïve Bayes**
- Confusion matrices

# Components of a Bayesian Model

- Identify the evidence, prior, posterior, and likelihood in the equation below

$$p(y = k|\mathbf{x}) = \frac{p(y = k)p(\mathbf{x}|y = k)}{p(\mathbf{x})}$$

# Components of a Bayesian Model

- Identify the evidence, prior, posterior, and likelihood in the equation below

$$p(y = k|\mathbf{x}) = \frac{p(y = k)p(\mathbf{x}|y = k)}{p(\mathbf{x})}$$

- Evidence:** this is the data (features) we actually observe, which we think will help us predict the outcome we're interested in

# Components of a Bayesian Model

- Identify the evidence, prior, posterior, and likelihood in the equation below

$$p(y = k|\mathbf{x}) = \frac{p(y = k)p(\mathbf{x}|y = k)}{p(\mathbf{x})}$$

- Prior:** without seeing any evidence (data), what is our prior believe about each outcome (intuition: what is the outcome in the population as a whole?)

# Components of a Bayesian Model

- Identify the evidence, prior, **posterior**, and likelihood in the equation below

$$p(y = k|\mathbf{x}) = \frac{p(y = k)p(\mathbf{x}|y = k)}{p(\mathbf{x})}$$

- Posterior**: this is the quantity we are actually interested in. *\*Given\** the evidence, what is the probability of the outcome?



# Components of a Bayesian Model

- Identify the evidence, prior, posterior, and likelihood in the equation below

$$p(y = k|\mathbf{x}) = \frac{p(y = k)p(\mathbf{x}|y = k)}{p(\mathbf{x})}$$

- Likelihood**: given an outcome, what is the probability of observing this set of features?

# Naive Bayes

$\vec{x}$ : features,  $y$ : label,  $y \in \{1, 2, 3 \dots K\}$  multi-class prediction.

$$\underbrace{P(y=k|\vec{x})}_{\text{posterior}} = \underbrace{P(y=k)}_{\text{prior}} \underbrace{\frac{P(\vec{x}|y=k)}{P(\vec{x})}}_{\text{evidence}} \text{likelihood.}$$

$S = \{\text{red, blue, green}\}$

Naive Bayes  
assumption.  $|S|=3$

$$P(\vec{x}|y=k) = P(\overset{\text{"A"}}{\underbrace{x_1, x_2, \dots, x_p}} | y=k)$$

$$= P(\overset{\text{B}}{\underbrace{x_2, x_3, \dots, x_p}} | y=k) P(\overset{\text{A}}{\underbrace{x_1}} | \overset{\text{B}}{\underbrace{x_2, x_3, \dots, x_p}}, y=k)$$

$$= P(x_3 \dots x_p | y=k) P(\cancel{x_2} | \cancel{x_3} \dots \cancel{x_p}, y=k) P(\cancel{x_1} | \cancel{x_2} \dots \cancel{x_p}, y=k)$$



## Naive Bayes model

$$p(y=k|\vec{x}) \propto p(y=k) \prod_{j=1}^P p(x_j|y=k)$$

↑  
proportional to

Prediction

$$\hat{y} = \operatorname{argmax}_{k \in \{1, \dots, K\}} \left[ p(y=k) \prod_{j=1}^P p(x_j|y=k) \right]$$

$$\operatorname{argmax}_{k \in \{1, 2, 3\}} \{0.5, 0.2, 0.8\}$$

$$\Rightarrow \boxed{\hat{y} = 3}$$

$$\Theta_k = \frac{N_k + 1}{n + K}$$

# of classes

How to estimate probabilities?

$\Theta_k$  = my approximation of  $p(y=k)$

$N_k$  = # of training examples with label  $k$ .

$$\Rightarrow \Theta_k = \frac{N_k}{n} \quad \Theta_1 = \frac{3}{7}$$

Laplace counts.

# Outline for September 26

- Reading Quiz
- Introduction to conditional probability
  - Clinical trials example
- Naïve Bayes
- Confusion matrices



# Confusion Matrix

true

	1	2
1	7	5
2	5	5

rows sum to 1

1	0
$\frac{1}{2}$	$\frac{1}{2}$

$$\text{accuracy} = \frac{80}{85} \star$$

$\Theta_{j,v,k}$  is estimating  $P(x_j=v|y=k)$

pred class

true class

	1	2	...	K
1	11			
2	11	1		
...				
K				

correct!

true	pred
Y	Y
✓ 2	1
✓ 1	1
✓ 1	1
✓ 2	2
✓ 2	1

logs!

$$\log(AB) = \log(A) + \log(B)$$

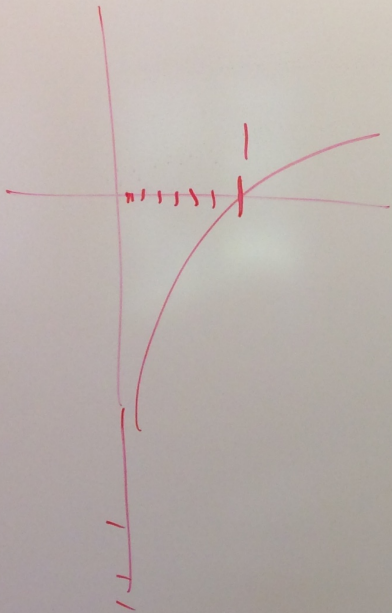
$$\Theta_k = \frac{N_{k+1}}{n+K}$$

$$\log(\Theta_k) = \log(N_{k+1}) - \log(n+K)$$

$P(\text{outlook}=\text{sun} | \text{tennis}=\text{yes})$



$$\log\left(\frac{A}{B}\right) = \log A - \log B$$



$$p(y=k) \prod_{j=1}^n p(x_j | y=k)$$

Tech  
TO  
danah



# Computer Science

Confusion Matrix

	predicted label		
	1	2	3
true label	1	11	
	2	1	1
	3	1	1

	true	pred
✓	Y	$\hat{Y}$
✓	2	3
✓	1	2
✓	3	1
✓	3	3
✓	2	2
✓	1	2
✓	⋮	⋮

	P	
	1	2
T	75	0
	5	5

75 with label 1  
10 with label 2

accuracy

$$\frac{80}{85} \star$$

Normalize.

$\frac{75}{85} = 1$	$\frac{0}{85} = 0$
$\frac{5}{85} = \frac{1}{17}$	$\frac{5}{85} = \frac{1}{17}$

rows sum to 1

Tech & Justice  
TODAY 4:30  
danah boyd



$$p(y=1|\vec{x}_{\text{test}}) \propto \frac{4}{75}$$

$$p(y=2|\vec{x}_{\text{test}}) \propto \frac{5}{54}$$

$$\log \left( p(y=k) \prod_{j=1}^P p(x_j | y=k) \right)$$

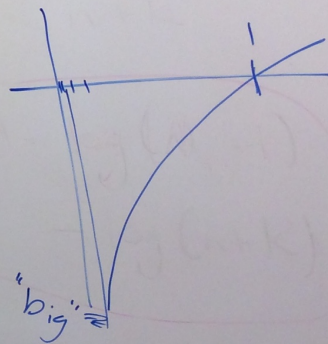
↑  
known value v

$$\log(p(y=k)) + \sum_{j=1}^P \log(p(x_j | y=k))$$

$$\log \Theta_k = \log \left( \frac{N_k + 1}{n + K} \right)$$

$$= \log(N_k + 1) - \log(n + K)$$

$$\log \Theta_{j,v,k}$$



"j"

$$P(\text{outlook} = \text{sun} | \text{tennis} = \text{yes}) \approx P(\underbrace{x_j = v}_{\text{testing}} | y = k)$$

params

$s_1 = 1$   $s_2 = 8$   $s_3 = 5(-)$   
 $s(2, -5)$   
 $f(s)$   
 $f_2(s, t)$   
 $f_2(s_1, t_1)$   
 $f_2(s_1, s_1)$

things available - hypothesis space

M.D. 4d

view 1.2 map examples