CS 260: Foundations of Data Science

Prof. Thao Nguyen

Fall 2024

HAVERFORD COLLEGE

Materials by Sara Mathieson

Admin

- **Midterm 1**
  - due at the *beginning* of class on Wednesday
- **Lab 4** grades & feedback posted on Moodle
- **Lab 5** posted
  - due Monday after fall break (Oct 21)

Midterm 1 Notes

- Timed exam: **3 hour limit**. DO NOT open the exam until you are ready to take it for 3 hours!
- You may use one letter page (front and back) "study sheet", handwritten, created by you
- Outside of your "study sheet" and calculator, **no other notes or resources**
- As per the Honor Code, all work must be your own

Informal Quiz (discuss with a partner)

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

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

$$P(A, B) =$$

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

Informal Quiz (discuss with a partner)

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

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

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

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 today

- Intro to Bayesian models
- Naïve Bayes algorithm

Outline for today

- Intro to Bayesian models
- Naïve Bayes algorithm

Components of a Bayesian Model

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

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

Components of a Bayesian Model

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

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

- **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

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

- **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

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

- **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

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

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

Examples

- Computing the probability an email message is **spam**, given the **words** of the email

- Another example: what is the probability of **Trisomy 21** (Down Syndrome), given the amount of sequencing of each chromosome?

Bayesian Model for Trisomy 21 (T_{21})

Input data are read counts for each chromosome (1,2,…,n):

$$q_1, q_2, \cdots, q_n = \vec{q}$$

Bayesian Model for Trisomy 21 (T_{21})

Input data are read counts for each chromosome (1,2,…,n):

$$q_1, q_2, \cdots, q_n = \vec{q}$$

Goal:

$$\mathbb{P}(T_{21}|\vec{q}\,) = \frac{\mathbb{P}(\vec{q}\,|T_{21}) \cdot \mathbb{P}(T_{21})}{\mathbb{P}(\vec{q}\,)}$$

$$= \frac{\mathbb{P}(\vec{q}\,|T_{21}) \cdot \mathbb{P}(T_{21})}{\mathbb{P}(\vec{q}\,|T_{21}) \cdot \mathbb{P}(T_{21}) + \mathbb{P}(\vec{q}\,|T_{21}^C) \cdot \mathbb{P}(T_{21}^C)}$$

Bayesian Model for Trisomy 21 (T_{21})

Input data are read counts for each chromosome (1,2,…,n):

$$q_1, q_2, \cdots, q_n = \vec{q}$$

Goal:

Prior probability of T_{21}

$$\mathbb{P}(T_{21}|\vec{q}\,) = \frac{\mathbb{P}(\vec{q}\,|T_{21}) \cdot \mathbb{P}(T_{21})}{\mathbb{P}(\vec{q}\,)}$$

$$= \frac{\mathbb{P}(\vec{q}\,|T_{21}) \cdot \mathbb{P}(T_{21})}{\mathbb{P}(\vec{q}\,|T_{21}) \cdot \mathbb{P}(T_{21}) + \mathbb{P}(\vec{q}\,|T_{21}^C) \cdot \mathbb{P}(T_{21}^C)}$$

Prior:

$P(T_{21})$

Maternal Age	Trisomy 21	All Trisomies
20	1 in 1,667	1 in 526
21	1 in 1,429	1 in 526
22	1 in 1,429	1 in 500
23	1 in 1,429	1 in 500
24	1 in 1,250	1 in 476
25	1 in 1,250	1 in 476
26	1 in 1,176	1 in 476
27	1 in 1,111	1 in 455
28	1 in 1,053	1 in 435
29	1 in 1,000	1 in 417
30	1 in 952	1 in 384
31	1 in 909	1 in 384
32	1 in 769	1 in 323
33	1 in 625	1 in 286
34	1 in 500	1 in 238
35	1 in 385	1 in 192
36	1 in 294	1 in 156
37	1 in 227	1 in 127
38	1 in 175	1 in 102
39	1 in 137	1 in 83
40	1 in 106	1 in 66
41	1 in 82	1 in 53
42	1 in 64	1 in 42
43	1 in 50	1 in 33
44	1 in 38	1 in 26
45	1 in 30	1 in 21
46	1 in 23	1 in 16
47	1 in 18	1 in 13
48	1 in 14	1 in 10
49	1 in 11	1 in 8

Outline for today

- Recap Bayesian models
- Naïve Bayes algorithm

Real-world example of Naïve Bayes

“A Comparison of Event Models for Naive Bayes Text Classification” (5649 citations!)

http://www.kamalnigam.com/papers/multinomial-aaaiws98.pdf

Goal: text classification (classify documents into topics based on the words as features)

Real-world example of Naïve Bayes

- Single document $\vec{x} = [x_1, x_2, \dots, x_p]^T$
- Multi-class response $y \in \{1, 2, \dots, K\}$

- Goal: Classification $\hat{y} = argmax_{k=1,\dots,K}\, p(y = k|\vec{x})$

Bayesian Model

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

can ignore

Naïve Bayes example

$$p(\vec{x}|y=k) = p(\underbrace{x_1}_{A}, \underbrace{x_2, x_3, \dots, x_p}_{B}|y=k)$$

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

$$= p(\overbrace{\underbrace{x_2}_{C}, \underbrace{x_3, \dots, x_p}_{D}}^{B}|y=k)\, p(\overbrace{x_1}^{A}|\overbrace{x_2, \dots x_p,}^{B} y=k)$$

$$= p(x_3, \dots, x_p|y=k)\, p(x_2|x_3, \dots, x_p, y=k)$$
$$p(x_1|x_2, \dots x_p, y=k)$$

Naïve Bayes assumption

Conditional Independence: "feature j is independent from all other features given label k"

$$p(x_1, x_2|y) = p(x_1|y)p(x_2|x_1, y)$$

x_1 = 4 legs

x_2 = fur

y = cat

assume $p(x_2|x_1, y) = p(x_2|y)$

$$\Rightarrow p(x_1, x_2|y) = p(x_1|y)p(x_2|y)$$

Naïve Bayes example

$$p(\vec{x}|y=k) = p(x_p|y=k)p(x_{p-1}|y=k)\dots p(x_2|y=k)\,p(x_1|y=k)$$

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

Naïve Bayes Model

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

↑ proportional to

Obtaining p(y=k) & p(x_j|y=k)

Estimate based on training data

- θ_k = estimate for p(y=k)
- $\theta_{k,j,v}$ = estimate for p(x_j=v|y=k)

feature j value v class k

Let N_k = # of examples with label k, we could define $\theta_k = \frac{N_k}{n}$

What happens if N_k = 0?

Laplace smoothing

- Technique to handle zero probability
- $\theta_k = \frac{N_k+1}{n+K}; \quad \sum \theta_k = \sum \frac{N_k+1}{n+K} = \frac{1}{n+K}(n+K)$

- Similarly, let $N_{k,j,v}$ = # of examples with feature j = value v and class label k

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

# of feature values for feature j

Handout 11

Say we have two tests for a specific disease. Each test (features f_1, f_2) can come back either positive "pos" or negative "neg", and the true underlying condition of the patient is represented by y ($y = 1$ is "healthy" and $y = 2$ is "disease"). We observe this training data where $n = 7$ and $p = 2$:

$\boldsymbol{x}$	f_1	f_2	y
$\boldsymbol{x}_1$	pos	neg	1
$\boldsymbol{x}_2$	pos	pos	2
$\boldsymbol{x}_3$	pos	neg	2
$\boldsymbol{x}_4$	neg	neg	1
$\boldsymbol{x}_5$	pos	neg	2
$\boldsymbol{x}_6$	neg	neg	1
$\boldsymbol{x}_7$	neg	pos	2

1. To estimate the probability $p(y = k)$, for $k = 1, 2, \cdots, K$, we will use the formula:

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

   where N_k is the count ("Number") of data points where $y = k$. Compute θ_1 and θ_2. What would θ_1 and θ_2 be if we in fact had *no* training data?

$\vec{x}$	f_1	f_2	y
$\vec{x}_1$	pos	neg	1
$\vec{x}_2$	pos	pos	2
$\vec{x}_3$	pos	neg	2
$\vec{x}_4$	neg	neg	1
$\vec{x}_5$	pos	neg	2
$\vec{x}_6$	neg	neg	1
$\vec{x}_7$	neg	pos	2

$$\Theta_1 = \frac{3+1}{7+2}$$

$$\frac{4}{9}$$

$$\Theta_2 = \frac{5}{9}$$

Handout 11

2. To estimate the probabilities $p(x_j = v|y = k)$ for all features j, values v, and class label k, we will use the formula:

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

where $N_{k,j,v}$ is the count of data points where $y = k$ and $x_j = v$, and $|f_j|$ is the number of possible values that f_j (feature j) can take on. Fill in the following tables with these θ values.

$y = 1$	pos	neg
f_1		
f_2		

$y = 2$	pos	neg
f_1		
f_2		

f_1	f_2	y
p	n	1
p	p	2
p	n	2
n	n	1
p	n	2
n	n	1
n	p	2

X

likelihood

$P(\vec{x} \mid y=1)$

feature values

y=1	p	n	
f_1	$\frac{1+1}{3+2}$	$\frac{2+1}{3+2}$	$\frac{3}{5}$
f_2	$\frac{0+1}{3+2}$	$\frac{3+1}{3+2}$	
	$\frac{1}{5}$		

names

y=2	p	n
f_1	$\frac{4}{6}$	$\frac{2}{6}$
f_2	$\frac{3}{6}$	$\frac{3}{6}$

prior

$$\theta_1 = \frac{3+1}{7+2} = \frac{4}{9}$$

$$\theta_2 = \frac{4+1}{7+2} = \frac{5}{9}$$

add to 1