

CS 260: Foundations of Data Science

Prof. Sara Mathieson

Spring 2025



HAVERFORD
COLLEGE

Admin

- **Lab 2** grades posted
- In lab today: check-ins about **Lab 4**
- **This week**: Naïve Bayes
- **Next week**: review
- Tuesday March 4 in-class: **Midterm 1**

Midterm 1 Notes

- In-class exam (85 min in this room)
- You may use a one page (front and back) “study sheet”, handwritten, created by you
- Outside of your “study sheet”, **no other notes or resources**
- As per the Honor Code, all work must be your own

Outline

- Intro to Bayesian models
- Naïve Bayes algorithm
- Handout 9

Outline

- Intro to Bayesian models
- Naïve Bayes algorithm
- Handout 9

Informal Quiz (discuss with a partner)

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

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

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

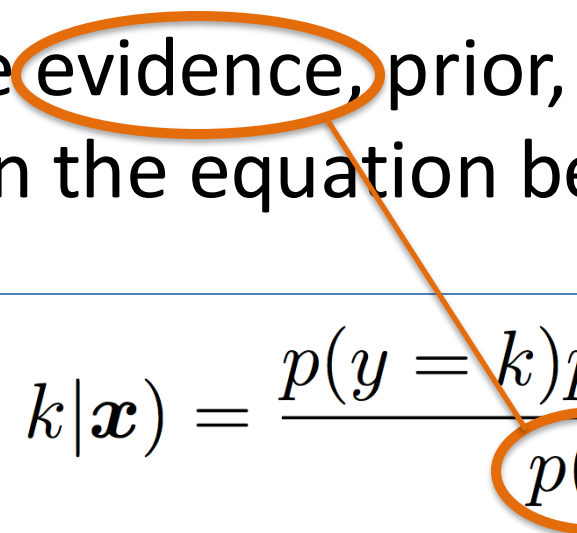
Components of a Bayesian Model

- 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

$$p(y = k|\mathbf{x}) = \frac{p(y = k)p(\mathbf{x}|y = k)}{p(\mathbf{x})}$$
A blue rectangular box contains the equation. An orange oval highlights the term $p(\mathbf{x})$ in the denominator. An orange line connects this oval to the word 'evidence' in the text above.

- 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?

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):

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

Bayesian Model for Trisomy 21 (T_{21})

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

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

Goal:

$$\begin{aligned}\mathbb{P}(T_{21}|\vec{q}) &= \frac{\mathbb{P}(\vec{q} | T_{21}) \cdot \mathbb{P}(T_{21})}{\mathbb{P}(\vec{q})} \\ &= \frac{\mathbb{P}(\vec{q} | T_{21}) \cdot \mathbb{P}(T_{21})}{\mathbb{P}(\vec{q} | T_{21}) \cdot \mathbb{P}(T_{21}) + \mathbb{P}(\vec{q} | T_{21}^C) \cdot \mathbb{P}(T_{21}^C)}\end{aligned}$$

Bayesian Model for Trisomy 21 (T_{21})

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

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

Goal:

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

Prior probability of T_{21}

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

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

- Intro to Bayesian models
- Naïve Bayes algorithm
- Handout 9

Real-world example of Naïve Bayes

“A Comparison of Event Models for Naive Bayes Text Classification” (6000+ citations!)

<http://www.cs.cmu.edu/~knigam/papers/multinomial-aaaiws98.pdf>

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

95 topics (i.e. $K=95$)

Naive Bayes

single example: $\vec{x} = [x_1, x_2, \dots, x_p]^T$ ^{i.e. word counts}

goal multi-class classification

label: $y \in \{1, 2, \dots, K\}$

code: $\{0, 1, 2, \dots, K-1\}$

idea Bayesian model

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

prediction

compute
for $k=1, 2 \dots K$

$$\hat{y} = \underset{k=1, 2 \dots K}{\operatorname{argmax}} p(y=k | \vec{x})$$

python

$$\operatorname{argmax} \{ [7, 2, \boxed{11}, 5, 8] \}$$

$$\operatorname{argmax} = 2$$

$$\boxed{K=3}$$

likelihood

$$p(y=1 | \vec{x}) = 0.3$$

$$p(y=2 | \vec{x}) = 0.1$$

$$p(y=3 | \vec{x}) = 0.6$$

$$\hat{y} = 3$$

(math)

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

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

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

assume assume

Naive
Bayes
assumption

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

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

Product, (like \sum for sum)

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

Naive Bayes assumption
feature x_j is independent
of all other features
given class label k

Naive Bayes model

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

proportional
to

~~$p(\vec{x})$~~

$$\hat{y} = \underset{k=1 \dots K}{\operatorname{argmax}}$$

Conditional independence example

Conditional independence

$$p(x_1, x_2 | y) = p(x_1 | y) p(x_2 | \cancel{x_1}, y)$$

$x_1 = 4 \text{ legs}$

$x_2 = \text{fur}$

$y = \text{cat}$

$$\Downarrow \quad \Downarrow \text{ assume} \\ p(x_1 | y) \cdot P(x_2 | y)$$

What are $p(y=k)$ & $p(x_j | y=k)$?

estimate based on training data

(A) θ_k = estimate of $p(y=k)$, prior

(B) $\theta_{k,j,v}$ = estimate of $p(x_j=v | y=k)$

class \downarrow
feature \uparrow value \uparrow

ex: $p(x_{\text{outlook}} = \text{Sun} | \text{tennis} = \text{yes})$

$\theta_{\text{yes, outlook, Sun}}$

THEOREM

(A) $N_k = \#$ examples with label k
training

$$\Theta_k = \frac{N_k + 1}{n + K} \quad \text{Laplace counts}$$

total # training data
classes

ex:

1 = healthy

2 = disease

$$N_1 = 875$$

$$N_2 = 125$$

$$n = 1000$$

$$\Theta_1 = \frac{875 + 1}{1000 + 2}$$

$$\Theta_2 = \frac{125 + 1}{1000 + 2}$$

add in just
in case $N_k = 0$

③ $N_{k,j,v}$ = # examples with class k
and $x_j = v$

$$\Theta_{k,j,v} = \frac{N_{k,j,v} + 1}{N_k + |f_j|} \approx \frac{p(x_j=v, y=k)}{p(y=k)} = p(x_j=v | y=k)$$

of values for feature j

$f_j = \{ \text{sun, rain, overcast} \}$

$|f_j| = 3$

magnitude

Outline

- Intro to Bayesian models
- Naïve Bayes algorithm
- Handout 9

Handout 9

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 $p = 2$:

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

Handwritten calculations on a chalkboard:

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

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

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

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

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?

Handout 9

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		

2

$y=1$	pos	neg
f_1	$\frac{1+1}{3+2}$	$\frac{2+1}{3+2}$
f_2	$\frac{0+1}{3+2}$	$\frac{3+1}{3+2}$

$y=2$	pos	neg
f_1	$\frac{4}{6}$	$\frac{2}{6}$
f_2	$\frac{3}{6}$	$\frac{3}{6}$

f_1	f_2	y
pos	neg	1
pos	pos	2
pos	neg	2
neg	neg	1
pos	neg	2
neg	neg	1
neg	pos	2

purpose of Laplace!

no zero probabilities
(because multiplication)

Quote of the week

“The highest form of creativity is creativity under constraint.” –
Unknown