CS 260: Foundations of Data Science

Prof. Sara Mathieson

Fall 2023

HAVERFORD
COLLEGE

Admin

- **Midterm 1**
  - due at the *beginning* of class on Thursday
- **Lab today:** continue midterm review
  - Attendance optional but recommended!
- **Zoom appointment slots** tomorrow (Wed)
  - Email me *today* to arrange
- **Lab 3 grades** should be up tomorrow
- **Lab 5** released tomorrow
  - due Wednesday after fall break

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 a one 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

Feedback forms (thank you!)

General workload/difficult

1 x

2 x

3 xxxxxxxxxxxx

4 xxxxxxxxx

5

Feedback forms (thank you!)

- Different office hours times, zoom?
  - Will try to shift a bit later on Monday and add a Friday zoom when I can
- Diversity of opinions about lab deadlines
  - Will generally try to keep earlier in the week
- More prep for the labs in class
  - For Naïve Bayes we'll discuss implementation
- Half and half more slides vs. more board
- Lab instructions could be clearer

Outline for October 10

- Recap Bayesian models
- Naïve Bayes algorithm
- Thurs: algorithmic bias

Outline for October 10

- Recap Bayesian models

- Naïve Bayes algorithm

- Thurs: algorithmic bias

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?

$P(A, B) =$ 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) $p(y \mid \vec{x})$]

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?

Outline for October 10

- Recap Bayesian models

- Naïve Bayes algorithm

- Thurs: algorithmic bias

Real-world example of Naïve Bayes

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

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

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

Naive Bayes

(word counts)

Single example $\vec{x} = [x_1, x_2 \ldots x_p]^T$

multi-class response $y \in \{1, 2, \ldots, K\}$

$\boxed{\text{code: } 0, 1, \ldots, K-1}$

<u>goal</u>: multi-class classification (i.e. $\hat{y}$)

Bayesian model

$$\overbrace{p(y=k \mid \vec{x})}^{\text{posterior}} = \frac{\overbrace{p(y=k)}^{\text{prior}}\ \overbrace{p(\vec{x} \mid y=k)}^{\star}}{\underbrace{p(\vec{x})}_{\cancel{\text{ignore?}}\ \text{same for all } k}}$$

prediction

$$\boxed{\hat{y} = \underset{k=1,\ldots,K}{\operatorname{argmax}}\ p(y=k \mid \vec{x})}$$

r all

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

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

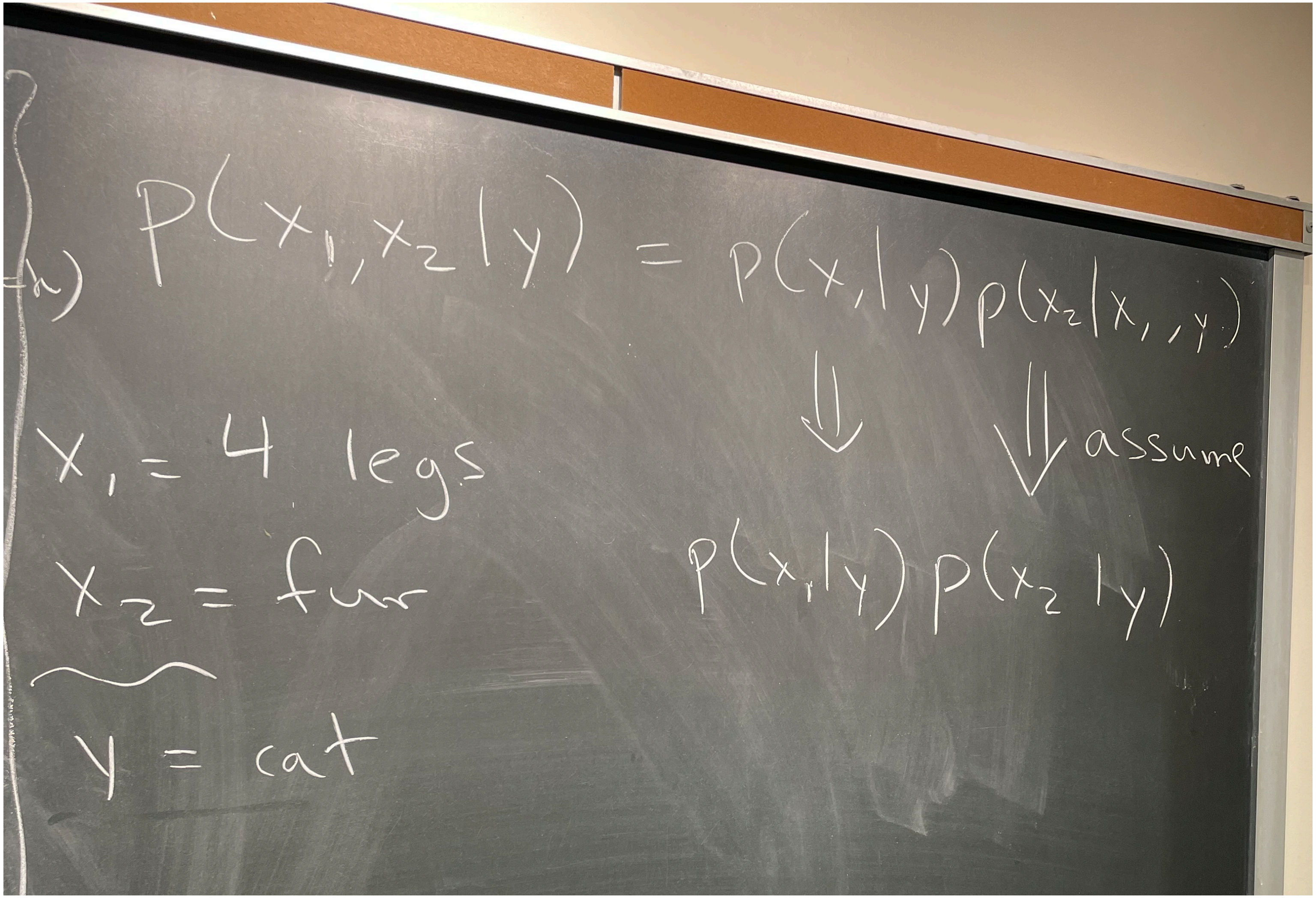
$$= p(\overbrace{\underbrace{x_2}_{C}, \underbrace{x_3 \cdots x_p}_{D}}^{B} \mid y=k)\, p(\overset{A}{x_1} \mid \overbrace{x_2 \cdots x_p}^{B}, y=k)$$

$$= p(x_3, \cdots x_p \mid y=k)\, p(x_2 \mid \cancel{x_3 \cdots x_p}, y=k)\, \underbrace{p(x_1 \mid \cancel{x_2 \cdots x_p}, y=k)}_{\text{assumption} \Rightarrow\ p(x_1 \mid y=k)}$$

conditional indenpendence assumption
(Naive Bayes)

"feature j is independent from all other features <u>given</u> label k"

Conditional independence example



$$p(x_1, x_2 \mid y) = p(x_1 \mid y)\, p(x_2 \mid x_1, y)$$

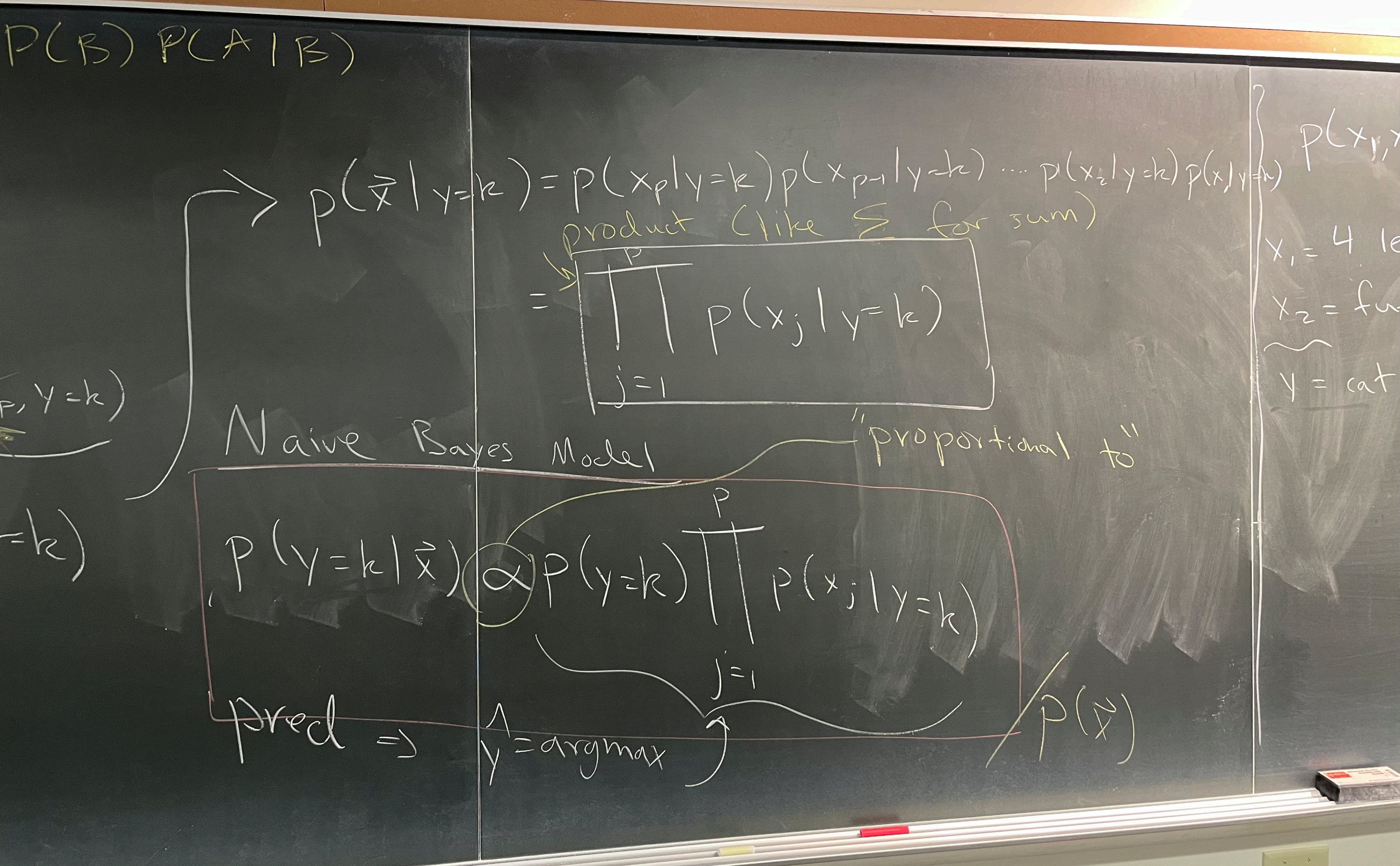
$$\Downarrow \qquad \Downarrow \text{ assume}$$

$$p(x_1 \mid y)\, p(x_2 \mid y)$$

x_1 = 4 legs

x_2 = fur

y = cat



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

$p,\ y=k)$

$=k)$

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

product (like Σ for sum)

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

Naive Bayes Model

"proportional to"

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

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

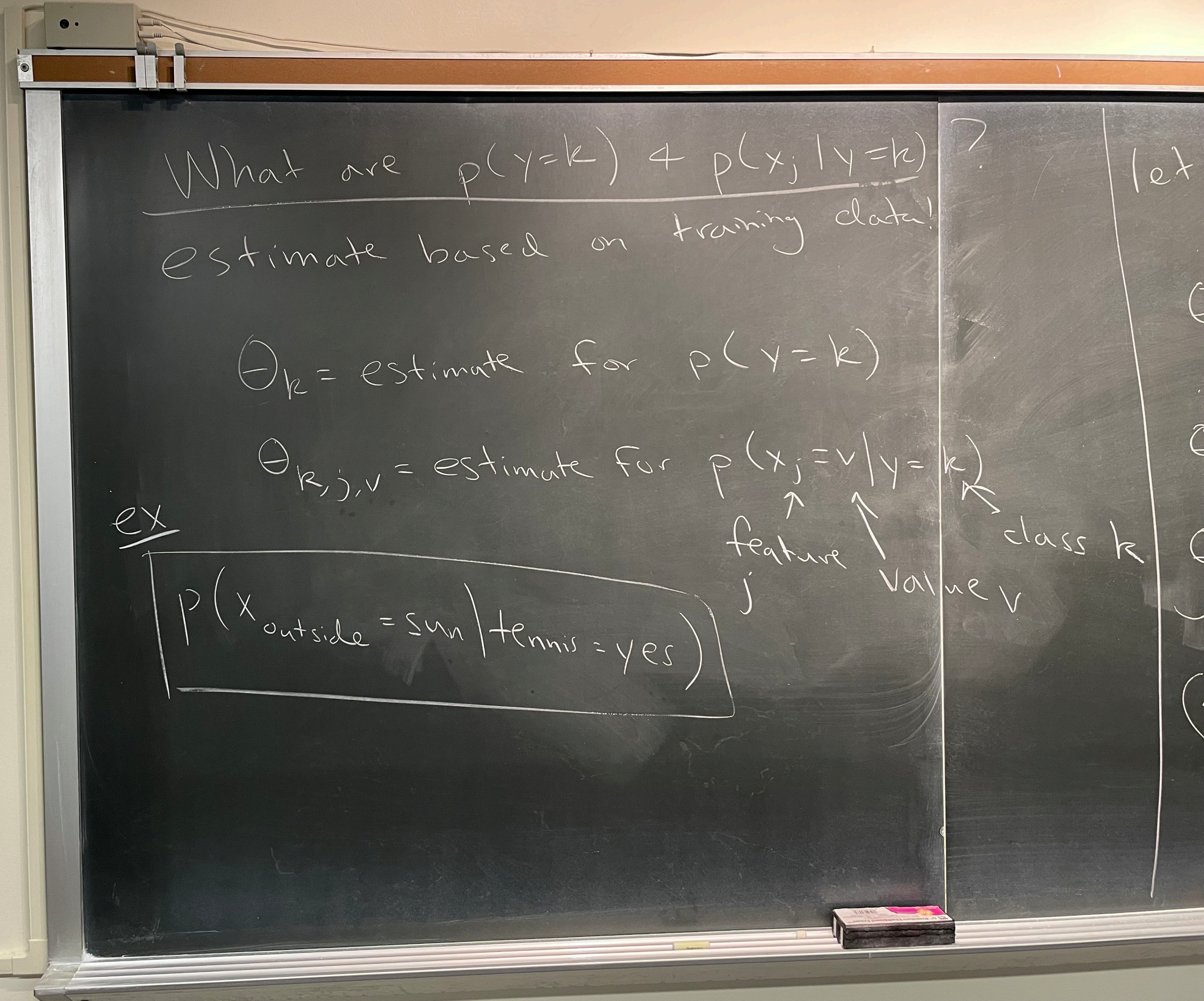
pred ⇒ $\hat{y} = \text{argmax}$

$p(x_1, x$

$x_1 = 4$ le

$x_2 =$ fu

$y =$ cat



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

estimate based on training data!

Θ_k = estimate for $p(y=k)$

$\Theta_{k,j,v}$ = estimate for $p(\underset{\substack{\uparrow\\ \text{feature}\\ j}}{x_j} = \underset{\substack{\uparrow\\ \text{value } v}}{v} \,|\, y = \underset{\substack{\uparrow\\ \text{class } k}}{k})$

ex

$$\boxed{p(x_{outside} = sun \,|\, tennis = yes)}$$

let

s k

let N_k = # examples with label k

$$\theta_k = \frac{N_k}{n}$$

1 = healthy

2 = disease

$$\theta_1 = \frac{875}{1000}$$

$$\theta_2 = \frac{125}{1000}$$

$N_1 = 875$

$N_2 = 125$

$n = 1000$

add to 1

(probability distribution)

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

← LaPlace count

implementation

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

let $N_{kj,v}$ = # of examples with feature j = value v & class label k

A

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

# of featur values for featur j

$N(\text{tennis}=\text{yes}) = 7$

$N_{\text{yes, outlook, sun}} = 4$

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

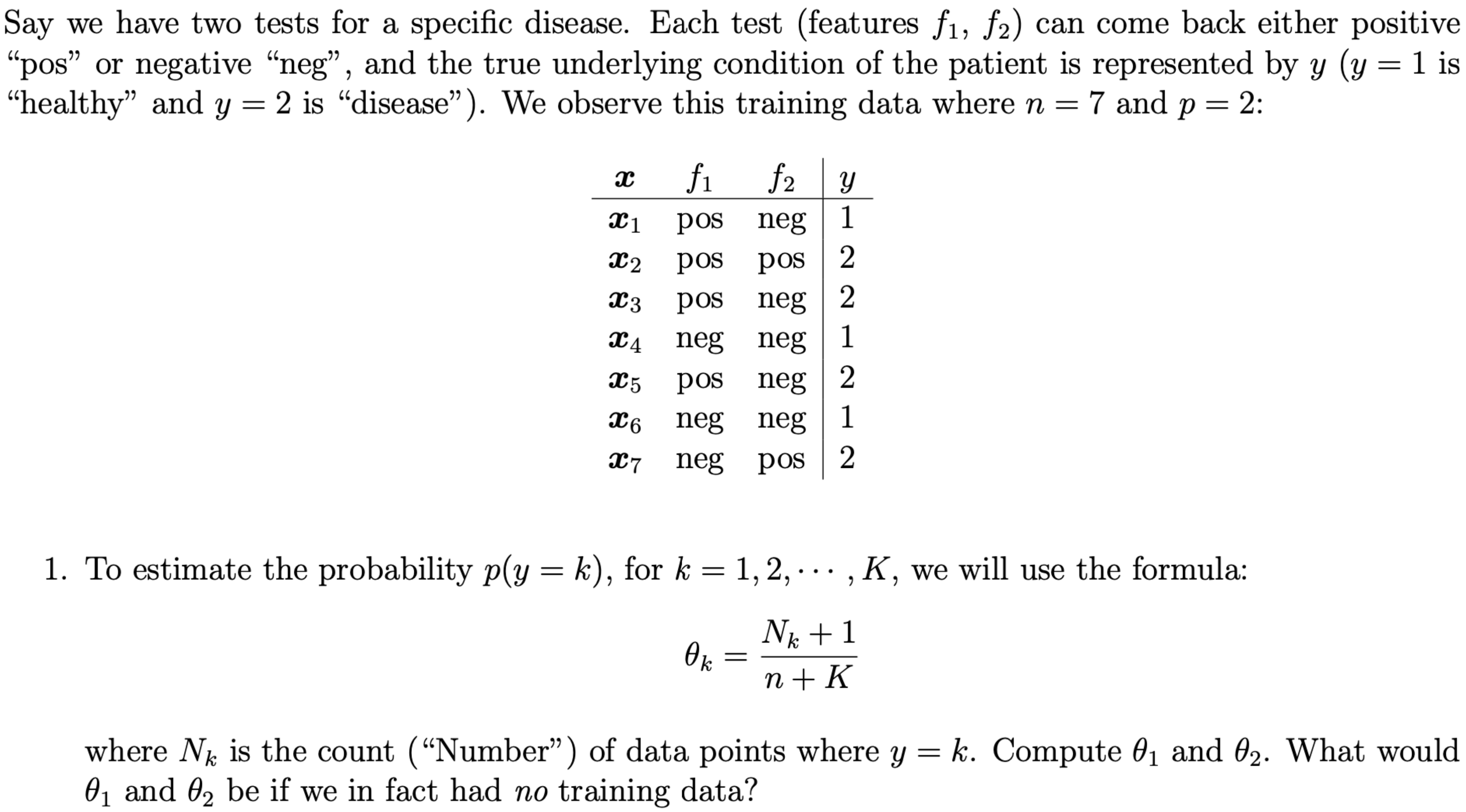
Naive Bayes Model

"proportional to"

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

pred ⇒ $\hat{y} = \text{argmax}$

Handout 11



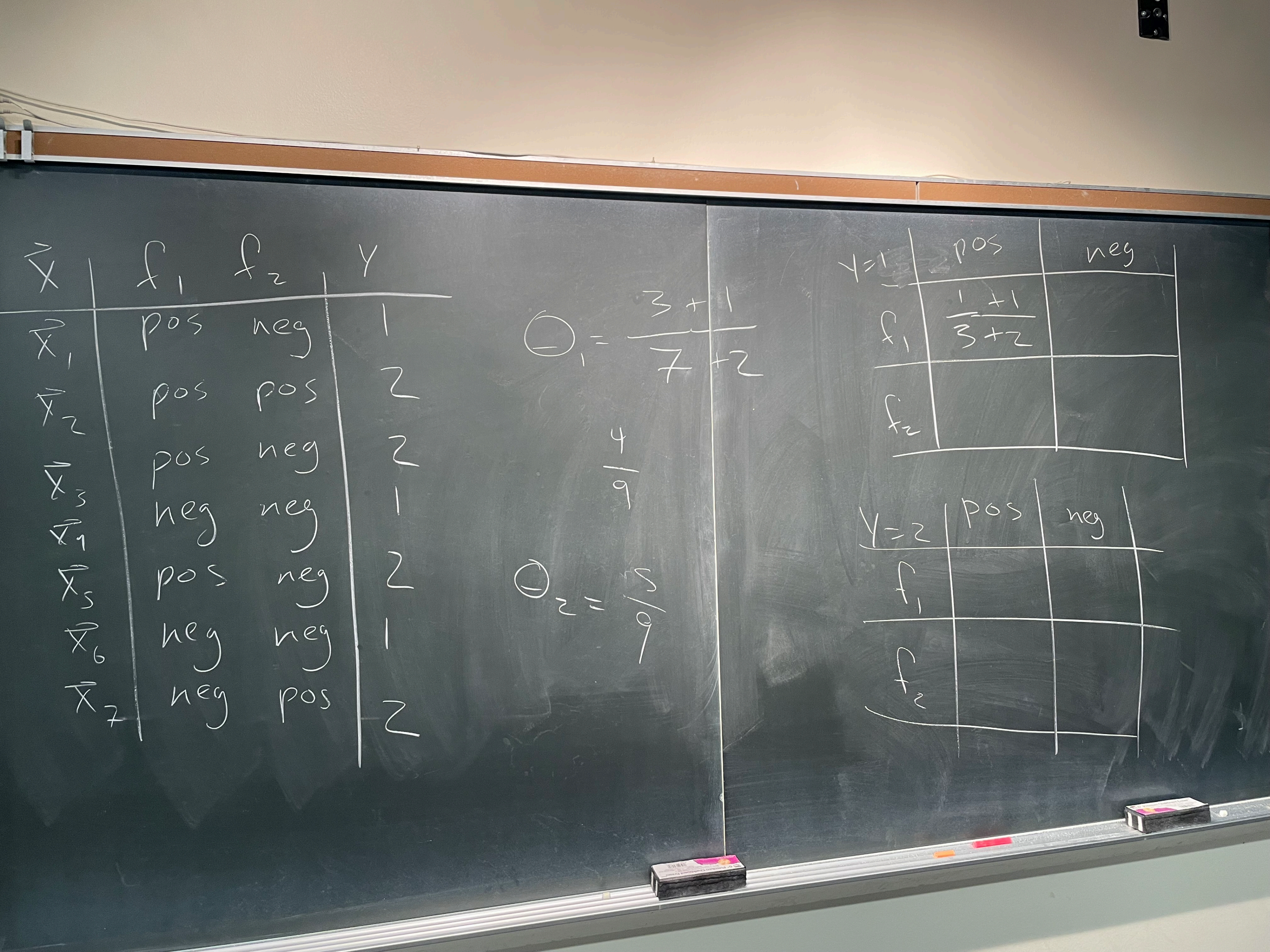
Say we have two tests for a specific disease. Each test (features f_1, f_2) can come back either positive "pos" or negative "neg", and the true underlying condition of the patient is represented by y ($y = 1$ is "healthy" and $y = 2$ is "disease"). We observe this training data where $n = 7$ and $p =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

$\boldsymbol{x}$	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

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

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

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

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

Y=2	pos	neg
f_1		
f_2		

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		