

# Lecture 24 – More Naive Bayes



DSC 40A, Winter 2024

# Announcements

- ▶ HW7 due tonight
  - ▶ I recommend you to attempt the second extra credit, so far we only had 11 submissions (including me), could be a good opportunity to earn easy EC.
- ▶ Midterm 2 next Wednesday.
- ▶ The senior capstone showcase is on Friday, March 15th in the Price Center East Ballroom. The DSC seniors will be presenting posters on their capstone projects. Come and ask them questions; if you're a DSC major, this will be your one day!
  - ▶ RSVP at [hdsishowcase.com](https://hdsishowcase.com)

# Agenda

- ▶ Naive Bayes with smoothing.
- ▶ Application — text classification.

# Naive Bayes with smoothing

# Recap: Naive Bayes classifier

*firmness, g-b, class*

- ▶ We want to predict a **class**, given certain **features**.

*ripe/unripe*

- ▶ Using Bayes' theorem, we write

$$P(\text{class}|\text{features}) = \frac{P(\text{class}) \cdot P(\text{features}|\text{class})}{P(\text{features})}$$

*ripe/unripe*

- ▶ For each class, we compute the numerator using the **naive assumption of conditional independence of features given the class**.
- ▶ We estimate each term in the numerator based on the training data.
- ▶ We predict the class with the **largest numerator**.
  - ▶ Works if we have multiple classes, too!

# Example: avocados, again

Proportional  
To

color	softness	variety	ripeness
bright green	firm	Zutano	unripe
green-black	medium	Hass	ripe
purple-black	firm	Hass	ripe
green-black	medium	Hass	unripe
purple-black	soft	Hass	ripe
bright green	firm	Zutano	unripe
green-black	soft	Zutano	ripe
purple-black	soft	Hass	ripe
green-black	soft	Zutano	ripe
green-black	firm	Hass	unripe
purple-black	medium	Hass	ripe

Naive  
Bayes Assumption

You have a soft green-black Hass avocado. Based on this data, would you predict that your avocado is ripe or unripe?



$$P(\text{ripe} | \text{soft, g-b, Hass}) \propto P(\text{ripe}) \cdot P(\text{soft} | \text{ripe}) \cdot P(\text{g-b} | \text{ripe}) \cdot P(\text{Hass} | \text{ripe})$$

$$\frac{7}{11} \cdot \frac{4}{7} \cdot \frac{3}{7} \cdot \frac{5}{7}$$

$$P(\text{unripe} | \text{soft, g-b, Hass}) \propto P(\text{unripe}) \cdot P(\text{soft} | \text{unripe}) \cdot P(\text{g-b} | \text{unripe}) \cdot P(\text{Hass} | \text{unripe})$$

$$= \frac{4}{11} \cdot \frac{0}{4} \dots = 0$$

# Uh oh...

- ▶ There are no soft unripe avocados in the data set.
- ▶ The estimate  $P(\text{soft}|\text{unripe}) \approx \frac{\# \text{ soft unripe avocados}}{\# \text{ unripe avocados}}$  is 0. 
- ▶ The estimated numerator,  
 $P(\text{unripe}) \cdot P(\text{soft, green-black, Hass}|\text{unripe}) = P(\text{unripe}) \cdot P(\text{soft}|\text{unripe}) \cdot P(\text{green-black}|\text{unripe}) \cdot P(\text{Hass}|\text{unripe})$ ,  
is also 0. 
- ▶ But just because there isn't a soft unripe avocado in the data set, doesn't mean that it's impossible for one to exist!
- ▶ **Idea:** Adjust the numerators and denominators of our estimate so that they're never 0.

# Smoothing

	ripe	unripe
Soft		
medium		
firm		

- Without smoothing:

Add to 1

$$P(\text{soft}|\text{unripe}) \approx \frac{\# \text{ soft unripe}}{\# \text{ soft unripe} + \# \text{ medium unripe} + \# \text{ firm unripe}}$$

$$P(\text{medium}|\text{unripe}) \approx \frac{\# \text{ medium unripe}}{\# \text{ soft unripe} + \# \text{ medium unripe} + \# \text{ firm unripe}}$$

$$P(\text{firm}|\text{unripe}) \approx \frac{\# \text{ firm unripe}}{\# \text{ soft unripe} + \# \text{ medium unripe} + \# \text{ firm unripe}}$$

- With smoothing:

$$P(\text{soft}|\text{unripe}) \approx \frac{\# \text{ soft unripe} + 1}{\# \text{ soft unripe} + 1 + \# \text{ medium unripe} + 1 + \# \text{ firm unripe} + 1}$$

$$P(\text{medium}|\text{unripe}) \approx \frac{\# \text{ medium unripe} + 1}{\# \text{ soft unripe} + 1 + \# \text{ medium unripe} + 1 + \# \text{ firm unripe} + 1}$$

$$P(\text{firm}|\text{unripe}) \approx \frac{\# \text{ firm unripe} + 1}{\# \text{ soft unripe} + 1 + \# \text{ medium unripe} + 1 + \# \text{ firm unripe} + 1}$$

→ Add 1 to top, add 3 to bottom

- When smoothing, we add 1 to the count of every group whenever we're estimating a conditional probability.



# Example: avocados, with smoothing

Without Smoothing:  $\frac{4}{7}$

With Smoothing:  $\frac{5}{10}$

color	softness	variety	ripeness
bright green	firm	Zutano	unripe
green-black	medium	Hass	ripe
purple-black	firm	Hass	ripe
green-black	medium	Hass	unripe
purple-black	soft	Hass	ripe
bright green	firm	Zutano	unripe
green-black	soft	Zutano	ripe
purple-black	soft	Hass	ripe
green-black	soft	Zutano	ripe
green-black	firm	Hass	unripe
purple-black	medium	Hass	ripe

You have a soft green-black Hass avocado. Using Naive Bayes, **with smoothing**, would you predict that your avocado is ripe or unripe?

$$\begin{aligned}
 P(\text{ripe} | \text{soft}, \text{g-b}, \text{Hass}) &\propto P(\text{ripe}) \cdot P(\text{soft} | \text{ripe}) \cdot P(\text{g-b} | \text{ripe}) \cdot P(\text{Hass} | \text{ripe}) \\
 &= \frac{7}{11} \cdot \frac{5}{10} = \frac{3+1}{7+3} \cdot \frac{5+1}{7+2}
 \end{aligned}$$

$$\begin{aligned}
 P(\text{unripe} | \text{soft}, \text{g-b}, \text{Hass}) &\propto P(\text{unripe}) \cdot P(\text{soft} | \text{unripe}) \cdot P(\text{g-b} | \text{unripe}) \cdot P(\text{Hass} | \text{unripe}) \\
 &= \frac{4}{11} \cdot \frac{0+1}{4+3} \cdot \frac{2+1}{4+3} \cdot \frac{2+1}{4+2}
 \end{aligned}$$

# Text classification

# Text classification

- ▶ Text classification problems include:
  - ▶ Sentiment analysis (e.g. positive and negative customer reviews).
  - ▶ Determining genre (news articles, blog posts, etc.).
  - ▶ Spam filtering.

# Spam filtering

<input type="checkbox"/>	☆	»	Azazie	LAST CHANCE FOR THE SALE 🕒 - ENDS TONIGHT! View this email in your browser BRIDE...
<input type="checkbox"/>	☆	»	Team Riipen	Riipen_The future of work is changing, and so are we. - Discover the reimagined Riipen, m...
<input type="checkbox"/>	☆	»	Shipping_Pending	You have (2) packages waiting for delivery - View this email in your browser Express Servic...
<input type="checkbox"/>	☆	»	Assemblymember.Boer.	Tasha's Take: Remember and Honor - From Assemblywoman Tasha Boerner Dear Janine, A...
<input type="checkbox"/>	☆	»	Volvo Cars USA	The Scandinavian design behind your Volvo EX90 - Where aerodynamics and aesthetics m...

- ▶ **Our goal:** given the body of an email, determine whether it's **spam** or **ham** (not spam).
- ▶ **Question:** How do we come up with **features**?

↳ Words.

# Features

## Idea:

- ▶ Choose a **dictionary** of  $d$  words.
- ▶ Represent each email with a **feature vector**  $\vec{x}$ :

$$\vec{x} = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \dots \\ x^{(d)} \end{bmatrix}$$

*Handwritten red annotations: a circle around  $x^{(1)}$  and an arrow pointing to the word "Prince".*

where

- ▶  $x^{(i)} = 1$  if word  $i$  is present in the email, and
- ▶  $x^{(i)} = 0$  otherwise.

This is called the **bag-of-words** model. This model ignores the frequency and meaning of words.

# Concrete example

- ▶ Dictionary: “prince”, “money”, “free”, and “just”.
- ▶ Dataset of 5 emails (red are spam, green are ham):
  - ▶ “I am the prince of UCSD and I demand money.”
  - ▶ “Tapioca Express: redeem your free Thai Iced Tea!”
  - ▶ “DSC 10: free points if you fill out CAPEs!”
  - ▶ “Click here to make a tax-free donation to the IRS.”
  - ▶ “Free career night at Prince Street Community Center.”

Prince	$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$
money					
free					
just					
	Email 1	Email 2	Email 3	Email 4	Email 5

# Naive Bayes for spam classification

$$P(\text{class} \mid \text{features}) = \frac{P(\text{class}) \cdot P(\text{features} \mid \text{class})}{P(\text{features})}$$

Spam vs. Ham

- ▶ To classify an email, we'll use Bayes' theorem to calculate the probability of it belonging to each class:
  - ▶  $P(\text{spam} \mid \text{features})$ .
  - ▶  $P(\text{ham} \mid \text{features})$ . ] which one is larger
- ▶ We'll predict the class with a larger probability.

# Naive Bayes for spam classification

$$P(\text{class} \mid \text{features}) = \frac{P(\text{class}) \cdot P(\text{features} \mid \text{class})}{P(\text{features})}$$

- ▶ Note that the formulas for  $P(\text{spam} \mid \text{features})$  and  $P(\text{ham} \mid \text{features})$  have the same denominator,  $P(\text{features})$ .
- ▶ Thus, we can find the larger probability just by comparing numerators:
  - ▶  $P(\text{spam}) \cdot P(\text{features} \mid \text{spam})$ .
  - ▶  $P(\text{ham}) \cdot P(\text{features} \mid \text{ham})$ .

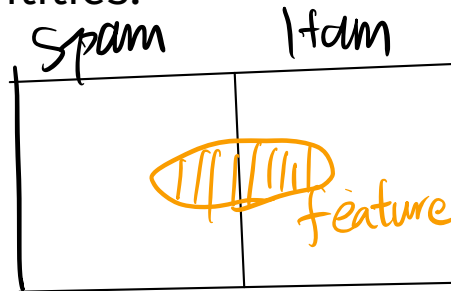


# Naive Bayes for spam classification

## Discussion Question

We need to determine four quantities:

1.  $P(\text{features} \mid \text{spam})$ .
2.  $P(\text{features} \mid \text{ham})$ .
3.  $P(\text{spam})$ .
4.  $P(\text{ham})$ .



Which of these probabilities should add to 1?

- a) 1, 2
- b) 3, 4
- c) Both (a) and (b).
- d) Neither (a) nor (b).

# Estimating probabilities with training data

- ▶ To estimate  $P(\text{spam})$ , we compute

$$P(\text{spam}) \approx \frac{\# \text{ spam emails in training set}}{\# \text{ emails in training set}}$$

- ▶ To estimate  $P(\text{ham})$ , we compute

$$P(\text{ham}) \approx \frac{\# \text{ ham emails in training set}}{\# \text{ emails in training set}}$$

- ▶ What about  $P(\text{features} \mid \text{spam})$  and  $P(\text{features} \mid \text{ham})$ ?

# Assumption of conditional independence

- ▶ Note that  $P(\text{features} \mid \text{spam})$  looks like

$$P(x^{(1)} = 0, x^{(2)} = 1, \dots, x^{(d)} = 0 \mid \text{spam})$$

*money is included.*

*Price not included*

- ▶ Recall: the key assumption that the Naive Bayes classifier makes is that **the features are conditionally independent given the class.**

- ▶ This means we can estimate  $P(\text{features} \mid \text{spam})$  as

$$P(x^{(1)} = 0, x^{(2)} = 1, \dots, x^{(d)} = 0 \mid \text{spam})$$

$$= P(x^{(1)} = 0 \mid \text{spam}) \cdot P(x^{(2)} = 1 \mid \text{spam}) \cdot \dots \cdot P(x^{(d)} = 0 \mid \text{spam})$$

# Concrete example

- ▶ Dictionary: “prince”, “money”, “free”, and “just”.
- ▶ Dataset of 5 emails (red are spam, green are ham):
  - ▶ **“I am the prince of UCSD and I demand money.”**
  - ▶ **“Tapioca Express: redeem your free Thai Iced Tea!”**
  - ▶ **“DSC 10: free points if you fill out CAPEs!”**
  - ▶ **“Click here to make a tax-free donation to the IRS.”**
  - ▶ **“Free career night at Prince Street Community Center.”**

Training  
Dataset

Prince	$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$
money					
free					
just					
	Email 1	Email 2	Email 3	Email 4	Email 5

# Concrete example

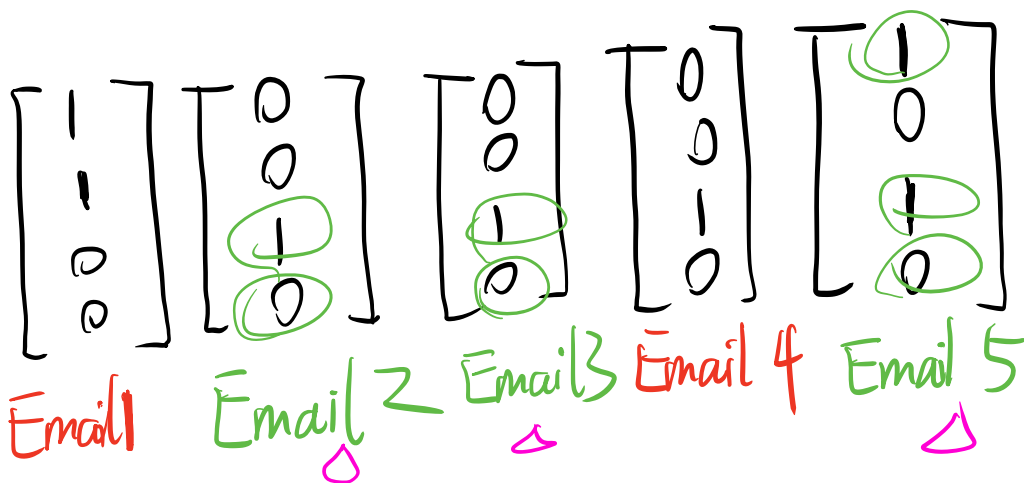
- ▶ New email to classify: "Download a free copy of the Prince of Persia."

Prince	$\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}$
money					
free					
just					
	Email 1	Email 2	Email 3	Email 4	Email 5

$$\begin{aligned}
 P(\text{spam} \mid \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}) &\propto P(\text{spam}) \cdot P(\begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \mid \text{spam}) \\
 &= P(\text{spam}) \cdot P(x^{(1)}=1 \mid \text{spam}) \cdot P(x^{(2)}=0 \mid \text{spam}) \\
 &\quad \cdot P(x^{(3)}=1 \mid \text{spam}) \cdot P(x^{(4)}=0 \mid \text{spam}) \\
 &= \frac{2}{5} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{2}{2} = \frac{1}{20}
 \end{aligned}$$

$$\begin{aligned}
 P(\text{Ham} \mid \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}) &\propto P(\text{Ham}) \cdot P\left(\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \mid \text{Ham}\right) \\
 &= P(\text{Ham}) \cdot P(X^{(1)}=1 \mid \text{Ham}) \cdot P(X^{(2)}=0 \mid \text{Ham}) \cdot P(X^{(3)}=1 \mid \text{Ham}) \\
 &= \frac{3}{5} \cdot \frac{1}{3} \cdot \frac{3}{3} \cdot \frac{3}{3} \cdot \frac{2}{3} = \frac{1}{5}
 \end{aligned}$$

Prince  
money  
free  
just

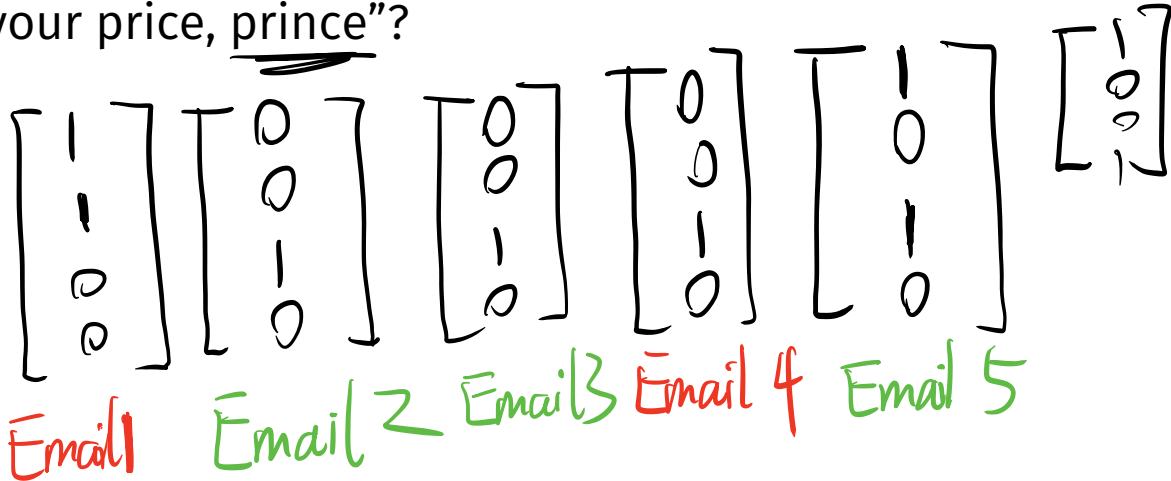


Ham > Spam  
 $\frac{1}{5} > \frac{1}{20}$

# Uh oh...

- ▶ What happens if we try to classify the email "just what's your price, prince"?

Prince  
money  
free  
just



$$P(\text{spam}) \cdot P(x^{(1)}=1 | \text{spam}) \cdot P(x^{(2)}=0 | \text{spam}) \\ \cdot P(x^{(3)}=0 | \text{spam}) \cdot P(x^{(4)}=1 | \text{spam}) = 0$$

0

# Smoothing

- ▶ **Without** smoothing:

$$P(x^{(i)} = 1 \mid \text{spam}) \approx \frac{\# \text{ spam containing word } i}{\# \text{ spam containing word } i + \# \text{ spam not containing word } i}$$

- ▶ **With** smoothing:

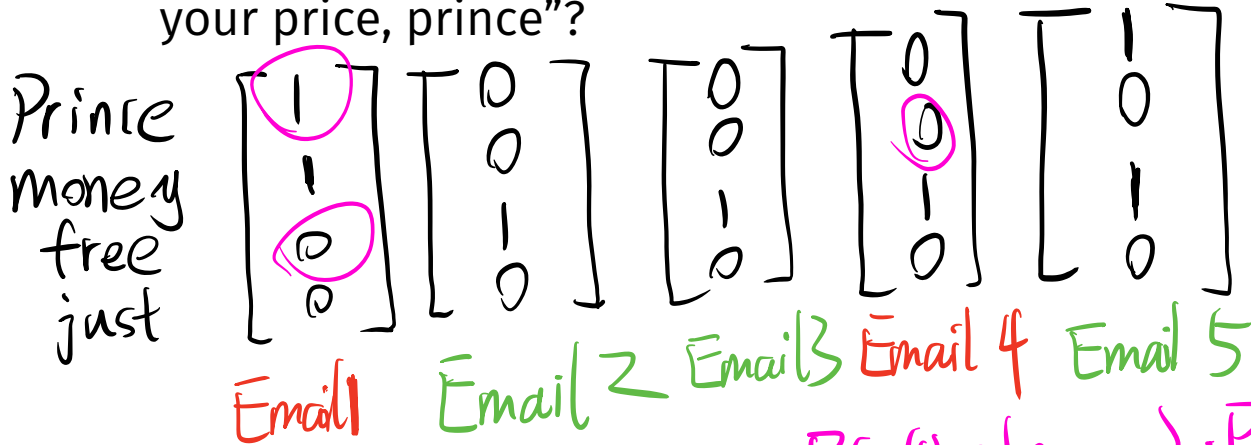
$$P(x^{(i)} = 1 \mid \text{spam}) \approx \frac{(\# \text{ spam containing word } i) + 1}{(\# \text{ spam containing word } i) + 1 + (\# \text{ spam not containing word } i) + 1}$$

- ▶ When smoothing, we add 1 to the count of every group whenever we're estimating a conditional probability.



# Concrete example with smoothing

- ▶ What happens if we try to classify the email “just what’s your price, prince”?



$$P(\text{spam} | \begin{bmatrix} 1 \\ | \\ 0 \\ 0 \end{bmatrix}) \propto P(\text{spam}) \cdot P(x^{(1)}=1 | \text{spam}) \cdot P(x^{(2)}=0 | \text{spam}) \cdot P(x^{(3)}=0 | \text{spam}) \cdot P(x^{(4)}=1 | \text{spam})$$

$$= \frac{2}{5} \cdot \frac{2}{4} \cdot \frac{2}{4} \cdot \frac{2}{4} \cdot \frac{1}{4}$$



# Modifications and extensions

- ▶ **Idea:** Use pairs (or longer sequences) of words rather than individual words as features.
  - ▶ This better captures the dependencies between words.
  - ▶ It also leads to a much larger space of features, increasing the complexity of the algorithm.

# Modifications and extensions

- ▶ **Idea:** Use pairs (or longer sequences) of words rather than individual words as features.
  - ▶ This better captures the dependencies between words.
  - ▶ It also leads to a much larger space of features, increasing the complexity of the algorithm.
- ▶ **Idea:** Instead of recording whether each word appears, record how many times each word appears.
  - ▶ This better captures the importance of repeated words.

# Summary

# Summary, next time

- ▶ Smoothing gives a way to make better predictions when a feature has never been encountered in the training data.
- ▶ The Naive Bayes classifier can be used for text classification, using the bag-of-words model.
- ▶ **Next time:** measuring performance of classifiers using precision and recall.