Lecture 18

More Naïve Bayes, Review

DSC 40A, Spring 2024

Announcements

only 2 questions!

- Homework 8 is due on **Thursday**. You cannot use slip days on it.
- The Final Exam is on Saturday from 8-11AM.
  - You will be assigned a seat, either in Center Hall 212 or 214.
  - 180 minutes, on paper, no calculators or electronics, but **you are allowed to bring two double-sided index cards (4 inches by 6 inches) of notes that you write by hand.**
- We will have two review sessions. In each of them, the first hour will be a mock exam **which you will take silently on paper**; we will take up the problems in the second half.
  - Tuesday, June 4th, 5-7PM (empirical risk minimization and linear algebra).
  - Thursday, June 6th, 5-7PM (gradient descent and probability).
  - Also, Friday, June 7th, 4-9PM: study session in HDSI 123.

≈ 43% right now

- If at least 90% of the class fills out both the End-of-Quarter Survey and SETs by 8AM on Saturday, then the entire class will have 2% of extra credit added to their overall grade.

Agenda

- Text classification.
  - Practical demo.
- Review.
  - Old exam problems.

Recap: The Naïve Bayes classifier

want to predict the most likely class

- We want to predict a class, given certain features.
- Using Bayes' Theorem, we write:

e.g. ripe/not ripe

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

← conditional independence assumption

↳ e.g. (Hass, firm, green-black)

- For each class, we compute the numerator using the **naïve 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!

$$\mathbb{P}(\text{Hass, firm, green-black}\,|\,\text{ripe}) = \mathbb{P}(\text{Hass}|\text{ripe}) \cdot \mathbb{P}(\text{firm}|\text{ripe}) \cdot \mathbb{P}(\text{g-b}|\text{ripe})$$

Question 🤔

Answer at q.dsc40a.com

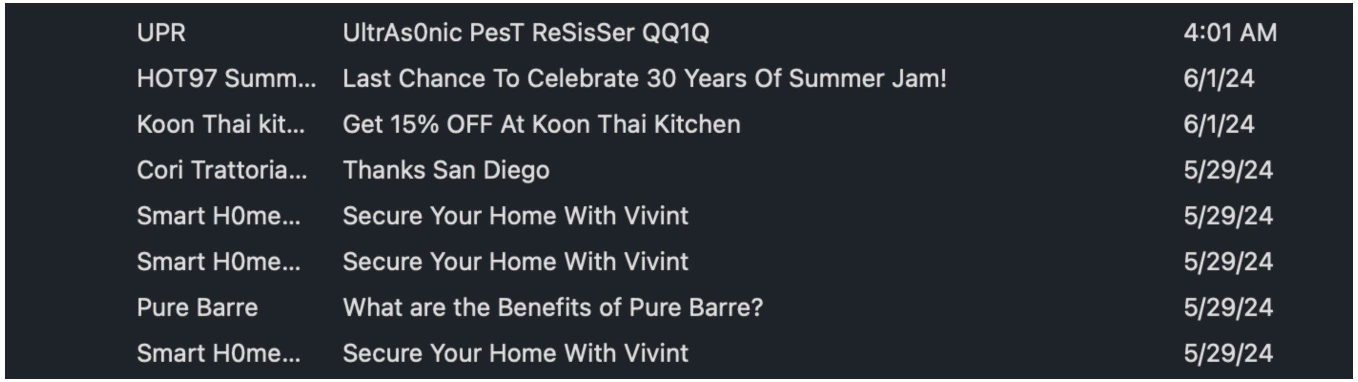
Remember, you can always ask questions at q.dsc40a.com!

If the direct link doesn't work, click the "🤔 Lecture Questions" link in the top right corner of dsc40a.com.

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 is a common text classification problem:



UPR	UltrAsOnic PesT ReSisSer QQ1Q	4:01 AM
HOT97 Summ...	Last Chance To Celebrate 30 Years Of Summer Jam!	6/1/24
Koon Thai kit...	Get 15% OFF At Koon Thai Kitchen	6/1/24
Cori Trattoria...	Thanks San Diego	5/29/24
Smart H0me...	Secure Your Home With Vivint	5/29/24
Smart H0me...	Secure Your Home With Vivint	5/29/24
Pure Barre	What are the Benefits of Pure Barre?	5/29/24
Smart H0me...	Secure Your Home With Vivint	5/29/24

- **Goal**: Given the body of an email, determine whether it's spam or ham (not spam).
- **Question**: What information do we use to make these predictions? What **features**?

Text features

$d = 3$

e.g. (1) "hOme" (2) "free" (3) "sale"

$\vec{x} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$ "come sale your hOme for sale"

Idea:

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

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

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 (orange are spam, blue are ham):
    - 1 "I am the prince of UCSD and I demand money."
    - 2 "Tapioca Express: redeem your free Thai Iced Tea!"
    - 3 "DSC 10: free points if you fill out SETs!"
    - 4 "Click here to make a tax-free donation to the IRS."
    - 5 "Free career night at Prince Street Community Center."

$$\vec{x}_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

	prince	money	free	just	spam?
1	1	1	0	0	spam
2	0	0	1	0	ham
3	0	0	1	0	ham
4	0	0	1	0	spam
5	1	0	1	0	ham

Naïve Bayes for spam classification

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

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

Naïve Bayes for spam classification

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

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

$\mathbb{P}(\text{the} | \text{spam}) + \mathbb{P}(\overline{\text{the}} | \text{spam}) = 1$

Question 🤔

Answer at q.dsc40a.com

- $\mathbb{P}(\text{features} \mid \text{spam})$.
- $\mathbb{P}(\text{features} \mid \text{ham})$.
- $\mathbb{P}(\text{spam})$.
- $\mathbb{P}(\text{ham})$.

Which of these probabilities should add to 1?

- ~~A. 1, 2~~ $\Rightarrow \mathbb{P}(\text{the} | \text{spam}) + \mathbb{P}(\text{the} | \text{ham}) > 1$
- [✓ B. 3, 4] $\Rightarrow \text{ham} = \overline{\text{spam}}$
- ~~C. Both (a) and (b).~~
- D. Neither (a) nor (b).

Estimating probabilities with training data

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

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

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

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

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

Assumption of conditional independence

does not have word 1,
does have word 2

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

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

- Recall: the key assumption that the Naïve Bayes classifier makes is that **the features are conditionally independent given the class.**
- This means we can estimate $\mathbb{P}(\text{features} \mid \text{spam})$ as:

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

Concrete example

- Dictionary: "prince", "money", "free", and "just".
- Dataset of 5 emails (**orange are spam**, **blue 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 SETs!"**
  - **"Click here to make a tax-free donation to the IRS."**
  - **"Free career night at Prince Street Community Center."**

Concrete example

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

	prince	money	free	just	spam?
1	1	1	0	0	spam
2	0	0	1	0	ham
3	0	0	1	0	ham
4	0	0	1	0	spam
5	1	0	1	0	ham

prince = 1
money = 0
free = 1
just = 0

$$\mathbb{P}(spam \mid features) \propto \mathbb{P}(spam) \cdot \mathbb{P}(features \mid spam)$$

$$= \mathbb{P}(spam) \cdot \mathbb{P}(prince=1 \mid spam) \cdot \mathbb{P}(money=0 \mid spam) \cdot \mathbb{P}(free=1 \mid spam) \cdot \mathbb{P}(just=0 \mid spam)$$

$$= \frac{2}{5} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{2}{2} = \boxed{\frac{1}{20}}$$

	prince	money	free	just	spam?
1	1	1	0	0	spam
2	0	0	1	0	ham
3	0	0	1	0	ham
4	0	0	1	0	spam
5	1	0	1	0	ham

prince = 1
money = 0
free = 1
just = 0

$$\mathbb{P}(\text{ham} \mid \text{features}) \propto \mathbb{P}(\text{ham}) \cdot \mathbb{P}(\text{features} \mid \text{ham})$$

$$= \mathbb{P}(\text{ham}) \cdot \mathbb{P}(\text{prince}=1 \mid \text{ham}) \cdot \mathbb{P}(\text{money}=0 \mid \text{ham}) \cdot \mathbb{P}(\text{free}=1 \mid \text{ham}) \cdot \mathbb{P}(\text{just}=0 \mid \text{ham})$$

$$= \frac{3}{5} \cdot \frac{1}{3} \cdot \frac{3}{3} \cdot \frac{3}{3} \cdot \frac{3}{3}$$

$$= \boxed{\frac{1}{5}}$$

Since $\frac{1}{5} > \frac{1}{20}$, we predict $\boxed{\text{ham}}$.

Uh oh...

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

	prince	money	free	just	spam?
1	1	1	0	0	spam
2	0	0	1	0	ham
3	0	0	1	0	ham
4	0	0	1	0	spam
5	1	0	1	0	ham

prince = 1
money = 0
free = 0
just = 1

$$\mathbb{P}(\text{spam} \mid \text{features}) \propto \mathbb{P}(\text{spam}) \cdot \mathbb{P}(\text{prince}=1 \mid \text{spam}) \cdot \mathbb{P}(\underset{=0}{\text{money}} \mid \text{spam}) \cdot \mathbb{P}(\underset{=0}{\text{free}} \mid \text{spam}) \cdot \mathbb{P}(\underset{=1}{\text{just}} \mid \text{sp})$$

$$= \frac{0}{2} = \boxed{0}$$

⇒ whole product is 0.

Smoothing

- **Without** smoothing:

$$\mathbb{P}(x^{(i)} = 1 \mid \text{spam}) \approx \frac{\#\text{ spam containing word } i}{\underbrace{\#\text{ spam containing word } i + \#\text{ spam not containing word } i}_{\text{total \# of spam emails}}}$$

- **With** smoothing:

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

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

$+2$

In general, add to the denominator the number of possible values for the feature

Concrete example with smoothing

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

	prince	money	free	just	spam?
1	1	1	0	0	spam
2	0	0	1	0	ham
3	0	0	1	0	ham
4	0	0	1	0	spam
5	1	0	1	0	ham

prince = 1
money = 0
free = 0
just = 1

$$\mathbb{P}(\text{spam} \mid \text{features}) \propto \mathbb{P}(\text{spam}) \cdot \mathbb{P}(\text{prince}=1 \mid \text{spam}) \cdot \mathbb{P}(\underset{=0}{\text{money}} \mid \text{spam}) \cdot \mathbb{P}(\underset{=0}{\text{free}} \mid \text{spam}) \cdot \mathbb{P}(\underset{=1}{\text{just}} \mid \text{sp})$$

$$= \frac{2}{5} \cdot \frac{1+1}{2+2} \cdot \frac{1+1}{2+2} \cdot \frac{1+1}{2+2} \cdot \frac{0+1}{2+2}$$

$$= \frac{\cancel{2}}{5} \cdot \frac{1}{\cancel{2}} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{4} = \boxed{\frac{1}{80}}$$

← bigger!

	prince	money	free	just	spam?
1	1	1	0	0	spam
2	0	0	1	0	ham
3	0	0	1	0	ham
4	0	0	1	0	spam
5	1	0	1	0	ham

prince = 1
money = 0
free = 0
just = 1

$$\mathbb{P}(\text{ham} \mid \text{features}) \propto \mathbb{P}(\text{ham}) \cdot \mathbb{P}(\text{prince}=1 \mid \text{ham}) \cdot \mathbb{P}(\underset{=0}{\text{money}} \mid \text{ham}) \cdot \mathbb{P}(\underset{=0}{\text{free}} \mid \text{ham}) \cdot \mathbb{P}(\underset{=1}{\text{just}} \mid \text{ham})$$

$$= \frac{3}{5} \cdot \frac{1+1}{3+2} \cdot \frac{3+1}{3+2} \cdot \frac{0+1}{3+2} \cdot \frac{0+1}{3+2}$$

$$= \frac{3}{5} \cdot \frac{2}{5} \cdot \frac{4}{5} \cdot \frac{1}{5} \cdot \frac{1}{5} = \frac{24}{3125} \quad \text{vs.} \quad \boxed{\frac{1}{80}}$$

bigger!

$\Rightarrow$ $\boxed{\text{predict spam}}$

$$\frac{1}{80} = \frac{40}{3200}$$

bigrams, more generally:
N-grams

DSC 80

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.

That's all the new content we have!

Let's now try out Naïve Bayes in code. Follow along here.

Review

We're done with new content! Let's work through some old exam problems.

Fall 2021 Final Exam, Problem 10

Suppose you're given the following probabilities:

- $\mathbb{P}(A|B) = \frac{2}{5}$.
- $\mathbb{P}(B|A) = \frac{1}{4}$.
- $\mathbb{P}(A|C) = \frac{2}{3}$.

$$\mathbb{P}(A|C) = \mathbb{P}(A) = \frac{2}{3}$$

Part 1: If A and C are independent, what is $\mathbb{P}(B)$?

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

$$= \frac{\frac{\overset{1}{\cancel{2}}}{3}\cdot\frac{1}{\cancel{4}_2}}{\frac{2}{5}} = \frac{1}{6}\cdot\frac{5}{2} = \boxed{\frac{5}{12}}$$

Fall 2021 Final Exam, Problem 10

Suppose you're given the following probabilities:

- $\mathbb{P}(A|B) = \frac{2}{5}$. $\Rightarrow \mathbb{P}(A) = \frac{2}{5}$
- $\mathbb{P}(B|A) = \frac{1}{4}$. $\Rightarrow \mathbb{P}(B) = \frac{1}{4}$
- $\mathbb{P}(A|C) = \frac{2}{3}$.

Part 2: Suppose A and C are not independent, and now suppose that $\mathbb{P}(A|\bar{C}) = \frac{1}{5}$. Given that A and B are independent, what is $\mathbb{P}(C)$?

$$\mathbb{P}(C) = \mathbb{P}(C \cap A) + \mathbb{P}(C \cap \bar{A}) = \mathbb{P}(A) \cdot \mathbb{P}(C|A) + \mathbb{P}(\bar{A}) \cdot \mathbb{P}(C|\bar{A})$$

$$\mathbb{P}(A) = \mathbb{P}(A \cap C) + \mathbb{P}(A \cap \bar{C}) = \mathbb{P}(C)\,\mathbb{P}(A|C) + \mathbb{P}(\bar{C}) \cdot \mathbb{P}(A|\bar{C})$$

Let $p = \mathbb{P}(C)$

$$\frac{2}{5} = p \cdot \frac{2}{3} + (1-p) \cdot \frac{1}{5} \Rightarrow \boxed{\text{solve for } p!}$$

Spring 2023 Midterm Exam 2, Problem 6.2

The events A and B are mutually exclusive, or disjoint. More generally, for **any** two disjoint events A and B, show how to express $P(\bar{A}|(A \cup B))$ in terms of $P(A)$ and $P(B)$ **only**.

Fall 2021 Final Exam, Problem 8

Billy brings you back to Dirty Birds, the restaurant where he is a waiter. He tells you that Dirty Birds has 30 different flavors of chicken wings, 18 of which are ‘wet’ (e.g. honey garlic) and 12 of which are ‘dry’ (e.g. lemon pepper).

Each time you place an order at Dirty Birds, you get to pick 4 different flavors. The order in which you pick your flavors does not matter.

Part 1: How many ways can we select 4 flavors in total?

Part 2: How many ways can we select 4 flavors in total such that we select an equal number of wet and dry flavors?

Part 3: Billy tells you he’ll surprise you with 4 different flavors, randomly selected from the 30 flavors available. What’s the probability that he brings you at least one wet flavor and and least one dry flavor?

Part 4: Suppose you go to Dirty Birds once a day for 7 straight days. Each time you go there, Billy brings you 4 different flavors, randomly selected from the 30 flavors available. What’s the probability that on at least one of the 7 days, he brings you all wet flavors or all dry flavors? (Note: All 4 flavors for a particular day must be different, but it is possible to get the same flavor on multiple days.)

Fall 2021 Final Exam, Problem 9

In this question, we'll consider the phone number 6789998212 (mentioned in Soulja Boy's 2008 classic, "Kiss Me thru the Phone").

Part 1: How many permutations of 6789998212 are there?

Part 2: How many permutations of 6789998212 have all three 9s next to each other?

Part 3: How many permutations of 6789998212 end with a 1 and start with a 6?

Part 4: How many different 3 digit numbers with unique digits can we create by selecting digits from 6789998212?