Lecture 18

More Naïve Bayes, Review

DSC 40A

Agenda

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

Recap: The Naïve Bayes classifier

Predict the Class that had highest $\mathbb{P}$

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

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

X_S : (type, softness, color)

Assumption Cond. Ind.

Ans. ripe or not ripe

- 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!

$\mathbb{P}(\text{HASS}, \text{firm}, \text{G.B}|\text{ripe}) = \mathbb{P}(\text{H}|\text{ripe}) \cdot \mathbb{P}(\text{F}|\text{ripe}) \cdot \mathbb{P}(\text{G.B}|\text{ripe})$

Question 🤔

Take a moment to pause and reflect...

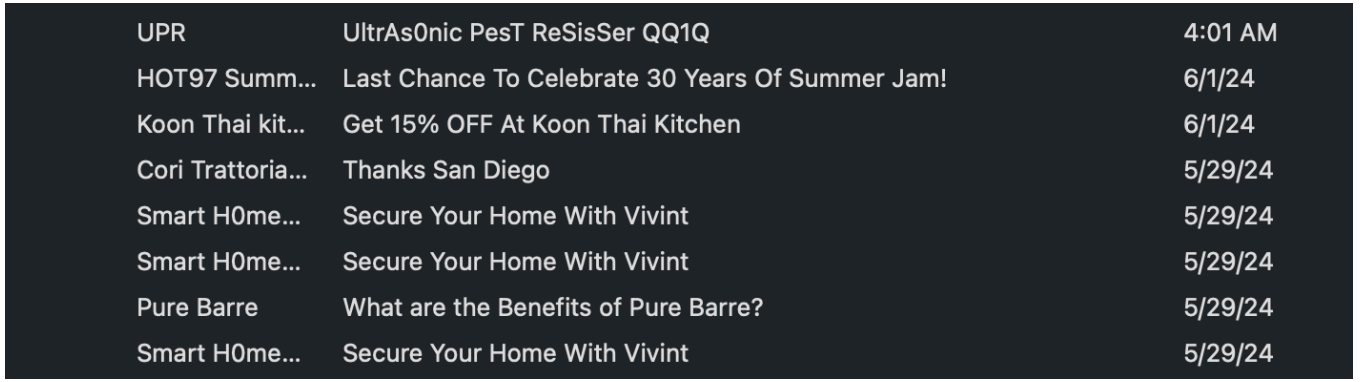
If you have any questions please post online to our forms/Q&A site.

Course staff will answer them ASAP!

Text classification

Text classification

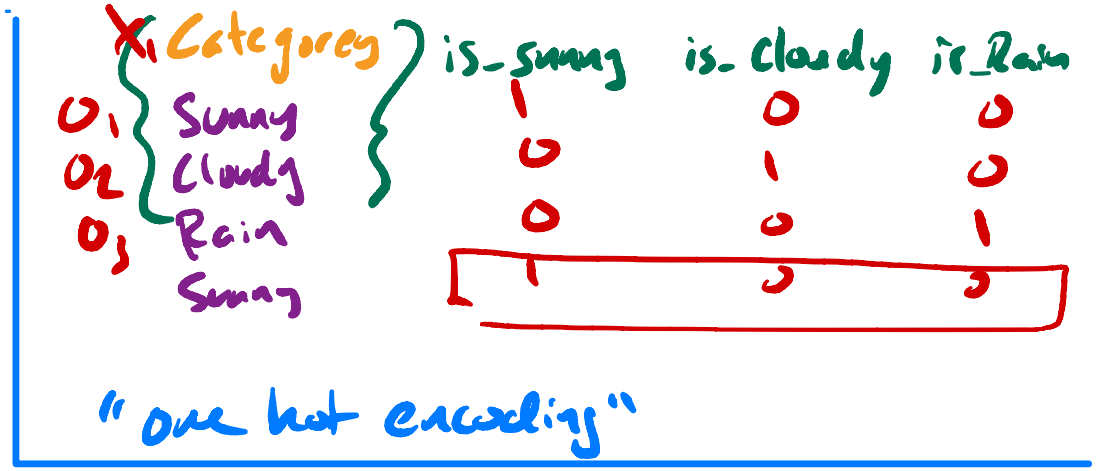
- Text classification problems include:
  - Sentiment analysis (e.g. positive and negative customer reviews).
  - Determining genre (news articles, blog posts, etc.).
- Spam filtering is a common text classification problem:



UPR	UltrAs0nic 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



Idea:

{"hOme", "free", "sale"}
(1) (2) (3)

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

Email subject

"Act fast now! Free items available, no sale"
(1) (3)

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

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

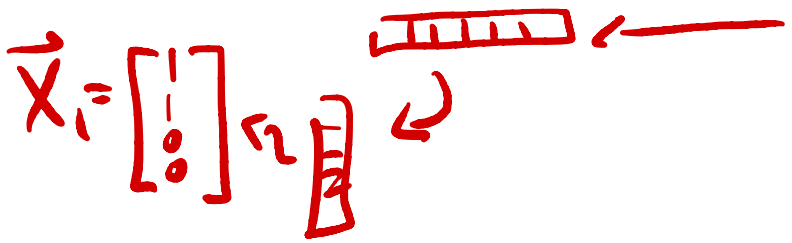
dx1

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."



	x_1 Prince	x_2 money	x_3 free	x_4 just	is 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:
  - $\mathbb{P}(\text{spam} \mid \text{features}) \propto \mathbb{P}(\text{spam}) \cdot \mathbb{P}(\text{features} \mid \text{spam})$.
  - $\mathbb{P}(\text{ham} \mid \text{features}) \propto \mathbb{P}(\text{ham}) \cdot \mathbb{P}(\text{features} \mid \text{ham})$.

(Proportional to)

Question 🤔

Take a moment to pause and reflect...

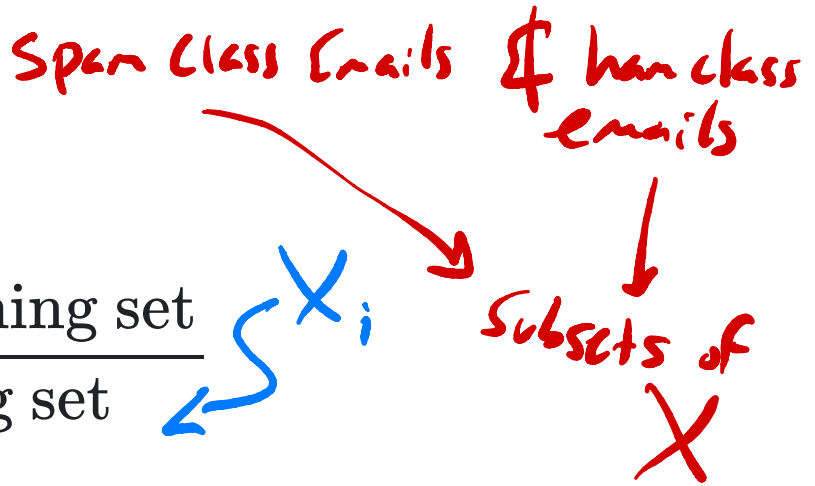
1. $\mathbb{P}(\text{features} \mid \text{spam})$.
2. $\mathbb{P}(\text{features} \mid \text{ham})$.
3. $\mathbb{P}(\text{spam})$.
4. $\mathbb{P}(\text{ham})$.

Which of these probabilities should add to 1?

- ~~A. 1, 2~~ => P("word"|spam) + P("word"|ham) = ~~1~~ > 1 e.g. "the"
- B. 3, 4 => P(ham) = ¬spam = 1 ¬ or c complement
- ~~C. Both (a) and (b).~~ Also... P("word"|spam) + P("word"^{c}|spam) = 1
- ~~D. Neither (a) nor (b).~~

same word

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

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

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

this training example, labeled as "spam" has word 2, but not word 1

- 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:

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

Concrete example

Spam: 1
Not spam or HAM: 0

- 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." 0
  - "Tapioca Express: redeem your free Thai Iced Tea!" 1
  - "DSC 10: free points if you fill out SETs!" 1
  - "Click here to make a tax-free donation to the IRS." 0
  - "Free career night at Prince Street Community Center." 1

$\vec{y} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 1 \end{bmatrix}$

y_s

Concrete example

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

d=
Prince = 1
money = 0
free = 1
just = 0

$\begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}$

	Prince	money	free	just	is 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

looking at spam first

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

$$= P(\text{spam}) \cdot P(\text{Prince}=1 \mid \text{spam}) \cdot P(M=0 \mid S) \cdot P(F=1 \mid S) \cdot P(J=0 \mid S)$$

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

	Prince	money	free	just	is 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

$d =$

Prince = 1

money = 0

free = 1

just = 0

$\begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}$

$$P(\text{ham}) \cdot P(P=1 \mid H) \cdot P(M=0 \mid H) \cdot P(F=1 \mid H) \cdot P(J=0 \mid H)$$

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

$\frac{1}{20}$ for SPAM & $\frac{1}{5}$ for HAM = Looks like we are getting a FREE Copy of P.O.P.

Uh oh...

S = spam
F = features

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

	x_1 Prince	x_2 money	x_3 free	x_4 just	is 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

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

$$P(S|f) \propto P(S) \cdot P(p=1|s) \cdot P(M=0|s) \cdot P(F=0|s) \cdot P(j=1|s)$$

$P(j=1|s) = 0$

$\Rightarrow$ whole Product = 0

Smoothing

- **Without** smoothing:

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

- **With** smoothing:

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

total # of SPAM

+2 "for each feature"

- 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"?

	Prince	money	free	just	is 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

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

$$P(S|f) \propto P(S) \cdot P(p=1|s) \cdot P(M=0|s) \cdot P(F=0|s) \cdot P(j=1|s)$$

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

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

	x_1 Prince	x_2 money	x_3 free	x_4 just	is 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

$$P(H|F) \propto P(H) \cdot P(P=1|H) \cdot P(M=0|H) \cdot P(F=0|H) \cdot P(J=1|H)$$

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

$$= \frac{3}{5} \cdot \frac{2}{5} \cdot \frac{4}{5} \cdot \frac{1}{5} \cdot \frac{1}{5} = \frac{24}{3125}$$ Compared to $\frac{1}{80}$

$\frac{24}{3125}$ → HAM

$\frac{1}{80}$ → SPAM

Predict SPAM

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.

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 $\mathbb{P}(B)$?

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

$\frac{\frac{2}{3} \cdot \frac{1}{4}}{\frac{2}{5}} = \frac{\frac{1}{6}}{\frac{2}{5}}$

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

$\hookrightarrow \frac{1}{6} \cdot \frac{5}{2} = \frac{5}{12}$

Fall 2021 Final Exam, Problem 10

Suppose you're given the following probabilities:

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

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 $\mathbb{P}(C)$?

$\mathbb{P}(A) = \frac{2}{5}$

$\mathbb{P}(B) = \frac{1}{4}$

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

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

$\frac{2}{5} = p \cdot \frac{2}{3} + (1-p) \cdot \frac{1}{5}$ · Solve for p