Lecture 24 – Naive Bayes

DSC 40A, Fall 2022 @ UC San Diego

Mahdi Soleymani, with help from **many others**

Agenda

- Naive Bayes.
- Naive Bayes in practice — text classification.
- Practical demo.

Naive Bayes

Naive Bayes classifier

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

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

- 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!

na·ive

/nīˈēv/

adjective

(of a person or action) showing a lack of experience, wisdom, or judgment.
"the rather naive young man had been totally misled"

- (of a person) natural and unaffected; innocent.
  "Andy had a sweet, naive look when he smiled"

  Similar: innocent | unsophisticated | artless | ingenuous | inexperienced

- of or denoting art produced in a straightforward style that deliberately rejects sophisticated artistic techniques and has a bold directness resembling a child's work, typically in bright colors with little or no perspective.

Example: comic characters

ALIGN	SEX	COMPANY
Bad	Male	Marvel
Neutral	Male	Marvel
Good	Male	Marvel
Bad	Male	DC
Good	Female	Marvel
Bad	Male	DC
Good	Male	DC
Bad	Male	Marvel
Good	Female	Marvel
Bad	Female	Marvel

My favorite character is a male Marvel character. Using Naive Bayes, would we predict that my favorite character is bad, good, or neutral?

ALIGN	SEX	COMPANY
Bad	Male	Marvel
Neutral	Male	Marvel
Good	Male	Marvel
Bad	Male	DC
Good	Female	Marvel
Bad	Male	DC
Good	Male	DC
Bad	Male	Marvel
Good	Female	Marvel
Bad	Female	Marvel

Example: comic characters

ALIGN	SEX	COMPANY
Bad	Male	Marvel
Neutral	Male	Marvel
Good	Male	Marvel
Bad	Male	DC
Good	Female	Marvel
Bad	Male	DC
Good	Male	DC
Bad	Male	Marvel
Good	Female	Marvel
Bad	Female	Marvel

My other favorite character is a female Marvel character. Using Naive Bayes, would we predict that my favorite character is bad, good, or neutral?

ALIGN	SEX	COMPANY
Bad	Male	Marvel
Neutral	Male	Marvel
Good	Male	Marvel
Bad	Male	DC
Good	Female	Marvel
Bad	Male	DC
Good	Male	DC
Bad	Male	Marvel
Good	Female	Marvel
Bad	Female	Marvel

Uh oh...

- There are no neutral female characters in the data set.
- The estimate $P(\text{female}|\text{neutral}) \approx \frac{\text{\# female neutral characters}}{\text{\# neutral characters}}$ is 0.
- The estimated numerator,
  $P(\text{neutral}) \cdot P(\text{female, Marvel}|\text{neutral}) =$
  $P(\text{neutral}) \cdot P(\text{female}|\text{neutral}) \cdot P(\text{Marvel}|\text{neutral})$,
  is also 0.
- But just because there isn't a neutral female character in the data set, doesn't mean they don't exist!
- **Idea:** Adjust the numerators and denominators of our estimate so that they're never 0.

Smoothing

- **Without** smoothing:

$$P(\text{female}|\text{neutral}) \approx \frac{\#\ \text{female neutral}}{\#\ \text{female neutral} + \#\ \text{male neutral}}$$

$$P(\text{male}|\text{neutral}) \approx \frac{\#\ \text{male neutral}}{\#\ \text{female neutral} + \#\ \text{male neutral}}$$

- **With** smoothing:

$$P(\text{female}|\text{neutral}) \approx \frac{\#\ \text{female neutral} + 1}{\#\ \text{female neutral} + 1 + \#\ \text{male neutral} + 1}$$

$$P(\text{male}|\text{neutral}) \approx \frac{\#\ \text{male neutral} + 1}{\#\ \text{female neutral} + 1 + \#\ \text{male neutral} + 1}$$

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

Example: comic characters

Using smoothing, let's determine whether Naive Bayes would predict a female Marvel character to be bad, good, or neutral.

ALIGN	SEX	COMPANY
Bad	Male	Marvel
Neutral	Male	Marvel
Good	Male	Marvel
Bad	Male	DC
Good	Female	Marvel
Bad	Male	DC
Good	Male	DC
Bad	Male	Marvel
Good	Female	Marvel
Bad	Female	Marvel

Recap: Naive Bayes classifier

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

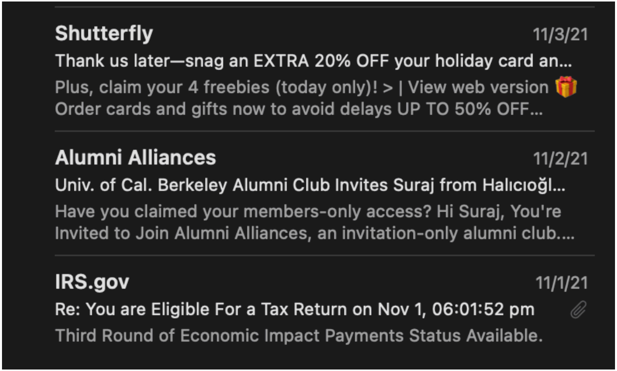
$$P(\text{class}|\text{features}) = \frac{P(\text{class}) \cdot P(\text{features}|\text{class})}{P(\text{features})}$$

- 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!

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.**
- **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?

Features

Idea:

- Choose a **dictionary** of d words, e.g. "prince", "money", "free"...

- Represent each email with a **feature vector** $\vec{x}$:

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

  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.

Concrete example

- Dictionary: "prince", "money", "free", and "xxx".
- 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 40A: free points if you fill out CAPEs!"**
  - **"Click here to make a tax-free donation to the IRS."**
  - **"Free COVID-19 tests at Price Center."**

Naive Bayes for spam classification

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

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

- 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

To answer, go to `menti.com` and enter 7053 7461.

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

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

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

Assumption of conditional independence

- Note that P(features | spam) looks like

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

- 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(features | spam) as

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

Concrete example

- Dictionary: "prince", "money", "free", and "xxx".
- 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 40A: free points if you fill out CAPEs!"**
  - **"Click here to make a tax-free donation to the IRS."**
  - **"Free COVID-19 tests at Prince Center."**

Concrete example

- New email to classify: “Download a free copy of the Prince of Persia.”‘

Uh oh...

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

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.
  - **Don't** smooth the estimates of unconditional probabilities (e.g. P(spam)).

Concrete example with smoothing

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

Practical demo

Follow along with the demo by clicking the **code** link on the course website next to Lecture 24.

Summary

Summary

- The Naive Bayes classifier works by estimating the numerator of $P(\text{class}|\text{features})$ for all possible classes.

- It uses Bayes' theorem:

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

- It also uses a simplifying assumption, that features are conditionally independent given a class:

$$P(\text{features}|\text{class}) = P(\text{feature}_1|\text{class}) \cdot P(\text{feature}_2|\text{class}) \cdot \ldots$$

- The Naive Bayes classifier can be used for text classification, using the bag-of-words model.