Discrete Naïve Bayes

Previously on CSCI 4622

Probabilistic Classification:

Train model on training data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$

Goal: Given new data $\mathbf{x}$, predict it's label y

y belongs to some class

For each class c, estimate probability that $\mathbf{x}$ belongs to class

$$p(y = c \mid \mathbf{x})$$

Assign $\mathbf{x}$ to the class that gives the highest probability

$$\hat{y} = \arg\max_c p(y = c \mid \mathbf{x})$$

Discriminative vs Generative Classifiers

How do we model/estimate these conditional probabilities?

Discriminative Classifiers:

- Model the conditional relationship between data and label directly: $p(y \mid \mathbf{x})$
- **Example**: Logistic Regression models this relationship as $p(y \mid \mathbf{x}) = \text{sigm}(\boldsymbol{\beta}^T \mathbf{x})$

Generative Classifiers:

- Model the joint probability distribution
- Make assumptions about the relationship between $\mathbf{x}$ and y
- Make assumptions about the data itself

Naïve Bayes

Naïve Bayes is a popular off-the-shelf classifier

- Known for being simple but very effective in some cases
- Very popular in text learning
- High Bias method
- For small training sets, outperforms many more sophisticated methods
- Usually a good thing to try right off the bat, just to see if it works

Motivation: SPAM Classification

Suppose you have the following (small) set of labeled emails (with stop-words removed):

HAM	SPAM	SPAM	SPAM	HAM
work	nigeria	fly	money	fly
buy	opportunity	buy	buy	home
money	viagra	nigeria	fly	nigeria

Let $\mathbf{x}$ be a Bag-of-Words feature vector representing frequency counts of words in vocab V

We want to estimate the probabilities $p(y = \text{SPAM} \mid \mathbf{x})$ and $p(y = \text{HAM} \mid \mathbf{x})$

We then assign the email $\mathbf{x}$ to class for which the probability is bigger

Bayesian Classifiers

We can estimate these probabilities from training data using **Bayes Rule**

$$p(y \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid y)p(y)}{p(\mathbf{x})}$$

Bayesian Classifiers

We can estimate these probabilities from training data using **Bayes Rule**

$$p(y \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid y)p(y)}{p(\mathbf{x})}$$

Bayesian methods give special names to each of these probabilities

$$\text{posterior} = \frac{\text{class-conditional likelihood} \times \text{prior}}{\text{evidence}}$$

Let's look of each of these one-by-one

Bayesian Classifiers

$$p(y \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid y)p(y)}{p(\mathbf{x})} \qquad \textbf{posterior} = \frac{\text{class-likelihood} \times \text{prior}}{\text{evidence}}$$

The Posterior Probability:

General Interpretation: The probability that a particular object belongs to a particular class given its observed features

Concretely: The probability an email is SPAM given the words in the email

Bayesian Classifiers

$$p(y \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid y)p(y)}{p(\mathbf{x})} \qquad \text{posterior} = \frac{\textbf{class-likelihood} \times \text{prior}}{\text{evidence}}$$

The Class-Conditional Likelihood:

General Interpretation: Given a class $y = c$, the probability that $\mathbf{x}$ is observed

Concretely: Given assumptions about the nature of SPAM email, the probability that we observe *this* particular email

Example: $p(\mathbf{x} = [\text{buy, viagra}] \mid SPAM)$

Bayesian Classifiers

$$p(y \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid y)p(y)}{p(\mathbf{x})} \qquad \text{posterior} = \frac{\text{class-likelihood} \times \textbf{prior}}{\text{evidence}}$$

The Class Prior:

General Interpretation: The probability that any object belongs to class $y = c$

Concretely: The probability that any new email is SPAM/HAM

Bayesian Classifiers

$$p(y \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid y)p(y)}{p(\mathbf{x})} \qquad \text{posterior} = \frac{\text{class-likelihood} \times \text{prior}}{\textbf{evidence}}$$

The Evidence:

General Interpretation: The probability we encounter data $\mathbf{x}$ independent of class

Or in concrete terms: The probability that of observing a particular email in general

Bayesian SPAM Classification

$$p(y \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid y)p(y)}{p(\mathbf{x})} \qquad \text{posterior} = \frac{\text{class-likelihood} \times \text{prior}}{\text{evidence}}$$

The Plan:

- Estimate and compare $p(y = \text{SPAM} \mid \mathbf{x})$ and $p(y = \text{HAM} \mid \mathbf{x})$ from data via Bayes
- Predict label for $\mathbf{x}$ based on which probability estimate is larger

$$\frac{p(\mathbf{x} \mid \text{SPAM}) \cdot p(\text{SPAM})}{p(\mathbf{x})} \quad \text{vs} \quad \frac{p(\mathbf{x} \mid \text{HAM}) \cdot p(\text{HAM})}{p(\mathbf{x})}$$

Question: Does the denominator add any information to this decision?

Bayesian SPAM Classification

$$p(y \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid y)p(y)}{p(\mathbf{x})} \qquad \text{posterior} = \frac{\text{class-likelihood} \times \text{prior}}{\text{evidence}}$$

The Plan:

- Estimate and compare $p(y = \text{SPAM} \mid \mathbf{x})$ and $p(y = \text{HAM} \mid \mathbf{x})$ from data via Bayes
- Predict label for $\mathbf{x}$ based on which probability estimate is larger

$$p(\mathbf{x} \mid \text{SPAM}) \cdot p(\text{SPAM}) \quad \text{vs} \quad p(\mathbf{x} \mid \text{HAM}) \cdot p(\text{HAM})$$

Answer: Nope, so we compare numerators. Note, no longer probabilities. Think of as **scores**

The Naïve Bayes Assumption

OK, so we need a way to estimate these SPAM and HAM scores from the data for email $\mathbf{x}$

$$p(\text{SPAM} \mid \mathbf{x}) \propto p(\mathbf{x} \mid \text{SPAM}) \cdot p(\text{SPAM}) \quad \text{vs} \quad p(\mathbf{x} \mid \text{HAM}) \cdot p(\text{HAM}) \propto p(\text{HAM} \mid \mathbf{x})$$

Consider again the following training data.

How can we estimate $p(\mathbf{x} = [\text{buy, viagra}] \mid SPAM)$?

HAM	SPAM	SPAM	SPAM	HAM
work	nigeria	fly	money	fly
buy	opportunity	buy	buy	home
money	viagra	nigeria	fly	nigeria

The Naïve Bayes Assumption

OK, so we need a way to estimate these SPAM and HAM scores from the data for email $\mathbf{x}$

$$p(\text{SPAM} \mid \mathbf{x}) \propto p(\mathbf{x} \mid \text{SPAM}) \cdot p(\text{SPAM}) \quad \text{vs} \quad p(\mathbf{x} \mid \text{HAM}) \cdot p(\text{HAM}) \propto p(\text{HAM} \mid \mathbf{x})$$

Consider again the following training data.

How can we estimate $p(\mathbf{x} = [\text{buy, viagra}] \mid SPAM)$?

JOINT

Joint probabilities are hard to estimate from data. Here is where we make our **naïve** assumption

Naïve Bayes Assumption: The features of $\mathbf{x}$ are **independent** given the class label y

The Naïve Bayes Assumption

Naïve Bayes Assumption: The features of $\mathbf{x}$ are **independent** given the class label y

Conditional Independence Example: You have two (possibly different) weighted coins. Pick a coin at random and flip it three times.

$$p(\mathbf{x} = [\mathrm{H}, \mathrm{H}, \mathrm{T}] \mid C_1) \stackrel{?}{=} p(H \mid C_1) \cdot p(H \mid C_1) \cdot p(T \mid C_1)$$

Note: Flips are independent given knowledge of the choice of coin, but **NOT** without.

$$p(\mathbf{x} = [\mathrm{H}, \mathrm{H}, \mathrm{T}]) \neq p(H) \cdot p(H) \cdot p(T)$$

For 1 coin

C_1, C_2, C_3

The Naïve Bayes Assumption

Naïve Bayes Assumption: The features of $\mathbf{x}$ are **independent** given the class label y

The Crux: What does this assumption mean for us and Email classification?

$$p(\mathbf{x} = [\text{buy, viagra}] \mid SPAM) = p(\text{buy} \mid \text{SPAM}) \cdot p(\text{viagra} \mid \text{SPAM})$$

This means that we can estimate class-conditional probabilities of an email by multiplying class-conditional probabilities of individual words

This is much easier to do!

Question: Is this assumption valid?

$p([\text{PEANUT, BUTTER}] \mid \text{HAM})$

The Naïve Bayes Assumption

Naïve Bayes Assumption: The features of $\mathbf{x}$ are **independent** given the class label y

The Crux: What does this assumption mean for us and Email classification?

$$p(\mathbf{x} = [\text{buy, viagra}] \mid SPAM) = p(\text{buy} \mid \text{SPAM}) \cdot p(\text{viagra} \mid \text{SPAM})$$

This means that we can estimate class-conditional probabilities of an email by multiplying class-conditional probabilities of individual words

This is much easier to do!

Answer: Probably not! But it makes things this method tractable, so we make it anyway

Example: Consider $p(\mathbf{x} = [\text{peanut, butter}] \mid HAM)$

Naïve Bayes

OK, so we need a way to estimate these SPAM and HAM scores from the data for email $\mathbf{x}$

$$p(\text{SPAM} \mid \mathbf{x}) \propto p(\mathbf{x} \mid \text{SPAM}) \cdot p(\text{SPAM}) \quad \text{vs} \quad p(\mathbf{x} \mid \text{HAM}) \cdot p(\text{HAM}) \propto p(\text{HAM} \mid \mathbf{x})$$

To estimate $p(\text{term} \mid \text{Class})$ we just have to count instances of term in Class in training data

$$\hat{p}(\text{term} \mid \text{Class}) = \frac{\#\text{ instances of term in Class}}{\#\text{ total } \mathbf{words} \text{ in Class}}$$

Note: Here by **words** we include repeated instances of all terms

p(term | SPAM)

Naïve Bayes

To estimate $p(\text{term} \mid \text{Class})$ we just have to count instances of term in Class in training data

$$\hat{p}(\text{term} \mid \text{Class}) = \frac{\#\text{ instances of term in Class}}{\#\text{ total } \mathbf{words} \text{ in Class}}$$

Example: Estimate $p(\text{buy} \mid \text{SPAM})$ from the training data

$\hat{p}(\text{buy} \mid \text{SPAM}) = \frac{2}{9}$

HAM	SPAM	SPAM	SPAM	HAM
work	nigeria	fly	money	fly
buy	opportunity	buy	buy	home
money	viagra	nigeria	fly	nigeria

9

Naïve Bayes

To estimate $p(\text{term} \mid \text{Class})$ we just have to count instances of term in Class in training data

$$\hat{p}(\text{term} \mid \text{Class}) = \frac{\#\text{ instances of term in Class}}{\#\text{ total } \mathbf{words} \text{ in Class}}$$

Example: Estimate $p(\text{buy} \mid \text{HAM})$ from the training data

HAM	SPAM	SPAM	SPAM	HAM
work	nigeria	fly	money	fly
buy	opportunity	buy	buy	home
money	viagra	nigeria	fly	nigeria

$\hat{p}(\text{buy} \mid \text{HAM}) = \frac{1}{6}$

Naïve Bayes

OK, so we can estimate $p(\mathbf{x} \mid y)$ by computing term probabilities and multiplying them together

We still need to estimate the class priors $p(y)$

Two ways to do this:

- Ask an expert what the priors should be (e.g. experts say 80% of all email is SPAM)
- Estimate class priors from the data

$$\hat{p}(\text{Class}) = \frac{\#\text{ emails from Class}}{\#\text{ total emails in training data}}$$

Naïve Bayes

To estimate $p(\text{Class})$ we just have to count emails in training data

$$\hat{p}(\text{Class}) = \frac{\#\text{ emails from Class}}{\#\text{ total emails in training data}}$$

Example: Estimate $p(\text{HAM})$ and $p(\text{SPAM})$ from the training data

HAM	SPAM	SPAM	SPAM	HAM
work	nigeria	fly	money	fly
buy	opportunity	buy	buy	home
money	viagra	nigeria	fly	nigeria

$\hat{p}(\text{HAM}) = 2/5$

$\hat{p}(\text{SPAM}) = 3/5$

Naïve Bayes

Putting it all together:

Example: Compute the HAM score for $\mathbf{x} = [\text{work}, \text{nigeria}]$

HAM	SPAM	SPAM	SPAM	HAM
work	nigeria	fly	money	fly
buy	opportunity	buy	buy	home
money	viagra	nigeria	fly	nigeria

$$p(\text{HAM} \mid \mathbf{x}) \propto p(\mathbf{x} \mid \text{HAM})\, p(\text{HAM}) =$$

$$\hat{p}(\text{work} \mid \text{HAM}) \cdot \hat{p}(\text{nigeria} \mid \text{HAM}) \cdot \hat{p}(\text{HAM}) = \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{2}{5} = \frac{1}{90}$$

Naïve Bayes

Putting it all together:

Example: Compute the SPAM score for $\mathbf{x} = [\text{work, nigeria}]$

HAM	SPAM	SPAM	SPAM	HAM
work	nigeria	fly	money	fly
buy	opportunity	buy	buy	home
money	viagra	nigeria	fly	nigeria

$$p(\text{SPAM} \mid \mathbf{x}) \propto p(\mathbf{x} \mid \text{SPAM})p(\text{SPAM}) =$$

$$\hat{p}(\text{work} \mid \text{SPAM}) \cdot \hat{p}(\text{nigeria} \mid \text{SPAM}) \cdot \hat{p}(\text{SPAM}) = \quad = \frac{0}{9} \cdot \frac{2}{9} \cdot \frac{3}{5} = 0$$

$\frac{0}{9} \cdot \frac{2}{9} \cdot \frac{3}{5}$

A Not-So-Little Hitch

We found that $p(\text{SPAM} \mid \mathbf{x} = [\text{work, nigeria}]) = 0$

This should bother you …

The problem is that work did not show up in the SPAM training set

From the spammer's perspective, including an uncommon word not in the training set would lead to $p(\text{SPAM} \mid \mathbf{x}) = 0$ and the filter would never catch it

It would be nice if we could ensure that no emails resulted in a zero SPAM or HAM score …

Additive Laplace Smoothing

We want to avoid zero probability scores when a term isn't in the training set

Silly Idea: Add 1 to all of the term frequency counts

$$\hat{p}(\text{term} \mid \text{Class}) \stackrel{?}{=} \frac{\#\text{ instances of term in Class} + 1}{\#\text{ total } \mathbf{words} \text{ in Class}}$$

But... it would be nice if $\hat{p}(\text{term} \mid \text{Class})$ behaved like a probability distribution

We're going to estimate a probability for every term in vocabulary V

We can restore the sum-to-one nature of these estimates by adding $|V|$ to denominator

Additive Laplace Smoothing

We want to avoid zero probability scores when a term isn't in the training set

$$\hat{p}(\text{term} \mid \text{Class}) = \frac{\#\text{ instances of term in Class} + 1}{\#\text{ total } \mathbf{words} \text{ in Class} + |V|}$$

What about the ridiculous case when we have no SPAM or no HAM documents in training set?

We can do additive smoothing for the class prior estimates as well

$$\hat{p}(\text{Class}) = \frac{\#\text{ emails from Class} + 1}{\#\text{ total emails in training data} + |C|}$$

where |C| is the number of classes (2 for SPAM vs HAM)

Naïve Bayes

$|V| = 11111111 = 8$

Example: Re-compute the SPAM score for $\mathbf{x} = [\text{work}, \text{nigeria}]$ with Laplace smoothing

HAM	SPAM	SPAM	SPAM	HAM
work	nigeria	fly	money	fly
buy	opportunity	buy	buy	home
money	viagra	nigeria	fly	nigeria

$$p(\text{HAM} \mid \mathbf{x}) \propto p(\mathbf{x} \mid \text{HAM}) p(\text{HAM}) =$$

$$\hat{p}(\text{work} \mid \text{HAM}) \cdot \hat{p}(\text{nigeria} \mid \text{HAM}) \cdot \hat{p}(\text{HAM}) =$$

$$\frac{1+1}{6+8} \cdot \frac{1+1}{6+8} \cdot \frac{2+1}{5+2}$$

$$\frac{2}{14} \cdot \frac{2}{14} \cdot \frac{3}{7} = \boxed{\frac{3}{7^3}}$$

$$= 0.0087$$

Naïve Bayes

Example: Re-compute the HAM score for $\mathbf{x} = [\text{work}, \text{nigeria}]$ with Laplace smoothing

HAM	SPAM	SPAM	SPAM	HAM
work	nigeria	fly	money	fly
buy	opportunity	buy	buy	home
money	viagra	nigeria	fly	nigeria

$$p(\text{SPAM} \mid \mathbf{x}) \propto p(\mathbf{x} \mid \text{SPAM}) p(\text{SPAM}) =$$

$$\hat{p}(\text{work} \mid \text{SPAM}) \cdot \hat{p}(\text{nigeria} \mid \text{SPAM}) \cdot \hat{p}(\text{SPAM}) =$$

$$\frac{0+1}{9+8} \cdot \frac{2+1}{9+8} \cdot \frac{3+1}{5+2}$$

$$\frac{1}{17} \cdot \frac{3}{17} \cdot \frac{4}{7} \cdot$$

$$\frac{12}{17 \cdot 17 \cdot 7} = .0059$$

Naïve Bayes

Example: How would Naïve Bayes classify the email $\mathbf{x} = [\text{work}, \text{nigeria}]$?

HAM	SPAM	SPAM	SPAM	HAM
work	nigeria	fly	money	fly
buy	opportunity	buy	buy	home
money	viagra	nigeria	fly	nigeria

Naïve Bayes

- The Naïve Bayes classifier is a probabilistic classifier
- We compute the posterior score of message $\mathbf{x}$ belonging to class c as

$$p(y = c \mid \mathbf{x}) \propto \hat{p}(c) \prod_k \hat{p}(x_k \mid c)$$

- Predicted class is the one with the highest posterior score

$$\hat{y} = \arg\max_c p(y = c \mid \mathbf{x}) = \arg\max_c \hat{p}(c) \prod_k \hat{p}(x_k \mid c)$$

Numerical Hiccup

For long messages, have to multiply a lot of probabilities together

$$\hat{y} = \arg\max_c \hat{p}(c) \prod_k \hat{p}(x_k \mid c)$$

Probabilities are numbers less than 1. If you multiply enough of them, you could get **underflow**

Fix: Compute and store the **log** of the term-class score estimates

Recall: The log of a product is the sum of the logs: $\log(ab) = \log(a) + \log(b)$

Predicted class becomes: $\hat{y} = \arg\max_c \log \hat{p}(c) + \sum_k \log \hat{p}(x_k \mid c)$

Training and Prediction Complexity

How much does Naïve Bayes cost to train?

How much does Naïve Bayes cost to make a prediction?

If-Time: Generative Models

Why is Naïve Bayes called a **generative** model?