

Lecture 14: Bayesian learning

CSI 5v93: Introduction to machine learning

Baylor University
Computer Science Department

Dr. Greg Hamerly
<http://cs.baylor.edu/~hamerly/>

CSI 5v93: Introduction to machine learning, Lecture 14 – p. 1/18

Announcements

- Homework 4 assigned soon

CSI 5v93: Introduction to machine learning, Lecture 14 – p. 2/18

Questions?

Comments on homework 3

Overall, great work!

- nice tables and graphics!
- many good self-posed questions about “why” and picking out interesting aspects
- good explanations of experiments and results

Some constructive remarks:

- color can be helpful in graphs (send electronically if you can't print them)
- good to ask difficult questions, but answer them carefully
- explanation, graphics, and results should be first – not code
- please put all parts of the document into \LaTeX , including graphs and tables and code – make the document uniform
- be careful of using % in \LaTeX – make sure you use $\%$, otherwise you will use %, which is the comment character!

Comparing to the base rate

One thing that no one did on the vowel task: report and compare to the “base rate”.

- the base error rate is the error rate if you simply predict one class (the largest population class) every time
- the classes were evenly distributed in test and training data
- base error rate was $1 - 1/11 = 100\% - 9.1\% = 90.9\%$

So even though LDA and linear regression don't do great on this task, they do much better than 90.9% error!

Bayesian learning and naive Bayes

- Bayes' rule and modelling
- naive Bayes
- smoothing
- binning continuous variables
- applications to text
- shaping probabilities

See handouts from Mitchell as primary source (also see section 6.6.3 in your book)

Bayes' rule

$$\Pr(M|X) = \frac{\Pr(X|M) \Pr(M)}{\Pr(X)}$$

Breaking it apart:

- M – model
- X – data
- $\Pr(M|X)$ – probability of the model given the data
- $\Pr(X|M)$ – probability of the data given the model
- $\Pr(M)$ – prior probability of the model
- $\Pr(X)$ – prior probability of the data

Using Bayes' rule for classification

$$\Pr(M|X) = \frac{\Pr(X|M) \Pr(M)}{\Pr(X)}$$

How do we use this to classify an input x ?

Different methods, depending on our assumptions:

- MAP – Maximum A Posteriori
- ML – Maximum Likelihood

Naive Bayes classifier

A naive Bayes classifier is based on Bayes' rule.

It's a simple, effective tool for learning from data.

It's called *naive* because of the... **Naive (conditional) independence assumption**

Assumption: input variable i is independent of input variable j , *given the class*. This is called *conditional independence*.

In probability terms, if input x has features x_i and x_j :

$$\Pr(x_i, x_j | c) = \Pr(x_i | c) \Pr(x_j | c)$$

What does this mean?

Back to Bayes' rule

$$\Pr(M|X) = \frac{\Pr(X|M) \Pr(M)}{\Pr(X)}$$

The naive assumption affects $\Pr(X|M)$, which can be expanded as:

$$\begin{aligned} \Pr(X|M) &= \Pr(X_1|M) \Pr(X_2|M) \cdots \Pr(X_d|M) \\ &= \prod_{i=1}^d \Pr(X_i|M) \end{aligned}$$

So instead of having a multivariate model for d -dimensional data, we have d univariate models.

Naive Bayes probabilities

Starting again with Bayes' rule:

$$\Pr(M|X) = \frac{\Pr(X|M) \Pr(M)}{\Pr(X)}$$

and substituting our new definition:

$$\Pr(X|M) = \prod_{i=1}^d \Pr(X_i|M)$$

we get the new probability:

$$\Pr(M|X) = \frac{\Pr(M) \prod_{i=1}^d \Pr(X_i|M)}{\Pr(X)}$$

Naive Bayes classification

MAP classification:

$$\begin{aligned} m_{MAP} &= \arg \max_{m \in M} \Pr(M = m|X) \\ &= \arg \max_{m \in M} \frac{\Pr(M = m) \prod_{i=1}^d \Pr(X_i|M = m)}{\Pr(X)} \\ &= \arg \max_{m \in M} \Pr(M = m) \prod_{i=1}^d \Pr(X_i|M = m) \end{aligned}$$

ML classification:

$$\begin{aligned} m_{ML} &= \arg \max_{m \in M} \Pr(X|M = m) \\ &= \arg \max_{m \in M} \prod_{i=1}^d \Pr(X_i|M = m) \end{aligned}$$

Typical naive Bayes application: discrete data

Suppose that our data is a *text document*. How would you model this?

In other words, we need a probability model for a document:

$$\Pr(X = \text{"John is a computer scientist. . ."}|M)$$

If we have this model $\Pr(X|M)$, then we can use Bayes' rule to do classification.

A bag-of-words model

Suppose that we model a document as a collection (“bag”) of words, without order.

For example, the following are equivalent documents from the bag-of-words model:

- document a = “John is a computer scientist”
- document b = “scientist a computer John is”
- document c = “computer a is John scientist”

Then we need $\Pr(X = a|M)$.

A bag-of-words model

How do we model $\Pr(X = \text{"John is a computer scientist..."}|M)$?

Imagine that we limit the vocabulary – to say, 10 words. Only these words are permitted: at computer scientist is c++ john baylor a he where.

Then using the bag-of-words model, this is a 10-dimensional problem. Each dimension can have a 1 (the word is present in the document) or 0 (the word is not present in the document).

The document “John is a computer scientist” would have the feature vector:

at	computer	scientist	is	c++	john	baylor	a	he	where
0	1	1	1	0	1	0	1	0	0

We ignore duplicate words for now.

So now we have a discrete representation, but not yet a probability model.

The bag-of-words model

For discrete data, we represent it using a non-parametric model, which means counting.

Counting probabilities:

$$\Pr(X = x) = \frac{\text{count}(X = x)}{n}$$

where $\text{count}(\alpha)$ means to count all records where α is true, and n is the total number of records considered in the training data.

So to model a document, we need to learn a probability such as above – the probability of the individual words appearing together in that document.

Modelling the joint bag-of-words probability

Document a = “John is a computer scientist”

Classes: EMAIL or SPAM

$$\Pr(X = a | \text{EMAIL}) = \frac{\text{count}(X = a \cap \text{EMAIL})}{\text{count}(\text{EMAIL})}$$

To calculate this, we would have to count all documents in the class EMAIL, and look for documents that have the words “john, is, a, computer, scientist”.

Likewise for SPAM:

$$\Pr(X = a | \text{SPAM}) = \frac{\text{count}(X = a \cap \text{SPAM})}{\text{count}(\text{SPAM})}$$

2-minute journal

Please write a response to the following on a piece of paper and hand it in immediately. Please make it anonymous (no names). Write about:

- major points you learned today
- areas not understood or requiring clarification