

Lecture 17: 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 17 – p. 1/12

Questions?

CSI 5v93: Introduction to machine learning, Lecture 17 – p. 2/12

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)

Smoothing zero probabilities

A standard way to deal with zero probabilities is to eliminate them by smoothing.

The simplest way to smooth probabilities is by adding a smoothing constant λ . For example,

$$\Pr(X = x|M) = \frac{\text{count}(X = x \cap M) + \lambda}{\text{count}(M) + k\lambda}$$

Where $\lambda > 0$ and k is the number of different values of X that have been observed.

Two ways to model words with naive Bayes

1. Each word in the vocabulary is a dimension with 0/1 value (the way I presented), therefore we have $|V|$ probability functions.
2. There is one attribute that represents all words (1 attribute with $|V|$ values), therefore we have 1 probability function. This is Mitchell's method.

With the first method, we consider the probability that a word is in a document, and the probability that it is not in the document:

$$\Pr(\text{"quick quick fox"}|M) = \Pr(\text{"quick"}|M) \Pr(\text{"fox"}|M) \dots \Pr(\neg\text{"assume"}|M) \dots$$

Here each document is a string of 0/1 values, each having some probability. We count multiple words only once.

With the second method, we consider only the probabilities of words in the document:

$$\Pr(\text{"quick quick fox"}|M) = \Pr(\text{"quick"}|M) \Pr(\text{"quick"}|M) \Pr(\text{"fox"}|M)$$

Here each document is a list of words, we include multiple words, we don't include probabilities of words that don't occur in the document.

CSI 5v93: Introduction to machine learning, Lecture 17 – p. 5/12

Smoothing for two model types

Method 1 (counts occur per document; multiple occurrences in a document are ignored)

$$\Pr(\text{"fox"}|M) = \frac{\text{count}(\text{"fox"} \cap M) + \lambda}{\text{count}(M) + 2\lambda}$$
$$\Pr(\neg\text{"fox"}|M) = \frac{\text{count}(\neg\text{"fox"} \cap M) + \lambda}{\text{count}(M) + 2\lambda}$$

Method 2 (counts occur over all words)

$$\Pr(\text{"fox"}|M) = \frac{\text{count}(\text{"fox"} \cap M) + \lambda}{\text{count}(M) + |V|\lambda}$$

CSI 5v93: Introduction to machine learning, Lecture 17 – p. 6/12

Binning

Most of the time with naive Bayes practitioners use the non-parametric model.

- most general assumption
- requires most data for learning

Suppose that you have real-valued data – how do you model this with a non-parametric model that expects discrete values?

Answer: divide up the real space into discrete regions.

Binning methods

Equal-width binning – choose a uniform interval width, and cover all inputs with bins of that width.

Equal-frequency binning – choose a number of bins, and cover all inputs with that many bins, each bin having the same number of examples.

Advantages? Disadvantages?

Shaping probabilities: thresholds

When classifying an example x , we have the probability that it belongs to each class: $\Pr(M = m|X = x)$.

We classify to the class that gives the largest probability.

If the number of classes is 2, then we really have only one probability (since the two probabilities must sum to 1).

The threshold for classifying to class A versus class B is usually 50%. In other words, classify to A if $\Pr(A|X) \geq 50\%$.

If we want to be more sensitive to one class over the other, we can change this threshold. This is equivalent to modifying the priors.

Shaping probabilities

Usually with naive Bayes, the probabilities that it gives are extreme – they are very close to 0 or 1.

This causes the probabilities to be unreliable.

How can we deal with this?

- ranking probabilities
- shaping probabilities

Project discussion

- what is your dataset?
- what is the classification task?
- have you selected the features?
- task 2 – feature ranking

CSI 5v93: Introduction to machine learning, Lecture 17 – p. 11/12

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

CSI 5v93: Introduction to machine learning, Lecture 17 – p. 12/12