

Lecture 16: 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 16 – p. 1/18

Questions?

CSI 5v93: Introduction to machine learning, Lecture 16 – p. 2/18

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)

Naive Bayes classifier

A naive Bayes classifier is based on Bayes' rule.

Naive 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 :

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

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:

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

Using log-probabilities

If we multiply many small numbers (between 0 and 1), we will get a very small number. Using logarithms helps us fix this problem.

$$\begin{aligned} \Pr(X = a|M) &= \prod_{i=1}^d \frac{\text{count}(X_i = a_i \cap M)}{\text{count}(M)} \\ \log \Pr(X = a|M) &= \log \prod_{i=1}^d \frac{\text{count}(X_i = a_i \cap M)}{\text{count}(M)} \\ &= \sum_{i=1}^d \log \frac{\text{count}(X_i = a_i \cap M)}{\text{count}(M)} \\ &= \sum_{i=1}^d \left[\log \text{count}(X_i = a_i \cap M) - \log \text{count}(M) \right] \\ &= \left[\sum_{i=1}^d \log \text{count}(X_i = a_i \cap M) \right] - d \log \text{count}(M) \end{aligned}$$

Using log-probabilities

We can apply the log probabilities to the entire Bayes' rule:

$$\begin{aligned}\log \Pr(M|X) &= \log \frac{\Pr(X|M) \Pr(M)}{\Pr(X)} \\ &= \log \Pr(X|M) + \log \Pr(M) - \log \Pr(X) \\ &= \left[\sum_{i=1}^d \log \text{count}(X_i = a_i \cap M) \right] - d \log \text{count}(M) \\ &\quad + \log \Pr(M) - \log \Pr(X)\end{aligned}$$

Then the MAP classification is still the class M that gives the largest $\log \Pr(M|X)$ (similar for ML).

Prior probabilities – $\Pr(M)$

For the tasks we will use, we will model $\Pr(M)$ as the frequency of class M .

Therefore:

$$\begin{aligned}\Pr(M) &= \frac{\text{count}(M)}{n} \\ \log \Pr(M) &= \log \text{count}(M) - \log(n)\end{aligned}$$

where n is the total number of records in the training set, and $\text{count}(M)$ is the number of times that class M occurs in the training set.

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.

Example of smoothing zero probabilities

We observe the following frequencies in the training set:

$$\begin{aligned}\text{count}(\text{TEMP} = \text{HIGH} \cap \text{SUNNY}) &= 4 \\ \text{count}(\text{TEMP} = \text{MED} \cap \text{SUNNY}) &= 2 \\ \text{count}(\text{TEMP} = \text{LOW} \cap \text{SUNNY}) &= 0\end{aligned}$$

Then if $\lambda = 1$

	before smoothing	after smoothing
$\Pr(\text{TEMP} = \text{HIGH} \text{SUNNY})$	$= 4/6$	$\frac{4+1}{6+3} = 5/9$
$\Pr(\text{TEMP} = \text{MED} \text{SUNNY})$	$= 2/6$	$\frac{2+1}{6+3} = 3/9$
$\Pr(\text{TEMP} = \text{LOW} \text{SUNNY})$	$= 0/6$	$\frac{0+1}{6+3} = 1/9$

Naive Bayes probability densities

We have described the naive Bayes classifier as using non-parametric probability models (based on counts).

If we have knowledge about the data that it is parametric (e.g. Gaussian), we can use that instead.

Example

For example, modelling two attributes with naive Bayes for bouncy balls:

- attribute 1: color (red, green, blue, orange, yellow, white, black)
- attribute 2: bounce return (real number)

We could model “color” with a typical non-parametric counting density function.

We could model “bounce return” with a Gaussian distribution (assuming this is appropriate).

The joint probability under the naive assumption is still just the product of the two probability functions.

$$\Pr(\text{color}=\text{red} \cap \text{bounce}=90\%) = \Pr(\text{color}=\text{red}) \Pr(\text{bounce}=90\%)$$

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

CSI 5v93: Introduction to machine learning, Lecture 16 – p. 17/18

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 16 – p. 18/18