Computer Science
University of Warwick

Exercise 2: Calculating with Naive Bayes

Module CS909 March 8, 2019

1 Introduction

In this exercise you will get some practice at calculating with naive Bayes classification model. The work is based on the material in Lecture 6 - Classification: KNN, Naïve Bayes and Cost-Sensitive Learning.

You are expected to be able to:

- explain the structure of the naive Bayes model;
- compute prior and conditional probabilities from a dataset;
- apply the naive Bayes model to compute class posterior probabilities;
- modify the naive Bayes model calculations to cope with the presence of missing data;
- modify the naive Bayes model calculations using the Laplace estimator.

2 Theory

The theory underpinning the naive Bayes was covered in lectures, but there is a quick recap of the basic formulae. Note that you are expected to memorise these formulae for the exam.

The basic theorem underpinning inference is Bayes' theorem:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}. \tag{1}$$

In this equation, E denotes the *observed data* and H is the *hypothesis*. In our applications, E is the set of input values for a particular example and H is the output value (or class). Each term in the equation has a name:

prior $P(H)$ is the prior probability of the hypothesis without seeing any data. In our application, it is the probability of a class given by the proportion of examples from that class in the dataset.

likelihood $P(E|H)$ is the probability of observing a particular set of values, given that the example belongs to a particular class.

posterior $P(H|E)$ is the probability of the class given that some data values have been observed. It is the quantity we are trying to model.

evidence $P(E)$ is a normalising constant that ensures that the posterior probabilities of all the classes add up to one.

Bayes' theorem is quite general and can be applied to classification problems. The only term that is difficult to calculate is the likelihood $P(E|H)$. The naive Bayes model makes the simplification that this can be decomposed as follows:

$$P(E|H) = \prod_{i=1}^{n} P(E_i|H) = P(E_1|H) \times P(E_2|H) \times \cdots \times P(E_n|H). \tag{2}$$

In this equation E_i denotes the value of the ith input attribute. This is equivalent to saying that the attributes are *independent* given the class. While this is not likely to be exactly true, it turns out to be a surprisingly good approximation. The *unnormalised probabilities* are given by the numerator of Bayes' formula (which is the likelihood times the prior):

$$P(E|H) \times P(H) = P(E_1|H) \times P(E_2|H) \times \cdots \times P(E_n|H) \times P(H). \tag{3}$$

So, to 'train' the naive Bayes model, we need to compute the following:

1. The prior probabilities $P(C_j)$ for each class C_j.
2. The conditional probabilities $P(E_i = v_k|C_j)$ for each value v_k of each input attribute E_i.

To apply the model to a new data point, we use Equation (2) to combine the appropriate probabilities computed during the training phase. For the sake of simplicity of calculation, we shall restrict ourselves to nominal attributes, but the model can be applied to continuous data as well (usually with a Gaussian distribution for $P(E_i|H)$).

If the value of an attribute is missing, it is simply omitted from the calculations.

3 Worked Example

These formulae look more daunting than they are to work with. A few examples should clarify matters. We are going to use the following dataset which represents information about a cricket team (form, number of key players present, ground where the match was played, whether they batted first) for predicting the result.

Form	Key Players	Ground	Match Result
Variable	one	B	Won
Good	one	B	Won
Variable	neither	B	Won
Poor	one	B	Lost
Poor	neither	B	Lost
Good	both	B	Lost
Poor	one	A	Won
Poor	one	A	Won
Poor	neither	A	Won
Good	neither	A	Won
Variable	both	A	Won
Variable	both	A	Won
Good	one	A	Lost
Good	both	A	Lost

To perform the calculations we can use the built-in calculator in Windows. Launch the calculator from the Accessories group of Programs and select 'scientific' mode.

In the exam you can bring your own calculator provided that it is in the CASIO FX82, FX83 and FX85 series.

1. We start by computing the prior probabilities of the two classes. This is done by simply calculating the fraction of examples in each class. Denote the 'Won' class by W and the 'Lost' class by L. There are fourteen examples, so $P(W) = 9/14$ and $P(L) = \ldots$.

2. The next step is to compute the likelihood of each attribute given the two classes. We shall start with the Ground. To calculate $P(\text{Ground} = \text{A}|W)$ we need to work out the fraction of 'Won' examples for which the Ground is A. There are 9 Won examples, of which 6 have Ground A, so

$$P(\text{Ground} = \text{A}|W) = \frac{6}{9} = \frac{2}{3}.$$

   I advise you to leave the probabilities as fractions until you have to calculate some values with them with Equation (2). In a similar way

$$P(\text{Ground} = \text{B}|W) = \frac{3}{9} = \frac{1}{3}.$$

   Now you have a go at calculating the corresponding conditional probabilities for class L.

$$P(\text{Ground} = \text{A}|L) = \frac{\ldots}{\ldots}$$
$$P(\text{Ground} = \text{B}|L) = \frac{\ldots}{\ldots}$$

3. As you can see, the calculation is not difficult, but there are potentially quite a large number of probabilities to calculate. However, for the purposes of the exam, you only need to calculate conditional probabilities for the values in the test point (though you do need to do this for all the classes; the reason for this will be clear in a moment). Suppose that we are to classify the example

$$(\text{Good}, \text{both}, \text{A}) \tag{4}$$

This means that we need to calculate the following additional conditional probabilities:

$$
\begin{aligned}
P(\text{Form} = \text{Good}|W) &= \frac{\dots}{\dots} \\
P(\text{Form} = \text{Good}|L) &= \frac{\dots}{\dots} \\
P(\text{Players} = \text{both}|W) &= \frac{\dots}{\dots} \\
P(\text{Players} = \text{both}|L) &= \frac{\dots}{\dots}
\end{aligned}
$$

4. Now we are ready to compute the posterior probabilities. The first step is to compute the *unnormalised probabilities* which correspond to Equation (3). First let us work out the value for class W:

$$
\begin{aligned}
P(\text{Form=Good}|W) \times P(\text{Players=both}|W) \times P(\text{Ground=A}|W) \times P(W) &= \frac{2}{9} \times \frac{2}{9} \times \frac{2}{3} \times \frac{9}{14} \\
&\approx 0.02116 \\
P(\text{Form=Good}|L) \times P(\text{Players=both}|L) \times P(\text{Ground=A}|L) \times P(L) &= \frac{\dots}{\dots} \times \frac{\dots}{\dots} \times \frac{\dots}{\dots} \times \frac{\dots}{\dots} \\
&\approx \dots
\end{aligned}
$$

As with the calculations of information gain, you should write down the intermediate results to a reasonable number of significant figures (four is about the minimum). One advantage of leaving probability values as fractions for as long as possible is that it introduces less rounding error (the fractions are *exact*).

5. To complete the application of Bayes' theorem, Equation (1), we need to compute the evidence $P(E)$. This is done by normalising the posterior probabilities so that they add up to one. Normalisation is simple: add up the unnormalised probability values you have calculated and divide each by the sum.

$$
\begin{aligned}
P(E) &= 0.02116 + 0.03429 = 0.05545 \\
P(W|E) &= \frac{0.02116}{0.05545} \approx 0.3816 \\
P(L|E) &= \frac{0.03429}{0.05545} \approx 0.6184
\end{aligned}
$$

These values should add up to one, which they do. So given this set of values, it is approximately twice as likely that the team lost as they won.

6. There is one more issue to address: what happens if $P(E_i|H) = 0$? Strictly speaking, this makes the corresponding conditional probability equal to zero. In fact, the same can happen (for another attribute) for another class, which would lead to the farcical scenario of all classes having an equal posterior probability of zero!

   There is a way to address this problem: Laplace estimation. Consider the following data point:

$$
(\text{Variable}, \text{both}, \text{A}) \tag{5}
$$

   which is the same as in Equation (4) except that the form is 'variable'. So we need to compute two new likelihoods:

$$
\begin{aligned}
P(\text{Form=Variable}|W) &= \frac{4}{9} \\
P(\text{Form=Variable}|L) &= \frac{0}{5}
\end{aligned}
$$

To solve the problem for the Lost class, we consider all the conditional probabilities involving the Form variable. We know that $P(\text{Form=Good}|L) = 3/5$ and it is easy to calculate that $P(\text{Form=Poor}|L) = 2/5$. The Laplace estimator is based on the idea that we don't expect there to be extreme results (i.e. 0 or 1) for the conditional probabilities. This prior belief can be converted into a calculation method: we add one to the count of each outcome. We recompute these values by adding 1 to each numerator and 3 to the denominator (because there are three values of the Form attribute). This gives the following results:

$$\begin{aligned} P(\text{Form=Variable}|L) &= \frac{1}{8} \\ P(\text{Form=Good}|L) &= \frac{4}{8} \\ P(\text{Form=Poor}|L) &= \frac{3}{8} \end{aligned}$$

As a check, the probabilities still add up to 1. Note that the probabilites still have the same ordering, but they are closer together (less extreme). Once these values are calculated, they are combined in exactly the same way as before.

This might look like a completely unprincipled hack, but in fact it can be justified as a Bayesian approach to estimating the conditional probabilities (with some prior beliefs about what they should be). However, further consideration of this point is beyond the scope of this module.

When should you apply the Laplace estimator? Arguably, all the time, but for the purposes of the examination, when the question makes it clear that you should (either by telling you directly, or when there is a calculation where a conditional probability is zero and you are asked to suggest a way round the problem).

4 Exercises

1. Using the Laplace estimator for the Form variable (for both classes), calculate the class posterior probabilities for the example

$$(\text{Poor}, \text{one}, \text{B})$$

You should be able to reuse some conditional probabilities from the worked example.

$$\begin{aligned} P(\text{Form=Poor}|W) &= \frac{\ldots}{\ldots} \\ P(\text{Form=Poor}|L) &= \frac{\ldots}{\ldots} \\ P(\ldots \qquad\qquad\qquad) &= \frac{\ldots}{\ldots} \\ P(\ldots \qquad\qquad\qquad) &= \frac{\ldots}{\ldots} \\ P(\ldots \qquad) &= \frac{\ldots}{\ldots} \\ P(\ldots \qquad) &= \frac{\ldots}{\ldots} \end{aligned}$$

Then the unnormalised probabilities are

$$\begin{aligned} P(E|W)P(W) &= \frac{\ldots}{\ldots} \times \frac{\ldots}{\ldots} \times \frac{\ldots}{\ldots} \times \frac{\ldots}{\ldots} \approx \ldots \\ P(E|L)P(L) &= \frac{\ldots}{\ldots} \times \frac{\ldots}{\ldots} \times \frac{\ldots}{\ldots} \times \frac{\ldots}{\ldots} \approx \ldots \end{aligned}$$

Now compute the normalisation constant: I make it 0.06389 and the final probabilities

$$P(W|E) \approx 0.4969 \qquad P(L|E) \approx 0.5031.$$