COMS10003 : Dealing with Uncertainty

Probability II

Andrew Calway

December 16, 2014

Introduction

We continue our look at probability by considering one of its most important topics: **Bayes' Theorem** (or Law, or Rule). It plays a key part in enabling probability theory to be used in real applications and it is used in many areas of CS. Its importance is best summed up by this quote by Sir Harold Jeffreys that Bayes' Theorem "is to the theory of probability what Pythagoras's theorem is to geometry" (thanks to Wikipedia). We will also look at how it is used in one of the most simple but useful applications of probability theory - the **naive Bayes classifier**.

And from the same Wikipedia page, some history:"Bayes' theorem is named after Thomas Bayes (1701-1761), who first suggested using the theorem to update beliefs. His work was significantly edited and updated by Richard Price before it was posthumously read at the Royal Society. The ideas gained limited exposure until they were independently rediscovered and further developed by Laplace, who first published the modern formulation in his 1812 Thorie analytique des probabilit". So there you are.

For these notes I've made use of the two books from my bookcase and a web page that provides an excellent explanation of Bayes' Theorem:

Probability, Random Variables and Stochastic Processes by A.Papoulis, McGraw-Hill
Linear Algebra and Probability for Computer Science Applications by E.Davis, CRC Press
An Intuitive Explanation of Bayes' Theorem by Eliezer S. Yudkowsky
`http://yudkowsky.net/rational/bayes`.

An Example

It is best to start with an example, both to see how we use Bayes' Theorem and why we need it. It is borrowed from Eliezer Yudkowsky's web page and is a classic Bayes problem.

> 1% of women at age forty who participate in routine screening have breast cancer. 80% of women with breast cancer will get positive mammographies. 9.6% of women without breast cancer will also get positive mammographies. A

> woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer?

So the difficulty here is that both women with and without breast cancer may get a positive mammography, with different probabilities, and only a very small number of woman actually have cancer. The thing that we are trying to find is something that we saw in the last lecture - the conditional probability that a woman has cancer *given* that she has a positive mammography, i.e. $P(C|M)$, where C denotes cancer and M a positive mammography.

How to proceed? Let W denote the number of women of forty who participate in screening. Then the number of women with cancer who will have a positive mammography is $(80/100)(1/100)W$, i.e. 80% of 1% of W. Similarly, the number of women without cancer who will have a positive mammography is $(9.6/100)(99/100)W$, i.e. 9.6% of 99% of W.

Thus we can now compute the fraction of women who will have cancer if they have a positive mammography, which in turn corresponds to the conditional probability that we seek, i.e.

$$P(C|M) = \frac{\#\text{ with cancer and positive mammography}}{\text{total }\#\text{ having positive mammography}} \tag{1}$$

$$P(C|M) = \frac{(80/100)(1/100)W}{(80/100)(1/100)W + (9.6/100)(99/100)W} = \frac{80}{80 + 9.6 \times 99} \approx 7.8\% \tag{2}$$

This result may seem low but careful thought should convince you that it is correct. For example, if there were 1000 women being screened, then around 10 would have cancer and 8 of those would have a positive mammography. Around 990 would not have cancer and 95 of those (9.6%) would have a positive test. Thus 8 out of 103 having a positive mammography would have cancer, i.e. around 7.8%.

Bayes' Theorem

Note from the above equation that 80/100 is the probability that a woman with cancer will get a positive mamography. i.e. $P(M|C)$, and that 1/100 is the probabilty that a woman will have cancer, i.e. P(C). Similarly, 9.6/100 is the probability that a woman without cancer will get a positive mamography. i.e. $P(M|\neg C)$, and that 99/100 is the probabilty that a woman will not have cancer, i.e. $P(\neg C)$, where $\neg C$ denotes "not cancer".

Hence, after cancelling the Ws top and bottom, we can write the above equation as

$$P(C|M) = \frac{P(M|C)P(C)}{P(M|C)P(C) + P(M|\neg C)P(\neg C)} \tag{3}$$

If we now note that the denominator $P(M|C)P(C) + P(M|\neg C)P(\neg C)$ is the probability of a woman having a positive mamography, i.e. $P(M)$, then we get

$$P(C|M) = \frac{P(M|C)P(C)}{P(M)} \tag{4}$$

which is **Bayes' Theorem**. In short, it provides us with a means of getting at the probability of something that we don't know (whether someone has cancer) given an observation (that the test is positive) using information that we do know, i.e. the accuracy of the test ($P(M|C)$ and $P(M|\neg C)$) and our knowledge about the probability of having cancer ($P(C)$ and $P(\neg C)$).

To get a more general form, note that if we have multiple events E_i, $1 \leq i \leq n$, linked with another event A, then we have $P(A) = P(A|E_1)P(E_1)+P(A|E_2)P(E_2)\ldots+P(A|E_n)P(E_n)$ and Bayes' theorem is

$$P(E_i|A) = \frac{P(A|E_i)P(E_i)}{P(A)} = \frac{P(A|E_i)P(E_i)}{P(A|E_1)P(E_1) + P(A|E_2)P(E_2)\ldots + P(A|E_n)} \tag{5}$$

For example, A might be the outcome of a test and the E_i are events corresponding to people having and not having various deseases that the test can indicate.

Interpretating Bayes'

There are a number of different ways of interpreting Bayes' Theorem. In fact, I would recommend that you do some research on the different interpretations - start with the Wikipedia page and see where you get - `http://en.wikipedia.org/wiki/Bayes'_theorem`.

One useful interpretation is as follows. Before a mammography is conducted, the only knowledge we have is the probability of someone having cancer $P(C)$. This is known as the **prior**, i.e. it is *prior knowledge*. Once the test is conducted, then we have additional information in the form of the test outcome, i.e. positive or negative, which we can regard as **evidence**.

What Bayes' theorem tells us is how to combine this evidence with the prior to get an updated probability for cancer given the test outcome. In this interpretation, $P(M|C)$ is known as the **likelihood**, i.e. the *likelihood* that the test is positive given that someone has cancer, and $P(M)$ is the probability for the evidence. The updated probability $P(C|M)$ is then known as the **posterior**, i.e. the conditional probability that is assigned *after* the relevant evidence is taken into account.

So, in an informal way, and assuming we are talking about probabilities, we can write

$$posterior \quad = \quad \frac{likelihood \times prior}{evidence} \tag{6}$$

Another Example

This one is taken from the Wikipedia page and is attributed to an article in the New York Times in 2011.

> You are given the choice of 3 coins. Two are fair and one is unfair such that it always gives H when flipped. If you choose a coin and it gives HHH when flipped three times, what is the probability that it is the unfair coin?

We can tackle this using Bayes' Theorem to compute $P(U|HHH)$, i.e. the posterior probability of the selected coin being unfair U given that we flip HHH. From above

$$P(U|HHH) = \frac{P(HHH|U)P(U)}{P(HHH|U)P(U) + P(HHH|\neg U)P(\neg U)} \tag{7}$$

The probability of getting HHH with the unfair coin, $P(HHH|U)$, is clearly 1 since it always gives H when flipped and the probability of choosing the unfair coin $P(U)$ is 1/3. Similarly, $P(HHH|\neg U) = 1/2 \times 1/2 \times 1/2 = 1/8$ and $P(\neg U) = 2/3$, i.e. 2 out of 3 of the coins are fair, i.e. $\neg U$. This gives

$$P(U|HHH) = \frac{1 \times (1/3)}{1 \times (1/3) + (1/8) \times (2/3)} = \frac{4}{5} \tag{8}$$

which is pretty close to 1 but not 1, as we have to take into account the fact that (a) fair coins can give HHH, albeit with low probability, and (b) that there is a 2 in 3 chance that we'll pick a fair coin.

Random Variables

Before we move on to looking at the naive Bayes' classifier application, we first need to introduce a new entity - a **random variable (RV)**. We will look at these in much more detail later in the unit but for now we just need the definition.

A RV is a variable that can have different values taken from a range of possible values. The RV may be *discrete*, taking values from a finite list, or *continuous*, taking on any value within a given range. Here we will be interested in the former.

It is useful to think of RVs as linked to the outcomes of an experiment. On conducting an experiment we get an outcome E and to this outcome we associate a number $x(E)$. The sample space of the experiment then defines the range of our RV x and the different outcomes define the values that x can take on. Each outcome has an associated probability and so associated with x is a probability that it will take on a given value. For example, the probability that x will take on the value k, denoted $P(x = k)$, is determined by the probability that the corresponding outcome, E_k say, i.e. $x(E_k) = k$, will occur when the experiment is conducted.

In practice, we don't usually have such an experiment, we just assume that the values of x that we are observing were the result of an experiment - it just makes our interpretation of x easier (hopefully!). For example, let c be a RV which we use to represent the colour of an apple and it can take on 1 of 3 values - red, green or yellow. In this case our 'imaginary experiment' is the process of nature to produce the apple and a subsequent process to determine the colour of the apple. Each outcome will then give us a value for c.

Associated with a RV is therefore a finite or continuous range of probability values. This is known as its **probability distribution**, such as $P(x \equiv k)$ above or in our apple example, $P(c \equiv red)$. For discrete RVs it is a **probability mass function**, whilst for continuous RVs it is a **probability density function**. We will return to this later in the unit.

Classifiers

We can now take a look at our application. The context is *classification* - given some *attributes* (or *features*) of an unknown entity, can we *classify* it as belonging to one of a finite number of classes? A simple example will clarify things.

Imagine that you are given 3 boxes labelled "Gala" (G), "Golden Delicious" (GD) and "Granny Smith" (GS). You are then handed an apple and asked to put it in the appropriate box. If it is yellow, you put it in the GD box, if it's red, in the G box and if it's green, in the GS box. You are then acting as the classifier: assigning a class (G, GD,or GS) to an apple based on its colour attribute, i.e. green, yellow or red. In this case both our colour attribute and our class are then forms of finite RVs.

An obvious question is how you gained the knowledge that yellow apples were GD, etc. One possibility is that you learnt the relationship between colour and variety and that you did this by observation - you were shown lots of apples and told their variety. If you were observant then you could link colour and variety.

The above is what we often do when we design non-human classifiers. We construct a table of *training examples* showing the class and the attribute values of the entities that we are dealing with. For example, we could take 500 apples, note their variety and colour and then build a table with entries of the form $\{variety, colour\}$. From this table we can learn the relationship between colour and variety. This is known as *supervised classification*, i.e. we supervise the classifier by providing it with training data. This is in contrast with *unsupervised classification* which you may wish to read up on.

A simple probabilistic way of doing that is to calculate the probabilities of a variety (class) given a colour (attribute value). For example, if we are given the following table

$variety$	G	GD	G	GS	GS	G	GD	G	GD	GS
$colour$	R	G	Y	R	G	R	R	R	Y	G

and assume that all three apple varieties are equally likely, then by computing the frequencies of the $\{colour, variety\}$ pairs we can estimate the probabilities of varieties given colours, e.g. $P(var = G|col = R) \approx 3/5$, $P(var = GD|col = Y) \approx 1/2$, and so on. We could then use the probabilities to determine the varieties given the colour, e.g. if the colour is green (G)

$$variety = \arg\max_{var} P(var|col = G) \tag{9}$$

This is known as a *Maximum Likelihood* classifier - we find the *variety* that maximises the likelihood $P(var|col)$. But this is a simple example, not least in that there are only 3 classes and the classification is based on a single 3-value attribute, i.e. colour.

In real applications, we often have many classess and many attributes and it's the combinations of attributes which characterise a given class. Some examples are given below.

Character Recognition Given a hand written letter from the alphabet in the form of a binary image, classify the letter based on various geometric attributes extracted from the binary image.

Computer Vision Given an image of a face, identify the individual as one from n possible individuals in a database, using attributes such as eye colour, size of nose, distance between eyes, etc

Medical Diagnosis Given test results (attributes) for a patient, determine the likelihood that the patient has a particular desease.

Finance Given atributes such as employment status, assets, collateral, purpose of loan, etc, determine whether to approve the advance of a loan or not - a two class problem.

Text Classification Given a piece of text, such as an email, the attributes are the words in the text, and the task is to determine the class of the text, e.g. in the case of email filtering, whether it is spam, ordinary or urgent.

In these examples our simple frequency approach outlined above will hit problems. The main problem is that because there are many attributes associated with classes, it is very likely that a particular combination of attributes which needs to be classified will not appear in the training table (since there are so many possibilities) and so we are not able to compute $P(class|attributes)$ directly.

For example, if classes C depend on n attributes A_1, ... , A_n, then the likelihood of $C = c$ given a particularly combination of attribute values, i.e. $P(C = c|A_1 = a_1, \ldots, A_n = a_n)$, from a training table containing entries of the form $\{C = c, A_1 = a_1, \ldots, A_n = a_n\}$ is given by

$$\frac{\#\ \ \{C = c, A_1 = a_1, \ldots, A_n = a_n\}\ \text{ entries in }\ T}{\#\ \text{ entries containing }\ \{A_1 = a_1, \ldots, A_n = a_n\}} \tag{10}$$

If the table has no entries containing the combination $\{A_1 = a_1, \ldots, A_n = a_n\}$ this becomes $0/0$, which is not very helpful.

To get around this, we can make use of a method known as *naive Bayes classification* which makes use of Bayes' Theorem and assumptions about independence of attributes.

Naive Bayes Classification

There are two tricks in naive Bayes classification:

1. Use Bayes Theorem to estimate $P(C = c|A_1 = a_1, ..., A_n = a_n)$;

2. Assume the attributes A_1, ... , A_n are conditionally independent *given* a class, i.e. if we know the class, then knowing the value of one attribute gives no information about the likely value of another attribute.

So, from Bayes Theorem we have

$$P(C = c|A_1 = a_1, ..., A_n = a_n) = \frac{P(A_1 = a_1, \ldots, A_n = a_n|C = c)P(C = c)}{P(A_1 = a_1, \ldots, A_n = a_n)} \tag{11}$$

and using an assumption of conditional independence, the likelihood can be written as

$$\begin{aligned} &P(A_1 = a_1, \ldots, A_n = a_n|C = c) = \\ &\qquad P(A_1 = a_1|C = c)P(A_2 = a_2|C = c)\ldots P(A_n = a_n|C = c) \end{aligned} \tag{12}$$

The key point here is that if we have entries in our training table containing each of the attribute values, e.g. $A_i = a_i$, and the class c, then we can estimate each of the conditional probabilities $P(A_i = a_i|C = c)$. We are no longer relying on there being entries containing the exact combination $\{A_1 = a_1, \ldots, A_n = a_n\}$.

The assumption of conditional independence is a simplification and is unlikely to be a correct one in real applications. This is why the technique is known as 'naive'. However in many applications it has been shown to give surprisingly good results and has the advantage of being easy to implement, as the following example illustrates.

Yet Another Example[1]

> We collect emails over a period of time. 1000 were spam and 200 were not. We use these as our training examples and compute the individual occurrences of three words w_1, w_2 and w_3 in all the emails. The results are shown below

word	# spam containing *word*	# non-spam containing *word*
w_1	500	100
w_2	100	80
w_3	800	40

> The task is to determine whether an email containing the words w_1 and w_2 but **not** w_3 is more likely to be spam than non-spam.

Using the above formulation, we need to compute

[1] Many thanks to Peter Flach for this form of example which I have borrowed from his notes for the 2nd year unit Symbols, Patterns and Signals

$$P(spam|w_1, w_2, \neg w_3) = \frac{P(w_1|spam)P(w_2|spam)P(\neg w_3|spam)P(spam)}{P(w_1, w_2, \neg w_3)} \tag{13}$$

and similarly for the non-spam case. The likelihoods can be estimated from the results table:

$$\begin{aligned} P(w_1|spam) &= 500/1000 = 1/2 \\ P(w_2|spam) &= 100/1000 = 1/10 \\ P(\neg w_3|spam) &= 200/1000 = 1/5 \end{aligned} \tag{14}$$

$$\begin{aligned} P(w_1|\neg spam) &= 100/200 = 1/2 \\ P(w_2|\neg spam) &= 80/200 = 2/5 \\ P(\neg w_3|\neg spam) &= 160/200 = 4/5 \end{aligned} \tag{15}$$

and the probabilities of spam or non-spam emails are $P(spam) = 1000/1200 = 5/6$ and $P(\neg spam) = 200/1200 = 1/6$. As we are only trying to determine the largest out of $P(spam|w_1, w_2, \neg w_3)$ and $P(\neg spam|w_1, w_2, \neg w_3)$ we do not need to worry about estimating $P(w_1, w_2, \neg w_3)$ as it will be the same for both cases.

From the above we have

$$P(w_1|spam)P(w_2|spam)P(\neg w_3|spam)P(spam) = \frac{1}{2}\frac{1}{10}\frac{1}{5}\frac{5}{6} = \frac{5}{600} \tag{16}$$

and

$$P(w_1|\neg spam)P(w_2|\neg spam)P(\neg w_3|\neg spam)P(\neg spam) = \frac{1}{2}\frac{2}{5}\frac{4}{5}\frac{1}{6} = \frac{16}{600} \tag{17}$$

and hence the email is more than three times more likely to be non-spam than spam. The key point is that although on the face of it w_1 seems like a strong indicator of spam, it is also an equally strong indicator of non-spam, hence its presence in an email is not that useful. In contrast, the presence of w_2 and non-presence of w_3 are good indicators of non-spam email.