Lecture 1: Bayes theorem and Bayesian inference

Nowadays it is common to group probability and statistics together. However the two subjects developed at very different times. Statistics emerged as an important mathematical discipline in the nineteenth century, when governments began to attach importance to measurements of population, housing, disease and so on. Probability is much older, and has been studied as long ago as man took an interest in games of chance, probably pre-dating the Babylonian civilisation.

Our story starts with the famous theorem of the Rev. Thomas Bayes, published in 1763. We can only speculate why a man of the church should be interested in games of chance.

Bayes' theorem concerns the probability of two events S and D. If S and D are independent variables then by inspection we can write the scalar equation:

$$P(D\&S) = P(D) \times P(S)$$

In cases where S and D are not independent, knowing that one event has occurred gives us some information about the other, so we write:

$$P(D\&S) = P(D) \times P(S|D)$$

where $P(S|D)$ is the probability of S given the D has occurred. A good way of looking at this is to think of D standing for a disease and S for a symptom. Once we know that a patient has a particular disease we might, for example, assign a much higher value to the probability that he has the symptom. Since conjunction is commutative:

$$P(D\&S) = P(S) \times P(D|S) = P(D) \times P(S|D)$$

and by rearranging we get Bayes' Theorem:

$$P(D|S) = P(D) \times P(S|D)/P(S)$$

We can look on Bayes' Theorem as an Inference Mechanism. Suppose we are doing simple medical diagnosis. We have that:

- $P(D|S)$ is the probability of the disease given the symptom. This is what we wish to infer.
- $P(D)$ is the probability of the disease (within a population) this is a measurable quantity.
- $P(S|D)$ is the probability of the symptom given the disease. We can measure this from the case histories of the disease.
- $P(S)$ is the probability of the symptom in the population. We could again measure this with statistics, but fortunately we don't need to do so.

Note that D and S are discrete variables that can take different values. Suppose, for example, that the variable D has just two states (or possible values) d_t (meaning the patient has the disease) and d_f (meaning the patient does not have the disease), and the variable S has three states: s_1 meaning no presentation of symptom S, s_2 meaning mild symptom and s_3 meaning severe symptom. Then, through the fact that $P(D|S)$ is a probability we have that:

$$P(d_t|s_i) + P(d_f|s_i) = 1$$

regardless of the value of S. If we apply Bayes' theorem we get:

$$P(s_i|d_t)P(d_t)/P(s_i) + P(s_i|d_f)P(d_f)/P(s_i) = 1$$
$$P(s_i) = P(s_i|d_t)P(d_t) + P(s_i|d_f)P(d_f)$$

In other words, given the values for $P(S|D)$ and $P(D)$ we can calculate $P(S)$ for each state s_i of variable S. Thus we can write Bayes' Theorem as:

$$P(D|S) = \alpha \times P(D) \times P(S|D)$$

where $\alpha = 1/P(S)$ is a normalising constant ensuring that $P(D|S)$ sums to 1 for each state of S.
$P(D)$ is prior information, since we knew it before we made any measurements
$P(S|D)$ is likelihood information, since we gain it from measurement (eg. of symptoms).
In summary, Bayes' theorem can be considered to connect "prior" and "likelihood" information.

Bayesian Inference (in its most general form)

With the above definitions we can now write down a basic algorithm for Bayesian inference. Given a set of competing hypotheses which explain a data set, then, for each hypothesis:

1. Convert the prior and likelihood information in the data into probabilities
2. Multiply them together
3. Normalise the result to get the posterior probability of each hypothesis given the evidence

Select the most probable hypothesis.

Prior Knowledge

In some cases we obtain the prior probability from statistics. For example, we can calculate the prior probability as the number of instances of a disease divided by the number of patients presenting for treatment. However in many cases this is not possible - since the data isn't there, and there may also be prior knowledge in other forms.

Example from Computer Vision

Consider a program to determine whether an image contains a picture of a cat. Computer vision algorithms normally work by processing the image to identify relevant features. There are a lot of features that we could use to detect a cat, but lets start with something simple. Well write a program to extract circles from the image. If we find two adjacent circles then we will test them to see if they could be a pair of cats eyes. Lets assume that in the perfect cat the eyes have the same radius, and that they are separated by 2 times the diameter as illustrated in Figure 1.

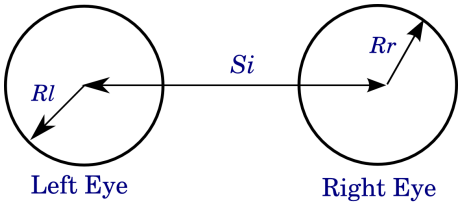

Figure 1: Prior Model of a Cat's eyes

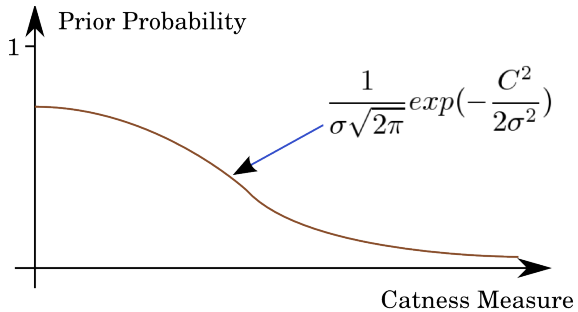

Figure 2: Heuristic association of a probability with a measure

For each pair of circles we find in the image we can calculate a catness measure which is:

$$Catness = |(R_l - R_r)/R_r| + |(S_i - 2 \times (R_l + R_r))/R_r|$$

The catness measure is zero for a perfect match to our model. We could choose some common sense way of changing our catness measure into a probability and we can make this look more respectable mathematically by choosing a distribution. Figure 2 shows how we could use a normal distribution for the purpose. Doing this we are making a subjective estimate of the probability.

An alternative strategy is to use objective methods. To do so we need to do some experimentation. In our example it would be necessary to make measurements from a large set of photographs. Every time we extract a pair of circles from a photograph, and calculate the catness measure, we also get an expert to tell us whether the extracted structure does represent a cat. From this we can construct a discrete distribution of the form shown in figure 3. For each bin (small range of the catness measure) we calculate the ratio of correctly identified cats to the total. To overcome experimental error, and to compact our representation we may try to fit an appropriate distribution. The figure shows the normal distribution, and the values of μ and σ are calculated to give the best fit to the data. These are called the maximum liklehood estimates of μ and σ.

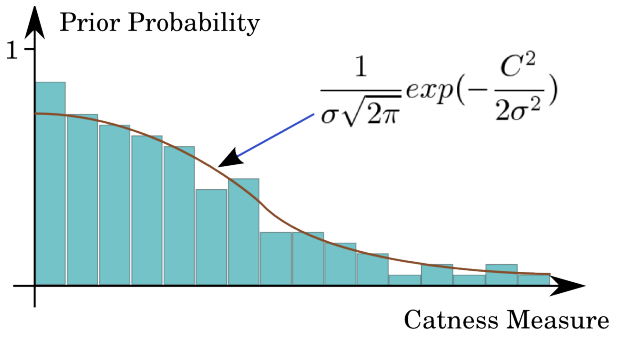

Figure 3: Discrete distribution of objective probabilities

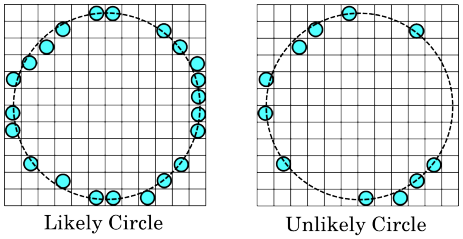

Figure 4: Likelihood Example from Computer Vision

Subjective vs. Objective Probabilities

There is a long standing debate as to whether the subjective or the objective approach is the most appropriate. Objective may seem more plausible at first, but does require lots of data and is prone to experimental error. We note in passing that some people say the Bayesian approach must be subjective. This is because, as mentioned above, there were no statistics when Bayes published the theorem, and so he would have considered a probability as purely subjective. However, we will use the term Bayesian to describe methods based on the calculus that develops out of Bayes theorem, regardless of whether the probabilities are estimated by subjective or objective methods.

Likelihood

Our prior probabilities represent long standing beliefs. They can be taken to be our established knowledge of the subject in question. When we process data and make measurements we create likelihood information. There may be uncertainty about each case, possibly due to experimental errors. In our case we can associate it with the computer vision algorithms. Our computer vision process could not just extract a circle, but also tell us how good a circle it is, for example by counting the pixels that contribute to it. Figure 4 shows one approach to determining the likelihood of a circle extracted by computer vision. Like the prior infor mation this can be converted into probabilities by some common sense procedure.

Summarising our inference rule for the cat we have:

$$P(C|I) = \alpha \times P(C) \times P(I|C)$$

$P(C)$ is the prior probability that two circles represent a cat, found by measuring catness and using prior knowledge to convert catness to probability.
$P(I|C)$ is the probability of the image information, given that two circles represent a cat. (Rather a round about way of expressing the idea that image information does actually represent two circles.) This is the likelihood information found during data measurements.

Should we use subjective or objective methods? (Many schools of thought exist) I favour the view that prior information should be subjective. It represents our belief about the domain we are considering. This is so even if data has made a substantial contribution to our belief. Likelihood information should be objective. It is a result of the data gathering from which we are going to make an inference. It makes some assessment of the accuracy of our data gathering. In practice either or both forms can be subjective or objective.