Bayesian RL Seminar

Chris Mansley

September 9, 2008

Bayes[1]

Basic Probability

One of the basic principles of probability theory, the chain rule, will allow us to derive most of the background material in Bayesian analysis. The chain rule allows us to take *joint probability*, $P(a, b)$, and write it as the product of the *conditional probability*, $P(a|b)$ and the *marginal probability*, $P(b)$ as follows

$$P(a, b) = P(a|b)P(b)$$

The selection of b as the marginal probability is arbitrary, so

$$P(a, b) = P(b|a)P(a)$$

is also true. From these two facts, we can show Bayes rule as follows

$$P(b|a) = \frac{P(a|b)P(b)}{P(a)}$$

Typically, this is presented in different terminology, as follows

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)}$$

where, $P(h)$ is the probability of hypothesis h being true before the data is seen or the belief that the agent holds before seeing the data, as known as the *prior probability*. It naturally follows that $P(d|h)$ is the probability of the data being d given the agent knows the hypothesis h, also know as the *likelihood*. By combining what the agent believes before seeing the data (the prior) with what the agent saw from the data (the likelihood), we can get the *posterior probability*, $P(h|d)$.

The denominator of the right side of the Bayes rule often is dropped because it is a normalizing factor. To see how this is true, we can expand $P(d)$ into

$$P(d) = \sum_h P(d, h)$$

[1]Most of the material in this section is from *Technical Introduction: A Primer on Probabilistic Inference* by Griffiths and Yuille

which just results from summing over all hypothesises, h, just as before in marginalization, now we can expand that joint probability just as before into

$$P(d, h) = P(d|h')P(h')$$

which gives us the final form of Bayes' formula

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h'} P(d|h')P(h')}$$

which is simple to see as a normalizing constant, because it divides by the sum over all possible hypotheses.

Two Hypotheses

If you had just two hypotheses, the hypothesis space would just be two possibilities, h_1 and h_2, with two priors, $P(h_1)$ and $P(h_2)$. We can compare the posteriors directly of these two likelihoods and priors, with *posterior odds* as follows

$$\frac{P(h_1|d)}{P(h_2|d)} = \frac{P(d|h_1)}{P(d|h_2)} \frac{P(h_1)}{P(h_2)}$$

The first part of the above ratio is called the *likelihood ratio* and the second part is called the *prior odds*. If we take the log of both sides, it is unsurprisingly called *log posterior odds*

$$\log \frac{P(h_1|d)}{P(h_2|d)} = \log \frac{P(d|h_1)}{P(d|h_2)} + \log \frac{P(h_1)}{P(h_2)}$$

This also illuminates how to think about one hypothesis is favored over another. It is simply a combination of prior beliefs plus contribution of the data.

Example

Lets assume that we have a coin that has a probability θ of turning up heads. We would like compute the probability of the data being described by a particular θ. For example,

$$P(d|\theta) = \theta^{N_H}(1-\theta)^{N_T}$$

is the likelihood (probability) that the data (in this case, counts of heads and tails) was generated by θ. This is an example of a binomial distribution.

Statistics could stop here and compute the *maximum likelihood estimate* (MLE) by estimating the fixed parameter of this stochastic model. We will now go through the motions of computing the MLE for this model.

$$\begin{aligned} P(d|\theta) &= \theta^{N_H}(1-\theta)^{N_T} \\ logP(d|\theta) &= \log \theta^{N_H} + \log(1-\theta)^{N_T} \\ logP(d|\theta) &= N_H \log \theta + N_T \log(1-\theta) \end{aligned}$$

$$
\begin{aligned}
\frac{d}{dx} log P(d|\theta) &= \frac{N_H}{\theta} + \frac{-N_T}{1-\theta} \\
0 &= \frac{N_H}{\theta} + \frac{-N_T}{1-\theta} \\
\frac{N_H}{\theta} &= \frac{N_T}{1-\theta} \\
\frac{N_H}{N_T} &= \frac{\theta}{1-\theta} \\
\frac{N_T}{N_H} &= \frac{1-\theta}{\theta} \\
\frac{N_T}{N_H} &= \frac{1}{\theta} - 1 \\
\frac{N_T + N_H}{N_H} &= \frac{1}{\theta} \\
\hat{\theta} &= \frac{N_H}{N_T + N_H}
\end{aligned}
$$

so this demonstrates that the maximum-likelihood estimate is quite similar to what we would expect, but there are issues.

Bayes and MAP

In the previous example, just flipping one head would give us a MLE of $\hat{\theta} = 1$ meaning we would predict there is zero possibility of the coin coming up tails. This may be reasonable for ten heads or even 30 heads, but not just one. This MLE also does not include any other information we have about coins (mainly by design). So, now we attack these problems using the Bayesian approach to this problem.

Let's assume that the parameter, θ, is a random variable. Then, we can apply Bayes theorem

$$P(\theta|d) = \frac{P(d|\theta)P(\theta)}{\int_0^1 P(d|\theta')P(\theta')d\theta'}$$

The subtle difference between this equation and earlier equations is that we just gained infinitely many different hypotheses in the form of θ. One consequence of this is the sum in the denominator has become an integral.

From this new Bayesian approach to the problem, we can take two approaches to choosing the best estimate for our parameter. The first is to choose the mode of the posterior. This is typically referred to as the *maximum a posterior* (MAP) estimate. The second way is to take the *posterior mean*, which is computed exactly the way an expectation is taken

$$\hat{\theta} = \int_0^1 \theta P(\theta|d) d\theta$$

This brings the natural question of what to make the prior. One initial thought would be to use the uniform distribution for $P(\theta)$ and in fact this was analyzed long ago.

Conjugate Priors

Obviously manipulating some of these distributions analytically will be difficult and performing certain integrations will be impossible. One of the ways to mitigate this effect is by way of conjugate priors. Depending on the likelihood distribution, for our example it was binomial, the prior can be choosen, so that posterior is of the same form as the prior. For example, the conjugate prior for the binomial distribution is the beta distribution. So, if your likelihood is of the form

$$P(d|\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

and your prior is of the form

$$P(\theta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

then your posterior will be of the form

$$P(\theta|d) = \frac{1}{B(\alpha+k,\beta+n-k)}\theta^{\alpha+k-1}(1-\theta)^{\beta+n-k-1}$$

Gaussian Process Regression [2]

We are going to look at one derivation of Gaussian process regression, the weight space view. There is another more popular and more powerful view called the function space view, which is slightly more conceptually different.

Lets assume that you start with the standard linear regression model with some Gaussian noise.

$$f(x) = x^T w$$

$$y = f(x) + \epsilon$$

where ϵ is the noise parameter that is Gaussian distributed

$$\epsilon \sim N(0, \sigma^2)$$

This gives us a likelihood distribution of the observations given the parameters, which we can assume each data point is independent, so we can write the likelihood as follows

$$P(y|X, w) = \Pi p(y_i|x_i, w)$$

[2] Material in this section was derived mainly from Gaussian Processes for Machine Learning by C. Rasmussen and C. Williams

where each individual distribution is Gaussian, which can be written as follows

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - x_i^T w)^2}{2\sigma^2}}$$

With a little algebra and Gaussian knowledge, we can rewrite the likelihood as

$$p(y|X, w) = N(X^T w, \sigma^2 I)$$

Now, we have a parameter,w, that we need to express a prior for. We happen to know that Gaussians are self-conjugate, so we decide to pick a Gaussian prior over our parameter,w

$$w \sim N(0, \Sigma_p)$$

We can now write the posterior probability as

$$p(w|X, y) = p(y|X, w)p(w)$$

we can condition on the data, X, everywhere, because the data is by definition independent of our prior. If you go through the algebric motions, we can find that the posterior is Gaussian (of course) of the form

$$p(w|X, y) \sim N(\frac{1}{\sigma^2} A^{-1} X y, A^{-1})$$

where $A = \sigma^{-2} X X^T + \Sigma_p^{-1}$

Another property of Gaussian distributions is that the mean is also the MAP estimate (or mode).