Notes on Bayesian Learning

Padhraic Smyth,
Department of Computer Science
University of California, Irvine

1 Introduction

In this set of notes we introduce a different approach to parameter estimation and learning: the Bayesian approach. We outline the concepts that form the basis for Bayesian thinking, discuss how these ideas can be applied to parameter estimation for various models, and conclude with a discussion of some of the broader aspects of Bayesian learning.

The key aspect of Bayesian estimation is that we treat unknown quantities, such as parameters θ as random variables that we can put distributions over, where these distributions reflect our uncertainty about their values. This is in contrast to the maximum likelihood approach for parameter estimation where an unknown parameter θ is treated as a unknown but fixed quantity: there is no notion in maximum likelihood of treating the parameter as a random variable.

As with maximum likelihood, in Bayesian estimation we begin by defining a likelihood function $p(D|\theta)$ for some observed data D conditioned on some unknown set of parameters θ. The likelihood function will be different for different problems depending on the nature of the data D (e.g., real-valued, discrete, multivariate, sequential, and so on) and depending on the nature of the modeling assumptions we make (e.g., Gaussian, Markov, and so on). The definition of the likelihood in Bayesian estimation is done in the same way as it is done for maximum likelihood estimation (i.e., by making some simplifying assumptions about a probabilistic "generative" model for how the observed data might have been generated).

Where Bayesian estimation differs is that it treats θ as a random variable. Specifically, we can apply Bayes rule to the problem of parameter estimation:

$$p(\theta|D) \; = \; \frac{p(D|\theta)p(\theta)}{p(D)} \tag{1}$$

This equation has introduced some new quantities we have not encountered before:

1. A **prior density** $p(\theta)$ on the unknown parameters θ. This is a density function that reflects our prior belief on the possible values of θ. For example, if we had prior knowledge for some particular problem that $0 \leq \theta \leq 10$, and that all values of θ in that range were equally likely, they we could defined $p(\theta)$ as the uniform density $p(\theta) = U[0, 10]$. The prior density $p(\theta)$ is (in principle at least) defined based only on our prior knowledge about the problem, before we see any data D.

2. A **posterior density** $p(\theta|D)$ on θ. Like the prior, this is another density function on θ, but now is a conditional density, conditioned on the data D. It reflects our posterior belief about θ, having seen the data D. For example, if our prior were $p(\theta) = U[0, 10]$, and θ represented the mean value for a Gaussian distribution, and we observed data with values $D = \{5.2, 4.8, 5.9, 5.1, 4.6\}$ we might expect our posterior density to more concentrated around the value 5 compared to the uniform prior from 0 to 10.

3. A **normalization term** $P(D)$: we can compute this by integrating over parameter space, i.e., $P(D) = \int p(\theta, D)d\theta = \int p(D|\theta)p(\theta)d\theta$. This normalization term $P(D)$ is sometimes referred to as the **marginal likelihood**:

We will not pay much attention to the normalization term $p(D)$ for now other than to note that the numerator term on the right $p(D|\theta)p(\theta)$ requires normalization in order to insure that $p(\theta|D)$. Ignoring the normalization term, we can write our earlier equation in a simple proportional form:

$$p(\theta|D) \propto p(D|\theta)\, p(\theta). \tag{2}$$

This is the classic definition of Bayesian learning. The posterior $p(\theta|D)$ is proportional to the likelihood $p(D|\theta)$ times the prior $p(\theta)$.

Bayesian estimation can be viewed as an information processing procedure: we begin with the prior, we update our prior information with observed data D, we combine the prior with a likelihood for the data, and we arrive at a posterior on θ given the data and the prior. This updating procedure corresponds to a very natural sequential notion of incorporating new information (data) by updating our prior beliefs (the prior) to arrive at a new posterior set of beliefs about the world (as represented by θ). Given this view it is not surprising that Bayesian estimation has found wide application in artificial intelligence and machine learning, particularly for problems such as computer vision, robotics, online learning, and more, where an agent is continually encountering new data D and needs to update its internal model of the world around it.

2 Example: Bayesian Estimation of a Bernoulli Parameter

Consider tossing a binary-valued (Bernoulli) random variable X that takes value 1 with probability θ and 0 with probability $1-\theta$, and where the value of the parameter θ is unknown and we wish to estimate it from data. This is the simple **Bernoulli model** we discussed in earlier notes. Now say we observe a data set D of outcomes where $D = \{x_1, \ldots, x_N\}$ and $x_i \in \{0, 1\}$ represents the outcome of the ith observation of X, and where we will assume that the observations are conditionally independent given θ.

2.1 Binomial Likelihood

As discussed earlier the binomial likelihood for this problem can be defined as

$$\begin{aligned} p(D|\theta) \;=\; L(\theta) \quad &= \quad \prod_{i=1}^{N} P(x_i|\theta) \\ &= \quad \theta^r (1-\theta)^{N-r} \end{aligned}$$

where r is the number of 1's observed in the data and $N - r$ is the number of 0's.

2.2 Beta Prior

Our parameter θ is bounded between 0 and 1. To define a prior on θ we need a density function that is (a) restricted to the interval $[0, 1]$ and (b) has some flexibility in its shape (so that we can specify different priors on θ by varying the parameters of the prior distribution).

A useful prior in this context is the Beta density, defined as

$$P(\theta) \;=\; Be(\theta; \alpha, \beta) \;=\; \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \tag{3}$$

where $0 < \theta < 1$. The two parameters of this density function are $\alpha > 0$ and $\beta > 0$ and $B(\alpha, \beta)$ is a normalization term to ensure that the density integrates to 1[1].

The parameters of the prior (α and β here) are referred to as **hyperparameters**. In our discussion below we will assume that these are fixed before we do our Bayesian analysis: in this classical Bayesian setup the data analyst fixed the values of the hyperparameters a priori, based on their prior knowledge, before looking at the data.

[1] $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$, $z > 0$, is the gamma function.

The mean of the Beta prior is

$$E[\theta] = \frac{\alpha}{\alpha + \beta}.$$

Since α and β are positive, the mean of the Beta prior is the relative size of α to the sum $\alpha + \beta$. And it turns out that the strength of our prior (how narrow it is around the mean) is proportional to $\alpha + \beta$. So, $Be(20, 20)$ is a much stronger (narrower) prior than $Be(1, 1)$, but they both have an expected value of 0.5 for our unknown parameter θ. A prior such as $Be(9, 1)$ has an expected value of 0.9, and a "strength" of 10, and so on. We can use α and β to express different prior beliefs about the value of θ.

2.3 Posterior Density for θ

Now that we have a prior (a Beta density) and a likelihood (binomial), we can use Bayes rule to derive an expression for the posterior density $p(\theta|D)$ which is what we are primarily interested in for a Bayesian estimation problem, i.e.,

$$\begin{aligned} p(\theta|D) &\propto p(D|\theta)p(\theta) \\ &\propto \theta^r(1-\theta)^{N-r}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{r+\alpha-1}(1-\theta)^{N-r+\beta-1} \\ &= \theta^{\alpha'-1}(1-\theta)^{\beta'-1} \end{aligned}$$

where $\alpha' = \alpha + r$ and $\beta' = \beta + N - r$. In our derivation above we dropped all the terms (from the likelihood and the prior) that don't involve θ.

Note that the final expression for $p(\theta|D)$ is itself in the form of a Beta density $Be(\alpha', \beta')$. Thus, we have the nice result that the Beta prior has been updated to a Beta posterior[2].

The posterior mean is defined as

$$E_{p(\theta|D)}[\theta] \;=\; \frac{\alpha'}{\alpha' + \beta'} \;=\; \frac{r + \alpha}{N + \alpha + \beta} \tag{4}$$

It is informative to compare this expression to the mean of the prior, $E[\theta] = \frac{\alpha}{\alpha+\beta}$. We see that the posterior mean can be seen as an updated version of the prior where we have (in effect) added r counts to α and $N - r$ counts to β.

In this sense, we can think of α and β in the prior as "pseudocounts": for example if $\alpha = 2$ and $\beta = 8$ we can think of this prior as representing 10 pseudocounts in total, 2 "successes" and 8 "failures." Looking

[2]Whenever a prior and posterior density are the same functional form, we say that the prior is **conjugate** to the likelihood (i.e., in this case the Beta prior is conjugate to the Bernoulli likelihood).

at $E_{p(\theta|D)}[\theta]$ above we can see that when $N < 10$ (in this example) that the information from the prior will tend to dominate the information from the data (in terms of how the posterior mean is defined), and as N gets bigger than 10, the prior has less influence on the posterior density for θ.

In particular, recall that $\hat{\theta}^{ML} = \frac{r}{N}$ is the maximum likelihood estimate of θ for a binomial likelihood. With a little algebra, we can rewrite the posterior mean as

$$\begin{aligned} E_{p(\theta|D)}[\theta] &= \gamma_N \frac{\alpha}{\alpha+\beta} + (1-\gamma_N)\frac{r}{N} \\ &= \gamma_N \text{ (prior mean) } + (1-\gamma_N) \text{ (ML estimate)} \end{aligned}$$

where $\gamma_N = \frac{\alpha+\beta}{\alpha+\beta+N}$. We see that the posterior mean of θ is a convex combination of the prior mean and the maximum likelihood estimate: if the prior mean is strong relative to the data ($\alpha + \beta >> N$ and $\gamma_N \to 1$) the prior will dominate the data (the likelihood). And for any fixed prior values, as $N \to \infty$, we have that $\gamma_N \to 0$, the posterior mean will converge to the maximum likelihood estimate $\frac{r}{N}$ and the prior is ignored. This is generally true in Bayesian estimation: as N gets large, the prior plays less of a role: and conversely, with small amounts of data the prior can play a significant role.

We have focused above on the posterior mean $E_{p(\theta|D)}[\theta]$, but it is also true that the variance ("width") of the posterior will tend to narrow as N increases. The variance of the posterior $p(\theta|D)$ reflects our uncertainty about θ as we see more data and in general this variance will get narrower and narrower as we get more data, eventually converging to a delta function centered at the maximum likelihood estimate.

We can also interpret the prior as providing a form of smoothing, i.e., denoting $\hat{\theta}^{MPE}$ be the **mean posterior estimate** of θ.

$$\hat{\theta}^{MPE} = E_{p(\theta|D)}[\theta] = \frac{r+\alpha}{N+\alpha+\beta} \tag{5}$$

The "raw" maximum likelihood estimate corresponds to $\frac{r}{N}$. Having positive α and β in the equation is a form of smoothing: it particular we can interpret $\hat{\theta}^{MPE}$ as a smoothed estimate of $\hat{\theta}^{ML}$, where the extreme cases of estimates of θ 0 or 1 are avoided in situations where $r = 0$ or $r = N$.

3 Different Types of Parameter Estimates

Given what we know so far about maximum likelihood and Bayesian estimation we can define a few different types of parameter estimates:

$$\begin{aligned} \text{Maximum Likelihood: } \hat{\theta}^{ML} &= \arg\max_\theta p(D|\theta) \\ \text{Maximum A Posteriori (MAP): } \hat{\theta}^{MAP} &= \arg\max_\theta p(\theta|D) = \arg\max_\theta p(D|\theta)p(\theta) \\ \text{Mean Posterior Estimate (MPE): } \hat{\theta}^{MPE} &= E_{p(\theta|D)}[\theta] \end{aligned}$$

The maximum a posteriori (MAP) and mean posterior estimate (MPE) differ only in the sense that one selects the mode of the posterior density (MAP) and the other selects the mean of the posterior (MPE). If the posterior density is symmetric, they will be the same, otherwise they will differ. Which one to use in practice may come down to how the estimates will be used later on, i.e., what the purpose is of computing the estimate.

The MAP estimate has the property that if the prior is flat, i.e., $p(\theta) \propto c$ where c is some constant, then the MLE and MAP estimates are the same, i.e., they both maximize the likelihood.

Note that all three of the parameter estimates defined above are **point estimates**, i.e., they only provide a single number to summarize what we have learned about θ given the data (and given the prior if we are being Bayesian). In contrast, the "fully Bayesian" approach is to use (or report) the full posterior $p(\theta|D)$. This makes sense in contexts such as public policy or medicine where a human will look at $p(\theta|D)$ and then make a decision—and if $p(\theta|D)$ is very wide (i.e., a lot of uncertainty) then the decision might be to go back and get more data before making a final decision. In contrast, in machine learning, θ is not just a single parameter but could consist of a set of thousands or even millions of parameters (e.g., all the weights in a deep neural network). In this situation it is often not practical computationally to work with the full posterior $p(\theta|D)$ and point estimates $\hat{\theta}$ (such as MAP or MPE estimates) are widely used.

4 Generalization to K Outcomes (Multinomial Model)

We can generalize from binary outcomes to K-ary outcomes. Say X is a discrete-variable taking values from 1 to K and we have IID data $D = \{x_1, \ldots, x_N\},\ x_i \in \{1, \ldots, K\}$, with the usual IID assumption. As an example x_i might be a randomly-selected word from a corpus of documents (e.g., all Web pages in Wikipedia) and K could be a large number (the vocabulary of words in our model), e.g., $K = 100,000$.

Let $\theta_k = p(x_i = k)$ be the probability of the kth outcome, with $\sum_{k=1}^{N} \theta_k = 1$ and let $\underline{\theta} = (\theta_1, \ldots, \theta_K)$ be the set of all of the θ's. The multinomial likelihood is defined as

$$L(\underline{\theta}) \ = \ p(\underline{\theta}|D) \ = \ \prod_{i=1}^{N} p(x_i|\underline{\theta}) \ = \ \prod_{k=1}^{K} \theta_k^{r_k} \tag{6}$$

where r_k is the number of times that $x_i = k$ in the data D, $x_i \in \{0, 1, 2, \ldots\}$, and $\sum_{k=1}^{K} r_k = N$. We recall from earlier that the maximum likelihood estimate of θ_k is $\hat{\theta}_k^{ML} = \frac{r_k}{N}$.

In order to pursue a Bayesian approach for estimation we need to define a prior for $\underline{\theta}$. For the multinomial likelihood the conjugate prior for $\underline{\theta}$ is the **Dirichlet** density:

$$p(\underline{\theta}) \quad = \quad \text{Dirichlet}(\underline{\alpha}), \quad \underline{\alpha} = (\alpha_1, \ldots, \alpha_K)$$

$$\propto \quad \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

(Note that a special case of the Dirichlet density is the Beta density with $K = 2, \alpha \rightarrow \alpha_1, \beta \rightarrow \alpha_2$.)

The Dirichlet density $p(\underline{\theta})$ is defined over the K-dimensional simplex, defined by $\theta_k > 0$, $\sum_{k=1}^{K} \theta_k = 1$. The mean of a Dirichlet is

$$E_{p(\underline{\theta})}[\theta_k] \; = \; \frac{\alpha_k}{\sum \alpha_k} \tag{7}$$

As with the Beta density, the relative size of each α_k to the sum of the α's reflects the location of the mean for θ_k, and the sum of the α's reflects how narrow the density function is.

A non-symmetric Dirichlet prior is one where the α_k's can be different to each other. For text, the α_k's could be proportional to frequency of words in English text (e.g., as measured from Web pages or a corpus of newspapers): we might then be interested in estimating the relative frequencies of words, θ_k in some new corpus where words might have different probabilities of occurring (e.g., in the biomedical research literature, or in a set of blogs).

Multiplying the Dirichlet prior and the multinomial likelihood it is straighforward to see that the posterior density for $\underline{\theta}$ given the data D is another Dirichlet:

$$p(\underline{\theta}|D) \; = \; \text{Dirichlet}(\alpha'_1, \ldots, \alpha'_K) \tag{8}$$

where $\alpha'_k = \alpha_k + r_k$ (again this is a generalization of the solution for the $K = 2$ case).

The posterior mean is

$$E_{p(\underline{\theta}|D)}[\theta_k] \; = \; \frac{r_k + \alpha_k}{N + \sum \alpha_k}. \tag{9}$$

This is directly analogous to what we found for the Beta binomial case, i.e.,

- Each α_k can be thought of as a "prior pseudocount" for outcome k,
- The sum $\sum \alpha_k$ represents the strength of the prior
- The posterior mean approaches $\hat{\theta}_k^{ML} = \frac{r_k}{N}$ as $N \rightarrow \infty$.

Additional notes on estimating parameters of Gaussians, predictive densities, and model selection, to follow soon...