MATH3871/5960: Bayesian Inference and Computation

Gerald Huang

Contents

Chapter 1

Introduction to Bayesian Statistics

1.1 The Bayesian paradigm

In statistical analysis and inference, we are primarily concerned with using data to determine an underlying distribution of a probability. Thus, our goal is to be able to estimate an unknown parameter θ using *observations* $x_1, x_2, \dots, x_n$ from our realisations $X_1, X_2, \dots X_n$. In classical (and frequentist) statistics, we usually treat θ as an unknown but fixed constant. In Bayesian inference however, we treat θ not as a constant, but rather as a *random variable.* This allows us to assign a probability distribution to θ. This is called the *prior* distribution, which models initial uncertainties about the parameter(s). We then use Bayes' theorem to describe the updated state of our parameter(s). This is called the *posterior* distribution. We now begin to develop these formulas.

1.2 Recap: Bayes' theorem

Let X and Y be random variables. Then the probability distribution of X conditioned on Y is given by

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)} = \frac{P(X \cap Y)}{P(Y)}.$$

Thomas Bayes proved a continuous version of Bayes' theorem.

Theorem 1.2.1. *Given two continuous random variables X, Y with conditional probability function $f(x \mid Y = y)$ and marginal distribution $g(y)$, the conditional distribution of Y given $X = x$ is given by*

$$g(y \mid X = x) = \frac{f(x \mid Y = y)g(y)}{\int_y f(x \mid Y = y)g(y)\, dy}.$$

We note that the denominator is just the marginal distribution of x since

$$\int_y f(x \mid Y = y)g(y)\, dy = \int_y f_{X,Y}(x, y)\, dy = f(x).$$

Since we consider θ as a random variable from the Bayesian point of view, we can directly apply Bayes' theorem to obtain

$$\pi(\theta \mid \mathbf{x}) = \frac{L(\theta; \mathbf{x})\pi(\theta)}{m(\mathbf{x})}.$$

Here, $\pi(\theta)$ is the *prior* distribution. After looking at a vector of observations $\mathbf{x}$, we obtain a *posterior* distribution $\pi(\theta \mid \mathbf{x})$. For completeness, we'll describe the remaining elements. The function $m(\mathbf{x})$ is the *marginal distribution* of $\mathbf{x}$ completely analogous to the continuous Bayes' theorem formula and $L(\theta; \mathbf{x})$ is the *likelihood function* which is given by

$$L(\theta; \mathbf{x}) = f(\mathbf{x} \mid \theta).$$

Chapter 2

Prior and posterior distributions

In the previous chapter, we introduced the notion of a *prior*. We will start to flesh out those ideas and look at ways of analysing the prior distribution. Recall that a prior distribution describes our knowledge about model parameters *before* we have seen the data. We then update that using the *posterior* which reflects the changes in information that we know about the model parameters. We now look at some types of priors that are used often in Bayesian inference.

2.1 Conjugate prior

To motivate the concept of a *conjugate prior*, we'll begin with an example. Suppose we have $X \sim \text{Bin}(n, \theta)$ and we want to make an inference on the unknown parameter $\theta \in [0, 1]$. We observe that the likelihood function is given by

$$\mathcal{L}(\theta; \mathbf{x}) = f(\mathbf{x} \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}.$$

Now, remember that we want to be able to make an inference on θ, which is the unknown parameter in our situation. So the binomial coefficient at the front won't affect our decision. Hence, it is enough to just look at the remaining parts that depend on θ. This is called the "core" of the distribution. The idea is that we want to find a suitable distribution for θ that allows us to say a lot about the posterior distribution of θ. By Bayes' theorem, recall that

$$\pi(\theta \mid \mathbf{x}) = \frac{\mathcal{L}(\theta; \mathbf{x})\pi(\theta)}{\pi(\mathbf{x})}.$$

Since $\pi(\mathbf{x})$ is not dependent on θ, we can treat it as a constant in our analysis. Hence, we say that

$$\pi(\theta \mid \mathbf{x}) \propto \mathcal{L}(\theta; \mathbf{x})\pi(\theta).$$

To make computation simple, we want to find a prior distribution for θ so that $\mathcal{L}(\theta; \mathbf{x})\pi(\theta)$ is easy to compute. Hence, we look for a distribution with a core (the part dependent on the unknown parameter) proportional to $\theta^x(1-\theta)^{n-x}$. One such example would be the Beta distribution; that is, let $\theta \sim \text{Be}(a, b)$. Then, by definition,

$$\pi(\theta) = \frac{\Gamma(a+b)}{\Gamma(a) \cdot \Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}.$$

Then the posterior distribution becomes

$$\begin{aligned} \pi(\theta \mid \mathbf{x}) &= \frac{\mathcal{L}(\theta; \mathbf{x})\pi(\theta)}{\pi(\mathbf{x})} \\ &= \left(\frac{\Gamma(a+b)}{\Gamma(a) \cdot \Gamma(b) \cdot \pi(\mathbf{x})} \right) \left(\theta^x (1-\theta)^{n-x} \right) \left(\theta^{a-1}(1-\theta)^{b-1} \right) \\ &\propto \theta^{x+a-1}(1-\theta)^{n-x+b-1}, \end{aligned}$$

which has the same core as the Beta distribution we started out with. So we see that $\pi(\theta \mid \mathbf{x})$ is proportional to a Beta distribution; that is,

$$\pi(\theta \mid \mathbf{x}) \propto \mathrm{Be}(x+a, n-x+b).$$

In other words,

$$\theta \mid \mathbf{x} \sim \mathrm{Be}(x+a, n-x+b).$$

This motivates the concept of a conjugate prior.

Definition 2.1.1 (Conjugate prior)**.** If $\mathcal{F}$ is a class of sampling distributions (likelihoods) $\mathcal{L}(\theta; \mathbf{x})$ and $\mathcal{P}$ is a class of prior distributions $\pi(\theta)$ for θ, then the class $\mathcal{P}$ is **conjugate** for $\mathcal{F}$ if $\pi(\theta \mid \mathbf{x}) \in \mathcal{P}$ for all $\mathcal{L}(\theta; \cdot) \in \mathcal{F}$ and $\pi(\cdot) \in \mathcal{P}$.

In other words, the posterior distribution stays in the same class of functions as the prior distribution. In our example, we saw that the posterior distribution was still Beta distributed. So, the Beta distribution is **conjugate** for the binomial model.

Example 2.1.2. Let $X \sim \text{Poisson}(\theta)$. Show that the Gamma distribution is a conjugate prior of X.

Proof. Let $X \sim \text{Poisson}(\theta)$. Then,

$$\begin{aligned}\mathcal{L}(\theta; \mathbf{x}) = f(\mathbf{x} \mid \theta) &= \frac{\theta^x e^{-\theta}}{x!} \\ &\propto \theta^x e^{-\theta}.\end{aligned}$$

Let $\theta \sim \text{Gamma}(a, b)$. Then,

$$\pi(\theta) = \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta}.$$

Then the posterior distribution is

$$\begin{aligned}\pi(\theta \mid \mathbf{x}) &\propto \mathcal{L}(\theta; \mathbf{x}) \pi(\theta) \\ &\propto \left(\theta^x e^{-\theta}\right) \left(\theta^{a-1} e^{-b\theta}\right) \\ &= \theta^{x+a-1} e^{-\theta(b+1)},\end{aligned}$$

which is the core of the Gamma distribution,

$$\pi(\theta \mid \mathbf{x}) \propto \text{Gamma}(x+a, b+1).$$

Hence, $\theta \mid \mathbf{x} \sim \text{Gamma}(x+a, b+1)$ and thus, the Gamma distribution is a suitable conjugate prior of the Poisson model. □