STA 360/602L: Module 2.3

Marginal likelihood and posterior prediction

Dr. Olanrewaju Michael Akande

Marginal likelihood

- Recall that the marginal likelihood is

$$L(y) = f(y_1, \ldots, y_n) = \int_\Theta f(y_1, \ldots, y_n|\theta)\pi(\theta)\mathrm{d}\theta = \int_\Theta L(\theta|y)\pi(\theta)\mathrm{d}\theta.$$

- For clarity, when dealing with a single y instead of $y_1, \ldots, y_n$, we can write

$$L(y) = f(y) = \int_\Theta f(y|\theta)\pi(\theta)\mathrm{d}\theta = \int_\Theta L(\theta|y)\pi(\theta)\mathrm{d}\theta.$$

- When this is the case, for example in the case of the binomial distribution, I will often write
  - the marginal likelihood as $L(y)$ or $f(y)$, and
  - the sampling (conditional) likelihood as $f(y|\theta)$ or $L(\theta|y)$.

Marginal likelihood

- What is the marginal likelihood for the beta-binomial?
- We have

$$\begin{aligned} L(y) &= \int_{\Theta} p(y|\theta)\pi(\theta)\mathrm{d}\theta \\ &= \int_0^1 \binom{n}{y} \theta^y (1-\theta)^{n-y} \frac{1}{B(a,b)} \theta^{a-1}(1-\theta)^{b-1} d\theta \\ &= \binom{n}{y} \frac{B(a+y,\, b+n-y)}{B(a,b)}, \end{aligned}$$

by the integral definition of the Beta function.

- Marginal likelihood for the beta-Bernoulli follows directly.
- Deriving the marginal likelihood for conjugate distributions is often relatively straightforward.

Prior predictive distribution

- We may care about making predictions before we even see any data.
- This is often useful as a way to see if the sampling distribution we have chosen is appropriate, after integrating out all unknown parameters.
- The prior predictive distribution is

$$
\begin{aligned}
p(y) &= \int_{\Theta} p(y, \theta)\, d\theta \\
&= \int_{\Theta} p(y|\theta) \cdot \pi(\theta)\, d\theta.
\end{aligned}
$$

- In some sense, the prior predictive distribution marginalizes the sampling distribution (for a single y) over the prior.
- When dealing with a single y instead of $y_1, \ldots, y_n$, this is just the marginal likelihood of the data.

Posterior predictive distribution

- We often care about making predictions for new data points, given the current pbserved data.
- For example, suppose $y_1, \ldots, y_n \stackrel{iid}{\sim} \text{Bernoulli}(\theta)$.
- We may wish to predict a new data point y_{n+1}.
- We can do so using the posterior predictive distribution $p(y_{n+1}|y_{1:n})$.
- Why are we not including the parameter in the posterior predictive distribution?

Posterior predictive distribution

- Recall that we have conditional independence of the y's given θ.
- So,

$$
\begin{aligned}
p(y_{n+1}|y_{1:n}) &= \int_\Theta p(y_{n+1}, \theta|y_{1:n})\, d\theta \\
&= \int_\Theta p(y_{n+1}|\theta, y_{1:n}) \cdot \pi(\theta|y_{1:n})\, d\theta \\
&= \int_\Theta p(y_{n+1}|\theta) \cdot \pi(\theta|y_{1:n})\, d\theta.
\end{aligned}
$$

- So, in some sense, the posterior predictive distribution marginalizes the sampling distribution over the posterior.

Posterior predictive distribution

- When we talk about the posterior predictive distribution for Bernoulli data, we are really referring to $p(y_{n+1} = 1 | y_{1:n})$ and $p(y_{n+1} = 0 | y_{1:n})$.

- Then,

$$\begin{aligned} p(y_{n+1} = 1 | y_{1:n}) &= \int_{\Theta} p(y_{n+1} = 1 | \theta) \cdot \pi(\theta | y_{1:n}) \, d\theta \\ &= \ldots \\ &= \ldots \end{aligned}$$

which simplifies to what? To be done on the board!

- What then is $p(y_{n+1} = 0 | y_{1:n})$?

- Posterior predictive pmf therefore takes the form

$$p(y_{n+1} | y_{1:n}) = \frac{a_n^{y_{n+1}} b_n^{1-y_{n+1}}}{a_n + b_n}; \quad y_{n+1} = 0, 1.$$

- What are a_n and b_n?

Going forward...

- From here on, we will often deal with multiple data points $y_1, \ldots, y_n$ frequently.

- To make that obvious, we will write the Bayes rule as one of the following

$$\pi(\theta|y_1, \ldots, y_n) = \frac{\pi(\theta) \cdot p(y_1, \ldots, y_n|\theta)}{\int_\Theta \pi(\tilde{\theta}) \cdot p(y_1, \ldots, y_n|\tilde{\theta}) \mathrm{d}\tilde{\theta}}$$

$$\pi(\theta|y_1, \ldots, y_n) = \frac{\pi(\theta) \cdot p(y_1, \ldots, y_n|\theta)}{p(y_1, \ldots, y_n)}$$

$$\pi(\theta|y) = \frac{\pi(\theta) \cdot L(\theta|y)}{L(y)},$$

where $y = (y_1, \ldots, y_n)$.

What's next?

Move on to the readings for the next module!