8 Fundementals of Bayesian Inference

Classical approaches studies thus far mostly rely on distributions of estimators and test statistics over hypothetical repeated samples. There is no conditioning on the observed data.
In contrast, Bayesian inference is based on the posterior distribution of the parameter of interest, given the observed data. These distributions are exact in finite samples, distributions are derived conditional on the observed data.
Classical hypothesis testing measures support for the data, $Pr(D|H_0)$, while Bayesian measures the support for the hypothesis in the data $Pr(H_0|D)$.

8.1 Bayes Rule

Definition 8.1.1: Bayes Rule

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

where $P(A|B)$ is the likelihood, $P(B)$ is the prior, and $P(A)$ is the marginal likelihood. $P(B|A)$ is the posterior.

A prior probability is an initial value of the probability of an event, we update this with data to get the posterior probability.
We can write this for parameters of a model, let $p(\theta)$ be the prior density on some unknown parameter θ. Let $p(\theta|y)$ be the posterior density of θ. The probability of observing y conditional on θ is given by the likelihood:

Definition 8.1.2: Likelihood Function

$$f(y|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$$

It is the conditional probability of observing the data given the parameter.

Bayes rule gives us

$$p(\theta|y) = \frac{f(y|\theta)p(\theta)}{f(y)}$$

However we note that the denominator is not a function of θ, so we can write

Definition 8.1.3: Posterior Distribution

$$\begin{aligned} p(\theta|y) &= \frac{f(y|\theta)p(\theta)}{f(y)} \\ &\propto f(y|\theta)p(\theta) \end{aligned}$$

8.2 Conjugate Priors

A class of priors is conjugate for a family of a likelihoods if both prior and posterior are in the same class for all data y.

Definition 8.2.1

If τ is a class of sampling distributions $p(y|\theta)$, and ω is a class of prior distributions for θ, then ω is conjugate for τ if:

$$p(\theta|y) \in \omega \quad \forall p(y|\theta) \in \tau, p(\theta) \in \omega$$

Intuitively what this means is that when we get some data, updating only involves updating the parameters of the distribution, not changing the distribution itself.

Example (Bernoulli distribution and Beta priors). Suppose we have a Bernoulli pdf we can write the likelihood in terms of the mean parameter as:

$$p(y|\theta) = \theta^y(1-\theta)^{1-y}$$

This suggests that a conjugate prior might be given by

$$p(\theta|\tau) \propto \theta^{\tau_1}(1-\theta)^{\tau_2}$$

This expression can be normalised if both τ's are greater than -1, and we get the beta distribution:

$$p(\theta|\alpha,\beta) = K(\alpha,\beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

where the normalisation constant can be found by solving:

$$\begin{aligned} 1 &= \int K(\alpha,\beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta \\ \Rightarrow K(\alpha,\beta) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \end{aligned}$$

If we multiply the beta density by the Bernoulli likelihood we obtain a beta density. Consider N *i.i.d.* Bernoulli trials, then the likelihood is:

$$\begin{aligned} p(\theta|y,\alpha,\beta) &\propto \left(\prod_{i=1}^{N}\theta^{y_i}(1-\theta)^{1-y_i}\right)\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{\sum_{i=1}^N y_i+\alpha-1}(1-\theta)^{N-\sum_{i=1}^N y_i+\beta-1} \\ &\sim \text{Beta}(\sum_{i=1}^{N} y_i+\alpha, N-\sum_{i=1}^{N} y_i+\beta) \end{aligned}$$

8.2.1 Beta distribution

Definition 8.2.2: Beta distribution

$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$

where Γ is the gamma function.

This is a super useful distribution, and it nests many other common distributions. For example, if $\alpha=\beta=1$ we get the uniform distribution. Indeed any Beta with $\alpha=\beta$ is symmetric.
When we have $\alpha>\beta$ ($\alpha<\beta$) the distribution is skewed to the left (right).

8.2.2 Examples

Definition 8.2.3: Diriclet distribution

$$\mathrm{f(x;}\ \alpha) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\prod_{i=1}^{k}x_i^{\alpha_i-1}$$

Question 1

Show that the Diriclet distribution is the conjugate prior for the multinomial distribution.

Solution:-

The likelihood for the multinomial distribution is given by:

$$p(y|\theta) = \theta_1^{\sum_{i=1}^{N}\mathbf{1}(y_i=1)}\theta_2^{\sum_{i=1}^{N}\mathbf{1}(y_i=2)}\dots\theta_k^{\sum_{i=1}^{N}\mathbf{1}(y_i=k)}$$

We write the Diriclet prior as:

$$p(\theta|\alpha) = K(\alpha)\prod_{i=1}^{k}\theta_i^{\alpha_i-1}$$

with

$$K(\alpha) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}$$

As before we can derive:

$$p(\theta|y,\alpha) \propto \left(\prod_{i=1}^{k}\theta_i^{\sum_{i=1}^{N}\mathbf{1}(y_i=i)}\right)\left(\prod_{i=1}^{k}\theta_i^{\alpha_i-1}\right) = \prod_{i=1}^{k}\theta_i^{\sum_{i=1}^{N}\mathbf{1}(y_i=i)+\alpha_i-1}$$

Question 2

Show that the gamma distribution is the conjugate prior for the poisson distribution.

Solution:-

The likelihood for the poisson distribution is given by:

$$p(y|\theta) = \frac{e^{-\theta}\theta^{y}}{y!}$$

We write the gamma prior as:

$$p(\theta|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}$$

As before we can derive:

$$\begin{aligned} p(\theta|y,\alpha,\beta) &= \left(\frac{e^{-\theta}\theta^{y}}{y!}\right)\left(\frac{\beta^{\alpha}}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}\right) \\ &\propto \theta^{\sum_{i=1}^{N}y_i+\alpha-1}e^{-\theta(N+\beta)} \\ &\sim \mathrm{Gamma}(\sum_{i=1}^{N}y_i+\alpha, N+\beta) \end{aligned}$$

8.2.3 Mean and variance of the posterior

Consider the bernoulli and beta example. The mean of a beta distribution is given by:

$$\mathbb{E}[\text{Beta}(\alpha, \beta)] = \frac{\alpha}{\alpha + \beta}$$

Thus the mean of the posterior is given by:

$$\begin{aligned}
\mathbb{E}[\theta|y] &= \frac{\sum_{i=1}^{N} y_i + \alpha}{N + \alpha + \beta} \\
&= \frac{N}{N + \alpha + \beta}\bar{y} + \frac{\alpha + \beta}{N + \alpha + \beta}\frac{\alpha}{\alpha + \beta} \\
&:= w\frac{\alpha}{\alpha + \beta} + (1 - w)\bar{y}
\end{aligned}$$

Clearly this is a weighted average of the prior mean and the sample mean. As $N \to \infty$ the weight on the prior mean goes to zero. That is, as $N \to \infty$ the mean of the posterior approaches the MLE estimate of θ.
We can also examine the variance, note that the variance of a beta distribution is

$$Var[\text{Beta}(\alpha, \beta)] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

Thus the posterior variance is:

$$Var[\theta|y] = \frac{(\sum_{i=1}^{N} y_i + \alpha)(N - \sum_{i=1}^{N} y_i + \beta)}{(N + \alpha + \beta)^2(N + \alpha + \beta + 1)}$$

We can see that $Var[\theta|y, \alpha, \beta] \to 0$, showing that the posterior distribution concentrates around the MLE as $N \to \infty$.

Claim 8.2.1. The posterior mean is a weighted average for both the gamma prior with a poisson likelihood and the Diriclet prior with a multinomial likelihood.

Proof. **Gamma prior with poisson likelihood:** Note that the mean of the gamma distribution is:

$$\mathbb{E}[\text{Gamma}(\alpha, \beta)] = \frac{\alpha}{\beta}$$

Thus the mean of the posterior is given by:

$$\begin{aligned}
\mathbb{E}[\theta|y] &= \frac{\sum_{i=1}^{N} y_i + \alpha}{N + \beta} \\
&= \frac{N}{N + \beta}\bar{y} + \frac{\beta}{N + \beta}\frac{\alpha}{\beta} \\
&:= w\frac{\alpha}{\beta} + (1 - w)\bar{y} \quad \text{where } w = \frac{\beta}{N + \beta}
\end{aligned}$$

Diriclet prior with multinomial likelihood: Note that the mean of the Diriclet distribution is:

$$\mathbb{E}[\text{Diriclet}(\alpha)] = \frac{\alpha_i}{\sum_{i=1}^{k} \alpha_i}$$

Thus the mean of the posterior is given by:

$$\begin{aligned}\mathbb{E}[\theta|y] &= \frac{\sum_{i=1}^{N} \mathbf{1}(y_i = i) + \alpha_i}{N + \sum_{i=1}^{k} \alpha_i} \\ &= \frac{N}{N + \sum_{i=1}^{k} \alpha_i}\bar{y} + \frac{\sum_{i=1}^{k} \alpha_i}{N + \sum_{i=1}^{k} \alpha_i}\frac{\alpha_i}{\sum_{i=1}^{k} \alpha_i} \\ &:= w\frac{\alpha_i}{\sum_{i=1}^{k} \alpha_i} + (1-w)\bar{y} \quad \text{where } w = \frac{\sum \alpha_i}{N + \sum \alpha_i}\end{aligned}$$

□

8.3 Exchangability

Skipped in 2024

8.4 Parameter Uncertainty

Suppose we want to make inference on a population parameter μ, from a classical perspective μ is fixed, and we construct a confidence interval through the sampling distribution of the random variable, say $\bar{y}$. We can make a statement about $\bar{y}$:

$$Pr(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \bar{y} \leq \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

the pivot this around $\bar{y}$ (inverting the test) to get a confidence interval for μ:

$$Pr(\bar{y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

Given that μ is fixed, the randomness comes from $\bar{y}$. The interpretation of this is that (1- α) of the time, the interval will contain μ.
Our confidence interval is thus

$$\bar{y} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

By contrast a Bayesian makes a probability statement based on the posterior of the parameter μ give the actual data: $Pr(\mu|y)$.

Definition 8.4.1: Posterior Odds

The posterior odds of one model vs another is given by:

$$\underbrace{\frac{Pr(M_1|y)}{Pr(M_2|y)}}_{\text{Posterior Odds}} = \underbrace{\frac{Pr(y|M_1)}{Pr(y|M_2)}}_{\text{Bayes Factor}} \times \underbrace{\frac{Pr(M_1)}{Pr(M_2)}}_{\text{Prior Odds}}$$

We use the Bayes factor (the relative likelihood) to update the prior odds to get the posterior odds.