Stat260: Bayesian Modeling and Inference **Lecture Date: February 1, 2010**

Lecture 3

Lecturer: Michael I. Jordan *Scribe: Joshua G. Schraiber*

1 Decision theory

Recall that decision theory provides a quantification of what it means for a procedure to be 'good'. This quantification comes from the **loss function**, $l(\theta, \delta(X))$. Frequentists and Bayesians use the loss function differently.

1.1 Frequentist interpretation, the risk function

In frequentist usage, the parameter θ is fixed and thus the data are averaged over. Letting $R(\theta, \delta)$ denote the **frequentist risk**, we have

$$R(\theta, \delta) = E_\theta l(\theta, \delta(X)). \tag{1}$$

This expectation is taken over the data X, with the parameter θ held fixed. Note that the data, X, is capitalized, emphasizing that it is a random variable.

Example 1 (Squared-error loss)**.** A commonly chosen loss function for parameter estimation is the squared error loss, defined by $l(\theta, \delta(X)) = (\theta - \delta(X))^2$. In this case, we have

$$\begin{aligned} R(\theta, \delta) &= E_\theta l(\theta, \delta(X)) \\ &= E_\theta(\theta - \delta(X))^2 \\ &= E_\theta(\theta - E_\theta\delta(X) + E_\theta\delta(X) - \delta(X))^2 \\ &= \underbrace{(\theta - E_\theta\delta(X))^2}_{\text{Bias}^2} + \underbrace{E_\theta(\delta(X) - E_\theta\delta(X))^2}_{\text{Variance}}. \end{aligned} \tag{2}$$

This result allows a frequentist to analyze the variance and bias of an estimator separately, and can be used to motivate frequentist ideas, e.g. minimum variance unbiased estimators.

1.2 Bayesian interpretation and posterior risk

Bayesian do not find the previous idea compelling, because it doesn't adhere to the conditionality principle by averaging over all possible data sets. Hence, in a Bayesian framework, we define the **posterior risk**, $\rho(x, \pi)$, based on the data x and a prior, π:

$$\rho(x, \pi) = \int l(\theta, \delta(x)) p(\theta|x) d\theta. \tag{3}$$

Note that the prior enters the equation when calculating the posterior density. Using the Bayes risk, we can define a bit of jargon.

Definition 2 (Bayes action)**.** The Bayes action, $\delta^*(x)$, is that value of $\delta(x)$ that minimizes the posterior risk.

Example 3 (Squared error loss)**.** Last time, we showed that under squared error loss, the Bayes action is

$$\delta^*(x) = E(\theta|x). \tag{4}$$

1.3 Hybrid ideas

Despite the tensions between frequentists and Bayesians, they occasionally steal ideas from each other.

Definition 4 (Bayes rule)**.** A Bayes rule is a function, δ_π, that minimizes

$$r(\pi, \delta) = \int R(\theta, \delta)\pi(\theta)d\theta, \tag{5}$$

where $R(\theta, \delta)$ is the frequentist risk. This averages the frequentist risk over a prior distribution of θ.

Definition 5 (Bayes risk)**.** $r(\pi) = r(\pi, \delta_\pi)$ This is just $r(\pi, \delta)$ with the Bayes rule plugged in.

While the "Bayes risk" is a frequentist concept (because it averages over X), the expression $r(\pi, \delta)$ can be interpreted in a different way.

$$\begin{aligned} r(\pi, \delta) &= \int\int l(\theta, \delta(x))p(x|\theta)dx\pi(\theta)d\theta \\ &= \int\int l(\theta, \delta(x))p(\theta|x)d\theta p(x)dx \\ &= \int \rho(x, \pi)p(x)dx. \end{aligned} \tag{6}$$

Note that the last equation is the posterior risk averaged over the marginal distribution of x. This implies that the Bayes rule can be obtained by taking the Bayes action for each particular x!

Another connection with frequentist theory include that finding a Bayes rule against the "worst possible prior" gives youa minimax estimator. While a Bayesian might not find this particularly interesting, it is useful from a frequentist perspective because it provides a way to compute the minimax estimator.

2 Priors

We will emphasize objective priors in this course. In order to derive useful objective priors, we will make use of many frequentist ideas. Table 1 lists several frequentist ideas and comments on their usefulness in developing Bayesian ideas.

2.1 Why is unbiasedness bad?

Definition 6 (Unbiased estimator)**.** An estimator, $\hat{\theta}$ is unbiased iff $E_\theta(\hat{\theta}) = \theta$.

This concept is entirely frequentist, because we are thinking of applying our estimation procedure to a large number of data sets, and then hoping that the average value of your estimation procedure reflects reality.

Concept	usefulness
Asymptotics	useful
Unbiasedness	not useful
Admissibility	useful
Minimax	questionably useful
Equivariance	useful when it exists

Table 1: Summarizes the utility of frequentist ideas to Bayesian thought

Example 7 (Estimating height)**.** Here, we will estimate the height of a daughter from the height of her mother. We assume that the mother's heigh, y is a known quantity and we wish to estimate the θ, the height of her daughter. The data are assumed to follow a bivariate normal distribution with $\mu_1 = \mu_2 = 160$cm, equal variances, and $\rho = .5$. The Bayes rule under squared error loss is the posterior mean, $E(\theta|y)$. This is also a natural estimator. Now, the conditional mean of an (x_1, x_2) bivariate normal distribution with equal variances is (and you should know this!)

$$E(x_2|x_1) = \mu_2 + \rho(x_1 - \mu_1). \tag{7}$$

We want to see if this estimator is unbiased. So, plugging into (7), we have

$$\begin{aligned} E_\theta E(\theta|Y) &= 160 + .5(E_\theta(Y) - 160) \\ &= 160 + .5(160 + .5(\theta - 160) - 160) \\ &= 160 + .25(\theta - 160) \end{aligned} \tag{8}$$

.

Clearly, (8) is not equal to θ. Suppose that we want to find an unbiased estimator; it seems pretty clear that we could do that if there were some clever cancelation in the previous derivation. It turns out that the unbiased estimator is

$$\hat{\theta} = 160 + 2(Y - 160) \tag{9}$$

But (9) is not a very good estimator. Imagine if the mother is quite taller, say 170cm. This predicts that her daughter will deviate from the mean by *twice as much* and we should estimate 180cm tall. But this defies our sense that exceptionally tall people are fluctuations, and that there should be some regression toward the mean.

Example 8 (Two heads in a row)**.** Imagine we flip a coin n times, so that the number of heads is binomial(n,θ). We observe r heads and we wish to estimate θ^2, the probability of observing two heads in a row. In this case the Uniformly Minimum Variance Unbiased Estimator (UMVUE) can be calculated, and is guaranteed to be the best unbiased estimator. It is

$$\hat{\theta^2} = \frac{r(r-1)}{n(n-1)}. \tag{10}$$

But when r = 1, this estimator suggests that there is *no probability* of throwing two heads in a row! What would be a good estimator? That's a good question!

2.2 Bayesian parametric models

For now we will consider parametric models, which means that the parameter θ is a fixed dimensional vector of numbers. Let $x \in \mathcal{X}$ be the observed data and $\theta \in \Theta$ be the parameter. Note that both $\mathcal{X}$ and Θ are

probability spaces. Now we define a few recurring objects:

$$p(x|\theta) \qquad \text{likelihood} \tag{11}$$

$$\pi(\theta) \qquad \text{prior} \tag{12}$$

$$p(x) = \int p(x|\theta)\pi(\theta)d\theta \qquad \text{marginal likelihood} \tag{13}$$

$$p(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{p(x)} \qquad \text{posterior probability} \tag{14}$$

$$p(x_{\text{new}}|x) = \int p(x_{\text{new}}|\theta)p(\theta|x)d\theta \qquad \text{predictive probability} \tag{15}$$

Most of Bayesian analysis is calculating these quantities in some way or another. Note that the definition of the predictive probability assumes exchangeability, but it can easily be modified if the data are not exchangeable.

As a helpful hint, note that for the posterior distribution

$$p(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{p(x)} \propto p(x|\theta)\pi(\theta) \tag{16}$$

and oftentimes it's best to not calculate the normalizing constant, $p(x)$, because you can recognize the form of $p(x|\theta)\pi(\theta)$ as a probability distribution you know. So don't normalize until the end!

The two questions we will consider for the rest of the class are

- How do we chose priors?
- How do we compute the aforementioned quantities?

Let's focus on priors for now.

2.3 How to choose priors

As mentioned, we will primarily focus on objective priors. Objective priors are typically obtained from the likelihood. Subjective priors are typically arrived at by a process involving interviews with domain experts and thinking really hard; in fact, there is arguably more philosophy and psychology in the study of subjective priors than mathematics.

We first focus on conjugate priors, which we will leave somewhat unclearly defined for now. The main justification for the use of conjugate priors is that they are computationally convenient and they have asymptotically desirably properties.

Definition 9 (Conjugate priors)**.** A family of priors such that, upon being multiplied by the likelihood, yields a posterior in the same family.

Example 10 (Conjugate priors)**.** All probability distributions!

Clearly this is an undesirably definition, but by working through an example we can get a better feel for what we mean.

Example 11 (Multinomial distribution)**.** We assume that we have n trials, k categories and x_i trials showed up in category i. We wish to estimate the parameters, θ_i, representing the proportion in category i. The likelihood is

$$p(x|\theta) = \binom{n}{x_1, \ldots, x_n} \theta_1^{x_1}\theta_2^{x_2}\cdots\theta_k^{x_k}. \tag{17}$$

Thus, we better choose a prior that looks like θ^α. For convenience, we will use $\alpha - 1$ in the exponent, giving

$$p(\theta|\alpha) \propto \theta_1^{\alpha_1 - 1}\theta_2^{\alpha_2 - 1} \cdots \theta_3^{\alpha_k - 1}. \tag{18}$$

Thus,

$$p(\theta|x, \alpha) \propto \theta_1^{x_1 + \alpha_1 - 1}\theta_2^{x_2 + \alpha_2 - 1} \cdots \theta_3^{x_k + \alpha_k - 1}, \tag{19}$$

which has the same form as (18). Now we can calculate the normalizing constant and that leads us to another little section.

3 Integrals you should know how to do

It turns out the conjugate prior for the multinomial distribution is the Dirichlet distribution, denoted $\mathrm{Dir}(\alpha)$. The normalizing constant for a $\mathrm{Dir}(\alpha)$ distribution can be calculated:

$$\int_{\theta \in \Theta} \theta_1^{\alpha_1 - 1}\theta_2^{\alpha_2 - 1} \cdots \theta_k^{\alpha_k - 1} d\theta = \frac{\prod_{i=1}^k \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^k \alpha_i)}, \tag{20}$$

where $\Gamma(x)$ is the usual gamma function, defined by

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt. \tag{21}$$

It's useful to note that $\Gamma(x + 1) = x\Gamma(x)$, so that in particular if $x \in \mathbb{N}$, we have $\Gamma(x + 1) = x!$. For a proof of the formula for the Dirichlet integral, see the online book chapters.

The beta distribution is a special case of the Dirichlet distribution with $k = 2$, so that

$$\int_0^1 \theta^{\alpha_1 - 1}(1 - \theta)^{\alpha_2 - 1} d\theta = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1 + \alpha_2)}. \tag{22}$$

The integral of the gamma distribution can be computed by integration by parts, giving

$$\int_0^\infty \theta^{\alpha - 1} e^{-\beta\theta} d\theta = \frac{\Gamma(\alpha)}{\beta^\alpha}. \tag{23}$$

The integral of a normal distribution is well known:

$$\int_{-\infty}^\infty e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx = \sqrt{2\pi}\sigma. \tag{24}$$

Example 12 (Mean of a beta distribution)**.** These formulas can be particularly useful for calculating moments. For instance, assume that $\theta \sim \mathrm{beta}(\alpha_1, \alpha_2)$ and calculate $E(\theta)$.

$$\begin{aligned}
E(\theta) &= \int_0^1 \theta \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \theta^{\alpha_1 - 1}(1 - \theta)^{\alpha_2 - 1} d\theta \\
&= \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^1 \theta^{\alpha_1 + 1 - 1}(1 - \theta)^{\alpha_2 - 1} d\theta && (25) \\
&= \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \frac{\Gamma(\alpha_1 + 1)\Gamma(\alpha_2)}{\Gamma(\alpha_1 + \alpha_2 + 1)} && (26) \\
&= \frac{\alpha_1}{\alpha_1 + \alpha_2} && (27)
\end{aligned}$$

To get from (25) to (26), we observed that the integral in (25) was an integral of a $\mathrm{beta}(\alpha_1 + 1, \alpha_2)$ and used (22). From (26) to (27), we used the fact that $\Gamma(x + 1) = x\Gamma(x)$. You should be able to do this off the top of your head.