Stat260: Bayesian Modeling and Inference **Lecture Date: January 26, 2010**

Lecture 2: Justification for Bayes

Lecturer: Michael I. Jordan *Scribe: Max F. Dama*

1 Motivation for Bayes 2: Statistical Decision Theory

Last time we saw Motivation for Bayes 1: de Finetti's Theorem. It says that if you assume the data is infinitely exchangeable, then there must exist an underlying parameter and prior.

Today we look at another motivation for the Bayesian approach: Statistical Decision Theory. Its origins go back to Von Neumann and Morgenstern's game theory, but the main character was Wald. In statistical decision theory, we formalize good and bad results with a *loss function*.

A loss function $l(\theta, \delta(x))$ is a function of $\theta \in \Theta$ a parameter or index, and $\delta(x)$ a decision based on the data $x \in \mathcal{X}$ (for example $\delta(x) = \hat{\theta}(x)$ is a familiar estimate of the sample mean). The loss function determines the penalty for predicting $\delta(x)$ if θ is the true parameter. To give some intuition, in the discrete case, we might use a 0-1 loss or in the continuous case, we might use the squared loss $l(\theta, \delta(x)) = (\theta - \delta(x))^2$. Notice that in general, $\delta(x)$ does not necessarily have to be an estimate of θ.

Loss functions provide a very good foundation for statistical decision theory. They are simply a function of the state of nature (θ) and a decision function ($\delta(\cdot)$). In order to compare procedures we need to calculate which procedure is best even though we cannot observe the true nature of the parameter space Θ and data $\mathcal{X}$. This is the main challenge of decision theory and the break between Frequentists and Bayesians.

Frequentist Risk

The frequentist risk is

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

where θ is held fixed and the expectation is taken over $\mathcal{X}$.

Figure 1 shows what it might look like with three different decisions over all the different states of Θ.

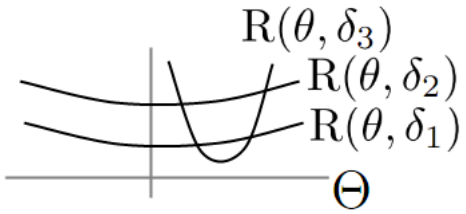

Figure 1: Frequentist Risk

Often one decision does not dominate the other everywhere as is the case with decisions δ_1, δ_2. The challenge is in saying whether, for example, δ_1 or δ_3 is better. In other words, *how should we aggregate over* Θ*?*

Frequentists have a few answers for deciding which is better:

1. Admissibility. A decision which is inadmissible is one that is dominated everywhere, for example δ_2 compared to δ_1 in Figure 1. It's easy to see how it would be easy to compare decisions if all but one were inadmissible. But usually they overlap so this criteria fails.

2. Restricted classes of procedures. For example, if we restrict our class of procedures being compared to just those which are *unbiased*, i.e. $E_\theta[\hat{\theta}] = \theta$, then the risk curves become like δ_1, δ_2 and a problematic overlapping curve like δ_3 becomes impossible. The existence of an optimal unbiased procedure is a nice Frequentist theory, but many good procedures are biased – for example Bayesian procedures are biased. More surprisingly, some unbiased procedures are actually inadmissible. For example, James Stein showed that the sample mean is an inadmissible estimate of the mean of a multivariate Gaussian in three dimensions.

   If we restrict our class of procedures to those which are *equivariant*, we also get nice properties. We do not go into detail here, but these are procedures with the same group theoretic properties as the data.

3. Minimax. In this approach we get around the problem by just looking at $\sup_\Theta \mathrm{R}(\theta, \delta)$. For example in Figure 2, δ_1 would be chosen over δ_2 because its maximum worst-case risk (the grey dotted line) is lower.

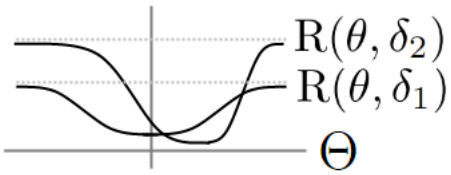

Figure 2: Minimax Frequentist Risk

   The minimax paradigm can be good when the maximum risk of the minimax procedure is low enough. Then it feels safer to use minimax procedure.

A Bayesian answer is to introduce a "weighting function" $p(\theta)$ to tell which part of Θ is important and integrate with respect to $p(\theta)$. Notice that we have made no appeal to de Finetti.

In some sense the Frequentist approach is the opposite of the Bayesian approach. However, sometimes an equivalent Bayesian procedure can be derived using a certain prior.

Before moving on, again note that $E_\theta[l(\theta, \delta(X))]$ is an expectation on $\mathcal{X}$, assuming fixed θ. A Bayesian would only look at x, the data you observed, not all possible $\mathcal{X}$. However the Frequentist insists on avoiding priors so that they can derive simple procedures that can be used everywhere i.e. software that can be used on multiple datasets. This is valid. An alternative definition of a frequentist is, "Someone who is happy to look at other data they could have gotten but didn't."

2 Motivation for Bayes 3: Principles

The Bayesain approach can also be motivated by a set of principles. Some books and classes start with a long list of axioms and principles conceived in the 1950s and 1960s. However, we will focus on three main principles.

Conditionality Principle: If an experiment concerning inference about θ is chosen from a collection of possible experiments independently, then any experiment not chosen is irrelevant to the inference.

For example, two different labs estimate the potency of drugs. Both have some error or noise in their measurements which can accurately estimate from past tests. Now we introduce a new drug. Then we test its potency at a randomly chosen lab. Should we use the noise level from the lab where it is tested or average over both? Intuitively we use the noise level from the lab where it was tested, but in some frequentist approaches, it is not always so straightforward.

Likelihood Principle: The relevant information in any inference about θ after x is observed is contained entirely in the likelihood function.

Remember the likelihood function $p(x|\theta)$ for fixed x is a function of θ not x.

For example in Bayesian approaches, $p(\theta|x) \propto p(x|\theta)p(\theta)$, so clearly inference about θ is based on the likelihood.

Another approach based on the likelihood principle is Fisher's maximum likelihood estimation. This approach can also be just justified by asymptotics.

In case this principle seems too indisputable, here is an example using hypothesis testing in coin tossing that shows how some reasonable procedures may not follow it. Let θ be the probability of a coin landing on heads and let $H_0 : \theta = 1/2$, $H_1 : \theta > 1/2$, and the observed data be 9 heads 3 tails. The likelihood is simply $\theta^9(1-\theta)^3 \propto p(x|\theta)$ with the data plugged in. Many non-Bayesian analyses would pick an experimental design that is reflected in $p(x|\theta)$, for example binomial (toss a coin 12 times) or negative binomial (toss a coin until you get 9 heads). However the two lead to different probabilities over the sample space $\mathcal{X}$. This results in different assumed tail probabilities and p-values.

Sufficiency Principle: If two different observations x, y are such that $T(x) = T(y)$ for sufficient statistic T, then inference based on x and y should be the same.

The sufficiency principle is the least controversial principle.

Theorem 1 (Birnbaum)**.** *Sufficiency + Conditionality* $\equiv$ *Likelihood*

Since sufficiency is accepted, conditionality $\equiv$ likelihood.

The Bayesian approach satisfies all of these principles.

3 Bayesian Decision Theory

Earlier we discussed the Frequentist approach to statistical decision theory. Now we discuss the Bayesian approach in which we condition on x and integrate over Θ (remember it was the other way around in the Frequentist approach).

The *posterior risk* is defined as $\underbrace{\rho(\pi, \delta(x))}_{vs.\mathrm{R}(\theta,\delta)} = \int l(\theta, \delta(x))p(\theta|x)d\theta$ where $p(\theta|x) \propto p(x|\theta)\pi(\theta)$.

The *Bayes action* $\delta^*(x)$ for any fixed x is the decision $\delta(x)$ that minimizes the posterior risk.

Example 1

$$
\begin{aligned}
l(\theta, \delta(x)) &= (\theta - \delta(x))^2 && (2)\\
\Rightarrow \rho &= \int (\theta - \delta(x))^2 p(\theta|x) d\theta && (3)\\
&= \delta(x)^2 - 2\delta(x) \int \theta p(\theta|x) d\theta + \int \theta^2 p(\theta|x) d\theta && (4)\\
\Rightarrow \frac{\partial \rho}{\partial \delta(x)} &= 2\delta(x) - 2 \int \theta p(\theta|x) d\theta = 0 && (5)\\
\Rightarrow \delta^*(x) &= \int \theta p(\theta|x) d\theta \quad \text{the posterior mean} && (6)
\end{aligned}
$$

Example 2 $l(\theta, \delta(x)) = |(\theta - \delta(x))| \Rightarrow$ the optimal decision will be to choose the posterior median.

Conclusion

Both Frequentists and Bayesians agree loss functions are good.
Frequentists integrate out the $\mathcal{X}$ part.
Bayesians integrate out the θ part.
By Fubini's Theorem, integrating over both gives you the misleadingly-named "Bayes Risk". More on this next class.