36-700: Probability and Mathematical Statistics I **Fall 2016**

Lecture 11: September 23

Lecturer: Siva Balakrishnan

11.1 Review and Outline

Last class we discussed:

- Bayes estimators
- Loss functions
- The risk of an estimator

In this lecture we will discuss how to evaluate and compare estimators. This is covered in the second half of Chapter 7 of C&B and Chapter 12 of Wasserman.

11.2 Comparing estimators

In the last lecture we saw that most often there was no uniformly dominant estimator, i.e. most often there was not an estimator that had smaller risk than every other estimator, everywhere.

Example: Let us consider the Bernoulli estimation problem: two natural estimators are the MLE:

$$\hat{p}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

and the Bayes estimator we defined previously:

$$\hat{p}_2 = \frac{\sum_{i=1}^{n} X_i + \alpha}{n + \alpha + \beta},$$

for some values α and β that we will specify soon. Again, suppose we consider the squared loss:

$$R(p, \hat{p}_1) = \frac{p(1-p)}{n}.$$

$$R(p, \hat{p}_2) = \text{Var}\left(\frac{\sum_{i=1}^{n} X_i + \alpha}{n + \alpha + \beta}\right) + \left(\mathbb{E}\frac{\sum_{i=1}^{n} X_i + \alpha}{n + \alpha + \beta} - p\right)^2.$$

In the second estimator if we choose $\alpha = \beta = \sqrt{n/4}$ we obtain that the risk is constant as a function of p, i.e.

$$R(p, \hat{p}_2) = \frac{n}{4(n + \sqrt{n})^2}.$$

We can compare these two estimators' risk functions but once again we see that neither estimator dominates the other. In such cases, we need other ways to compare estimators and to find "best" estimators.

Admissibility: One basic step in this direction is to weed out what are called *inadmissible estimators*. Particularly, it seems natural to disregard an estimator $\hat{\theta}_1$ if there is another estimator $\hat{\theta}_2$ such that,

$$R(\theta, \hat{\theta}_2) \leq R(\theta, \hat{\theta}_1),$$

for every $\theta \in \Theta$, and

$$R(\theta, \hat{\theta}_2) < R(\theta, \hat{\theta}_1),$$

for some $\theta \in \Theta$. Estimators like θ_1 are called *inadmissible* estimators. As one might expect there are often many admissible estimators so we need other ways to narrow our search further.

The essential idea is to try to summarize the risk function by a single parameter. There are two dominant ideas in this space:

1. Minimax risk: The minimax estimator $\hat{\theta}$ is one that minimizes the worst-case risk,i.e., it is one that satisfies:

   $$\sup_{\theta \in \Theta} R(\theta, \hat{\theta}(X)) = \min_{\theta'} \max_{\theta \in \Theta} \mathbb{E}_\theta L(\theta, \theta'(X)).$$

   More generally however the minimax risk idea suggests comparing two estimators on the basis of their worst-case risk. For many problems of interest we can actually find the minimax estimator (or at least one that achieves the minimax risk upto constants).

2. Bayes risk: The Bayes risk of an estimator is its risk averaged with respect to a prior, i.e., for some prior $\pi(\theta)$:

   $$R_\pi(\hat{\theta}) = \mathbb{E}_{\theta \sim \pi} R(\theta, \hat{\theta}(X)).$$

   The Bayes risk (similar to the worst-case risk) of an estimator is just a number and thus it is easy to compare two estimators. A natural estimator is one that minimizes the Bayes risk, and this is sometimes called the *Bayes rule* with respect to the prior π. This might seem as intractable as finding an optimal estimator by any other metric but it turns out that we can simplify the problem a bit.

Bayes Estimator & Bayes Rules: Suppose that $\theta \sim \pi$ and $X \sim f_\theta$. Observe that the above is defining the conditional distribution of X: we denote the marginal distribution as $m(X)$. Then the Bayes risk of an estimator is:

$$\begin{aligned}\int_\theta R(\theta, \hat{\theta}(X))\pi(\theta)d\theta &= \int_\theta \left[\int_X L(\theta, \hat{\theta}(X))f_\theta(X)dX\right]\pi(\theta)d\theta \\ &= \int_X \left[\int_\theta L(\theta, \hat{\theta})\pi(\theta|X)d\theta\right] m(X)dX.\end{aligned}$$

Now, the Bayes rule just minimizes this expression, so we can see that for every X our estimator is given by:

$$\hat{\theta}(X) = \arg\min_{\theta'} \int_\theta L(\theta, \theta'(X))\pi(\theta|X)d\theta.$$

The term on the RHS is called the posterior expected loss. So the Bayes rule is one that minimizes the posterior expected loss. A simple but important special case of this is that if L is the squared loss then the Bayes rule is a conditional expectation, i.e.,

$$\hat{\theta}(X) = \mathbb{E}[\theta|X].$$

The posterior mean is the same Bayes estimator we defined previously. For different loss functions however we obtain different Bayes rules: for instance if the loss function was ℓ_1 the Bayes rule would be the median of the posterior. Lets consider an example:

Example: Suppose we have the loss function: $L(\theta, \hat{\theta}) = w(\theta)(\theta - \hat{\theta})^2$, where $w(\theta) \geq 0$ is some positive weighting function. The Bayes rule minimizes the posterior risk,i.e.:

$$\hat{\theta}(X) = \arg\min_{\hat{\theta}} \mathbb{E}\left[w(\theta)(\theta - \hat{\theta})^2|X\right].$$

Taking the derivative and setting it to 0, we obtain that the Bayes rule is:

$$\hat{\theta}(X) = \frac{\mathbb{E}\left[\theta w(\theta)|X\right]}{\mathbb{E}\left[w(\theta)|X\right]},$$

so if we weight the squared loss, we obtain a weighted conditional expectation as the Bayes rule.

Another way to think about Bayes risk: An alternative way to think about the Bayes risk is as the average of the *loss* when we have the generative model: $\theta \sim \pi$ and $X \sim f_\theta$, i.e.,

$$R_\pi(\hat{\theta}) = \mathbb{E}L(\theta, \hat{\theta}(X)),$$

where the expectation is over both (θ, X).

As a point of comparison, the max-risk does not involve the choice of an arbitrary prior so in that sense has some advantages over the Bayes risk.

Example: Let us revisit the two Bernoulli estimators from the standpoint of maximum risk and Bayes risk. Suppose we take the uniform prior, then:

$$R_\pi(\hat{p}_1) = \int_p \frac{p(1-p)}{n} dp = \frac{1}{6n},$$
$$R_\pi(\hat{p}_2) = \frac{n}{4(n+\sqrt{n})^2},$$

so for large n the MLE has smaller Bayes risk. On the other hand the estimator $\hat{p}_2$ always has lower maximum risk.

11.3 Connections

In general, finding exactly minimax estimators is not easy but often we can find approximately minimax (or rate minimax) estimators. There are however, some important connections between Bayes rules, minimax estimators and admissible estimators that are worth knowing.

1. For any estimator the Bayes risk with respect to any prior π lower bounds its maximum risk.

2. Suppose that for some prior π, we have that the corresponding Bayes rule $\hat{\theta}_\pi$ has the property that:

   $$R(\theta, \hat{\theta}_\pi) \leq R_\pi(\hat{\theta}_\pi),$$

   for every θ. Then π is called a **least favorable prior** and $\hat{\theta}_\pi$ is minimax.

3. A simple consequence of the above is that if for some π we have that $R(\theta, \hat{\theta}_\pi)$ is a constant (as a function of θ) then $\hat{\theta}_\pi$ is a minimax estimator. In words, a Bayes rule with constant risk is minimax.

   Example: Recall the Binomial Bayes estimator with $\alpha = \beta = \sqrt{n}/2$. We saw that its MSE is:

   $$R(\hat{p}, p) = \frac{n}{4(n+\sqrt{n})^2}.$$

   Since the Bayes estimator has constant risk it is minimax.

   Example: Suppose we consider Bernoulli estimation with the loss function:

   $$L(p, \hat{p}) = \frac{(p-\hat{p})^2}{p(1-p)}.$$

This loss function is sometimes called the Fisher loss. We will see why in a future lecture.

Let us consider the MLE under this loss. We can see that its risk:

$$R(\hat{p}, p) = \frac{1}{n},$$

which is again constant over the parameter space. Can we conclude that the MLE is minimax? Unfortunately, we do not yet know if it is a Bayes estimator. It turns out that another lengthy calculation reveals that the MLE is actually the Bayes estimator under the uniform prior. So we can conclude that the MLE is minimax with respect to the Fisher loss. This is actually a pretty general phenomenon.

4. Bayes estimators with respect to certain "nice" priors are admissible. Formally, if the prior π is non-zero everywhere over the parameter space and if the risk of the Bayes rule is finite then the Bayes rule is admissible.

5. An estimator with constant risk that is admissible is also minimax.

6. Minimax estimators are not always admissible. This is easy to visualize.

   Perhaps the most stunning example of this is due to Charles Stein. This is sometimes called Stein's phenomenon or Stein's paradox. We will not go over this in detail but its worth knowing the paradox. The setting is that you observe many different quantities corrupted by Gaussian noise, i.e. we observe:

   $$y_i \sim N(\theta_i, 1),$$

   for $i \in \{1, \ldots, d\}$, and we would like to estimate the vector θ in MSE. The natural estimator of course is just the observed data, i.e.:

   $$\hat{\theta} = y.$$

   It turns out that this estimator is minimax optimal. What Stein showed was that this estimator is inadmissible if $d \geq 3$, in particular he showed that if you make every entry of y slightly smaller the resulting estimator dominates the MLE. The amount by which you make the entries smaller depends on the entire set of observations y. This is stunning because the different objects you are measuring are unrelated, yet if you measure more than two objects, you can improve on the obvious estimator by using information you got from *unrelated* measurements.