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

Lecture 10: September 21

Lecturer: Siva Balakrishnan

10.1 Review and Outline

Last class we discussed:

- Method-of-moments
- Maximum Likelihood Estimation
- (Briefly) Bayes Estimators

In this lecture we will go over Bayes estimators in more detail and then discuss some methods to evaluate estimators. This is the second half of Chapter 7 of C&B.

10.2 Bayes Estimators

The third general method to derive estimators is a bit different philosophically from the first two. At a high-level we need to understand the Bayesian approach (we will of course stay clear of any philosophical questions): in our approaches so far (so-called "frequentist" approaches) we assumed that there was a fixed but unknown true parameter, and that we observed samples drawn i.i.d from the population (whose density/mass function/distribution was indexed by the unknown parameter).

In the Bayesian approach we consider the parameter θ to be random. We have a *prior* belief of the distribution of the parameter, which we update after seeing the samples $X_1, \ldots, X_n$. We update our belief using Bayes rule and the updated distribution is known as the *posterior* distribution.

We denote the prior distribution by $\pi(\theta)$, and the posterior distribution as $\pi(\theta|X_1, \ldots, X_n)$. Using Bayes' rule we have that:

$$\pi(\theta|X_1, \ldots, X_n) = \frac{\pi(\theta)L(\theta|X_1, \ldots, X_n)}{\int_\theta \pi(\theta)L(\theta|X_1, \ldots, X_n)}.$$

The posterior distribution is a distribution over the possible parameter values. In this lecture we are focusing on point estimation so one common candidate is the posterior mean (i.e. the expected value of the posterior distribution).

Ignoring any philosophical questions, one can view this methodology as a way to generate candidate point estimators, by specifying reasonable prior distributions. In practice however this calculation is hard to do analytically, so we often end up specifying priors out of convenience rather than any real prior belief. Particularly, a convenient choice of prior distribution is one for which the posterior distribution belongs to the same family as the prior distribution: such priors are called *conjugate priors*.

Example 1: Binomial Bayes estimator: Suppose $X_1, \ldots, X_n \sim \text{Ber}(p)$. We will first need to define the Beta distribution: a RV has a Beta distribution with parameters α and β if its density on $[0, 1]$ is

$$f(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}.$$

For us it will be sufficient to ignore the normalizing part and just remember that the Beta density:

$$f(p) \propto p^{\alpha-1}(1-p)^{\beta-1}.$$

The mean of the Beta distribution is: $\alpha/(\alpha+\beta)$.

Let us denote

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

There are two candidate priors we will consider:

1. The flat/uninformative prior: $\pi(p) = 1$ for $0 \leq p \leq 1$. In this case, the posterior density is:

$$\begin{aligned} f(p|X_1, \ldots, X_n) &\propto p^{n\bar{X}}(1-p)^{n(1-\bar{X})} \\ &= p^{(n\bar{X}+1)-1}(1-p)^{(n(1-\bar{X})+1)-1}. \end{aligned}$$

   This is just a Beta$(n\bar{X}+1, n(1-\bar{X})+1)$ distribution. So our estimate (the posterior mean) would be:

$$\begin{aligned} \hat{p} &= \frac{n\bar{X}+1}{n+2} \\ &= \frac{n}{n+2}\bar{X} + \frac{2}{n+2}\frac{1}{2} \\ &= w\bar{X} + (1-w)\frac{1}{2}, \end{aligned}$$

   which can be viewed as a convex combination of the MLE and the prior mean 1/2.

2. The other common prior is the one that is conjugate to the bernoulli likelihood, i.e. the Beta prior. A similar calculation as the one above will show that if use $\pi(p) \sim \text{Beta}(\alpha, \beta)$, then the posterior distribution will also be a Beta distribution:

$$f(p|X_1, \ldots, X_n) \sim \text{Beta}(\alpha + n\bar{X}, \beta + n(1 - \bar{X})),$$

and our Bayes estimator would be:

$$\hat{p} = \frac{\alpha + n\bar{X}}{\alpha + \beta + n}.$$

Example 2: Gaussian Bayes estimator: Here we suppose that our prior belief is that the parameter has distribution $N(\mu, \tau^2)$ and we observe $X_1, \ldots, X_n$ drawn from $N(\theta, \sigma^2)$. We assume that σ^2, μ, τ^2 are all known.

A fairly involved calculation in this case (see for instance Problem 1 of Chapter 11 of Wasserman) will show that the posterior distribution is also a Gaussian with parameters:

$$\hat{\mu} = \frac{n\tau^2}{\sigma^2 + n\tau^2}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) + \frac{\sigma^2}{\sigma^2 + n\tau^2}(\mu)$$
$$\hat{\sigma}^2 = \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2},$$

which suggests the point estimate:

$$\hat{\mu} = \frac{n\tau^2}{\sigma^2 + n\tau^2}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) + \frac{\sigma^2}{\sigma^2 + n\tau^2}(\mu).$$

Which can once again be seen as a convex combination of our prior belief and the MLE, i.e.:

$$\hat{\mu} = w\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) + (1 - w)\mu,$$

for some $0 \leq w \leq 1$.

10.3 Evaluating Estimators

We have already discussed the MSE and the bias-variance decomposition. Today we will discuss *decision theory* more broadly.

The central idea in decision theory is that we want to minimize our *expected* loss.

Let us first try to understand the decision theoretic setup. We observe data X, where $X \sim f_\theta$, with $\theta \in \Theta$, and we make a decision, i.e. we select an action a.

In point estimation, the decision is just our guess of the parameter. In hypothesis testing situations our decision will instead be which of the hypotheses we believe to be true. Once we take an action we suffer a loss. The loss function in point estimation is roughly something that is large if a is far from θ and small if our guess is good, i.e., if a is close to θ.

Some very common loss functions are:

1. **Squared loss:** $L(a, \theta) = (a - \theta)^2$.
2. **Absolute loss:** $L(a, \theta) = |a - \theta|$.

There are however many other loss functions that we will encounter frequently. For instance, we often consider losses like:

$$L(a, \theta) = \frac{(a - \theta)^2}{|\theta| + 1},$$

which penalizes errors in estimation more for small values of θ than for large values. We can similarly design a loss function that penalizes errors more strongly for large values of θ.

Another important point is that there are cases when we do not really care about estimating the parameter well but rather just the distribution f_θ. This is true when we care about prediction in regression or in density estimation. In this case we could define the loss between θ and a in terms of the distributions f_θ and f_a. One canonical example:

Kullback-Leibler loss:

$$L(a, \theta) = \mathrm{KL}(f_\theta, f_a) = \mathbb{E}_{X \sim f_\theta} \log \left(\frac{f_\theta(X)}{f_a(X)} \right).$$

Once we have a loss function, and an estimator, we can assess the estimator via its expected loss. This expected loss is called the *risk* of the estimator. Suppose we consider an estimator $\hat{\theta}(X)$. Then we define:

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

In general, we do not a-priori know anything about the value of θ so we would like estimators with low risk for all parameters $\theta \in \Theta$. So ideally, we would like to find an estimator $\hat{\theta}$ such that for any other estimator θ' we have that:

$$R(\theta, \hat{\theta}(X)) \leq R(\theta, \theta')$$

for all values θ. Such estimators will most often not exist – why not?

Example: Suppose $X \sim N(\theta, 1)$, and we care about estimating θ in MSE. Consider two estimators: $\hat{\theta} = X$ and $\hat{\theta} = 0$. The risk of X is: $\mathbb{E}(X - \theta)^2 = 1$, while the risk of 0 is $\mathbb{E}\theta^2 = \theta^2$. So when $\theta < 1$, 0 is a better estimator than the estimator X. Neither estimator dominates the other.

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:

$$R(p, \hat{p}_1) = \frac{p(1-p)}{n}.$$
$$R(p, \hat{p}_2) = \mathrm{Var}\left(\frac{\sum_{i=1}^{n} X_i + \alpha}{n + \alpha + \beta}\right) + \left(\mathbb{E}\frac{\sum_{i=1}^{n} X_i + \alpha}{n + \alpha + \beta} - p\right)^2.$$

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.

A lot of statistical theory was developed from this decision theoretic starting point. At a high-level there are several different paradigms and ideas:

1. The notion of admissibility: With our decision theoretic mechanism in place we could attempt to weed out the really useless 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.

2. 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 various reasons this is one of the dominant paradigms for evaluating estimators. This is because for many problems of interest we can actually find the minimax estimator (or at least one that achieves the minimax risk upto constants).

3. 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 estimator. 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: 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 estimator 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 estimator 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 estimator is a conditional expectation, i.e.,

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

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. In the next lecture we will show that this estimator is actually minimax optimal.