LECTURE NOTES 10

1 Evaluating Estimators: Decision Theory

We have seen three methods to define esimators: the method of moments, maximum likelihood and Bayes estimators. No we will intoduce decision theory which is a way to evaluate how good an estimator is.

The basic idea is this. We observe data $X_1, \ldots, X_n \sim p_\theta$. We construct an estimator $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$. To evaluate $\widehat{\theta}$ we introduce a loss function $L(\widehat{\theta}, \theta)$. Some common loss functions are:

1. **Squared loss:** $L(\widehat{\theta}, \theta) = (\widehat{\theta} - \theta)^2$.
2. **Absolute loss:** $L(\widehat{\theta}, \theta) = |\theta - \theta|$.
3. **Kullback-Leibler loss:** $L(\widehat{\theta}, \theta) = \mathrm{KL}(\widehat{\theta}, \theta) \equiv \int p_\theta(u) \log\left(\frac{p_\theta(u)}{p_{\widehat{\theta}}(u)}\right) du$.

Next, we define the **risk function**:

$$R(\theta, \widehat{\theta}) = \mathbb{E}_\theta[L(\widehat{\theta}, \theta)] = \int L(\widehat{\theta}, \theta) p_\theta(x_1) p_\theta(x_2) \cdots p_\theta(x_n) dx_1 \cdots dx_n.$$

The most commonly used loss function is $L(\widehat{\theta}, \theta) = (\widehat{\theta} - \theta)^2$. The risk function $R = \mathbb{E}[(\widehat{\theta} - \theta)^2]$ is called the mean squared error, or MSE.

We want to choose an estimator $\widehat{\theta}$ whose risk function is small. At this point, it is not obvious how to compare risk functions.

Example: Suppose $X \sim N(\theta, 1)$, and we care about estimating θ in MSE. Consider two estimators: $\widehat{\theta} = X$ and $\widehat{\theta} = 0$. 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:

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

and the Bayes estimator we defined previously:

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

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

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

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.

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

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.

There are two common ways to define a 'best estimator.'

1. Minimax risk: The minimax estimator $\widehat{\theta}$ is one that minimizes the maximum risk: the minimax estimator $\widehat{\theta}$ satisfies

$$\sup_{\theta\in\Theta} R(\theta, \widehat{\theta}) = \inf_{\theta'} \sup_{\theta\in\Theta} R(\theta, \theta')$$

where the infimum is over all estimators.

2. Bayes estimator: We called the mean of a posterior distribution, the Bayes estimator. Now we intoduce a more general notion of Bayes estimator. Given a prior p define the **Bayes risk** of $\widehat{\theta}$ by

$$R_p(\widehat{\theta}) = \int R(\theta, \widehat{\theta}) p(\theta) d\theta.$$

The **Bayes estimator:** is the $\widehat{\theta}$ that minimizes $R_p(\widehat{\theta})$.

Let

$$\mathcal{L}(\theta) = p(x_1, \ldots, x_n|\theta) = \prod_i p_\theta(X_i)$$

denote the likelihood function. Note that

$$\begin{aligned}
\int R(\theta,\widehat{\theta})p(\theta)d\theta &= \int \left[\int L(\widehat{\theta},\theta)p_\theta(x_1)p_\theta(x_2)\cdots p_\theta(x_n)dx_1\cdots dx_n\right] p(\theta)d\theta \\
&= \int\int L(\widehat{\theta},\theta)p(x_1,\ldots,x_n|\theta)p(\theta)d\theta dx_1\cdots dx_n \\
&= \int\int L(\widehat{\theta},\theta)p(x_1,\ldots,x_n,\theta)d\theta dx_1\cdots dx_n \\
&= \int\int L(\widehat{\theta},\theta)p(\theta|x_1,\ldots,x_n)p(x_1,\ldots,x_n)d\theta dx_1\cdots dx_n \\
&= \int\int L(\widehat{\theta},\theta)p(\theta|x_1,\ldots,x_n)d\theta p(x_1,\ldots,x_n)dx_1\cdots dx_n \\
&= \int r(\theta|x_1,\ldots,x_n)p(x_1,\ldots,x_n)dx_1\cdots dx_n
\end{aligned}$$

where

$$r(\theta|x_1,\ldots,x_n) = \int L(\widehat{\theta},\theta)p(\theta|x_1,\ldots,x_n)d\theta$$

is called the posterior risk. To minimize the Bayes risk, it suffices to choose $\widehat{\theta}$ to minimize $r(\theta|x_1,\ldots,x_n) =$.

For example, suppose that $L = (\widehat{\theta}-\theta)^2$. Then

$$r(\theta|x_1,\ldots,x_n) = \int (\widehat{\theta}-\theta)^2 p(\theta|x_1,\ldots,x_n)d\theta.$$

Take the derivative w.r.t. $\widehat{\theta}$ and set it equal to 0. We get

$$\widehat{\theta} = \frac{\int \theta p(\theta|x_1,\ldots,x_n)d\theta}{\int p(\theta|x_1,\ldots,x_n)d\theta} = \mathbb{E}(\theta|x_1,\ldots,x_n)$$

which is the posterior mean.

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:

$$\begin{aligned}
R_p(\widehat{\theta}_1) &= \int \frac{\theta(1-\theta)}{n}d\theta = \frac{1}{6n}, \\
R_p(\widehat{\theta}_2) &= \frac{n}{4(n+\sqrt{n})^2},
\end{aligned}$$

so for large n the MLE has smaller Bayes risk.

On the other hand the estimator $\widehat{\theta}_2$ always has lower maximum risk. In the next lecture we will show that this estimator is actually minimax optimal.