STATS 300A: Theory of Statistics **Fall 2015**

Lecture 8 — October 15

Lecturer: Lester Mackey *Scribe: Hongseok Namkoong, Phan Minh Nguyen*

Warning: *These notes may contain factual and/or typographic errors.*

8.1 Bayes Estimators and Average Risk Optimality

8.1.1 Setting

We discuss the average risk optimality of estimators within the framework of Bayesian decision problems. As with the general decision problem setting the Bayesian setup considers a model $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Omega\}$, for our data X, a loss function $L(\theta, d)$, and risk $R(\theta, \delta)$. In the frequentist approach, the parameter θ was considered to be an unknown deterministic quantity. In the Bayesian paradigm, we consider a measure Λ over the parameter space which we call a prior. Assuming this measure defines a probability distribution, we interpret the parameter θ as an outcome of the random variable $\Theta \sim \Lambda$. So, in this setup both X and θ are random. Conditioning on $\Theta = \theta$, we assume the data is generated by the distribution $\mathbb{P}_\theta$. Now, the **optimality goal** for our decision problem of estimating $g(\theta)$ is the minimization of the **average risk**

$$r(\Lambda, \delta) = \mathbb{E}[L(\Theta, \delta(X))] = \mathbb{E}[\mathbb{E}[L(\Theta, \delta(X)) \mid X]].$$

An estimator δ which minimizes this average risk is a **Bayes estimator** and is sometimes referred to as being **Bayes**.

Note that the average risk is an expectation over both the random variables Θ and X. Then by using the tower property, we showed last time that it suffices to find an estimator δ which minimizes the posterior risk $\mathbb{E}[L(\Theta, \delta(X))|X = x]$ for almost every x. This estimator will generally depend on the choice of loss function, as we shall now see.

8.1.2 Examples of Bayes estimators

Example 1. Suppose the loss function is the absolute error loss. Our task is to find a δ that minimizes the posterior risk which in this case is given by

$$\mathbb{E}[|g(\Theta) - \delta(X)||X].$$

Thus, the minimizer $\delta_\Lambda(X)$, or the Bayes estimator, is any median of $g(\Theta)$ under the posterior distribution $\Theta|X$. This is given by a quantity c such that the following two properties hold:

$$\mathbb{P}(g(\Theta) \geq c|X) \geq 1/2$$
$$\mathbb{P}(g(\Theta) \leq c|X) \geq 1/2$$

We have seen that for the squared error loss function the Bayes estimator is given by the mean of the posterior distribution. In the next example we consider a generalization of the squared error loss.

Example 2. Suppose the loss function is of the form $L(\theta, d) = w(\theta)(d - g(\theta))^2$, where $w(\theta) \geq 0$. Here the function $w(\theta)$ can be thought of as a weight function. Then the Bayes estimator minimizes the posterior risk

$$\mathbb{E}[w(\Theta)(g(\Theta) - d)^2 \big| X = x]$$

with respect to d, where we are able to take d out of the conditional expectation because our estimator is a function only of X and so d is a constant for fixed $X = x$. Rewriting, we have

$$d^2\mathbb{E}[w(\Theta)|X = x] - 2d\mathbb{E}[w(\Theta)g(\Theta)|X = x] + \mathbb{E}[w(\Theta)g^2(\Theta)|X = x]$$

Now, this is nothing but a convex quadratic function in d. Taking derivatives w.r.t. d, we see that this function takes on its minimum when

$$2d\mathbb{E}[w(\Theta)|X = x] - 2\mathbb{E}[w(\Theta)g(\Theta)|X = x] = 0.$$

Hence, the Bayes estimator is given by

$$\delta_\Lambda(X) = \frac{\mathbb{E}[w(\Theta)g(\Theta)|X = x]}{\mathbb{E}[w(\Theta)|X = x]}.$$

Note that this is the ratio of the posterior mean of $w(\Theta)g(\Theta)$ and that of $w(\Theta)$. In particular, if $w \equiv 1$, our loss function is the usual squared error loss function and the above expression yields the Bayes estimator in this case to be the posterior mean of $g(\Theta)$ as we had seen before.

Example 3 (Binary Classification)**.** Our next example is inspired by a quintessential binary classification task, email spam filtering. Our goal is, given an incoming e-mail, to classify that email as either spam or non-spam (we will call this higher-quality email "ham" in the latter case). To model this, the parameter space is taken to be $\Omega = \{0, 1\}$, where 0 corresponds to a "ham" and a 1 represents "spam", and we suppose that the email X is drawn from either the ham distribution f_0 or the spam distribution f_1. Naturally, the decision space is $\mathcal{D} = \{0, 1\}$ as well: we predict either "ham" or "spam" for the incoming email. A natural loss function to consider in a such a binary classification problem set up is the 0-1 loss function[1]:

$$L(\theta, d) = \begin{cases} 0 & d = \theta \\ 1 & d \neq \theta \end{cases}.$$

We tackle this problem in Bayesian fashion by defining a prior distribution with $\pi(1) = p$ and $\pi(0) = 1 - p$ for some fixed $p \in [0, 1]$. The hyperparameter p is the probability assigned to an e-mail being spam before observing any data point. One way to select p based on prior

[1] This is not the only reasonable choice of loss. We may for instance want to impose different penalties for misclassifying true spam emails and misclassifying true ham emails.

knowledge is to set p equal to the proportion of previously received emails which were spam. Given this decision problem, the Bayes estimator minimizes the average risk

$$r(\pi, \delta) = \mathbb{E}[L(\Theta, \delta(X))] = \mathbb{P}(\delta(X) \neq \Theta) = 1 - \mathbb{P}(\delta(X) = \Theta).$$

We see that minimizing the average risk is equivalent to maximize the probability of correct classification $\mathbb{P}(\delta(X) = \Theta)$.

As a thought experiment, consider attempting to predict the label of an incoming email before it has been viewed. Since we have no data to condition on, our only (non-randomized) options are the constant estimators $\delta_1 \equiv 1$ and $\delta_0 \equiv 0$. The average risks for these are

$$r(\pi, \delta_1) = \pi(1)R(1, \delta_1) + \pi(0)R(0, \delta_1) = 1 - p$$

and

$$r(\pi, \delta_0) = \pi(1)R(1, \delta_0) + \pi(0)R(0, \delta_0) = p.$$

So δ_1 has smaller average risk when $p > 1/2$, and δ_0 has smaller average risk when $p < 1/2$.

After observing the data $X = x$, we make the estimation $\theta = 1$ if it has higher posterior probability $\mathbb{P}(\Theta = \delta(X)|X = x)$, which is proportional to the product of the likelihood times the prior. This gives the following two relations:

$$\begin{aligned}\mathbb{P}(\Theta = 1|X = x) \propto f_1(x) \cdot \pi(1) =& f_1(x)p \\ \mathbb{P}(\Theta = 0|X = x) \propto f_0(x) \cdot \pi(0) =& f_0(x)(1-p).\end{aligned}$$

Note that both of these posterior probabilities have the common normalizing constant $1/[f_1(x)p + f_0(x)(1-p)]$ so it is enough to consider just the numerators. So the Bayes estimator predicts 1 iff

$$\frac{f_1(x)p}{f_0(x)(1-p)} > 1,$$

i.e., iff

$$\frac{f_1(x)}{f_0(x)} > \frac{1-p}{p}.$$

The left-hand side of the above inequality is known as a likelihood ratio as it is the ratio of the likelihoods of X under the two values of θ in the parameter space. The right-hand side is known as the prior odds.

Example 4. Consider the binary classification setting again with likelihood $f_j = \text{Exp}(\lambda_j)$, for $j \in \{0, 1\}$, where λ_j are known rate parameters. We assume w.l.o.g. $\lambda_0 > \lambda_1$. Consider that data $X_1, \ldots, X_n$ (a batch of n emails) are drawn i.i.d. from f_θ. From the calculations above it follows that we estimate 1 iff

$$\frac{\lambda_1^n \exp\left(-\lambda_1 \sum_{i=1}^n x_i\right)}{\lambda_0^n \exp\left(-\lambda_0 \sum_{i=1}^n x_i\right)} > \frac{1-p}{p}.$$

This condition is equivalent to

$$-(\lambda_1 - \lambda_0) \sum_{i=1}^n x_i > \log(1-p) - \log(p) + n \log(\lambda_0/\lambda_1).$$

The above means we estimate 1 iff $\sum_{i=1}^n x_i > h$, where h depends on λ_1, λ_0, p, and n.

Note that in the example above, the Bayes estimator depends on the data only through the sufficient statistic $\sum_i X_i$. This is not surprising and is in general true for Bayes estimators. To see this, suppose a sufficient statistic $T(X)$ exists. Then the posterior density, which is proportional to the product of the likelihood and the prior can be written as

$$\begin{aligned} \text{posterior } \pi(\theta|x) &\propto f(x|\theta)\pi(\theta) \\ &\propto h(x)g_\theta(T(x))\pi(\theta) \quad \text{(by NFFC)} \\ &\propto g_\theta(T(x))\pi(\theta). \end{aligned}$$

Note that $h(x)$ does not involve θ, so it cancels out with the same term in the normalizing constant. Since the posterior depends on the data only through the sufficient statistic T, the same will be the case for the estimator minimizing the posterior risk. Keeping this fact in mind, we consider another Bayesian example.

Example 5 (Normal mean estimation)**.** Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\Theta, \sigma^2)$, with σ^2 known. Further, let $\Theta \sim \mathcal{N}(\mu, b^2)$, where μ, and b are fixed prior hyper-parameters. Thus, we have that, the posterior $\pi(\theta|x)$ is

$$\begin{aligned} \pi(\theta|x) &\propto \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x_i-\theta)^2\right) \frac{1}{\sqrt{2\pi}b} \exp\left(-\frac{1}{2b^2}(\theta-\mu)^2\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\theta)^2 - \frac{1}{2b^2}(\theta-\mu)^2\right) \\ &\propto \exp\left(\frac{1}{\sigma^2}\sum_{i=1}^n x_i\theta - \frac{n\theta^2}{2\sigma^2} - \frac{1}{2b^2}\theta^2 + \frac{\mu}{b^2}\theta\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{b^2}\right)\theta^2 + \left(\frac{n\bar{x}}{\sigma^2}+\frac{\mu}{b^2}\right)\theta\right). \end{aligned}$$

Thus the posterior density, $\pi(\theta|x)$ is that of an exponential family with sufficient statistics θ and θ^2, which implies that the posterior distribution is normal. Looking to the normal density, we know that the coefficient of θ in the last line above will be

$$\left(\frac{n\bar{x}}{\sigma^2}+\frac{\mu}{b^2}\right) = \frac{\text{mean}}{\text{variance}},$$

and the coefficient of θ^2 is

$$-\frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{b^2}\right) = -\frac{1}{2\cdot\text{variance}}.$$

So the posterior distribution is $\mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2)$ with

$$\tilde{\mu} = \frac{n\bar{x}/\sigma^2 + \mu/b^2}{n/\sigma^2+1/b^2}, \quad \tilde{\sigma}^2 = \frac{1}{n/\sigma^2+1/b^2}.$$

Under squared error loss, the Bayes estimator is the posterior mean, which can be written as

$$\tilde{\mu} = \frac{n/\sigma^2}{n/\sigma^2+1/b^2}\bar{x} + \frac{1/b^2}{n/\sigma^2+1/b^2}\mu.$$

We note that, the posterior mean is just a convex combination of the sample mean and prior mean. We further note that, for small values of n, the Bayes estimator gives significant weight to the prior mean, but as $n \to \infty$, we have $|\bar{X} - \tilde{\mu}| \to 0$ a.s. irrespective of the hyper-parameters μ and b^2, i.e. the data overwhelms the prior. However, for finite values of n, we note that as the coefficient of $\bar{X}$ is less than 1, and $\mathbb{E}\bar{X} = \theta$, we note that $\tilde{\mu}$ is a biased estimator unless $\mu = \theta$.

The fact that the unbiased estimator $\bar{X}$ from the example was not the Bayes estimator is a special case of a more general result:

Theorem 1 (TPE 4.2.3)**.** *If δ is unbiased for $g(\theta)$ with $r(\Lambda, \delta) < \infty$ and $\mathbb{E}[g(\Theta)^2] < \infty$ then δ is not Bayes under squared error loss*[2] *unless its average risk is zero i.e.,*

$$\mathbb{E}[(\delta(X) - g(\Theta))^2] = 0,$$

where the expectation is taken over X and Θ.

Proof. Let δ be an unbiased Bayes estimator under squared error loss satisfying the assumptions of the theorem. Then, we know that δ is the mean of the posterior distribution, *i.e.*,

$$\delta(X) = \mathbb{E}[g(\Theta)|X] \text{ a.s.} \tag{8.1}$$

Thus we have that,

$$\mathbb{E}[\delta(X)g(\Theta)] = \mathbb{E}[\mathbb{E}[\delta(X)g(\Theta)|X]]{=}\mathbb{E}[\delta(X)\mathbb{E}[g(\Theta)|X]] \overset{(a)}{=} \mathbb{E}[\delta^2(X)], \tag{8.2}$$

where (a) follows by substituting for $\mathbb{E}[g(\Theta)|X]]$, using (8.1).

We also have that,

$$\mathbb{E}[\delta(X)g(\Theta)] = \mathbb{E}[\mathbb{E}[\delta(X)g(\Theta)|\Theta]] = \mathbb{E}[g(\Theta)\mathbb{E}[\delta(X)|\Theta]] \overset{(b)}{=} \mathbb{E}[g^2(\Theta)], \tag{8.3}$$

where (b) follows since $\mathbb{E}[\delta(X)|\Theta] = g(\Theta)$ because δ is an unbiased estimator of $g(\theta)$.

We have that the average risk under squared error,

$$\begin{aligned} \mathbb{E}[(\delta(X) - g(\Theta))^2] &= \mathbb{E}[\delta^2(X)] - 2\mathbb{E}[\delta(X)g(\Theta)] + \mathbb{E}[g^2(\Theta)], \\ &= \mathbb{E}[\delta^2(X)] - \mathbb{E}[\delta(X)g(\Theta)] + \mathbb{E}[g^2(\Theta)] - \mathbb{E}[\delta(X)g(\Theta)], \\ &\overset{(c)}{=} \mathbb{E}[\delta^2(X)] - \mathbb{E}[\delta^2(X)] + \mathbb{E}[g^2(\Theta)] - \mathbb{E}[g^2(\Theta)], \\ &= 0, \end{aligned}$$

where (c) follows, by substituting for $\mathbb{E}[\delta(X)g(\Theta)]$ once from (8.2) and once from (8.3).

Thus, we have that $\mathbb{E}[(\delta(X) - g(\Theta))^2] = 0$, i.e., the average risk is zero, proving the claim. □

The take-away message from this theorem is that (finite risk) squared error Bayes estimators are only unbiased in the degenerate case where perfect estimation is possible, that is, when the Bayes risk is 0. In fact, we can use this to check if a given unbiased estimator is a squared-error Bayes estimator.

[2] Note that this theorem only applies to Bayes estimators under the squared error loss. Bayes estimators under other losses may be unbiased without achieving 0 average risk.

Example 6. Let $X_1, ..., X_n \sim \mathcal{N}(\theta, \sigma^2)$, with $\sigma^2 > 0$ known. Is $\bar{X}$ Bayes under squared error for some choice of prior distribution? We know that, $\mathbb{E}(\bar{X}|\theta) = \theta$, i.e., $\bar{X}$ is an unbiased estimator of θ. Further, we have that the average risk under squared error,

$$\mathbb{E}[(\bar{X} - \Theta)^2] = \frac{\sigma^2}{n} \neq 0,$$

which means that $\bar{X}$ is not the Bayes estimator under any prior distribution!