CS731 Spring 2011 Advanced Artificial Intelligence

Statistical Decision Theory

Lecturer: Xiaojin Zhu *jerryzhu@cs.wisc.edu*

Consider a parameter $\theta \in \Theta$. We observe data x sampled from the distribution parametrized by θ. Let $\hat{\theta} \equiv \hat{\theta}(x)$ be an estimator of θ based on data x. We are going to compare different estimators.

Let a *loss function* $L(\theta, \hat{\theta}) : \Theta \times \Theta \mapsto \mathbb{R}_+$ be defined. For example,

$$L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2 \quad (1)$$

$$L(\theta, \hat{\theta}) = \begin{cases} 0 & \theta = \hat{\theta} \\ 1 & \theta \neq \hat{\theta} \end{cases} \quad (2)$$

$$L(\theta, \hat{\theta}) = \int p(x;\theta) \log \left(\frac{p(x;\theta)}{p(x;\hat{\theta})} \right) dx \quad (3)$$

The *risk* $R(\theta, \hat{\theta})$ is the average loss, averaged over training sets sampled from the true θ:

$$R(\theta, \hat{\theta}) = \mathbb{E}_\theta[L(\theta, \hat{\theta}(x))] = \int p(x;\theta) L(\theta, \hat{\theta}(x)) dx \quad (4)$$

Recall that $\mathbb{E}_\theta$ means the expectation over x drawn from the distribution with fixed parameter θ, not the expectation over different θ.

Example 1 *Let $X \sim N(\theta, 1)$. Let $\hat{\theta}_1 = X$ and $\hat{\theta}_2 = 3.14$. Assume squared error loss. Then $R(\theta, \hat{\theta}_1) = 1$ (hint: variance), $R(\theta, \hat{\theta}_2) = \mathbb{E}_\theta(\theta - 3.14)^2 = (\theta - 3.14)^2$. (hint: no X here) Over the whole range of possible $\theta \in \mathbb{R}$, neither estimator consistently dominates.*

Example 2 *Let $X_1, \ldots, X_n \sim Bernoulli(\theta)$. Consider squared error loss. Let $\hat{\theta}_1 = \frac{\sum X_i}{n}$, the sample mean. Let $\hat{\theta}_2 = \frac{\alpha + \sum X_i}{\alpha + \beta + n}$ which is the "smoothed" estimate, i.e., the posterior mean under a $Beta(\alpha, \beta)$ prior. Let $\hat{\theta}_3 = X_1$, the first sample. Then, $R(\theta, \hat{\theta}_1) = \mathbb{V}(\frac{\sum X_i}{n}) = \frac{\theta(1-\theta)}{n}$ and $R(\theta, \hat{\theta}_3) = \mathbb{V}(X_1) = \theta(1-\theta)$. So $\hat{\theta}_3$ is out as a learning algorithm. But what about $\hat{\theta}_2$?*

$$\begin{aligned} R(\theta, \hat{\theta}_2) &= \mathbb{E}_\theta(\theta - \hat{\theta}_2)^2 & (5) \\ &= \mathbb{V}_\theta(\hat{\theta}_2) + (bias(\hat{\theta}_2))^2 & (6) \\ &= \frac{n\theta(1-\theta)}{(n+\alpha+\beta)^2} + \left(\frac{n\theta + \alpha}{n+\alpha+\beta} - \theta \right)^2 & (7) \end{aligned}$$

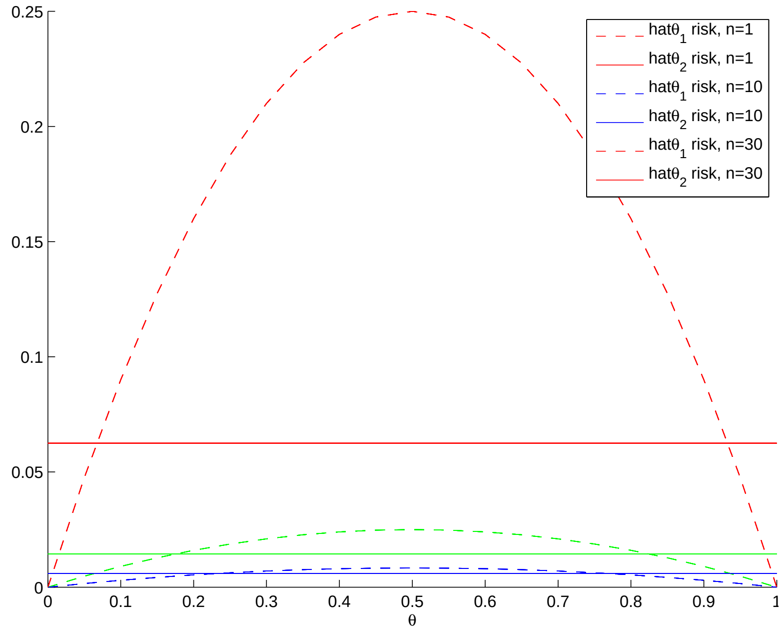
It is not difficult to show that one can make θ disappear from the risk (i.e., task insensitivity) by setting

$$\alpha = \beta = \sqrt{n}/2$$

with

$$R(\theta, \hat{\theta}_2) = \frac{1}{4(\sqrt{n}+1)^2}$$

It turns out this particular choice of α, β leads to a so-called minimax estimator $\hat{\theta}_2$, as we will show later. However, there is no dominance between $\hat{\theta}_1$ and $\hat{\theta}_2$ as the figure below shows:



The *maximum risk* is

$$R^{max}(\hat{\theta}) = \sup_{\theta} R(\theta, \hat{\theta}) \tag{8}$$

The *Bayes risk* under prior $f(\theta)$ is

$$R_f^{Bayes}(\hat{\theta}) = \int R(\theta, \hat{\theta}) f(\theta) d\theta. \tag{9}$$

Accordingly, two different criteria to define "the best estimator" (or the best learning algorithm) is the *Bayes rule* and the *minimax rule*, respectively. An estimator $\hat{\theta}^{Bayes}$ is a Bayes rule with respect to the prior f if

$$\hat{\theta}^{Bayes} = \arg\inf_{\hat{\theta}} \int R(\theta, \hat{\theta}) f(\theta) d\theta, \tag{10}$$

where the infimum is over all estimators $\hat{\theta}$. An estimator $\hat{\theta}^{minimax}$ that minimizes the maximum risk is a *minimax rule*:

$$\hat{\theta}^{minimax} = \arg\inf_{\hat{\theta}} \sup_{\theta} R(\theta, \hat{\theta}), \tag{11}$$

where again the infimum is over all estimators $\hat{\theta}$.

We list the following theorems without proof. For details see AoS p.197.

Theorem 1 *Let $f(\theta)$ be a prior, x a sample, and $f(\theta \mid x)$ the corresponding posterior. If $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ then the Bayes rule is the posterior mean:*

$$\hat{\theta}^{Bayes}(x) = \int \theta f(\theta \mid x) d\theta = \mathbb{E}(\theta \mid X = x). \tag{12}$$

If $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$ then the Bayes rule is the posterior median. If $L(\theta, \hat{\theta})$ is zero-one loss then the Bayes rule is the posterior mode.

Theorem 2 *Suppose that $\hat{\theta}$ is the Bayes rule with respect to some prior f. Suppose further that $\hat{\theta}$ has a constant risk: $R(\theta, \hat{\theta}) = c$ for all $\theta \in \Theta$. Then $\hat{\theta}$ is minimax.*

Example 3 *In example 2 we made the choice* $\alpha = \beta = \sqrt{n}/2$ *so that the risk* $R(\theta, \hat{\theta}_2) = \frac{1}{4(\sqrt{n}+1)^2}$ *is a constant. Also,* $\hat{\theta}_2$ *is the posterior mean and hence by Theorem 1 is a Bayes rule under the prior* $Beta(\sqrt{n}/2, \sqrt{n}/2)$*. Putting them together, by Theorem 2* $\hat{\theta}_2$ *is minimax.*