CS761 Spring 2013 Advanced Machine Learning

Statistical Decision Theory

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

Consider a parameter $\theta \in \Theta$.

★ *θ is the unobserved ground truth in machine learning. Θ is the hypothesis space (equivalently, the parameter space) that we consider.*

We observe data $D = (x_1, \ldots, x_n)$ sampled *i.i.d* from $p(x; \theta)$, the distribution parametrized by θ.
★ *D is a collection of iid random variables. Think of D as a particular training set.*

Let $\hat{\theta} \equiv \hat{\theta}(D)$ be an estimator of θ based on data D.
★ *$\hat{\theta}$ refers both to the procedure for producing an estimate from D (such as maximum likelihood or something else. This is the "learning algorithm" in machine learning), and the resulting value of this estimate (e.g., a vector for the mean). It should be clear from context which is which.*

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,

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

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

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

★ *The loss seems like the very basic "quality measure" between a learned parameter and the true parameter. However, keep in mind that $\hat{\theta}(D)$ is a random variable: it depends on the particular training set D. Thus $L(\theta, \hat{\theta})$ is a random variable, too.*

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

$$R(\theta, \hat{\theta}) = \mathbb{E}_D[L(\theta, \hat{\theta}(D))]. \tag{4}$$

$\mathbb{E}_D$ means the expectation over random training sets D.
★ *The risk is the "average training set" behavior of a learning algorithm when the world is θ. Trouble is, we don't know which θ the world is in.*

Example 1 *Let $D = X_1 \sim N(\theta, 1)$. Let $\hat{\theta}_1 = X_1$ and $\hat{\theta}_2 = 3.14$. Assume squared loss. Then $R(\theta, \hat{\theta}_1) = 1$ (hint: variance), $R(\theta, \hat{\theta}_2) = \mathbb{E}_D(\theta - 3.14)^2 = (\theta - 3.14)^2$. Over the range of possible $\theta \in \mathbb{R}$, neither estimator consistently dominates the other.*

★ *In machine learning terms, we have a smart learning algorithm $\hat{\theta}_1$ and a dumb one $\hat{\theta}_2$. However, we are saying that for some tasks $\theta \in (3.14 - 1, 3.14 + 1)$, the dumb algorithm is better.*

Example 2 *Let $X_1, \ldots, X_n \sim$ Bernoulli(θ). Consider squared 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" or regularized estimate, i.e., the posterior mean under a Beta(α, β)*

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 inadmissible as a learning algorithm (being dominated by* $\hat{\theta}_1$*). But what about* $\hat{\theta}_2$*?*

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

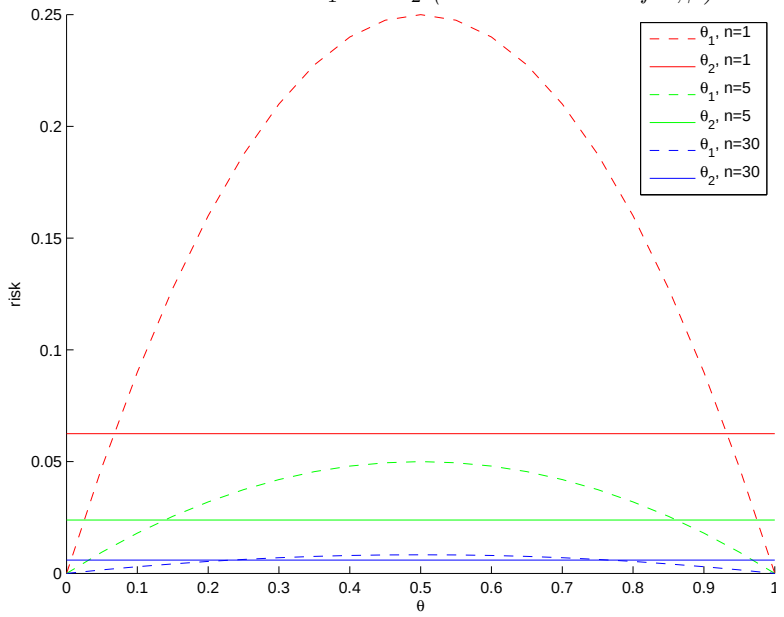
It is not difficult to show that one can make θ *disappear from the risk (i.e., task insensitivity) by setting*

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

with

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

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



★ *So, which learning algorithm (estimator) is better depends on your task (*θ*). This is definitely nasty and unsatisfactory. Like many before us, we will define-away the nastiness.*

The *maximum risk* is

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

★ *In example 2,* $R^{max}(\hat{\theta}_1) = 1/(4n) > R^{max}(\hat{\theta}_2)$*. However, as the figure shows, when* n *is large* θ_1 *is better except in a small region around* $\theta = 0.5$*.*

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

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

★ *There is no subjectivity in the maximum risk. There might be subjectivity in the Bayes risk depending on how the prior* f *over possible worlds is chosen.*

Accordingly, two different criteria to define "the best estimator" 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 $\hat{\theta}$.
★ *The Bayes rule does well in typical worlds overall. This requires a notation of what is typical, as defined by the prior f. It may not want to guard against a catastrophic world, as long as the chance of catastrophe is tiny.*

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 $\hat{\theta}$.
★ *The minimax rule is obsessed with guarding against the worst possible world.*

We list the following theorems without proof. For details see AoS p.197.
Theorem 1 *Let $f(\theta)$ be a prior, D a sample, and $f(\theta \mid D)$ the corresponding posterior. If $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ then the Bayes rule is the posterior mean:*

$$\hat{\theta}^{Bayes}(D) = \int \theta f(\theta \mid D) d\theta. \tag{12}$$

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.

★ *The Bayes rule is a point estimate, not a Bayesian approach.*

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.*