Lecture 2: Statistical Decision Theory (Part I)

Hao Helen Zhang

Spring, 2013

Outline of This Note

- Part I: Statistics Decision Theory
  - loss and risk
  - MSE and bias-variance tradeoff
  - Bayes risk and minimax risk
- Part II: Learning Theory for Supervised Learning
  - optimal learner
  - empirical risk minimization
  - restricted estimators

Statistical Inference

In statistical inference,

- we collect data $X_1, \cdots, X_n$, which follow the distribution $f(\mathbf{x}|\theta)$. Here $\theta \in \Theta$ is unknown parameter of interest;
- the goal of the inference is to estimate θ using the data.

Denote the estimator $\hat{\theta}(\mathbf{X})$, a function of data.

Three major types of inference:

- point estimator ("educated guess")
- confidence interval
- hypotheses testing

What is Statistical Decision Theory

Statistical decision theory is concerned with the problem of making decisions, in the presence of statistical knowledge which sheds light on the uncertainties involved in the problem.

- the uncertainties are presented by θ (scalar, vector, or matrix)

Examples:

- predicting the survival time of cancer patients
- deciding email or spam
- deciding whether the stock rate will rise or fall in a short term

Early works in decision theoy was extensively done by Wald (1950).

Loss Function

- Classical statistics is only directed towards the use of sampling information (data only) in making inferences about θ
- Decision theory combines the sampling information (data) with a knowledge of the consequences of our decisions.

A *loss function* is used to quantify the consequence that would be incurred for each possible decision for various possible values of θ.

$$L(\theta, \hat{\theta}(\mathbf{X})) : \quad \Theta \times \Theta \longrightarrow R.$$

This is known as *gains* or *utility* in economics and business. In decision theory, sometimes

- θ is called the *state of nature*, $\hat{\theta}(\mathbf{X})$ is called an *action*.

Examples of Loss Functions

- squared loss function: $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$
- absolute error loss: $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$
- L_p loss: $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|^p$
- 0-1 loss function: $L(\theta, \hat{\theta}) = I(\theta \neq \hat{\theta})$
- Kullback-Leibler loss: $L(\theta, \hat{\theta}) = \int \log \left(\frac{f(x|\theta)}{f(x|\hat{\theta})} \right) f(x|\theta) dx$

In general, we use a non-negative loss

$$L(\theta, \hat{\theta}) \geq 0, \quad \forall \theta, \hat{\theta}.$$

Risk Function

Intuitively, we prefer decision rules with small "expected (long-term average) loss" resulting from the use of $\hat{\theta}(\mathbf{X})$ repeatedly with varying $\mathbf{X}$. This leads to the *risk function* of a decision rule.

The **risk function** of an estimator $\hat{\theta}$ is

$$R(\theta, \hat{\theta}) = E_\theta[L(\theta, \hat{\theta})] = \int_{\mathcal{X}} L(\theta, \hat{\theta}(\mathbf{x})) f(\mathbf{x}|\theta) d\mathbf{x},$$

where $\mathcal{X}$ is the sample space (the set of possible outcomes) of $\mathbf{X}$.

Bias-Variance Decomposition of MSE

For the squared loss function, the risk is known as the *mean squared error* (MSE)

$$MSE = E_\theta\{[\theta - \hat{\theta}(\mathbf{X})]^2\}.$$

We show that MSE has the following decomposition:

$$\begin{aligned}
\text{MSE} &= \text{E}_\theta\{[\hat{\theta}(\mathbf{X}) - \theta]^2\} \\
&= \text{E}_\theta\{[\hat{\theta}(\mathbf{X}) - E_\theta(\hat{\theta}(\mathbf{X})) + E_\theta(\hat{\theta}(\mathbf{X})) - \theta]^2\} \\
&= \text{E}_\theta\{[\hat{\theta}(\mathbf{X}) - E_\theta(\hat{\theta}(\mathbf{X}))]^2\} + [E_\theta(\hat{\theta}(\mathbf{X})) - \theta]^2 \\
&= \text{Var}_\theta[\hat{\theta}(\mathbf{X})] + \text{Bias}_\theta^2[\hat{\theta}(\mathbf{X})].
\end{aligned}$$

This is known as bias-variance tradeoff.

Risk Comparison

How do we compare two estimators?

Given $\hat{\theta}_1$ and $\hat{\theta}_2$, if

$$R(\theta, \hat{\theta}_1) < R(\theta, \hat{\theta}_2), \quad \forall \theta \in \Theta,$$

we say $\hat{\theta}_1$ is the preferred estimator.

Ideally, we would like to use the decision rule $\hat{\theta}$ which minimizes the risk $R(\theta, \hat{\theta})$ for all values of θ. However,

- This problem has no solution, as it is possible to reduce the risk at a specific θ_0 to zero by making $\hat{\theta}$ equal to θ_0 for all $\mathbf{x}$.

Example 1

Let $X \sim N(\theta, 1)$. Consider two estimators:

- $\hat{\theta}_1 = X$
- $\hat{\theta}_2 = 3$.

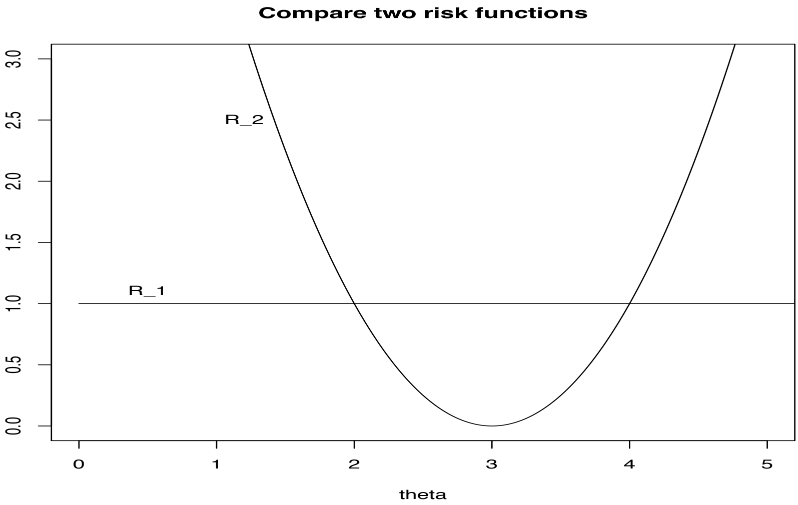
Using the squared error loss, direct computation gives

$$R(\theta, \hat{\theta}_1) = E_\theta(X - \theta)^2 = 1.$$
$$R(\theta, \hat{\theta}_2) = E_\theta(3 - \theta)^2 = (3 - \theta)^2.$$

Which has a smaller risk? Comparison:

- If $2 < \theta < 4$, then $R(\theta, \hat{\theta}_2) < R(\theta, \hat{\theta}_1)$,
- Otherwise, $R(\theta, \hat{\theta}_1) < R(\theta, \hat{\theta}_2)$.

Two risk functions cross. Neither estimator uniformly dominates the other.


Compare two risk functions
R_2
R_1
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
theta

Example 2: Binomial Risk

Let $X_1, \cdots, X_n \sim Bernoulli(p)$. Consider two estimators:

- $\hat{p}_1 = \bar{X}$ (Maximum Likelihood Estimator, MLE).
- $\hat{p}_2 = \frac{\sum_{i=1}^n X_i + \alpha}{\alpha + \beta + n}$ (Bayes estimator using a Beta(α, β) prior).

Using the squared error loss, direct calculation gives (Homework 1)

$$R(p, \hat{p}_1) = \frac{p(1-p)}{n}$$

$$R(p, \hat{p}_2) = \mathrm{V}_p(\hat{p}_2) + \mathrm{Bias}_p^2(\hat{p}_2) = \frac{np(1-p)}{(\alpha+\beta+n)^2} + \left(\frac{np+\alpha}{\alpha+\beta+n} - p\right)^2$$

Let $\alpha = \beta = \sqrt{n/4}$, we have

$$\hat{p}_2 = \frac{\sum_{i=1}^n X_i + \sqrt{n/4}}{n + \sqrt{n}}, \quad R(p, \hat{p}_2) = \frac{n}{4(n+\sqrt{n})^2}.$$

Best Decision Rule

In general, there exists no *uniformly best* estimator which simultaneously minimizes the risk for all values of θ. How to avoid this difficulty?

- One solution is to restrict the class of estimators by ruling out estimators that too strongly favor specific values of θ at the cost of neglecting other possible values.
- Commonly used classes of estimators:
  - Unbiased rules satisfy that $E_\theta[\hat{\theta}(\mathbf{X})] = \theta$.
  - Linear decision rules

BLUE (Best Linear Unbiased Estimator)

The data $(\mathbf{X}_i, Y_i)$ follows the model

$$Y_i = \sum_{j=1}^{K} \beta_j X_{ij} + \varepsilon_i, \quad i = 1, \cdots n,$$

- $\boldsymbol{\beta}$ is a vector of non-random unknown parameters, X_{ij} are "explanatory variables"
- ε_i's are random error terms following Gaussian-Markov assumptions: $E(\varepsilon_i) = 0, V(\varepsilon_i) = \sigma^2 < \infty$, and uncorrelated

The class of linear estimators consists of all $\widehat{\boldsymbol{\beta}}$ which is linear in Y.

Theorem: The ordinary least squares estimator (OLS) $\widehat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y}$ is best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$.

Maximum Risk and Bayes Risk

Alternatively, we can use a one-number summary of the risk function. Two cases:

- The **maximum risk** is

$$\bar{R}(\hat{\theta}) = \sup_{\theta \in \Theta} R(\theta, \hat{\theta}).$$

- The **Bayes risk** is

$$r_B(\pi, \hat{\theta}) = \int_{\Theta} R(\theta, \hat{\theta}) \pi(\theta) d\theta,$$

  where $\pi(\theta)$ is a prior for θ.

These two summaries suggest two different methods for deriving estimators: **Bayes** rule and **minimax** rule

Maximum Binomial Risk

Let $X_1, \cdots, X_n \sim Bernoulli(p)$. Under the squared error, we have

- $\hat{p}_1 = \bar{X}, \quad R(p, \hat{p}_1) = \frac{p(1-p)}{n}$.
- $\hat{p}_2 = \frac{\sum_{i=1}^n X_i + \sqrt{n/4}}{n+\sqrt{n}}, \quad R(p, \hat{p}_2) = \frac{n}{4(n+\sqrt{n})^2}$.

Compute the maximum risk

$$\begin{aligned} \bar{R}(\hat{p}_1) &= \max_{0 \le p \le 1} \frac{p(1-p)}{n} = \frac{1}{4n}. \\ \bar{R}(\hat{p}_2) &= \frac{n}{4(n+\sqrt{n})^2}. \end{aligned}$$

Based on the maximum risk, $\hat{\theta}_2$ is better than $\hat{\theta}_1$. However,

- When n is large, $R(p, \hat{p}_1)$ is smaller than $R(p, \hat{p}_2)$ except for a small region near $p = 1/2$. Many people prefer $\hat{p}_1$ to $\hat{p}_2$.
- Considering the worst-case risk only can be conservative.

Bayes Risk for Binomial Example

Assume the prior for θ is $\pi(p) = 1$. Then

$$\begin{aligned} r_B(\pi, \hat{p}_1) &= \int_0^1 R(p, \hat{p}_1) dp = \int_0^1 \frac{p(1-p)}{n} dp = \frac{1}{6n}, \\ r_B(\pi, \hat{p}_2) &= \int_0^1 R(p, \hat{p}_2) dp = \frac{n}{4(n+\sqrt{n})^2}. \end{aligned}$$

For $n \geq 20$, $r_B(\pi, \hat{p}_2) > r_B(\pi, \hat{p}_1)$, so $\hat{p}_1$ is better in terms of Bayes risk.

- This answer depends on the choice of prior.

Bayes Rule

A decision rule that minimizes the Bayes risk is called a **Bayes rule**. Formally,

- $\hat{\theta}$ is a Bayes rule with respect to the prior π if

$$r_B(\pi, \hat{\theta}) = \inf_{\tilde{\theta}} r_B(\pi, \tilde{\theta}),$$

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

Posterior Risk

Assume that $\mathbf{X} \sim f(\mathbf{x}|\theta)$ and $\theta \sim \pi(\theta)$. The marginal distribution of $\mathbf{X}$ is

$$m(\mathbf{x}) = \int f(\mathbf{x}|\theta)\pi(\theta)d\theta.$$

From Bayes theorem, the posterior density of θ given $\mathbf{x}$ is

$$\begin{aligned}\pi(\theta|\mathbf{x}) &= \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})} \\ &= \frac{f(\mathbf{x}|\theta)\pi(\theta)}{\int f(\mathbf{x}|\theta)\pi(\theta)d\theta}\end{aligned}$$

For any estimator $\hat{\theta}$, define its **posterior risk**

$$r(\hat{\theta}|\mathbf{x}) = \int L(\theta, \hat{\theta}(x))\pi(\theta|\mathbf{x})d\theta.$$

Bayes Rule Construction

Theorem: The Bayes risk $r_B(\pi, \hat{\theta})$ satisfies

$$r_B(\pi, \hat{\theta}) = \int r(\hat{\theta}|\mathbf{x}) m(\mathbf{x}) d\mathbf{x}.$$

- The posterior risk is a function only of $\mathbf{x}$ not a function of θ.
- If we choose $\hat{\theta}(\mathbf{x})$ to minimize the posterior risk, then we will minimize the integrand at every $\mathbf{x}$, and thus we minimize the Bayes risk and obtain the Bayes estimator.

Bayes Rule for Particular Loss Functions

Theorem:

- If $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$, then the Bayes estimator is

$$\hat{\theta}(\mathbf{x}) = \int \theta \pi(\theta|\mathbf{x}) d\theta = E(\theta|\mathbf{X} = \mathbf{x}).$$

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

Example: Normal

Let $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$ where σ^2 is known. Suppose we use a $N(a, b)$ prior for μ. The Bayes estimator with respect to the squared error loss is the posterior mean, which is

$$\hat{\theta}(\mathbf{X}) = \frac{b^2}{b^2 + \sigma^2/n}\bar{X} + \frac{\sigma^2/n}{b^2 + \sigma^2/n}a.$$

Minimax Rule

A decision rule that minimizes the maximum risk is called a **minimax rule**. Formally,

- $\hat{\theta}$ is minimax if

$$\sup_{\theta \in \Theta} R(\theta, \hat{\theta}) = \inf_{\tilde{\theta}} \sup_{\theta \in \Theta} R(\theta, \tilde{\theta}),$$

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