10-716: Advanced Machine Learning **Spring 2019**

Lecture 1: January 15

Lecturer: Pradeep Ravikumar *Scribes: Amir Alavi, Zhaoqi Cheng, Mark Cheung*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

1.1 Decision Principles

1.1.1 Basics

1. State of nature: $\theta \in \Theta$, where θ is also called the parameter and Θ is called the parameter space (all possible states of nature).
2. Statistical investigation performed to obtain information about θ, the outcome is called X, whose distribution depends on θ)
3. Infer state of nature
4. Take action/decision (in general, assumptions needed) denoted by a and $\mathbb{A}$ is the action space.

In general, $\mathbb{A} \neq \Theta$. Loss is a function of the parameter and action space i.e., $L : \Theta \times \mathbb{A} \to \mathbb{R}$

1.1.2 3 Sources of info

- Model-based $X \sim P_\Theta()$
- Prior Knowledge of θ
- Loss/gain from taking action (based on estimation)

1.1.2.1 Examples

- **E1** A lady claims to be able to tell whether milk was added before tea (or vice versa). In all 10 trials, she correctly determines which was poured first.
- **E2** A drunk friend claims to be able to predict the outcome of a flip of a fair coin. In all 10 trials, he is correct.

State of nature: θ: probability person answering correctly
Null hypothesis (person guessing) H_0: $\Theta = \frac{1}{2}$
Assuming 50% chance of guessing $X \sim \text{Bin}(10, \theta)$, $P_{H_0}(x) = 2^{-10}$, we would reject the hypothesis that the person guesses correctly. In the first example, it is not quite clear what to conclude. In the second example, we might think that this is a lucky streak.

E3 Drug Company deciding whether to sell the drug θ_1 is the probability drug is effective. Action will be to sell or don't sell. Overestimating θ_1 can lead to lawsuits and underestimating it can lead to lower profits.

Let θ_2 be proportion of the market the drug will capture. Since θ_2 is a proportion, we can define the parameter space as $\Theta = \{\theta_2 : 0 \leq \theta_2 \leq 1\}$ (state of nature). Action space is $A = [0, 1]$ (estimate of θ_2). The loss function might be:

$$L(\theta_2, a) = \begin{cases} 2(a - \theta_2) & \text{if } a > \theta_2; \\ \theta_2 - a & \text{if } \theta_2 \geq a. \end{cases} \tag{1.1}$$

There could be prior information about θ_2 e.g., $\pi(\theta_2) = \text{uniform}(0.7, 1)$.

E4 Investor deciding whether to buy bonds. The parameter space is $\Theta = \{\theta_1, \theta_2\}$ where θ_1 denotes when bond defaults and θ_2 denotes when bond does not default. Action space is $\mathbb{A} = \{a_1, a_2\}$, where a_1 denotes buying the bond and a_2 denotes not buying the bond. (Parameter space is not equal to the action space) We can plot the loss function using a table (called a loss matrix)

	buy	don't buy
default	1000	-300
no default	-500	-300

Since a gain is a negative loss, we want the loss to be as negative as positive. Prior information can be written as $\pi(\text{default}) = \pi(\theta_1) = 0.1$

1.2 Principles for minimizing loss

1.2.1 Conditional Bayesian Principle

Let π^* be the uncertainty over $\theta \in \Theta$ i.e. the posterior after seeing the data. Under Conditional Bayesian Principle, we select action a that minimize the expectation over the loss under π^*

$$\rho(\pi^*, a) = \mathbb{E}_{\theta \sim \pi^*} L(\theta, a). \tag{1.2}$$

1.2.2 Frequentist Principles

Instead of the Bayesian thinking which is conditioning on X and averaging θ, frequentists fix $\theta \in \Theta$ and average over x.

The decision rule is a function mapping domain of samples to domain of actions

$$\delta : X \to \mathcal{A}. \tag{1.3}$$

Here, we assuming the decision rule is deterministic.

E5 Suppose a biased coin comes up heads with probability θ when tossed. When the coin is tossed for n times, the number of heads follows Binomial distribution

$$X \sim Bin(n, \theta).$$

In this case, the decision rule may be

$$\delta(x) = x/n.$$

Frequentists define the risk $R(\theta, \delta)$ as the loss of δ over repeated applications, that is

$$R(\theta, \delta) = \mathbb{E}_{X \sim P(\cdot|\theta)} L(\theta, \delta(x)) = \int_X L(\theta, \delta(x)) f(x \mid \theta) dx \tag{1.4}$$

where $(x \mid \theta)$ is the density function.

1.2.3 Comparison

There are some issues with the frequentist thinking. First, it doesn't provide a way to find the decision $\delta(x)$. Besides, θ is unknown. Hence, the frequentist principle isn't quite actionable. On the contrary, Bayesian thinking is more actionable since it directly gives the best action a.

1.3 Notions of optimality

Although we don't have actionable avenues for picking the decision rule, we can still define some "certificates" of optimality.

Definition 1.1 *δ_1 is R-better than δ_2 when $R(\theta, \delta_1) \leq R(\theta, \delta_2)$ for all $\theta \in \Theta$ with strict inequality for some θ.*

Definition 1.2 *δ_1 is R-equivalent to δ_2 when $R(\theta, \delta_1) = R(\theta, \delta_2)$ for any $\theta \in \Theta$.*

Definition 1.3 *A decision rule δ is admissible if there exists no R-better decision rule. A decision rule δ is inadmissible if there does exist an R-better decision rule.*

E6 Suppose that our parameter is a single real value $\theta \in \mathbb{R}$, our samples are drawn from a Gaussian centered around our parameter $X \sim \mathcal{N}(\theta, 1)$, and that we are using the squared loss:

$$L(\theta, a) = (\theta - a)^2$$

Let our decision rule simply be a constant multiple of our data:

$$\delta_c(x) = cx \qquad c \in [0, \infty]$$

Thus our risk can be written as a function of θ and c:

$$\begin{aligned}
R(\theta, \delta_c(X)) &= \mathbb{E}_x \left[L\left(\theta, \delta_c(X)\right) \right] \\
&= \mathbb{E}_x (\theta - cX)^2 \\
&= \mathbb{E}_x (\theta - cX + c\theta - c\theta)^2 \\
&= \mathbb{E}_x \left(c(\theta - X) + \theta(1 - c) \right)^2 \\
&= c^2 \mathbb{E}_x (\theta - X)^2 + 2c\theta(1 - c)\mathbb{E}_x[\theta - X] + \theta^2 (1 - c)^2 \\
&= c^2 + \theta^2 (1 - c)^2
\end{aligned}$$

Now, we can do a case analysis of R over different values of c:

- if $c > 1$

  $R(\theta, \delta_c) > 1$

  But then $R(\theta, \delta_1) = \mathbb{E}_x(\theta - X)^2 = 1$

  $\Rightarrow R(\theta, \delta_1) < R(\theta, \delta_c)$ for $c > 1$, and δ_c is inadmissible!

- if $c \leq 1$

  Consider $c = 0$, $R(\theta, \delta_0) = \theta^2$

  Now we cannot say that a particular decision rule (restricted to $c \leq 1$) is strictly better than the other as we vary over the states of nature θ. In fact, for $c \leq 1$, all decision rules are admissible! This situation is depicted in Figure 1.1, where we see that none of the risk functions dominate the others for all states of nature.

Here we see a rather funny scenario where a basic estimator that just agrees with the data (δ_1) is admissible, but so is the clearly awful estimator of δ_0 which always estimates zero. Even though both of these estimators are admissible, we would have a clear preference for δ_1. This illustrates that just being admissible may not good enough, both from an actionable perspective (we don't have a clear procedure to produce admissible estimators on demand), and from a performance perspective.

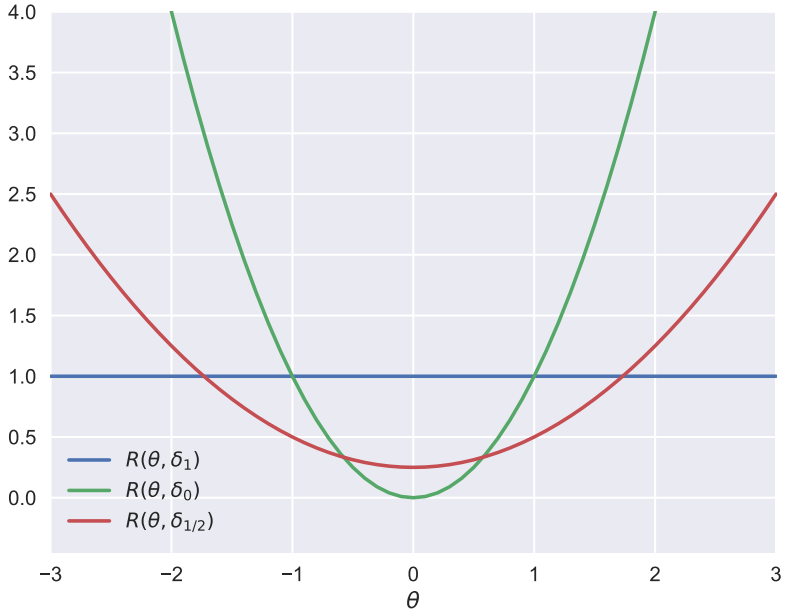

Figure 1.1: Risks for various decision rules for Example 6

1.3.1 Bayes Risk Principle

In order to talk about optimality, we must vary over the states of nature. One way of doing this is to take a prior distribution over θ, and then average over θ (take the expectation with respect to the prior distribution

on θ). This is called the Bayes risk. The Bayes risk of a decision rule δ, with respect to a prior distribution π on Θ, is defined as

$$r(\pi, \delta) = \mathbb{E}_{\theta \sim \pi} R(\theta.\delta) \tag{1.5}$$

Bayes decision rule minimize the risk by choosing the decision

$$\delta^{\pi} \in \arg\inf r(\pi, \delta). \tag{1.6}$$

1.4 Randomization

1.4.1 Randomized rules

A randomized decision rule $\delta^*(x, A)$ is the probability that an action in A (a subset of $\mathcal{A}$) will be chosen if x is observed. The loss function $L(\theta, \delta^*(x, \cdot))$ of the randomized rule δ^* is defined as

$$L(\theta, \delta^*(x, \cdot)) = \mathbb{E}_{a \sim \delta^*(x,\cdot)} L(\theta, a). \tag{1.7}$$

E7 (Mathching Pennies). You and your opponent are to simulaneously uncover a penny. If the two coins match, you win \$1 from your opponent. If the coins don't match, your opponent win \$1 from you. The actions which are available to you are a_1–choose heads, or a_2–choose tails. The possible states of nature are θ_1–the opponent's coin is a head, and θ_2–the opponent's coin is a tail. The loss matrix is

	a_1	a_2
θ_1	-1	1
θ_2	1	-1

The only certain way of preventing ultimate defeat is to choose a_1 and a_2 by some random mechanism. A natural way to do this is simply to choose a_1 and a_2 with probabilities, which is an example of randomized decision rule.

E8 Assume that a random variable $X \sim Bin(n, \theta)$ is observed, and that we want to choose between action a_0–accepting $\theta = \theta_0$ versus action a_1–accepting $\theta = \theta_1$ (where $\theta_0 > \theta_1$.

We can consider the randomized rules given by

$$\delta_j^*(x, a_1) = \begin{cases} 0 & \text{if } x > j, \\ p & \text{if } x = j, \\ 1 & \text{if } x < j, \end{cases}$$

and

$$\delta_j^*(x, a_0) = 1 - \delta_j^*(x, a_1).$$

Thus, if $x > j$ is observed, we will always pick θ_0. If $x = j$ is observed, a randomization will be performed, picking θ_0 with probability p and picking θ_1 with probability $1 - p$. Through proper choice of j and p, a most powerful test of the above form can be found for any given size α.

Let the loss function be

$$L(\theta_i, a_j) = \mathbb{I}(i \neq j).$$

The risk will be

$$R(\theta_0, \delta_j) = P_{\theta_0}(x < j) + pP_{\theta_1}(x = j).$$