DS 102 Data, Inference, and Decisions **Spring 2020**

Lecture 2: Frequentist and Bayesian Decision-Making

Lecturer: Moritz Hardt

1 Elements of Decision Theory

Statistical decision theory provides a framework for decision-making problems where there is uncertainty, and encompasses both Bayesian and frequentists ideas.

We assume that a pair (X, θ) of data X and paramater θ is generated from some joint distribution $\mathbb{P}(X, \theta)$, denoted $(X, \theta) \sim \mathbb{P}(X, \theta)$. We think of θ typically as some ground-truth parameter, state of reality, or label.

Example 1.1. Some possible (X, θ) pairs include:

- X is blood pressure of a patient and $\theta \in \{0, 1\}$ represents whether a patient is sick ($\theta = 1$) or not sick ($\theta = 0$).
- X is an image and $\theta \in \{\text{cat},\ \text{dog}\}$.
- X is annual temperature of town and θ is its altitude.

We define a **decision procedure** $\delta(X)$ that operates on the data to make a decision. Typically, $\delta(X)$ is trying to "guess" θ from X. To measure how good the guess of our procedure is, we take a loss function $\ell(\theta, \delta(X))$, which takes as input the ground-truth parameter, as well as our prediction.

Example 1.2. Some possible loss values include:

- zero/one loss $\ell(\theta, \delta(X)) = \mathbb{1}\{\theta \neq \delta(X)\}$
- squared loss $\ell(\theta, \delta(X)) = \frac{1}{2}(\theta - \delta(X))^2$

Many types of familiar statistical inference problems – including hypothesis testing and parameter estimation – fit into this framework for appropriate choices of θ, δ, and ℓ.

When we design our inference procedure $\delta(\cdot)$, both arguments of the loss function are unknown; θ because we don't know the ground truth, and $\delta(X)$ because we still don't have the data. The treatment of uncertainty regarding θ and X differs between Bayesian and frequentist approaches.

The frequentist approach assumes θ is deterministic, however unknown. In this view, we incur some average error for each of the possible "realities", where the average is taken over the randomness in X, given θ. The frequentist **risk** is defined as:

$$r(\theta) = \mathbb{E}_{X \sim \mathbb{P}(X|\theta)}[\ell(\theta, \delta(X))],$$

where the randomness in the argument of the expectation comes *only* from X.

In the Bayesian view, the parameter θ is also random, and we are interested in the average error given the data X. Here, the average is taken over the randomness in θ. This gives the Bayesian **posterior risk**:

$$\rho(X) = \mathbb{E}_{\theta\sim\mathbb{P}(\theta|X)}[\ell(\theta, \delta(X))].$$

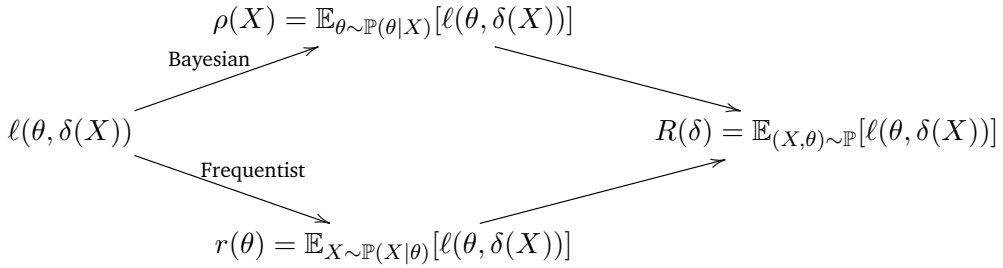
If we accept the Bayesian view of making θ random, then averaging the frequentist risk over θ, and averaging the Bayesian posterior risk over X gives the same number by Fubini's theorem:

$$\mathbb{E}_{\theta\sim\mathbb{P}(\theta|X)}[r(\theta)] = \mathbb{E}_{X\sim\mathbb{P}(X|\theta)}[\rho(X)].$$

The resulting quantity is called the **Bayes risk**

$$R(\delta) = \mathbb{E}_{(X,\theta)\sim\mathbb{P}}[\ell(\theta, \delta(X))].$$

We can summarize the situation in the following diagram.



1.1 Comments on Bayesian and Frequentist Thinking

In statistics, Bayesian and frequentist thinking comprise two major approaches to making decisions. Here we illustrate some of the key differences between the two.

In frequentism, we want to design a procedure that works "on average", or with high probability. We assume that we don't have just one data set, but rather we repeatedly draw data sets independently from a population.

For example, Frequentist hypothesis testing is captured by the Neyman-Pearson framework. There is an unknown ground truth which says whether the hypothesis is null (0) or non-null (1), and for a given significance level α (typically 0.05), the test should guarantee that a discovery is proclaimed with probability at most α, if the ground-truth reality is 0. This setting requires knowing what data we expect to see, if the reality is 0 or 1, respectively.

In Bayesian hypothesis testing, one additionally specifies with what probability the null and non-null occur. This allows us to compute the probability that a hypothesis is 0 (or 1) given the data. We declare a discovery if $\mathbb{P}(H = 0 \mid \text{data})$ is small enough.

Therefore, the frequentist perspective is an unconditional one, requiring inferential procedures to give good answers in repeated use. The Bayesian perspective is a conditional one, making inferences conditional on observed data, as opposed to all possible data one could have observed.

The Bayesian perspective is natural in the setting of a long-term project with a domain expert. For example, if one has a limited data set and works on a multi-year project with a biologist who has good prior knowledge biological phenomena, Bayesian thinking is a sensible approach. The frequentist approach is more "robust", and as such is natural in settings where we develop procedures that will be used repeatedly, in many different, possibly unexpected settings. One example is writing software that will be used by many people for various problems.

2 Risk Minimization

One common goal is to choose a decision rule $\delta(\cdot)$ which minimizes the Bayes risk $R(\delta)$. That is, we'd like to solve the minimization problem

$$\min_{\delta} R(\delta) = \min_{\delta} \mathbb{E}_{(X,\theta)\sim\mathbb{P}}[\ell(\theta, \delta(X))].$$

To solve this problem, we can think of sampling the pair (X, θ) in two steps:

1. Sample X from $\mathbb{P}(X)$, the marginal distribution of $\mathbb{P}$ on X
2. Sample $\theta \sim \mathbb{P}(\theta|X)$ from the posterior

Then, we can minimize the expectation, $\mathbb{E}_{(X,\theta)\sim\mathbb{P}}[\ell(\theta, \delta(X))]$, by minimizing it pairwise. The $\delta(\cdot)$ which solves this minimization problem is also called the Bayes decision rule or Bayes optimal rule.

Example 1.3. For the squared loss, we can solve the minimization problem

$$\min_{\delta} \mathbb{E}_{(X,\theta)\sim\mathbb{P}} \left[\frac{1}{2}(\theta - \delta(X))^2\right] = \min_{\delta} \mathbb{E}_{X\sim\mathbb{P}(X)} \left[\mathbb{E}_{\theta\sim\mathbb{P}(\theta|X)} \left[\frac{1}{2}(\theta - \delta(X))^2\right]\right]$$

by minimizing it pointwise. If we take $a = \delta(X)$, then by differentiating with respect to a and setting equal to zero,

$$\nabla f(a) = \mathbb{E}_{(X,\theta)\sim\mathbb{P}}[\theta - a] = 0 \implies \mathbb{E}_{(X,\theta)\sim\mathbb{P}}[\theta] = \mathbb{E}_{(X,\theta)\sim\mathbb{P}}[a].$$

Here, we applied linearity of expectation. Then, by plugging in for a,

$$\mathbb{E}_{X\sim\mathbb{P}(X)}[\delta(X)] = \mathbb{E}_{X\sim\mathbb{P}(X)}[\mathbb{E}_{\theta\sim\mathbb{P}(\theta|X)}[\theta]] \implies \delta(X) = \mathbb{E}_{\theta\sim\mathbb{P}(\theta|X)}[\theta]$$

by the definition of conditional expectation.