STATS 300A - Lecture 1

Michael Howes

October 27, 2021

1 The Big Picture

Sciences/engineered systems → data → statistics → inference.

Core questions:

(a) Modelling: how to trnslate domain question to a statistical or mathematical question framework.

(b) Methodology: what tools can I use? Which are appropriate?

(c) Analysis: how can I compare methods? How can we construct optimal methods?

Statistics is about limited resources. Things would be easy if we had enough time and money to collect unlimited data.

This course is primarily focus on (c) but we will touch on (a) and (b). The courses STATS305 A/B/C will teach more about (a) and (b). STATS300 A/B/C focuses on optimality in all its forms.

2 Decision Theory

Our framework will include the following.

(a) The *data* is a realisation of a r.v. (random variable) $X \in \mathcal{X}$. The set $\mathcal{X}$ is called the *sample space*. Often our data will take the form $X = (X_1, X_2, \ldots, X_n)$ where $X_i \in \mathbb{R}^d$ and each X_i is iid (independent and indentically distributed). When we have iid data, n will be called the *sample size*.

(b) The *statistical model* which is a family of probability distributions

$$\mathcal{P} = \{P_\theta : \theta \in \Omega\}.$$

We know that $X \sim P_\theta$ (X is distributed according to P_θ) for some $\theta \in \Omega$ but we do not know the exact value of θ. The set Ω is called the *parameter space* and θ is called a *parameter of interest*. The set Ω is may be finite dimensional (linear regression with Gaussian noise) or infinite dimensional (non-parametric methods).

Here is an example that we will return to at several points during this lecture. Our data is $X = (X_i)_{i=1}^n$ where X_i are iid and $X_i \sim \text{Bern}(\theta)$. Thus (X_i) are independent and $P_\theta(X_i = 1) = \theta$ and $P_\theta(X_i = 0) = 1 - \theta$. Here the sample space is $\{0,1\}^n$ and the parameter space is $[0,1]$. Each X_i is a *Bernoulli random variable* and models a binary outcome with θ chance of success.

(c) A *decision procedure* δ is a function from $\mathcal{X}$ to the *decision space* D. That is, δ takes some data $X \in \mathcal{X}$ to a decision $d \in D$.

Let's continue our Bernoulli example. If we want to estimate θ, then our decision space will be $D = \Omega = [0,1]$. An example of a decision procedure is the sample average $\delta(X) = \frac{1}{n}\sum_{i=1}^n X_i$.

We may be interested in something other than estimating the exact value of θ. For example, we may wish to test the hypothesis $\theta > \frac{1}{2}$. In this case we will have $D = \{\text{accept}, \text{reject}\}$. Decision theory thus provides a common framework for both estimation and hypothesis testing.

(d) The *loss function* is a function $L : \Omega \times D \to [0, \infty)$

$$L(\theta, d) = \text{The penalty incurred if } \theta \text{ is the true parameter and we take decision } d.$$

If Ω and D are both subsets of $\mathbb{R}$, then a common chouce of loss function is $L(\theta, d) = (\theta - d)^2$. This common loss function is called the square error loss.

(e) The *risk function* of a decision procedure δ is defined as

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

The notation $\mathbb{E}_\theta[f(X)]$ means that $X \sim P_\theta$ and we take the expectation of $f(X)$ with respect to P_θ. The risk is the average loss of choosing the decision procedure δ when θ is the true parameter. Note that the risk is a function of the unknown parameter θ.

Within this framework our goal is to compare decision procedures by comparing their respective risk functions. One way of comparing risk procedures is by using inadmissibility.

Definition 1. A decision procedure δ is *inadmissible* if there exists another decision procedure δ' such that

(a) For all $\theta \in \Omega$, $R(\theta, \delta') \leq R(\theta, \delta)$ and,

(b) There exists $\theta \in \Omega$ such that $R(\theta, \delta') < R(\theta, \delta)$.

In the case when a decision procedure is inadmissible we have no reason for using it. The procedure δ' is guaranteed to be at least as good as δ and sometimes it is strictly better. Unfortunately we are rarely in such a clear cut situation.

Consider again our Bernoulli example. For each k we have the estimator $\delta_k(X) = \frac{1}{k}\sum_{i=1}^k X_i$. We can study the square error loss $L(\theta, d) = (\theta - d)^2$. Note that $\mathbb{E}_\theta[\delta_k(X)] = \theta$ and thus $R(\theta, \delta_k) = \mathbb{E}_\theta[(\theta - \delta_k(X))^2]$ is equal to the variance of $\delta_k(X)$ which is $\frac{\theta(1-\theta)}{k}$. Thus for $k = 1, 2, \ldots, n-1$, the procedure δ_k is inadmissible since δ_n always has a lower risk (and if $\theta \neq 0, 1$, then the risk is strictly lower).

Consider now the constant estimator $\delta'(X) = \frac{1}{2}$. Then $(\theta - \delta(X))^2$ is a constant and hence $R(\theta, \delta') = \left(\theta - \frac{1}{2}\right)^2$. For values of θ close to $\frac{1}{2}$, δ' has a lower risk than δ_n. For other values of θ, δ_n has a lower risk than δ'.

The fact that the risk function is a function of θ means that it is difficult to compare two decision procedures by comparing their risk functions. The decision functions will often cross as in the example above. To make the comparison of decision functions more tractable, two board approaches are used. They are:

(a) Adding constraints on δ. For example:

i. We may decide to restrict our attention to *unbiased* decision procedures. For a fixed function g we may restrict our attention to decision procedures δ that satisfy $\mathbb{E}_\theta[\delta(X)] = g(\theta)$. Such a procedure is called *unbiased* for $g(\theta)$.

ii. We may restrict our attention to estimators that are invariant under certain symmetries of our problem. For example we may ask that $f(X + c) = f(X) + c$ for all $X \in \mathcal{X}$ and $c \in \mathbb{R}$. [Note: X will often be a matrix or a vector. Thus for $c \in \mathbb{R}$, the notation $X + c$ will be interpreted as adding c to every entry of X.]

(b) The second approach is to collapse $R(\theta, \delta)$ into a one dimensional quantity (as opposed to a function of θ). For example:

   i. A Bayes procedure minimizes
   $$\int_{\Omega} R(\theta, \delta) d\Lambda(\theta),$$
   where Λ is a probability distribution on Ω. Such a distribution is called a *prior distribution.*

   ii. A minimax procedure minimizes
   $$\sup_{\theta \in \Omega} R(\theta, \delta).$$
   By taking a supremum we are looking at the worst case scenario given a fixed δ. We will see a relationship between i. and ii. when we study "least favourable priors."

3 Sufficient statistics

Our goal is to investigate which parts of the data can be discarded.

Definition 2. A statistic $T(X)$ (T is a function of the sample X) is sufficient for the model $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ is for all t, the conditional distribution of $X|T = t$ does not depend on θ.

Consider again $X_n \sim \text{Bern}(\theta)$, X_i iid. The statistic $T(X) = \sum_{i=1}^n X_i$, is sufficient. To see why, note that

$$\begin{aligned} P_\theta(X = x|T(X) = t) &= \frac{P_\theta(X = x, T(X) = t)}{P_\theta(T(X) = t)} \\ &= \begin{cases} \frac{P_\theta(X=x)}{P_\theta(T=t)} & \text{if } \sum_{i=1}^n x_i = t, \\ 0 & else. \end{cases} \\ &= \begin{cases} \frac{\prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} & \text{if } \sum_{i=1}^n x_i = t, \\ 0 & else. \end{cases} \\ &= \begin{cases} \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} & \text{if } \sum_{i=1}^n x_i = t, \\ 0 & else. \end{cases} \\ &= \begin{cases} \frac{1}{\binom{n}{t}} & \text{if } \sum_{i=1}^n x_i = t, \\ 0 & else. \end{cases} \end{aligned}$$

Since this quantity does not depend on θ, the statistic $T(X)$ must be sufficient. In the above calculation we used the fact that the sum of iid Bernoulli random variables has a binomial distirbution.

The relationship between sufficency and data reduction may be unclear but the following Theorem shows we can restrict our attention to sufficient statistics. This is Theorem 6.1 in TPE.

Theorem 1. *If $X \sim P_\theta \in \mathcal{P}$ and T is sufficient, then for any decision procedure δ, there exists a randomised decision procedure $\widetilde{\delta}$ that depends only on $T(X)$ and a uniform random variable U that is independent of X such that $R(\theta, \delta) = R(\theta, \widetilde{\delta})$ for all $\theta \in \Omega$.*

Proof. Since $X|T = t$ does not depend on θ, we can use U to draw a random variable X' from the distribution $X|T = t$. Thus we can define

$$\widetilde{\delta}(T, U) = \delta(X').$$

It can be shown that the joint distribution of (X', T) and (X, T) are the same. Thus δ and $\widetilde{\delta}$ have the same risk function. □

The important step in this agrument is the fact that $X|T = t$ does not depend on θ. This allows us to construct X' without knowing what θ is. The details of how to use U to sample from an arbitrary distribution are omitted from this proof but can be found in the textbook.

The following theorem provides a method for finding sufficient statistics. It is called the Neyman-Fisher Factorisation Condition (NFFC). It is on page 15 of TSH.

Theorem 2. *Suppose each P_θ has a density $p(x|\theta)$ (wrt a common measure μ), then $T(X)$ is sufficient for X if and only if*

$$p(x|\theta) = g_\theta(T(x))h(x),$$

for all $x \in X$ and $\theta \in \Omega$.

Proof. To be done next lecture. □

Note that the factorisation is not unique. Indeed for any $c > 0$ we could set $g'_\theta(t) = cg_\theta(t)$ and $h'(x) = c^{-1}h(x)$.

Exercise 1. Suppose $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ are iid, that is

$$p(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \prod_{i=1}^{n} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}.$$

Find a two dimensional sufficient statistic for X where the parameter θ is equal to (μ, σ^2).