In this lecture\textsuperscript{1}, we will introduce some of the basic concepts of statistical decision theory, which will play crucial roles throughout the course.

1 Motivation

The process of inductive inference consists of

1. Observing a phenomenon
2. Constructing a model of that phenomenon
3. Making predictions using this model.

The main idea behind learning algorithms is to generalize past observations to make future predictions. Of course, given a set of observations, one can always find a function to exactly fit the observed data, or to find a probability measure that generates the observed data with probability one. However, without placing any restrictions on how the future is related to the past, the No Free Lunch theorem essentially says that it is impossible to generalize to new data. In other words, data alone cannot replace knowledge. Hence, a central theme of this course is

\[ \text{Generalization = Data + Knowledge.} \]

We can view statistical decision theory and statistical learning theory as different ways of incorporating knowledge into a problem in order to ensure generalization.

2 Decision Theory

2.1 Basic Setup

The basic setup in statistical decision theory is as follows: We have an outcome space \( X \) and a class of probability measures \( \{P_\theta : \theta \in \Theta\} \), and observations \( X \sim P_\theta, X \in X \). An example would be the Gaussian measure, defined in the above framework as follows:

\[
\frac{dP_\theta}{dx} = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}; \quad \theta = (\mu, \sigma^2).
\]

Our goal as a statistician is to estimate \( g(\theta) \), where \( g \) is an arbitrary function that is known to us and \( \theta \) is unknown. Thus, observe that our goal is more general than simply estimating \( \theta \), and the generality of our goal is motivated by the fact that the parametrization \( \theta \) can be arbitrary.

A useful perspective to view the goal of estimating \( g(\theta) \) is as a game between Nature and the statistician, where Nature chooses \( \theta \) and the statistician chooses a decision rule \( \delta(X) \). Note that a general decision rule may be randomized, i.e., for any realization of \( X = x \), \( \delta(x) \) produces an action \( a \in A \) following the probability

\textsuperscript{1}Some readings:
3. Abraham Wald (1949) “Statistical Decision Functions”.
distribution $D(da|x)$. Intuitively, after observing $x$, the decision rule chooses the action $a \in A$ randomly from the distribution $D(da|x)$. As a special case, for a deterministic decision rule $\delta(x)$, $D(da|x)$ is just a one-point probability mass at $\delta(x)$.

Moreover, we have a loss function $L(\theta, a) : \Theta \times A \to \mathbb{R}$ to quantify the loss between $\theta$ and any action $a \in A$, such that (usually we assume that $g(\theta)$ belongs to the action space $A$)

$$L(\theta, a) \geq 0, \forall a \in A, \theta \in \Theta$$  \hspace{1cm} (1)

$$L(\theta, g(\theta)) = 0, \forall \theta \in \Theta.$$  \hspace{1cm} (2)

Immediately, we see that our loss function $L(\theta, \delta(X))$ is inherently a random quantity. Note that we may have two sources of randomness: the observation $X$ is inherently random, and for a fixed $X$ the action produced by the decision rule $\delta(X) \sim D(da|X)$ may also be random.

Instead, we will define a risk function $R(\theta, \delta)$ in terms of our loss function, which will instead be deterministic.

**Definition 1 (Risk).** For general randomized decision rule $\delta(X)$:

$$R(\theta, \delta) \triangleq \mathbb{E}_\theta[L(\theta, \delta(X))] = \int\int L(\theta, a)D(da|x)dP_\theta(x). \hspace{1cm} (3)$$

When the decision rule $\delta(X)$ is deterministic, we have $R(\theta, \delta) = \int L(\theta, \delta(x))dP_\theta(x)$.

Intuitively, the risk $R(\theta, \delta)$ is the average loss incurred by using the decision procedure $\delta$ over many draws of data from $P_\theta$.

**2.2 Comparing Procedures**

Decision theory offers a principled way of analyzing the performance of different statistical procedures by evaluating and comparing their respective risks $R(\theta, \delta)$. For example, suppose we observe data $X \sim P_\theta$. Should we use maximum likelihood, method of moments, or some other procedure to estimate $g(\theta)$? Decision theory allows us to rule out certain inadmissible procedures.

**Definition 2 (Admissibility).** A decision rule $\delta(X)$ is called “inadmissible” if there exists some decision rule $\delta'(X)$ such that

1. $R(\theta, \delta') \leq R(\theta, \delta)$ for all $\theta \in \Theta$, and
2. there exists $\theta_0 \in \Theta$ such that $R(\theta_0, \delta') < R(\theta_0, \delta)$.

Such a rule $\delta'$ is said to dominate $\delta$. If there does not exist any rule $\delta'(X)$ which dominates $\delta(X)$, $\delta(X)$ is called “admissible”.

Having ruled out inadmissible decision rules, one might seek to find the best rule, namely the rule that uniformly minimizes the risk. However, in most cases of interest, there is no such uniformly best decision rule. For instance, if Nature chooses $\theta_0$, then the trivial procedure $\delta(X) \triangleq g(\theta_0)$ achieves the minimum possible risk $R(\theta_0, \delta(X)) = 0$. Thus, in most cases $\delta$ is admissible. However, such a decision rule is intuitively unreasonable and, further, can have arbitrarily high risk for $\theta \neq \theta_0$.

There are two main approaches to restrict the order of the game to avoid this triviality, namely, we can restrict the statistician, or we can assume more knowledge about Nature.

**2.3 Restrict the Statistician**

Let $D$ be the set of all possible decision rules. One way to restrict the statistician is to force her to choose a decision rule from some subset $D' \subseteq D$. How should we choose $D'$? Some possibilities are:
1. Take $D'$ to be the set of unbiased decisions, namely, the decisions $\delta(X)$ such that $E_{\theta}[\delta(X)] = g(\theta)$, $\forall \theta \in \Theta$.

2. Equivariance (which we will not discuss as much in class): As an example, if $P_\theta = f(X - \theta)$, where $\theta \in \mathbb{R}$, $f(X) \triangleq \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, and our observations $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_\theta$, then we will choose our decision function $\delta$ to be such that $\delta(X_1 + c, X_2 + c, \ldots, X_n + c) = \delta(X_1, \ldots, X_n) + c$, where $c$ is a constant. The reason for why we might want to choose such a decision function $\delta$ is because, for equivariant estimators, the variance of the estimator does not depend on $\theta$.

2.4 Assume more knowledge about Nature

The downside of restricting the statistician is that we may be ignoring good decision rules $\delta$. Instead, an alternative approach is to assume more knowledge about Nature. Here, we run into the classic Bayesian vs. Frequentist debate, which can be viewed as two different ways of assuming more knowledge about Nature.

1. Bayes: We assume that $\theta$ is drawn from a probability density, namely, $\theta \sim \lambda(\theta)$. For instance, for the problem of disease detection, our prior $\lambda(\theta)$ could be the population disease density.

Thus, in the Bayesian setting, our new risk function becomes

$$r(\delta) = \int R(\theta, \delta) \lambda(\theta) d\theta.$$ (4)

As a result, to minimize the risk, the optimization problem becomes:

$$\min_\delta r(\delta),$$ (5)

which is a well-posed optimization problem only over $\delta$ (and can be solved in many cases, though it can be computationally intractable to compute).

It is known that if $\theta$ is finite dimensional, our observations $X_1, \ldots, X_n \sim P_\theta$, and $\lambda(\theta)$ is supported everywhere, then as $n \to \infty$, $\lambda(\theta)$ has no effect on inference asymptotically. However, there are a couple catches here to notice. The first catch is that here we are assuming $n \to \infty$, but if $n$ is finite, then $\lambda(\theta)$ will influence the decision rule rather significantly. The second catch is that we are assuming $\theta$ is finite dimensional, which does not hold generally in semiparametric and nonparametric models, e.g., if we are trying to estimate a function.

2. Frequentist: The frequentist approach is to not impose a prior, which leads to a natural choice to guard against the worst case risk using a minimax approach. Namely, the risk in the minimax setting becomes

$$r(\delta) = \max_\theta R(\theta, \delta),$$ (6)

and our optimization problem is as usual,

$$\min_\delta r(\delta).$$ (7)

3 Data Reduction

Not all data is relevant to a particular decision problem. Indeed, the irrelevant data can be discarded and replaced with some statistic $T(X^n)$ of the data without hurting performance. We make this precise below.
3.1 Sufficient Statistics

**Definition 3** (Markov chain). Random variables \(X, Y, Z\) are said to form a Markov Chain if \(X\) and \(Z\) are conditionally independent given \(Y\). In particular, the joint distribution can be written as

\[
p(X, Y, Z) = p(X)p(Y | X)p(Z | Y) = p(Z)p(Y | Z)p(X | Y).
\]

**Definition 4** (Sufficiency). A statistic \(T(X)\) is “sufficient” for the model \(P = \{P_\theta : \theta \in \Theta\}\) if and only if the following Markov chains hold for any distribution on \(\theta\):

1. \(\theta - T(X) - X\)
2. \(\theta - X - T(X)\)

One useful interpretation of the first condition is, if we know \(T(X)\), we can generate \(X\) without knowing \(\theta\) (because \(X\) and \(\theta\) are independent conditioned on \(T(X)\)). The second condition usually trivially holds, since in most cases \(T(X)\) is a deterministic function of \(X\). However, a subtle but important point about the second condition is that the statistic \(T(X)\) could be a random function of \(X\). In this case, the second condition implies that given \(X\), \(T(X)\) does not depend on the randomness of \(\theta\).

3.2 Neyman-Fisher Factorization Criterion

Often checking the definition of sufficiency can be difficult. In cases where the model distributions admit densities, we can give a characterization that is easier to work with.

We first give technical conditions necessary for the existence of density functions. Let \((\Omega, \mathcal{F})\) be a measurable space. Recall \(A \in \mathcal{F}\) is null for a measure \(\mu\) on \((\Omega, \mathcal{F})\) if \(\mu(B) = 0\) for every \(B \in \mathcal{F}\) with \(B \subseteq A\).

**Definition 5** (Absolute continuity). Suppose \(\mu\) and \(\nu\) are measures on \((\Omega, \mathcal{F})\). We call \(\nu\) is absolutely continuous with respect to \(\mu\) if every null set of \(\mu\) is also a null set of \(\nu\). We write \(\nu \ll \mu\).

**Definition 6** (Dominated statistical model). The statistical model \((X, \{P_\theta : \theta \in \Theta\})\) is dominated by measure \(\mu\) if, for all \(\theta \in \Theta\), \(P_\theta \ll \mu\).

**Theorem 1** (Neyman-Fisher Factorization Criterion). Suppose \(P = \{P_\theta : \theta \in \Theta\}\) is dominated by a \(\sigma\)-finite measure \(\mu\). In particular, \(P_\theta\) has a density \(p_\theta = \frac{dP_\theta}{d\mu}\) for all \(P_\theta \in P\). Then \(T(X)\) is sufficient iff there exists functions \(g_\theta\) and \(h\) such that

\[
p_\theta(x) = g_\theta(T(x))h(x).
\]

**Example 2.** Let \(X_1, X_2, \ldots, X_n\) be i.i.d. Poisson(\(\lambda\)) and \(\theta = \lambda\). Then the joint distribution is

\[
p_\theta(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{X_i}}{X_i!} = e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^{n} X_i}}{X_1! \cdots X_n!} = g_\theta(T(X_1, \ldots, X_n))h(X_1, \ldots, X_n),
\]

where \(T(X) = \sum_{i=1}^{n} X_i\) is sufficient by the Neyman-Fisher Factorization Criterion.

\[\text{2Note that this definition of Sufficiency is slightly different from the conventional frequentist definition, and is along the lines of the Bayes definition. These two definitions are equivalent under mild conditions. Cf.[1]}\]
3.3 Rao-Blackwell

In the decision theory framework, sufficient statistics provide a reduction of the data without loss of information. In particular, any risk that can be achieved using a decision rule based on $X$ can also be achieved by a decision rule based on $T(X)$, as the following theorem makes precise.

**Theorem 3.** Suppose $X \sim P \in \mathcal{P}$ and $T$ is sufficient for $P$. For all decision rules $\delta(X)$ achieving risk $R(\theta, \delta(X))$, there exists a decision rule $\delta'(T(X))$ that achieves risk

$$R(\theta, \delta'(T(X))) \leq R(\theta, \delta(X)) \quad \text{for all } \theta \in \Theta.$$ 

**Sketch of Proof** Given $T(X)$, we can sample a new dataset $X'$ from the from the conditional distribution $p(X \mid T(X))$. By sufficiency (the first condition in Definition 4), the conditional distribution $p(X \mid T(X))$ doesn’t depend on $\theta$, and hence we can sample $X'$ without knowing $\theta$. We then define a randomized procedure

$$\delta'(T(X)) \triangleq \delta(X') \overset{d}{=} \delta(X).$$

Since $\delta(X)$ and $\delta'(T(X))$ have the same distribution, they also have the same risk function. 

It is rarely necessary to regenerate a dataset from sufficient statistics. Rather, in the case of convex losses, it is possible to obtain a non-randomized decision rule that matches or improves the performance of the original rule using sufficient statistics alone.

**Definition 7** (Strict convexity). A function $f : C \to \mathbb{R}$ is strictly convex if $C$ is a convex set and

$$f(tx + (1-t)y) < tf(x) + (1-t)f(y)$$

for all $x, y \in C$, $x \neq y$, $t \in (0,1)$.

**Example 4.** For any $\theta$, the function $L(\theta, a) = (a - \theta)^2$ is strictly convex with respect to $a$ on $\mathbb{R}$.

**Theorem 5** (Rao-Blackwell). Assume $L(\theta, a)$ is strictly convex in $a$. Then

$$R(\theta, \mathbb{E}[(\delta(X) \mid T(X))]) < R(\theta, \delta(X)),$$

unless $\mathbb{E}[(\delta(X) \mid T(X))] = \delta(X)$ almost everywhere.

Rao-Blackwell is stronger than the previous theorem because, when the loss function is strictly convex, it is possible to improve upon the original decision rule using only a reduced version of the data $T(X)$ and using a deterministic rule.

**References**