Statistical Decision Theory: an Introduction

These notes go into more detail on statistical decision theory. This is a theoretical branch of statistics where our goal is to understand and quantify the errors we make.

Remember our setup: we have some **unknown quantity** θ that we're interested in. We collect data x. Our data are random, and come from the distribution $p(x|\theta)$. We use the data to reason about θ. Often, we'll want to use the data to compute an estimate for θ but sometimes, we may want to do something slightly different. In order to describe "the thing we do to the data", we'll use the notation $\delta(x)$. This represents the result of applying some procedure δ to the data. For example, δ might be the sample average of many data points, or the result of logistic regression. The obvious next question is: how do we choose which procedure δ to use? We'll decide by quantifying how "good" each δ is, and then trying to choose the "best" one.

"Good" is a very abstract notion: to quantify it, we'll need a quantitative measure of how good (or to be more precise, how bad) our procedure δ is. We'll call this a **loss function**. Notationally, we'll write $\ell(\theta, \delta(x))$ to represent the loss associated with the outcome $\delta(x)$ if the true value is θ. To summarize:

Variable/notation	What it means
θ	$\text{unknown quantity/quantities of interest: parameter(s)}$
x	observed data
$p(x\|\theta)$	$\text{probability distribution for data } x \text{ (depends on } \theta)$
$\delta(x)$	$\text{decision or result computed from } x\text{, often an estimate of } \theta$
$\ell(\delta(x), \theta)$	$\text{loss (badness) for output } \delta(x) \text{ and true parameter(s) } \theta$

Examples

That's a very abstract definition: let's make it more concrete with a few examples.

Binary decision

Let's return to our binary decision-making setting. In that case:

- Our unknown parameter θ is binary, and corresponds to reality, which we've been calling R.
- Our data x were whatever we used to compute the p-value (e.g., group sample means, etc.).
- The decision δ is a binary decision, which we've been calling D.
- We haven't really talked much about $p(x|\theta)$, since we've been working with $\delta(x)$ directly.
- Our loss function depends on the problem we're solving. Since both the inputs are binary, we can write the loss in a 2x2 table that looks exactly like the ones we've seen before:

	D = $\delta(x)$ = 0	D = $\delta(x)$ = 1
$R = \theta = 0$	TN loss	FP loss
$R = \theta = 1$	FN loss	TP loss

If both kinds of error (false positive and false negative) are equally bad, we can use the simplest loss function, the **0-1 loss**:

$$\ell(\delta(x), \theta) = \begin{cases} 0 & \text{if } \theta = \delta(x) \\ 1 & \text{if } \theta \neq \delta(x) \end{cases}$$

Suppose we have a situation where a false positive is five times worse than a false negative. How would you write the loss function?

ℓ_2 loss

Now, suppose our parameterθ is continuous, and $\delta(x)$ is our estimate of the parameter from the data. To make things a little more concrete, x could be the heights of people in a random sample, and θ could be the average height of people in a population. In this case, our loss shouldn't just be right vs wrong: we should use a loss function that's lower when we're close. You've probably already seen one before: the squared error loss, also known as the ℓ_2 **loss**:

$$\ell(\delta(x), \theta) = \big(\delta(x) - \theta\big)^2$$

We'll analyze the ℓ_2 loss a little more later.

Known and unknown

At this point, you may be wondering: if θ is unknown, how can we ever compute the loss function? It's important to keep in mind that when we apply $\delta(x)$ on real data, we don't know θ. But right now, we're building up some machinery to help us analyze different procedures. In other words, we're trying to get to a place where we can answer questions like "what procedures are most likely to give us estimates that are close to θ?"

Fixed and random

The loss function is a function of $\delta(x)$, which is the procedure result for particular data x, and the particular parameter θ. This isn't particularly useful to us: we'd like to understand how the loss does "on average". But in order to compute any kind of averages, we need to decide what's random and what's fixed. This is an important fork in the road: we can either take the Bayesian or the frequentist route. Let's examine what happens if we try each one:

Frequentist loss analysis

In the frequentist world, we assume that our unknown θ is **fixed**. The data are the only random piece. So, we're going to look at the average across different possibilities for the data x. Since the data comes from the distribution $p(x|\theta)$, which depends on θ, we should expect that this "averaging" will produce something that depends on θ. We'll call our average the **frequentist risk**:

$$\begin{aligned} R(\theta) &= E_{x|\theta}\left[\ell(\delta(x), \theta)\right] \\ &= \begin{cases} \displaystyle\sum_x \ell(\delta(x), \theta) p(x|\theta) & \text{if } x \text{ discrete} \\ \displaystyle\int \ell(\delta(x), \theta) p(x|\theta) d\theta & \text{if } x \text{ continuous} \end{cases} \end{aligned}$$

The frequentist risk is a function of θ. It tells us: for a particular value of θ, how poorly does the procedure δ do if we average over all possible datasets?

Bayesian loss analysis

In the Bayesian world, we assume that our unknown θ is **random**. Since we observe a particular dataset x, we'll be a lot more interested in the randomness in θ than the randomness in x. So, we'll condition on the particular dataset we got, and look at the average across different possibilities for the unknown parameter θ. We'll call our average the **Bayesian posterior risk**:

$$\begin{aligned}\rho(x) &= E_{\theta|x}\left[\ell(\delta(x), \theta)\right] \\ &= \begin{cases} \sum_{\theta} \ell(\delta(x), \theta) p(\theta|x) & \text{if } \theta \text{ discrete} \\ \int \ell(\delta(x), \theta) p(\theta|x) d\theta & \text{if } \theta \text{ continuous} \end{cases}\end{aligned}$$

The Bayesian risk is a function of x. It tells us: given that we observed a particular dataset x, how poorly does the procedure δ do, averaged over all possible values of the parameter θ?

Comparing frequentist and Bayesian risk

Operationally, both of these look kind of similar: we're averaging the loss with respect to some conditional probability distribution. But conceptually, they're very different: the frequentist risk fixes the parameter, and averages over all the data; the Bayesian posterior risk fixes the data, and averages over all parameters.

Example: frequentist risk for ℓ_2 loss and the bias-variance tradeoff

Let's work through an example compuing the frequentist risk using the ℓ_2 loss. We'll find that the result can give us some important insights.

$$\begin{aligned} R(\theta) &= E_{x|\theta}\left[\ell(\delta(x), \theta)\right] && (1) \\ &= E_{x|\theta}\Big[\big(\delta(x) - \theta\big)^2\Big] && (2) \end{aligned}$$

To make the math work out later, we'll add and subtract the term $E_{x|\theta}[\delta(x)]$. But first, let's think about what this term means: it's the average value of the procedure δ: in other words, for a particular θ, it tells us what value of $\delta(x)$ we should expect to get, averaged across different possible values of x.

$$\begin{aligned} R(\theta) &= E_{x|\theta}\Big[\big(\delta(x) - \theta\big)^2\Big] && (3) \\ &= E_{x|\theta}\Big[\big(\delta(x) \overbrace{-E_{x|\theta}[\delta(x)] + E_{x|\theta}[\delta(x)]}^{=0} - \theta\big)^2\Big] && (4) \\ &= E_{x|\theta}\Big[\big(\underbrace{\delta(x) - E_{x|\theta}[\delta(x)]}_{\text{prediction minus avg. prediction}} + \underbrace{E_{x|\theta}[\delta(x)] - \theta}_{\text{avg. prediction minus true value}} \big)^2\Big] && (5) \end{aligned}$$

To make the math a little easier to read, we'll write $\delta = \delta(x)$ and $\bar{\delta} = E_{x|\theta}[\delta(x)]$:

$$
\begin{aligned}
R(\theta) &= E_{x|\theta}\Big[\big(\delta(x) - E_{x|\theta}[\delta(x)] + E_{x|\theta}[\delta(x)] - \theta\big)^2\Big] && (6)\\
&= E_{x|\theta}\Big[\big(\delta - \bar{\delta} + \bar{\delta} - \theta\big)^2\Big] && (7)\\
&= E_{x|\theta}\Big[\big(\delta - \bar{\delta}\big)^2 + \underbrace{2\big(\delta - \bar{\delta}\big)\big(\bar{\delta} - \theta\big)}_{=0} + \big(\bar{\delta} - \theta\big)^2\Big] && (8)\\
&= E_{x|\theta}\Big[\big(\delta - \bar{\delta}\big)^2\Big] + E_{x|\theta}\Big[\big(\bar{\delta} - \theta\big)^2\Big] && (9)\\
&= \underbrace{E_{x|\theta}\Big[\big(\delta - \bar{\delta}\big)^2\Big]}_{\text{variance of } \delta(x)} + \big(\underbrace{\bar{\delta} - \theta}_{\text{bias of } \delta(x))}\big)^2 && (10)
\end{aligned}
$$

We've shown that for the ℓ_2 loss, the frequentist risk is the sum of two terms, called the **variance** and the square of the **bias**.

The **variance**, $E_{x|\theta}\Big[\big(\delta(x) - E_{x|\theta}[\delta(x)]\big)^2\Big]$, answers the question: as the data varies, how far away will δ be from its average value? In general, if your procedure δ is very sensitive to variations in the data, your variance will be high.

The **bias**, $E_{x|\theta}[\delta(x)] - \theta$, answers the question: how far is the average value of δ from the true parameter θ? In general, if your procedure δ does a good job of capturing the complexity of predicting θ, your bias will be low.

When trying to reduce the risk (average loss), most methods try to reduce the variance and/or the bias. Many methods for estimation and prediction try to deal with the tradeoff between variance and bias: ideally we'd like both to be as small as possible, but we often need to accept a little more of one in order to make big reductions in the other.

Bayes risk

The two risks above are obtained by taking the expectation with respect to either the data x or the parameter θ. What if we take the expectation with respect to both? The **Bayes risk** is exactly that:

$$
\begin{aligned}
R(\delta) &= E_{x,\theta}[\ell(\delta(x), \theta)]\\
&= E_{\theta}[R(\theta)]\\
&= E_{x}[R(x)]
\end{aligned}
$$

where the last two inequalities follow from Fubini's theorem (i.e., that we can do the integrals for the expectations in either order and get the same result). The Bayes risk is a single number that summarizes the procedure δ. The name is somewhat misleading: it isn't really Bayesian or frequentist.