2

Decision making

The goal in decision theory is to distinguish between two alternatives under uncertainty. In this course, we'll focus on modeling uncertainty with probability and understanding the algorithmic implications of this view point. The cultural underpinnings of the field choose to start with an *optimization-based* view of decision theory. Moreover, issues of which data we collect and how we will represent it will shape our decision rules. In particular, we will see that when we have full knowledge of a probabilistic model of the world, the optimal decision rule will amount to computing a real-valued function of the collected data, and making decisions based on whether this function is greater than or less than zero. This sets the stage for the subsequent chapters on what is now called machine learning: making near-optimal decisions from data alone, without probabilistic models of the environment.

Our core setup supposes we have two alternative hypotheses H_0 and H_1. Given some measured information, our goal is to decide whether H_0 or H_1 is true. For example, H_1 could indicate that a patient has a specific bone fracture and H_0 could be the absence of a fracture. Or H_1 could be the condition that an email is a spam message and H_0 the condition that it is not.

The key is to assume a probabilistic view of the world in which both the hypotheses and the available information follow a probability distribution. Under this probabilistic view of the world, we assume that H_0 and H_1 each have some *a priori* (or *prior*) probabilities:

$$p_0 = \mathbb{P}[H_0 \text{ is true}] \qquad p_1 = \mathbb{P}[H_1 \text{ is true}]$$

In order to determine which hypothesis is true, we acquire some data that hopefully allows us to make an informed guess. We'll always model available data as being a random vector X with support in $\mathbb{R}^d$. Its distribution depends on whether H_0 or H_1 is true. In other words, we assume two different probability models for the data, one for each hypothesis. Notationally, we specify the probability density at a point $x \in \mathbb{R}^d$ under the hypothesis H_i for $i \in \{0,1\}$ as $p(x \mid H_i \text{ is true})$.

This density function is called a *generative model* or *likelihood function* for each scenario. The fact that we write the likelihood function

in the same way as a conditional density is intentional. Indeed, we will soon introduce a joint distribution over data and hypothesis whose conditional distribution with respect to the hypothesis coincides with the likelihood function.

Example: Apples or oranges

As a simple taxonomy example, consider the case that we are presented with a piece of fruit, and we know it is either an apple or an orange. Our observation X consists of a set of observable *features* of the fruit, including perhaps its color, its weight, its sugar content. If the fruit is an apple, we'd expect its color to range between green and red. For an orange, we'd expect the colors to vary from orange to red. From this set of features, our goal will be to decide whether we have an apple or an orange in front of us.

Example: Is there a needle in my haystack?

For a simple example with more mathematical formalism, suppose that when H_0 is true we observe a scalar $X = \omega$ where ω is unit-variance, zero mean gaussian noise $\omega \sim \mathcal{N}(0,1)$. Then suppose when H_1 is true, we would observe $X = s + \omega$ for some scalar s. That is, the conditional densities are

$$p(X \mid H_0 \text{ is true}) = \mathcal{N}(0,1)$$
$$p(X \mid H_1 \text{ is true}) = \mathcal{N}(s,1)\,.$$

Recall that the gaussian distribution of mean μ and variance σ^2 is given by the density $\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$.

When s has large magnitude, it would be obvious whether H_0 or H_1 were true. For example, suppose $s = 10$ and we observed $X = 11$. Under H_0, the probability that the observation greater than 10 is on the order of 10^{-23}, and hence we'd likely think we're in alternative H_1. However, if s were very close to zero, distinguishing between the two alternatives is rather challenging. We can think of a small signal s that we're trying to detect as a *needle in a haystack*.

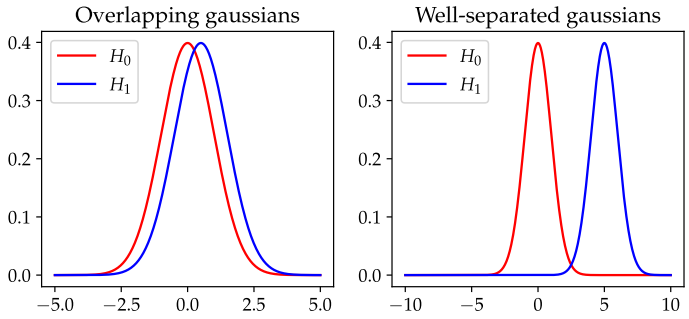

Figure 1: Illustration of shifted gaussians

Bayesian binary hypothesis testing

Our core approach to all statistical decision making will be to formulate an appropriate optimization problem for which the decision rule is the optimal solution. That is, we will optimize over *algorithms*, searching for functions that map data to decisions and predictions. We will define an appropriate notion of the cost associated to each decision, and attempt to construct decision rules that minimize the expected value of this cost. As we will see, choosing this optimization framework has many immediate consequences.

Labels

In order to neatly write decision problems in the language of mathematical optimization, we distinguish between boolean and integer values for the hypotheses. We associate real number valued *labels* with hypotheses. Labels will be denoted by Y, and are almost always integer valued in this book. However, we will describe cases where it is useful to have vector-valued and real-valued labels when the need arises. A particularly common example we will use in this chapter is to set the label $Y = 1$ when H_1 is true and $Y = 0$ when H_0 is true. Another common convention is to instead set $Y = -1$ when H_0 is true. A practitioner is free to choose whatever label mapping is most convenient for their application.

Loss functions and risk

Define the *loss function* associated with the decision of declaring H_i is true when H_j is true as $\ell(i, j)$. This loss could be symmetric, though often times we want it to be asymmetric. For instance, we may prefer false positives—where we declare H_1 is true even though H_0 is true—to missed detections—-where we declare H_0 is true even though H_1 is true.

A decision rule $\widehat{Y}(X)$ takes the variable X as input and returns a discrete decision, either 0 to indicate selection of H_0 or 1 for the selection of H_1.

Definition 1. *We define the* risk *associated with $\widehat{Y}$ to be*

$$R[\widehat{Y}] := \mathbb{E}[\ell(\widehat{Y}(X), Y)] .$$

Here, the expectation is taken jointly over X and Y.

Keep in mind that $Y = i$ corresponds to the case that H_i is true. Our goal is to determine which decision rule minimizes the risk.

Let's get a sense for how we can minimize risk. In order to minimize the risk, theoretically speaking, we need to solve an *infinite*

dimensional optimization problem over binary-valued functions. That is, *for every* x, we need to find a binary assignment. Fortunately, the infinite dimension here turns out to not be a problem analytically once we make use of the law of iterated expectation. Note

$$\begin{aligned}\mathbb{E}[\ell(\widehat{Y}(X), Y)] &= \mathbb{E}\left[\mathbb{E}\left[\ell(\widehat{Y}(X), Y) \mid X\right]\right] \\ &= \int \mathbb{E}[\ell(\widehat{Y}(X), Y) \mid X = x] p(x)\, \mathrm{d}x\,.\end{aligned}$$

In this last expression, we have a nonnegative combination of terms, and we have no constraints on the decision rule $\widehat{Y}$. Hence, to minimize, we can treat $\widehat{Y}(x)$ individually for each fixed value of x. Indeed, when we fix x,

$$\begin{aligned}\mathbb{E}[\ell(0, Y) \mid X = x] &= \ell(0,0)\,\mathbb{P}[Y = 0 \mid X = x] + \ell(0,1)\,\mathbb{P}[Y = 1 \mid X = x] \\ \mathbb{E}[\ell(1, Y) \mid X = x] &= \ell(1,0)\,\mathbb{P}[Y = 0 \mid X = x] + \ell(1,1)\,\mathbb{P}[Y = 1 \mid X = x]\,.\end{aligned}$$

And hence, the optimal assignment for this x is to pick whichever of these two expressions is smaller. Collecting terms and rearranging, we find the decision rule

$$\widehat{Y}(x) = \mathbb{1}\left\{\mathbb{P}[Y = 1 \mid X = x] \geq \frac{\ell(1,0) - \ell(0,0)}{\ell(0,1) - \ell(1,1)}\ \mathbb{P}[Y = 0 \mid X = x]\right\}.$$

That is, the optimal rule is to declare H_1 is true when the probability of H_1 given the data x is sufficiently larger than the probability of H_0 given the same data.

Now, we have not explicitly modeled $\mathbb{P}[Y = i \mid X]$, but we can compute these using Bayes rule:

$$\mathbb{P}[Y = i \mid X = x] = \frac{p(x \mid H_i \text{ is true})\,\mathbb{P}[H_i \text{ is true}]}{p(x)}\,.$$

Plugging this expression into our decision rule, we find a decision rule computable only using the probabilities of the hypotheses and our conditional data observation models

$$\widehat{Y}(x) = \mathbb{1}\left\{\frac{p(x \mid H_1 \text{ is true})}{p(x \mid H_0 \text{ is true})} \geq \frac{p_0(\ell(1,0) - \ell(0,0))}{p_1(\ell(0,1) - \ell(1,1))}\right\}.$$

This rule is called a *likelihood ratio test* (LRT). The *likelihood ratio* is simply the ratio of the likelihoods:

$$\mathcal{L}(x) := \frac{p(x \mid H_1 \text{ is true})}{p(x \mid H_0 \text{ is true})}\,.$$

If we fix

$$\eta = \frac{p_0(\ell(1,0) - \ell(0,0))}{p_1(\ell(0,1) - \ell(1,1))}\,,$$

then the decision rule that minimizes the risk is

$$\widehat{Y}(x) = \mathbb{1}\{\mathcal{L}(x) \geq \eta\}\,.$$

A LRT naturally partitions the sample space in two:

$$\mathcal{X}_0 = \{x \in \mathcal{X} \colon \mathcal{L}(x) \leq \eta\}$$
$$\mathcal{X}_1 = \{x \in \mathcal{X} \colon \mathcal{L}(x) > \eta\} \ .$$

The sample space $\mathcal{X}$ then becomes the disjoint union of $\mathcal{X}_0$ and $\mathcal{X}_1$. Since we only need to identify which set x belongs to, we can use any function $h : \mathcal{X} \to \mathbb{R}$ which gives rise to the same threshold rule. As long as $h(x) \leq t$ whenever $\mathcal{L}(x) \leq \eta$ and vice versa, these functions give rise to the same partition into $\mathcal{X}_0$ and $\mathcal{X}_1$. So, for example, if g is any monotonically increasing function, then the decision rule

$$\widehat{Y}_g(x) = \mathbb{1}\{g(\mathcal{L}(x)) \geq g(\eta)\}$$

is equivalent to using $\widehat{Y}(x)$. In particular, it's popular to use the logarithmic decision rule

$$\widehat{Y}_{\log}(x) = \mathbb{1}\{\log p(x \mid H_1 \text{ is true}) - \log p(x \mid H_0 \text{ is true}) \geq \log(\eta)\} \, ,$$

as it is often more convenient or mathematically stable to work with logarithms of likelihoods.

This discussion shows that there are an *infinite number of functions* which give rise to the same binary decision rule. Hence, we don't need to know the conditional densities exactly and can still compute the optimal decision rule. For example, suppose the true partitioning of the real line under an LRT is

$$\mathcal{X}_0 = \{x \colon x \geq 0\} \quad \text{and} \quad \mathcal{X}_1 = \{x \colon x < 0\} \, .$$

Setting the threshold to $t = 0$, the functions $h(x) = x$ or $h(x) = x^3$ give the same decision rule, as does any odd function which is positive on the right half line.

Example: needle in a haystack revisited

Let's return to our needle in a haystack example with

$$p(X \mid H_0 \text{ is true}) = \mathcal{N}(0, 1)$$
$$p(X \mid H_1 \text{ is true}) = \mathcal{N}(s, 1) \, ,$$

and assume $p_1 = 10^{-6}$—a very rare needle. Suppose that if we declare H_0 to be true, we do not pay a cost. If we declare H_1 to be true but are wrong, we incur a cost of 100. But if we guess H_1 and it is actually true, we actually gain a reward of $1,000,000$. That is

$$\ell(0,0) = 0 \, , \ \ell(0,1) = 0 \, , \ \ell(1,0) = 100 \, , \text{ and } \ell(1,1) = -1,000,000 \, .$$

What is the LRT for this problem? Here, it's considerably easier to work with logarithms:

$$\log(\eta) = \log\left(\frac{(1-10^{-6})\cdot 100}{10^{-6}\cdot 10^{6}}\right) \approx 4.61$$

Now,

$$\log p(x \mid H_1 \text{ is true}) - \log p(x \mid H_0 \text{ is true}) = -\frac{1}{2}(x-s)^2 + \frac{1}{2}x^2 = sx - \frac{1}{2}s^2$$

Hence, the optimal decision rule is to declare H_1 if $sx > \frac{1}{2}s^2 + \log(\eta)$ and H_0 otherwise. The optimal rule here is *linear*. Moreover, the rule divides the space into two open intervals. In the case when s is positive.

$$\mathcal{X}_0 = \{x \colon x \leq \tfrac{1}{2}s + s^{-1}\log(\eta)\}\,.$$

Also note here that while the entire real line lies in the union of $\mathcal{X}_0$ and $\mathcal{X}_1$, it is exceptionally unlikely to ever see an x larger than $|s|+5$. Hence, even if our decision rule were incorrect in these regions, the risk would still be nearly optimal as these terms have almost no bearing on our expected risk!

Canonical cases of likelihood ratio tests

A folk theorem of statistical decision theory states that essentially all optimal rules are equivalent to likelihood ratio tests. While this isn't *always* true, most of the rules used in practice do end up being equivalent to LRTs. In the next section, we'll present a mathematical framework that lets us see how far LRTs can take us. But before that, we can already show that the well known *maximum likelihood* and *maximum a posteriori* decision rules are both LRTs.

Maximum a posteriori decision rule

The expected error of a decision rule is the expected number of times we declare H_0 (resp. H_1) when H_1 (resp. H_0) is true. Minimizing the error is equivalent to minimizing the risk with cost $\ell(0,0) = \ell(1,1) = 0$, $\ell(1,0) = \ell(0,1) = 1$. The optimum decision rule is hence a likelihood ratio test. In particular,

$$\widehat{Y}(x) = \mathbb{1}\{\mathcal{L}(x) \leq \tfrac{p_0}{p_1}\}\,.$$

Using Bayes rule, one can see that this rule is equivalent to

$$\widehat{Y}(x) = \arg\max_i \mathbb{P}[Y = i \mid X = x]\,.$$

The expression $\mathbb{P}[Y \mid X]$ is called the *posterior probability* of Y given X. And this rule is hence referred to as the *maximum a posteriori* (MAP) decision rule.

Maximum likelihood decision rule

As we discussed above, the expressions $p(x \mid H_i)$ are often called the *likelihood* of the data x given the hypothesis H_i. A maximum likelihood decision rule would set

$$\widehat{Y}(x) = \arg\max_i p(x \mid H_i) .$$

This is completely equivalent to the LRT when $p_0 = p_1$ and the costs are $\ell(0,0) = \ell(1,1) = 0$, $\ell(1,0) = \ell(0,1) = 1$. Hence, the maximum likelihood rule is equivalent to the MAP rule with a uniform prior on the hypothesis.

That both of these popular rules ended up reducing to LRTs is no accident. In the next lecture, we will show that LRTs are almost always the optimal solution of optimization-driven decision theory.

Types of errors and successes

Let $\widehat{Y}(x)$ denote any decision rule mapping into $\{0,1\}$. Denote $\mathcal{X}_i = \{x \colon \widehat{Y}(x) = i\}$ for $i = 0$ or 1. In this section we define some popular notions of error and success.

1. **True Positive Rate:** $\text{TPR} = \mathbb{P}[\widehat{Y}(X) = 1 \mid H_1 \text{ is true}]$. Also known as *power*, *sensitivity*, *probability of detection*, or *recall*.
2. **False Negative Rate:** $\text{FNR} = 1 - \text{TPR}$. Also known as *type II error* or *probability of missed detection*.
3. **False Positive Rate:** $\text{FPR} = \mathbb{P}[\widehat{Y}(X) = 1 \mid H_0 \text{ is true}]$. Also known as *size* or *type I error* or *probability of false alarm*.
4. **True Negative Rate** $\text{TNR} = 1 - \text{FPR}$, the probability of accepting the null hypothesis given the null hypothesis. This is also known as *specificity*.

There are other quantities that are also of interest in statistics and machine learning:

1. **Precision:** $P[H_1 \text{ is true} \mid \widehat{Y}(X) = 1]$. This is equal to $(p_1\text{TPR})/(p_0\text{FPR} + p_1\text{TPR})$.
2. **F1-score:** F_1 is the harmonic mean of precision and recall. We can write this as
$$F_1 = \frac{2\text{TPR}}{1 + \text{TPR} + \frac{p_0}{p_1}\text{FPR}}$$
3. **False discovery rate:** False discovery rate (FDR) is equal to the expected ratio of the number of false positives to the total number of positives.

In the case where both hypotheses are equally likely, precision, F_1, and FDR are also only functions of FPR and TPR. However, these

quantities explicitly account for *class imbalances*: when there is a significant skew between p_0 and p_1, such measures are often preferred.

TPR and FPR are competing objectives. We'd like TPR as large as possible and FPR as small as possible. We can write Bayesian decision theory as a problem of optimizing a balance between TPR and FPR:

$$R[\widehat{Y}] := \mathbb{E}[\ell(\widehat{Y}(X), Y)] = \alpha \mathrm{FPR} - \beta \mathrm{TPR} + \gamma\,,$$

where α and β are nonnegative and γ is some constant. For all such α, β, and γ, the risk-minimizing decision rule is an LRT.

Other cost functions might try to balance TPR versus FPR in other ways. Which of these give rise to LRTs? Each value of η in an LRT gives a pair $(\mathrm{TPR}, \mathrm{FPR})$. The curve traced out in the TPR-FPR plane by varying η from negative to positive infinity is called the *receiver operating characteristic* (ROC) of the likelihood ratio test.

What pairs of $(\mathrm{TPR}, \mathrm{FPR})$ are achievable using LRTs? Clearly we can always achieve $(0, 0)$ and $(1, 1)$ with $\eta = \pm\infty$. What happens in between these values?

Example: the needle one more time

Consider again the *needle in a haystack* example, where $p(x \mid H_0) = \mathcal{N}(0, \sigma^2)$ and $p(x \mid H_1) = \mathcal{N}(s, \sigma^2)$ with s a positive scalar. The optimal decision rule is to declare H_1 when X is greater than $\gamma := \frac{s}{2} + \frac{\sigma^2 \log \eta}{s}$. Hence we have

$$\mathrm{TPR} = \int_\gamma^\infty p(x \mid H_1)\,\mathrm{d}x = \tfrac{1}{2}\,\mathrm{erfc}\left(\frac{\gamma - s}{\sqrt{2}\sigma}\right)$$
$$\mathrm{FPR} = \int_\gamma^\infty p(x \mid H_0)\,\mathrm{d}x = \tfrac{1}{2}\,\mathrm{erfc}\left(\frac{\gamma}{\sqrt{2}\sigma}\right).$$

For fixed s and σ, the ROC curve $(\mathrm{FPR}(\gamma), \mathrm{TPR}(\gamma))$ only depends on the *signal to noise ratio* (SNR), s/σ. For small SNR, the ROC curve is close to the FPR = TPR line. For large SNR, TPR approaches 1 for all values of FPR.

The Neyman-Pearson Lemma

The Neyman-Pearson Lemma, a fundamental lemma of decision theory, will be an important tool for us to establish three important facts. First, it will be a useful tool for understanding the geometric properties of ROC curves. Second, it will demonstrate another important instance where an optimal decision rule is a likelihood ratio test. Third, it introduces the notion of probabilistic decision rules.

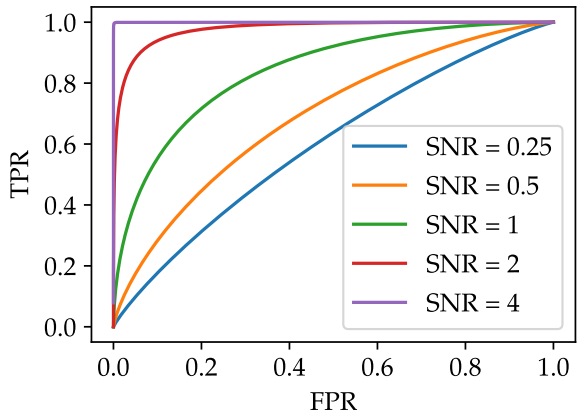

Figure 2: The ROC curves for various signal to noise ratios in the needle in the haystack problem.

Suppose we want to maximize the probability of detection subject to the false alarm rate being upper bounded by some fixed number. That is, we aim to solve the optimization problem:

$$\begin{array}{ll} \text{maximize} & \text{TPR} \\ \text{subject to} & \text{FPR} \leq \alpha \end{array}$$

Let's optimize over *probabilistic decision rules*. A probabilistic decision rule Q returns H_1 with probability $Q(x)$ and H_0 with probability $1 - Q(x)$. With such rules, we can rewrite our optimization problem as:

$$\begin{array}{ll} \text{maximize}_Q & \int Q(x) p(x \mid H_1 \text{ is true}) \, \mathrm{d}x \\ \text{subject to} & \int Q(x) p(x \mid H_0 \text{ is true}) \, \mathrm{d}x \leq \alpha \\ & \forall x \colon Q(x) \in [0, 1] \end{array}$$

Lemma 1. ***Neyman-Pearson Lemma.*** *The optimal probabilistic decision rule that maximizes* TPR *with an upper bound on* FPR *is a deterministic likelihood ratio test.*

Even in this constrained setup, allowing for more powerful probabilistic rules, we can't escape the inevitability of LRTs. The Neyman-Pearson Lemma has many interesting consequences in its own right that we will discuss momentarily. But first, let's see why the lemma is true.

The key insight is that the optimization problem does not depend on the prior probabilities of H_0 and H_1. Hence, we can choose priors and a loss function to construct a Bayesian Hypothesis testing problem for which an LRT is optimal. The optimality condition for this chosen problem will imply the lemma.

Proof. Let η be the threshold for an LRT such that the decision rule

$$Q_\eta(x) = \mathbb{1}\{\mathcal{L}(x) > \eta\}$$

has FPR $= \alpha$. Let β denote the true positive rate of Q_η. Note that Q_η is optimal for the risk minimization problem where we minimize the probability of error, keep the likelihood functions the same, but adjust the priors to be

$$\widehat{p}_0 = \frac{\eta}{1+\eta} \qquad\qquad \widehat{p}_1 = \frac{1}{1+\eta}\,.$$

That is, Q_η is the MAP rule for this choice of $\widehat{p}_0$ and $\widehat{p}_1$. Now let Q be any other decision rule with $\mathrm{FPR}(Q) \leq \alpha$. We have

$$\begin{aligned}\widehat{p}_0\alpha + \widehat{p}_1(1-\beta) &\leq \widehat{p}_0\mathrm{FPR}(Q) + \widehat{p}_1(1-\mathrm{TPR}(Q)) \\ &\leq \widehat{p}_0\alpha + \widehat{p}_1(1-\mathrm{TPR}(Q))\,,\end{aligned}$$

which implies $\mathrm{TPR}(Q) \leq \beta$. This in turn means that Q_η maximizes TPR for all rules with FPR $\leq \alpha$, proving the lemma.

□

Properties of ROC curves

A specific randomized decision rule that is useful for analysis combines two other rules. Suppose decision rule one yields $(\mathrm{FPR}^{(1)}, \mathrm{TPR}^{(1)})$ and the second rule achieves $(\mathrm{FPR}^{(2)}, \mathrm{TPR}^{(2)})$. If we flip a biased coin and use rule one with probability p and rule 2 with probability $1-p$, then this yields a randomized decision rule with $(\mathrm{FPR}, \mathrm{TPR}) = (p\mathrm{FPR}^{(1)} + (1-p)\mathrm{FPR}^{(2)}, p\mathrm{TPR}^{(1)} + (1-p)\mathrm{TPR}^{(2)})$. Using this rule lets us prove several properties of ROC curves.

Proposition 1. $(0,0)$ *and* $(1,1)$ *are on the ROC curve.*

This proposition follows because $(0,0)$ is achieved when $\eta = \infty$. $(1,1)$ is achieved when $\eta = -\infty$.

Proposition 2. TPR $\geq$ FPR.

To see why this proposition is true, fix some $\alpha > 0$. Using a randomized rule, we can achieve a decision rule with TPR $=$ FPR $= \alpha$. But the Neyman-Pearson LRT with FPR constrained to be less than or equal to α achieves a probability of detection greater than or equal to the randomized rule.

Proposition 3. *The ROC curve is concave.*

Suppose $(\mathrm{FPR}(\eta_1), \mathrm{TPR}(\eta_1))$ and $(\mathrm{FPR}(\eta_2), \mathrm{TPR}(\eta_2))$ are achievable. Then

$$(t\mathrm{FPR}(\eta_1) + (1-t)\mathrm{FPR}(\eta_2), t\mathrm{TPR}(\eta_1) + (1-t)\mathrm{TPR}(\eta_2))$$

is achievable by a randomized test. Fixing FPR $\leq t\mathrm{FPR}(\eta_1) + (1-t)\mathrm{FPR}(\eta_2)$, we see that the optimal Neyman-Pearson LRT achieves $\mathrm{TPR} \geq \mathrm{TPR}(\eta_1) + (1-t)\mathrm{TPR}(\eta_2)$.

Area under the ROC curve

A popular summary statistic for evaluating the quality of a decision function is the area under its associated ROC curve. This is commonly abbreviated as AUC. In the ROC curve plotted in the previous section, as the SNR increases, the AUC increases. However, AUC does not tell the entire story. Here we plot two ROC curves with the same AUC.

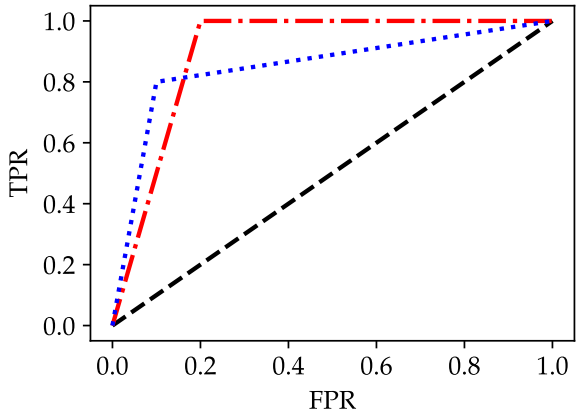

Figure 3: Two ROC curves with the same AUC. Note that if we constrain FPR to be less than 10%, for the blue curve, TPR can be as high as 80% whereas it can only reach 50% for the red.

If we constrain FPR to be less than 10%, for the blue curve, TPR can be as high as 80% whereas it can only reach 50% for the red. AUC should be always viewed skeptically: the shape of an ROC curve is always more informative than any individual number.

Looking ahead: what if we don't know the models?

This chapter examined how to make decisions when we have access to known probabilistic models about both data and priors about the distribution of hypotheses. The ubiquitous solution to decision problems is a likelihood ratio test. But note we first derived something even simpler: a posterior ratio test. That is, we could just compare the probability of H_1 given our data to the probability of H_0 given our data, and decide on H_1 if its probability was sufficiently larger than that of H_0. Comparing likelihoods or posteriors are equivalent up to a rescaling of the decision threshold.

What if we don't have a probabilistic model of how the data is generated? There are two natural ways forward: Either estimate $p(X \mid H)$ from examples or estimate $\mathbb{P}[Y \mid X]$ from examples. Estimating likelihood models is a challenge as, when X is high dimensional, estimating $p(X \mid H)$ from data is hard in both theory and practice. Estimating posteriors on the other hand seems more promising. Esti-

mating posteriors is essentially like populating an excel spreadsheet and counting places where many columns are equal to one another.

But estimating the posterior is also likely overkill. We care about the likelihood or posterior ratios as these completely govern our decision rules. It's possible that such ratios are easier to estimate than the quantities themselves. Indeed, we just need to find *any function* f where $f(x) \geq 0$ if $p(x \mid H_1 \text{ is true})/p(x \mid H_0 \text{ is true}) \geq \eta$ and $f(x) \leq 0$ if $p(x \mid H_1 \text{ is true})/p(x \mid H_0 \text{ is true}) \leq \eta$. So if we only have samples

$$S = \{(x_1, y_1), \ldots (x_n, y_n)\}$$

of data and their labels, we could try to minimize the sample average

$$R_S(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

with respect to f. This approach is called *empirical risk minimization* (ERM) and forms the basis of most contemporary ML and AI systems. We will devote a the next several chapters of this text to understanding the ins and outs of ERM.

Decisions that discriminate

Binary decision rules always draw a boundary between one group in the population and its complement. Some are labeled *accept*, others are labeled *reject*. When decisions have serious consequences for the individual, however, this decision boundary is not just a technical artifact. Rather it has moral and legal significance.

Many decisions entail a life changing event for the individual. The decision could grant access to a major opportunity, such as college admission, or deny access to a vital resource, such as a social benefit.

The decision maker often has access to data that encode an individual's status in socially salient groups relating to race, ethnicity, gender, religion, or disability status. These and other categories that have been used as the basis of adverse treatment, oppression, and denial of opportunity in the past and in many cases to this day.

Some see formal or algorithmic decision making as a neutral mathematical tool. However, numerous scholars have shown how formal models can perpetuate existing inequities and cause harm. In her book on this topic, Ruha Benjamin warns of

> *the employment of new technologies that reflect and reproduce existing inequities but that are promoted and perceived as more objective or progressive than the discriminatory systems of a previous era.*[1]

[1] Benjamin, *Race After Technology* (Polity, 2019).

Even though the problems of inequality and injustice are much broader than one of formal decisions, we already encounter an important and challenging facet within the narrow formal setup of this

chapter. Specifically, we are concerned with decision rules that *discriminate* in the sense of creating an unjustified basis of differentiation between individuals.

A concrete example is helpful. Suppose we want to accept or reject individuals for a job. Suppose we have a perfect estimate of the number of hours an individual is going to work in the next 5 years. We decide that this a reasonable measure of productivity and so we accept every applicant where this number exceeds a certain threshold. On the face of it, our rule might seem neutral. However, on closer reflection, we realize that this decision rule systematically disadvantages individuals who are more likely than others to make use of their parental leave employment benefit that our hypothetical company offers. We are faced with a conundrum. On the one hand, we trust our estimate of productivity. On the other hand, we consider taking parental leave *morally irrelevant* to the decision we're making. It should not be a disadvantage to the applicant. After all that is precisely the reason why the company is offering a parental leave benefit in the first place.

The simple example shows that statistical accuracy alone is no safeguard against discriminatory decisions. It also shows that ignoring *sensitive attributes* is no safeguard either. So what then is *discrimination* and how can we avoid it? This question has occupied scholars from numerous disciplines for decades. There is no simple answer. Before we go into attempts to formalize discrimination in our statistical decision making setting, it is helpful to take a step back and reflect on what the law says.

Legal background in the United States

The legal frameworks governing decision making differ from country to country, and from one domain to another. We take a glimpse at the situation in the United States, bearing in mind that our description is incomplete and does not transfer to other countries.

Discrimination is not a general concept. It is concerned with socially salient categories that have served as the basis for unjustified and systematically adverse treatment in the past. United States law recognizes certain *protected categories* including race, sex (which extends to sexual orientation), religion, disability status, and place of birth.

Further, discrimination is a domain specific concept concerned with important opportunities that affect people's lives. Regulated domains include credit (Equal Credit Opportunity Act), education (Civil Rights Act of 1964; Education Amendments of 1972), employment (Civil Rights Act of 1964), housing (Fair Housing Act), and

public accommodation (Civil Rights Act of 1964). Particularly relevant to machine learning practitioners is the fact that the scope of these regulations extends to marketing and advertising within these domains. An ad for a credit card, for example, allocates access to credit and would therefore fall into the credit domain.

There are different legal frameworks available to a plaintiff that brings forward a case of discrimination. One is called *disparate treatment*, the other is *disparate impact*. Both capture different forms of discrimination. Disparate treatment is about purposeful consideration of group membership with the intention of discrimination. Disparate impact is about unjustified harm, possibly through indirect mechanisms. Whereas disparate treatment is about *procedural fairness*, disparate impact is more about *distributive justice*.

It's worth noting that anti-discrimination law does not reflect one overarching moral theory. Pieces of legislation often came in response to civil rights movements, each hard fought through decades of activism.

Unfortunately, these legal frameworks don't give us a formal definition that we could directly apply. In fact, there is some well-recognized tension between the two doctrines.

Formal non-discrimination criteria

The idea of formal non-discrimination (or *fairness*) criteria goes back to pioneering work of Anne Cleary and other researchers in the educational testing community of the 1960s.[2]

[2] Hutchinson and Mitchell, "50 Years of Test (Un) Fairness: Lessons for Machine Learning," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, 49–58.

The main idea is to introduce a discrete random variable A that encodes membership status in one or multiple protected classes. Formally, this random variable lives in the same probability space as the other covariates X, the decision $\widehat{Y} = \mathbb{1}\{R > t\}$ in terms of a score R, and the outcome Y. The random variable A might coincide with one of the features in X or correlate strongly with some combination of them.

Broadly speaking, different statistical fairness criteria all equalize some group-dependent statistical quantity across groups defined by the different settings of A. For example, we could ask to equalize acceptance rates across all groups. This corresponds to imposing the constraint for all groups a and b:

$$\mathbb{P}[\widehat{Y} = 1 \mid A = a] = \mathbb{P}[\widehat{Y} = 1 \mid A = b]$$

Researchers have proposed dozens of different criteria, each trying to capture different intuitions about what is *fair*. Simplifying the landscape of fairness criteria, we can say that there are essentially three fundamentally different ones of particular significance:

- Acceptance rate $\mathbb{P}[\widehat{Y} = 1]$
- Error rates $\mathbb{P}[\widehat{Y} = 0 \mid Y = 1]$ and $\mathbb{P}[\widehat{Y} = 1 \mid Y = 0]$
- Outcome frequency given score value $\mathbb{P}[Y = 1 \mid R = r]$

The meaning of the first two as a formal matter is clear given what we already covered. The third criterion needs a bit more motivation. A useful property of score functions is *calibration* which asserts that $\mathbb{P}[Y = 1 \mid R = r] = r$ for all score values r. In words, we can interpret a score value r as the propensity of positive outcomes among instances assigned the score value r. What the third criterion says is closely related. We ask that the score values have the same meaning in each group. That is, instances labeled r in one group are equally likely to be positive instances as those scored r in any other group.

The three criteria can be generalized and simplified using three different conditional independence statements.

Table 1: Non-discrimination criteria

Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

Each of these applies not only to binary prediction, but any set of random variables where the independence statement holds. It's not hard to see that independence implies equality of acceptance rates across groups. Separation implies equality of error rates across groups. And sufficiency implies that all groups have the same rate of positive outcomes given a score value.[3]

Researchers have shown that any two of the three criteria are *mutually exclusive* except in special cases. That means, generally speaking, imposing one criterion forgoes the other two.[4]

Although these formal criteria are easy to state and arguably natural in the language of decision theory, their merit as measures of discrimination has been subject of an ongoing debate.

[3] Barocas, Hardt, and Narayanan, *Fairness and Machine Learning* (fairmlbook.org, 2019).

[4] Kleinberg, Mullainathan, and Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," in *Innovations in Theoretical Computer Science*, 2017; Chouldechova, "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments," in *Fairness, Accountability, and Transparency*, 2016.

Merits and limitations of a narrow statistical perspective

The tension between these criteria played out in a public debate around the use of risk scores to predict *recidivism* in pre-trial detention decisions.

There's a risk score, called COMPAS, used by many jurisdictions in the United States to assess *risk of recidivism* in pre-trial bail decisions.[5] Judges may detain defendants in part based on this score.

Investigative journalists at ProPublica found that Black defendants

[5] Recidivism refers to a person's relapse into criminal behavior. In the United States, a defendant may either be detained or released on a bail prior to the trial in court depending on various factors.

face a higher false positive rate, i.e., more Black defendants labeled *high risk* end up not committing a crime upon release than among White defendants labeled *high risk*.[6] In other words, the COMPAS score fails the separation criterion.

A company called Northpointe, which sells the proprietary COMPAS risk model, pointed out in return that Black and White defendants have equal recidivism rates *given* a particular score value. That is defendants labeled, say, an '8' for *high risk* would go on to recidivate at a roughly equal rate in either group. Northpointe claimed that this property is desirable so that a judge can interpret scores equally in both groups.[7]

The COMPAS debate illustrates both the merits and limitations of the narrow framing of discrimination as a classification criterion.

On the hand, the error rate disparity gave ProPublica a tangible and concrete way to put pressure on Northpointe. The narrow framing of decision making identifies the decision maker as responsible for their decisions. As such, it can be used to interrogate and possibly intervene in the practices of an entity.

On the other hand, decisions are always part of a broader system that embeds structural patterns of discrimination. For example, a measure of recidivism hinges crucially on existing policing patterns. Crime is only found where policing activity happens. However, the allocation and severity of police force itself has racial bias. Some scholars therefore find an emphasis on statistical criteria rather than structural determinants of discrimination to be limited.

[6] Angwin et al., "Machine Bias," *ProPublica*, May 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentenci

[7] Dieterich, Mendoza, and Brennan, "COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity," 2016, https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html.

Chapter notes

This is the unique chapter in the book where the overwhelming consensus is that the material is settled. The theory we covered in this chapter is also called *detection theory* rather than decision theory. Detection theory has not changed much at all since the 1950s and is essentially considered a "solved problem." Neyman and Pearson invented the likelihood ratio test[8] and later proved their lemma showing it to be optimal for maximizing true positive rates while controlling false positive rates.[9] Wald followed this work by inventing general Bayes risk minimization in 1939.[10] Wald's ideas were widely adopted during World War II for the purpose of interpreting RADAR signals which were often very noisy. Much work was done to improve RADAR operations, and this led to the formalization that the output of a RADAR system (the receiver) should be a likelihood ratio, and a decision should be made based on an LRT. Our proof of Neyman-Pearson's lemma came later, and is due to Bertsekas and Tsitsiklis (See Section 9.3 of *Introduction to*

[8] Neyman and Pearson, "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part i," *Biometrika*, 1928, 175–240.

[9] Neyman and Pearson, "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society of London. Series A* 231, no. 694–706 (1933): 289–337.

[10] Wald, "Contributions to the Theory of Statistical Estimation and Testing Hypotheses," *The Annals of Mathematical Statistics* 10, no. 4 (1939): 299–326.

Probability[11]).

Our current theory of detection was fully developed by Peterson, Birdsall, and Fox in their report on optimal signal detectability.[12] Peterson, Birdsall, and Fox may have been the first to propose Receiver Operating Characteristics as the means to characterize the performance of a detection system, but these ideas were contemporaneously being applied to better understand psychology and psychophysics as well.[13]

Statistical Signal Detection theory was adopted in the pattern recognition community at a very early stage. Chow proposed using optimal detection theory,[14] and this led to a proposal by Highleyman to approximate the risk by its sample average.[15] This transition from population risk to "empirical" risk gave rise to what we know today as machine learning.

There is a large amount of literature now on the topic of fairness and machine learning. For a general introduction to the problem and dangers associated with algorithmic decision making not limited to discrimination, see the books by Benjamin,[16] Broussard,[17] Eubanks,[18] Noble,[19] and O'Neil.[20] The technical material in our section on discrimination follows Chapter 2 in the textbook by Barocas, Hardt, and Narayanan.[21]

[11] Bertsekas and Tsitsiklis, *Introduction to Probability*, 2nd ed. (Athena Scientific, 2008).

[12] Peterson, Birdsall, and Fox, "The Theory of Signal Detectability," *Transactions of the IRE* 4, no. 4 (1954): 171–212.

[13] Tanner Jr. and Swets, "A Decision-Making Theory of Visual Detection," *Psychological Review* 61, no. 6 (1954): 401.

[14] Chow, "An Optimum Character Recognition System Using Decision Functions," *IRE Transactions on Electronic Computers*, no. 4 (1957): 247–54.

[15] Highleyman, "Linear Decision Functions, with Application to Pattern Recognition," *Proceedings of the IRE* 50, no. 6 (1962): 1501–14.

[16] Benjamin, *Race After Technology*.

[17] Broussard, *Artificial Unintelligence: How Computers Misunderstand the World* (MIT Press, 2018).

[18] Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (St. Martin's Press, 2018).

[19] Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (NYU Press, 2018).

[20] O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Broadway Books, 2016).

[21] Barocas, Hardt, and Narayanan, *Fairness and Machine Learning*.

References

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. "Machine Bias." *ProPublica*, May 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019.

Benjamin, Ruha. *Race After Technology*. Polity, 2019.

Bertsekas, Dimitri P., and John N. Tsitsiklis. *Introduction to Probability*. 2nd ed. Athena Scientific, 2008.

Broussard, Meredith. *Artificial Unintelligence: How Computers Misunderstand the World*. MIT Press, 2018.

Chouldechova, Alexandra. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments." In *Fairness, Accountability, and Transparency*, 2016.

Chow, Chi-Keung. "An Optimum Character Recognition System Using Decision Functions." *IRE Transactions on Electronic Computers*, no. 4 (1957): 247–54.

Dieterich, William, Christina Mendoza, and Tim Brennan. "COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity," 2016. https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html.

Eubanks, Virginia. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, 2018.

Highleyman, Wilbur H. "Linear Decision Functions, with Application to Pattern Recognition." *Proceedings of the IRE* 50, no. 6 (1962): 1501–14.

Hutchinson, Ben, and Margaret Mitchell. "50 Years of Test (Un) Fairness: Lessons for Machine Learning." In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 49–58, 2019.

Kleinberg, Jon M., Sendhil Mullainathan, and Manish Raghavan. "Inherent Trade-Offs in the Fair Determination of Risk Scores." In *Innovations in Theoretical Computer Science*, 2017.

Neyman, Jerzy, and Egon S. Pearson. "On the Problem of the Most Efficient Tests of Statistical Hypotheses." *Philosophical Transactions of the Royal Society of London. Series A* 231, no. 694–706 (1933): 289–337.

———. "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part i." *Biometrika*, 1928, 175–240.

Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, 2018.

O'Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books, 2016.

Peterson, W. Wesley, Theodore G. Birdsall, and W. C. Fox. "The Theory of Signal Detectability." *Transactions of the IRE* 4, no. 4 (1954): 171–212.

Tanner Jr., Wilson P., and John A. Swets. "A Decision-Making Theory of Visual Detection." *Psychological Review* 61, no. 6 (1954): 401.

Wald, Abraham. "Contributions to the Theory of Statistical Estimation and Testing Hypotheses." *The Annals of Mathematical Statistics* 10, no. 4 (1939): 299–326.