

2

Fundamentals of prediction

Prediction is the art and science of leveraging patterns found in natural and social processes to conjecture about uncertain events. We use the word *prediction* broadly to refer to statements about things we don't know for sure *yet*, including but not limited to the outcome of future events.

Machine learning is to a large extent the study of algorithmic prediction. Before we can dive into machine learning, we should familiarize ourselves with prediction. Starting from first principles, we will motivate the goals of prediction before building up to a statistical theory of prediction.

We can formalize the goal of prediction problems by assuming a population of N instances with a variety of attributes. We associate with each instance two variables, denoted X and Y . The goal of prediction is to conjecture a plausible value for Y after observing X alone. But when is a prediction good? For that, we must quantify some notion of the quality of prediction and aim to optimize that quantity.

To start, suppose that for each variable X we make a deterministic prediction $f(X)$ by means of some prediction function f . A natural goal is to find a function f that makes the fewest number of incorrect predictions, where $f(X) \neq Y$, across the population. We can think of this function as a computer program that reads X as input and outputs a prediction $f(X)$ that we hope matches the value Y . For a fixed prediction function, f , we can sum up all of the errors made on the population. Dividing by the size of the population, we observe the average (or mean) error rate of the function.

Minimizing errors

Let's understand how we can find a prediction function that makes as few errors as possible on a given population in the case of binary prediction, where the variable Y has only two values.

Consider a population of Abalone, a type of marine snail with colorful shells featuring a varying number of rings. Our goal is to predict the sex, male or female, of the Abalone from the number of rings on the shell. We can tabulate the population of Abalone by counting for each possible number of rings, the number of male and female instances in the population.

From this way of presenting the population, it is not hard to compute the predictor that makes the fewest mistakes. For each value on the X-axis,

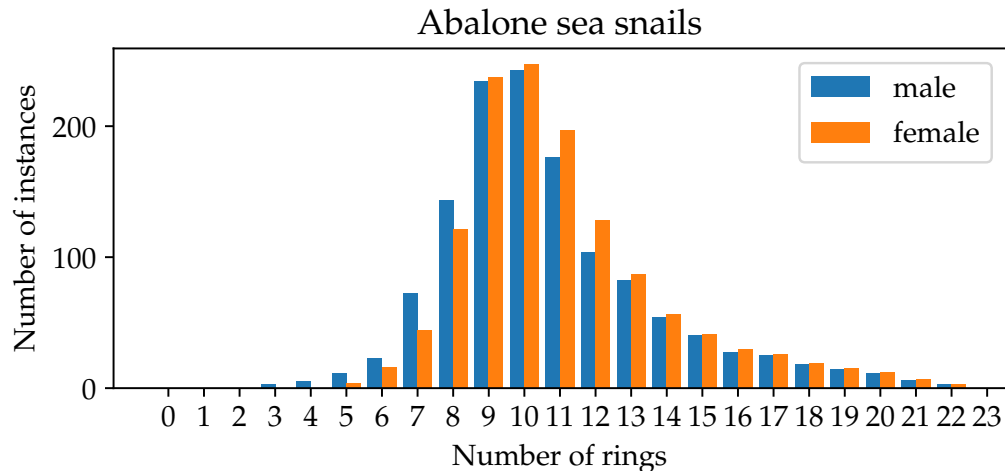


Figure 1: Predicting the sex of Abalone sea snails

we predict “female” if the number of female instances with this X-value is larger than the number of male instances. Otherwise, we predict “male” for the given X-value. For example, there’s a majority of male Abalone with seven rings on the shell. Hence, it makes sense to predict “male” when we see seven rings on a shell. Scrutinizing the figure a bit further, we can see that the best possible predictor is a *threshold function* that returns “male” whenever the number of rings is at most 8, and “female” whenever the number of rings is greater or equal to 9.

The number of mistakes our predictor makes is still significant. After all, most counts are pretty close to each other. But it’s better than random guessing. It uses whatever there is that we can say from the number of rings about the sex of an Abalone snail, which is just not much.

What we constructed here is called the *minimum error rule*. It generalizes to multiple attributes. If we had measured not only the number of rings, but also the length of the shell, we would repeat the analogous counting exercise over the two-dimensional space of all possible values of the two attributes.

The minimum error rule is intuitive and simple, but computing the rule exactly requires examining the entire population. Tracking down every instance of a population is not only intractable. It also defeats the purpose of prediction in almost any practical scenario. If we had a way of enumerating the X and Y value of all instances in a population, the prediction problem would be solved. Given an instance X we could simply look up the corresponding value of Y from our records.

What’s missing so far is a way of doing prediction that does not require

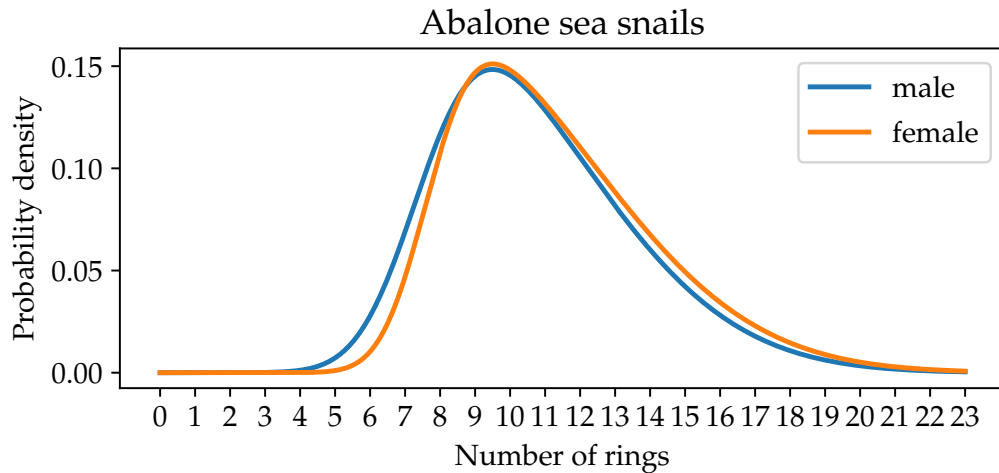


Figure 2: Representing Abalone population as a distribution

us to enumerate the entire population of interest.

Modeling knowledge

Fundamentally, what makes prediction without enumeration possible is *knowledge* about the population. Human beings organize and represent knowledge in different ways. In this chapter, we will explore in depth the consequences of one particular way to represent populations, specifically, as *probability distributions*.

The assumption we make is that we have knowledge of a probability distribution $p(x, y)$ over pairs of X and Y values. We assume that this distribution conceptualizes the “typical instance” in a population. If we were to select an instance uniformly at random from the population, what relations between its attributes might we expect? We expect that a uniform sample from our population would be the same as a sample from $p(x, y)$. We call such a distribution a *statistical model* or simply *model* of a population. The word *model* emphasizes that the distribution isn’t the population itself. It is, in a sense, a sketch of a population that we use to make predictions.

Let’s revisit our Abalone example in probabilistic form. Assume we know the distribution of the number of rings of male and female Abalone, as illustrated in the figure.

Both follow a skewed normal distribution described by three parameters each, a location, a scale, and a skew parameter. Knowing the distribution is to assume that we know these parameters. Although the specific numbers

won't matter for our example, let's spell them out for concreteness. The distribution for male Abalone has location 7.4, scale 4.48, and skew 3.12, whereas the distribution for female Abalone has location 7.63, scale 4.67, and skew 4.34. To complete the specification of the joint distribution over X and Y , we need to determine the relative proportion of males and females. Assume for this example that male and female Abalone are equally likely.

Representing the population this way, it makes sense to predict "male" whenever the probability density for male Abalone is larger than that for female Abalone. By inspecting the plot we can see that the density is higher for male snails up until 8 rings at which point it is larger for female instances. We can see that the predictor we derive from this representation is the same threshold rule that we had before.

We arrived at the same result without the need to enumerate and count all possible instances in the population. Instead, we recovered the minimum error rule from knowing only 7 parameters, three for each conditional distribution, and one for the balance of the two classes.

Modeling populations as probability distributions is an important step in making prediction algorithmic. It allows us to represent populations succinctly, and gives us the means to make predictions about instances we haven't encountered.

Subsequent chapters extend these fundamentals of prediction to the case where we don't know the exact probability distribution, but only have a random sample drawn from the distribution. It is tempting to think about machine learning as being all about *that*, namely what we do with a sample of data drawn from a distribution. However, as we learn in this chapter, many fundamentally important questions arise even if we have full knowledge of the population.

Prediction from statistical models

Let's proceed to formalize prediction assuming we have full knowledge of a statistical model of the population. Our first goal is to formally develop the minimum error rule in greater generality.

We begin with binary prediction where we suppose Y has two alternative values, 0 and 1. Given some measured information X , our goal is to conjecture whether Y equals zero or one.

Throughout we assume that X and Y are random variables drawn from a joint probability distribution. It is convenient both mathematically and conceptually to specify the joint distribution as follows. We assume that Y has *a priori* (or *prior*) probabilities:

$$p_0 = \mathbb{P}[Y = 0], \quad p_1 = \mathbb{P}[Y = 1]$$

That is, we assume we know the proportion of instances with $Y = 1$ and $Y = 0$ in the population. We'll always model available information as being a random vector X with support in \mathbb{R}^d . Its distribution depends on whether Y is equal to zero or one. In other words, there are two different statistical models for the data, one for each value of Y . These models are the conditional probability densities of X given a value y for Y , denoted $p(x | Y = y)$. This density function is often called a *generative model* or *likelihood function* for each scenario.

Example: signal versus noise

For a simple example with more mathematical formalism, suppose that when $Y = 0$ we observe a scalar $X = \omega$ where ω is unit-variance, zero mean Gaussian noise $\omega \sim \mathcal{N}(0, 1)$. Recall that the Gaussian distribution of mean μ and variance σ^2 is given by the density $\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$.

Suppose when $Y = 1$, we would observe $X = s + \omega$ for some scalar s . That is, the conditional densities are

$$\begin{aligned}p(x | Y = 0) &= \mathcal{N}(0, 1), \\p(x | Y = 1) &= \mathcal{N}(s, 1).\end{aligned}$$

The larger the shift s is, the easier it is to predict whether $Y = 0$ or $Y = 1$. For example, suppose $s = 10$ and we observed $X = 11$. If we had $Y = 0$, the probability that the observation is greater than 10 is on the order of 10^{-23} , and hence we'd likely think we're in the alternative scenario where $Y = 1$. However, if s were very close to zero, distinguishing between the two alternatives is rather challenging. We can think of a small difference s that we're trying to detect as a *needle in a haystack*.

Prediction via optimization

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.

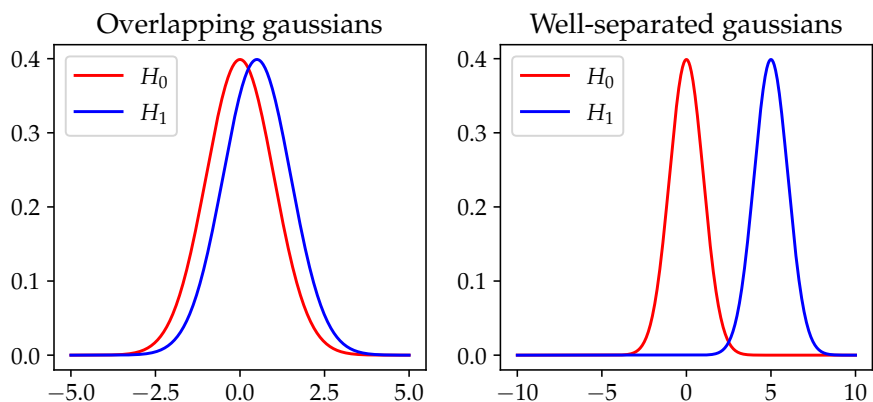


Figure 3: Illustration of shifted Gaussians

Predictors and labels

A *predictor* is a function $\hat{Y}(x)$ that maps an input x to a prediction $\hat{y} = \hat{Y}(x)$. The prediction \hat{y} is also called a *label* for the point x . The target variable Y can be both real valued or discrete. When Y is a discrete random variable, each different value it can take on is called a *class* of the prediction problem.

To ease notation, we take the liberty to write \hat{Y} as a shorthand for the random variable $\hat{Y}(X)$ that we get by applying the prediction function \hat{Y} to the random variable X .

The most common case we consider through the book is binary prediction, where we have two classes, 0 and 1. Sometimes it's mathematically convenient to instead work with the numbers -1 and 1 for the two classes.

In most cases we consider, labels are scalars that are either discrete or real-valued. Sometimes it also makes sense to consider vector-valued predictions and target variables.

The creation and encoding of suitable labels for a prediction problem is an important step in applying machine learning to real world problems. We will return to it multiple times.

Loss functions and risk

The final ingredient in our formal setup is a *loss function* which generalizes the notion of an error that we defined as a mismatch between prediction and target value.

A *loss function* takes two inputs, \hat{y} and y , and returns a real number $loss(\hat{y}, y)$ that we interpret as a quantified loss for predicting \hat{y} when the target is y . A loss could be negative in which case we think of it as a reward.

A prediction error corresponds to the loss function $loss(\hat{y}, y) = \mathbb{1}\{\hat{y} \neq y\}$ that indicates disagreement between its two inputs. Loss functions give us modeling flexibility that will become crucial as we apply this formal setup throughout this book.

An important notion is the expected loss of a predictor taken over a population. This construct is called *risk*.

Definition 1. We define the risk associated with \hat{Y} to be

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

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

Now that we defined risk, our goal is to determine which decision rule minimizes risk. Let's get a sense for how we might go about this.

In order to minimize 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.

Lemma 1. We claim that the optimal predictor is given by

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

This rule corresponds to the intuitive rule we derived when thinking about how to make predictions over the population. For a fixed value of the data $X = x$, we compare the frequency of which $Y = 1$ occurs to which $Y = 0$ occurs. If this frequency exceeds some threshold that is defined by our loss function, then we set $\hat{Y}(x) = 1$. Otherwise, we set $\hat{Y}(x) = 0$.

Proof. To see why this rule is optimal, we make use of the law of iterated expectation:

$$\mathbb{E}[loss(\hat{Y}(X), Y)] = \mathbb{E} \left[\mathbb{E} [loss(\hat{Y}(X), Y) \mid X] \right].$$

Here, the outer expectation is over a random draw of X and the inner expectation samples Y conditional on X . Since there are no constraints on the predictor \hat{Y} , we can minimize the expression by minimizing the inner expectation independently for each possible setting that X can assume.

Indeed, for a fixed value x , we can expand the expected loss for each of the two possible predictions:

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

The optimal assignment for this x is to set $\hat{Y}(x) = 1$ whenever the second expression is smaller than the first. Writing out this inequality and rearranging gives us the rule specified in the lemma. \square

Probabilities of the form $\mathbb{P}[Y = y \mid X = x]$, as they appeared in the lemma, are called *posterior* probability.

We can relate them to the likelihood function via Bayes rule:

$$\mathbb{P}[Y = y \mid X = x] = \frac{p(x \mid Y = y)p_y}{p(x)},$$

where $p(x)$ is a density function for the marginal distribution of X .

When we use posterior probabilities, we can rewrite the optimal predictor as

$$\hat{Y}(x) = \mathbb{1} \left\{ \frac{p(x \mid Y = 1)}{p(x \mid Y = 0)} \geq \frac{p_0(\text{loss}(1,0) - \text{loss}(0,0))}{p_1(\text{loss}(0,1) - \text{loss}(1,1))} \right\}.$$

This rule is an example of a likelihood ratio test.

Definition 2. *The likelihood ratio is the ratio of the likelihood functions:*

$$\mathcal{L}(x) := \frac{p(x \mid Y = 1)}{p(x \mid Y = 0)}$$

A likelihood ratio test (LRT) is a predictor of the form

$$\hat{Y}(x) = \mathbb{1}\{\mathcal{L}(x) \geq \eta\}$$

for some scalar threshold $\eta > 0$.

If we denote the optimal threshold value

$$\eta = \frac{p_0(\text{loss}(1,0) - \text{loss}(0,0))}{p_1(\text{loss}(0,1) - \text{loss}(1,1))}, \quad (1)$$

then the predictor that minimizes the risk is the likelihood ratio test

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

A LRT naturally partitions the sample space in two regions:

$$\begin{aligned} \mathcal{X}_0 &= \{x \in \mathcal{X} : \mathcal{L}(x) \leq \eta\} \\ \mathcal{X}_1 &= \{x \in \mathcal{X} : \mathcal{L}(x) > \eta\}. \end{aligned}$$

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 $h : \mathcal{X} \rightarrow \mathbb{R}$ 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 predictor

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

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

$$\hat{Y}_{\log}(x) = \mathbb{1}\{\log p(x | Y = 1) - \log p(x | Y = 0) \geq \log(\eta)\},$$

as it is often more convenient or numerically 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 predictor. Hence, we don't need to know the conditional densities exactly and can still compute the optimal predictor. For example, suppose the true partitioning of the real line under an LRT is

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

Setting the threshold to $t = 0$, the functions $h(x) = x$ or $h(x) = x^3$ give the same predictor, 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

$$\begin{aligned} p(X | Y = 0) &= \mathcal{N}(0, 1), \\ p(X | Y = 1) &= \mathcal{N}(s, 1), \end{aligned}$$

and assume that the prior probability of $Y = 1$ is very small, say, $p_1 = 10^{-6}$. Suppose that if we declare $\hat{Y} = 0$, we do not pay a cost. If we declare $\hat{Y} = 1$ but are wrong, we incur a cost of 100. But if we guess $\hat{Y} = 1$ and it is actually true that $Y = 1$, we actually gain a reward of 1,000,000. That is $loss(0, 0) = 0$, $loss(0, 1) = 0$, $loss(1, 0) = 100$, and $loss(1, 1) = -1,000,000$.

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,

$$\log p(x | Y = 1) - \log p(x | Y = 0) = -\frac{1}{2}(x - s)^2 + \frac{1}{2}x^2 = sx - \frac{1}{2}s^2$$

Hence, the optimal predictor is to declare

$$\hat{Y} = \mathbb{1} \left\{ sx > \frac{1}{2}s^2 + \log(\eta) \right\} .$$

The optimal rule here is *linear*. Moreover, the rule divides the space into two open intervals. While the entire real line lies in the union of these two intervals, it is exceptionally unlikely to ever see an x larger than $|s| + 5$. Hence, even if our predictor were incorrect in these regions, the risk would still be nearly optimal as these terms have almost no bearing on our expected risk!

Maximum a posteriori and maximum likelihood

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, many important prediction rules end up being equivalent to LRTs. Shortly, we'll see an optimization problem that speaks to the power of LRTs. But before that, we can already show that the well known *maximum likelihood* and *maximum a posteriori* predictors are both LRTs.

The expected error of a predictor is the expected number of times we declare $\hat{Y} = 0$ (resp. $\hat{Y} = 1$) when $\hat{Y} = 1$ (resp. $\hat{Y} = 0$) is true. Minimizing the error is equivalent to minimizing the risk with cost $loss(0, 0) = loss(1, 1) = 0$, $loss(1, 0) = loss(0, 1) = 1$. The optimum predictor is hence a likelihood ratio test. In particular,

$$\hat{Y}(x) = \mathbb{1} \left\{ \mathcal{L}(x) \geq \frac{p_0}{p_1} \right\} .$$

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

$$\hat{Y}(x) = \arg \max_{y \in \{0,1\}} \mathbb{P}[Y = y | X = x] .$$

Recall that the expression $\mathbb{P}[Y = y | X = x]$ is called the posterior probability of $Y = y$ given $X = x$. And this rule is hence referred to as the *maximum a posteriori* (MAP) rule.

As we discussed above, the expression $p(x | Y = y)$ is called the *likelihood* of the point x given the class $Y = y$. A maximum likelihood rule would set

$$\hat{Y}(x) = \arg \max_y p(x | Y = y) .$$

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

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

Types of errors and successes

Let $\hat{Y}(x)$ denote any predictor mapping into $\{0, 1\}$. Binary predictions can be right or wrong in four different ways summarized by the *confusion table*.

Table 1: Confusion table

	$Y = 0$	$Y = 1$
$\hat{Y} = 0$	true negative	false negative
$\hat{Y} = 1$	false positive	true positive

Taking expected values over the populations give us four corresponding *rates* that are characteristics of a predictor.

1. **True Positive Rate:** $TPR = \mathbb{P}[\hat{Y}(X) = 1 \mid Y = 1]$. Also known as *power, sensitivity, probability of detection, or recall*.
2. **False Negative Rate:** $FNR = 1 - TPR$. Also known as *type II error or probability of missed detection*.
3. **False Positive Rate:** $FPR = \mathbb{P}[\hat{Y}(X) = 1 \mid Y = 0]$. Also known as *size or type I error or probability of false alarm*.
4. **True Negative Rate:** $TNR = 1 - FPR$, the probability of declaring $\hat{Y} = 0$ given $Y = 0$. This is also known as *specificity*.

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

1. **Precision:** $P[Y = 1 \mid \hat{Y}(X) = 1]$. This is equal to $(p_1 TPR) / (p_0 FPR + p_1 TPR)$.
2. **F1-score:** F_1 is the harmonic mean of precision and recall. We can write this as

$$F_1 = \frac{2TPR}{1 + TPR + \frac{p_0}{p_1}FPR}$$

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 labels 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 think of risk minimization as optimizing a balance between TPR and FPR:

$$R[\hat{Y}] := \mathbb{E}[\text{loss}(\hat{Y}(X), Y)] = \alpha \text{FPR} - \beta \text{TPR} + \gamma,$$

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

Other cost functions might try to balance TPR versus FPR in other ways. Which pairs of (FPR, TPR) are achievable?

ROC curves

True and false positive rate lead to another fundamental notion, called the *receiver operating characteristic (ROC) curve*.

The ROC curve is a property of the joint distribution (X, Y) and shows for every possible value $\alpha \in [0, 1]$ the best possible true positive rate that we can hope to achieve with any predictor that has false positive rate α . As a result the ROC curve is a curve in the FPR-TPR plane. It traces out the maximal TPR for any given FPR. Clearly the ROC curve contains values $(0, 0)$ and $(1, 1)$, which are achieved by constant predictors that either reject or accept all inputs.

We will now show, in a celebrated result by Neyman and Pearson, that the ROC curve is given by varying the threshold in the likelihood ratio test from negative to positive infinity.

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 predictor is a likelihood ratio test. Third, it introduces the notion of probabilistic predictors.

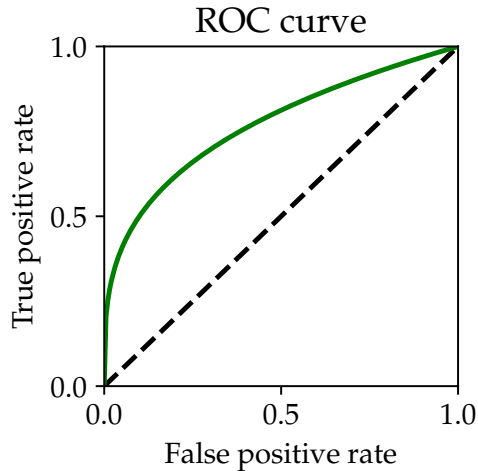


Figure 4: Example of an ROC curve

Suppose we want to maximize true positive rate subject to an upper bound on the false positive rate. That is, we aim to solve the optimization problem:

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

Let's optimize over *probabilistic predictors*. A probabilistic predictor Q returns 1 with probability $Q(x)$ and 0 with probability $1 - Q(x)$. With such rules, we can rewrite our optimization problem as:

$$\begin{aligned} & \text{maximize}_Q && \mathbb{E}[Q(X) \mid Y = 1] \\ & \text{subject to} && \mathbb{E}[Q(X) \mid Y = 0] \leq \alpha \\ & && \forall x: Q(x) \in [0, 1] \end{aligned}$$

Lemma 2. Neyman-Pearson Lemma. *Suppose the likelihood functions $p(x|y)$ are continuous. Then the optimal probabilistic predictor 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 likelihood ratio tests. 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 for any LRT, we can find a loss function for which it is optimal. We will prove the lemma by constructing such a problem, and using the associated condition of optimality.

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

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

has $\text{FPR} = \alpha$. Such an LRT exists because we assumed our likelihoods were continuous. Let β denote the TPR of Q_η .

We claim that Q_η is optimal for the risk minimization problem corresponding to the loss function

$$\text{loss}(1,0) = \frac{\eta p_1}{p_0}, \text{loss}(0,1) = 1, \text{loss}(1,1) = 0, \text{loss}(0,0) = 0.$$

Indeed, recalling Equation 1, the risk minimizer for this loss function corresponds to a likelihood ratio test with threshold value

$$\frac{p_0(\text{loss}(1,0) - \text{loss}(0,0))}{p_1(\text{loss}(0,1) - \text{loss}(1,1))} = \frac{p_0 \text{loss}(1,0)}{p_1 \text{loss}(0,1)} = \eta.$$

Moreover, under this loss function, the risk of a predictor Q equals

$$\begin{aligned} R[Q] &= p_0 \text{FPR}(Q) \text{loss}(1,0) + p_1 (1 - \text{TPR}(Q)) \text{loss}(0,1) \\ &= p_1 \eta \text{FPR}(Q) + p_1 (1 - \text{TPR}(Q)). \end{aligned}$$

Now let Q be any other predictor with $\text{FPR}(Q) \leq \alpha$. We have by the optimality of Q_η that

$$\begin{aligned} p_1 \eta \alpha + p_1 (1 - \beta) &\leq p_1 \eta \text{FPR}(Q) + p_1 (1 - \text{TPR}(Q)) \\ &\leq p_1 \eta \alpha + p_1 (1 - \text{TPR}(Q)), \end{aligned}$$

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

Properties of ROC curves

A specific randomized predictor that is useful for analysis combines two other rules. Suppose predictor one yields $(\text{FPR}^{(1)}, \text{TPR}^{(1)})$ and the second rule achieves $(\text{FPR}^{(2)}, \text{TPR}^{(2)})$. 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 predictor with $(\text{FPR}, \text{TPR}) = (p\text{FPR}^{(1)} + (1 - p)\text{FPR}^{(2)}, p\text{TPR}^{(1)} + (1 - p)\text{TPR}^{(2)})$. Using this rule lets us prove several properties of ROC curves.

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

Proof. This proposition follows because the point $(0,0)$ is achieved when the threshold $\eta = \infty$ in the likelihood ratio test, corresponding to the constant 0 predictor. The point $(1,1)$ is achieved when $\eta = 0$, corresponding to the constant 1 predictor. \square

The Neyman-Pearson Lemma gives us a few more useful properties.

Proposition 2. *The ROC must lie above the main diagonal.*

Proof. To see why this proposition is true, fix some $\alpha > 0$. Using a randomized rule, we can achieve a predictor with $\text{TPR} = \text{FPR} = \alpha$. But the Neyman-Pearson LRT with FPR constrained to be less than or equal to α achieves true positive rate greater than or equal to the randomized rule. \square

Proposition 3. *The ROC curve is concave.*

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

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

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

Example: the needle one more time

Consider again the *needle in a haystack* example, where $p(x | Y = 0) = \mathcal{N}(0, \sigma^2)$ and $p(x | Y = 1) = \mathcal{N}(s, \sigma^2)$ with s a positive scalar. The optimal predictor is to declare $\hat{Y} = 1$ when X is greater than $\gamma := \frac{s}{2} + \frac{\sigma^2 \log \eta}{s}$. Hence we have

$$\begin{aligned} \text{TPR} &= \int_{\gamma}^{\infty} p(x | Y = 1) dx = \frac{1}{2} \operatorname{erfc} \left(\frac{\gamma - s}{\sqrt{2}\sigma} \right) \\ \text{FPR} &= \int_{\gamma}^{\infty} p(x | Y = 0) dx = \frac{1}{2} \operatorname{erfc} \left(\frac{\gamma}{\sqrt{2}\sigma} \right). \end{aligned}$$

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

Area under the ROC curve

Oftentimes in information retrieval and machine learning, the term ROC curve is overloaded to describe the achievable FPR-TPR pairs that we get by varying the threshold t in any predictor $\hat{Y}(x) = \mathbb{1}\{R(x) > t\}$. Note such curves must lie below the ROC curves that are traced out by the optimal

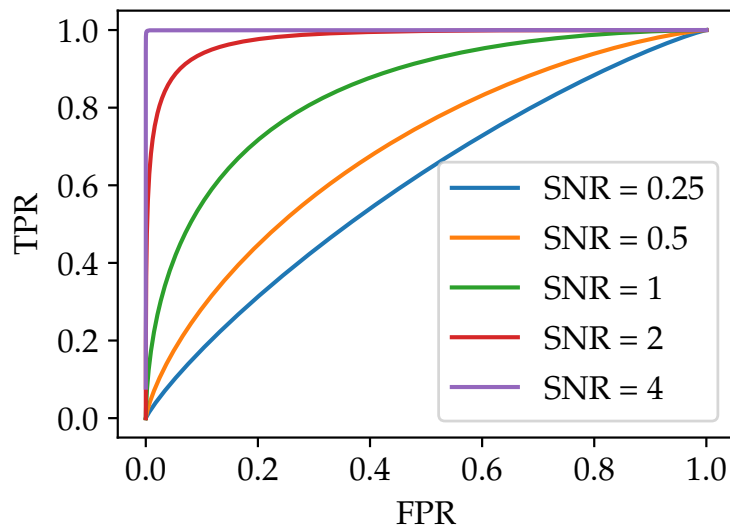


Figure 5: The ROC curves for various signal to noise ratios in the needle in the haystack problem.

likelihood ratio test, but may approximate the true ROC curves in many cases.

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.

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.

Decisions that discriminate

The purpose of prediction is almost always decision making. We build predictors to guide our decision making by acting on our predictions. 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.

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,

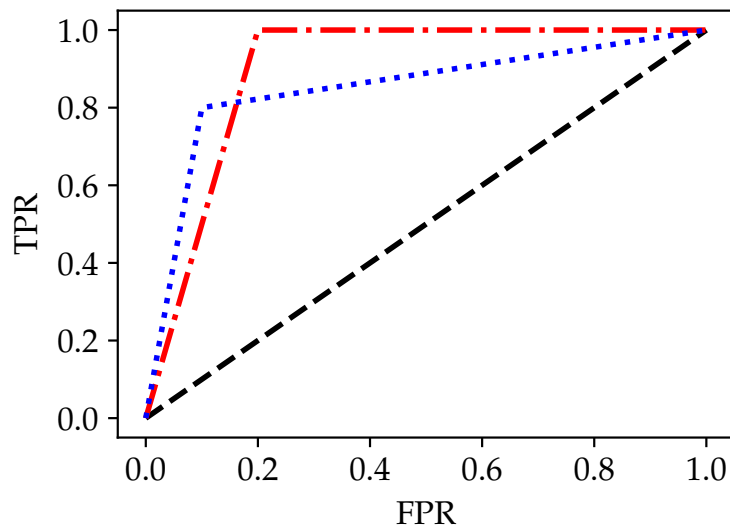


Figure 6: Two ROC curves with the same AUC. Note that if we constrain FPR to be less than 10%, for the blue curve, TPR can be as high as 80% whereas it can only reach 50% for the red.

however, this decision boundary is not just a technical artifact. Rather it has moral and legal significance.

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

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

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 $\hat{Y} = \mathbb{1}\{R > t\}$ 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 :

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

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 $\mathbb{P}[\hat{Y} = 1]$
- Error rates $\mathbb{P}[\hat{Y} = 0 \mid Y = 1]$ and $\mathbb{P}[\hat{Y} = 1 \mid Y = 0]$
- 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 2: 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.^{4,5}

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.

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. Recidivism refers to a person's relapse into criminal behavior. In the United States, a defendant may either be detained or released on bail prior to the trial in court depending on various factors. Judges may detain defendants in part based on this score.

Investigative journalists at ProPublica found that Black defendants 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.

Chapter notes

The theory we covered in this chapter is also called *detection theory* and *decision theory*. Similarly, what we call a predictor throughout has various different names, such as *decision rule* or *classifier*.

The elementary detection theory covered in this chapter has not changed much at all since the 1950s and is essentially considered a "solved problem". Neyman and Pearson invented the likelihood ratio test⁸ 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*

*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.

Of course, how decisions and predictions are applied and interpreted remains an active research topic. 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,¹ Broussard,¹⁶ Eubanks,¹⁷ Noble,¹⁸ and O’Neil.¹⁹ The technical material in our section on discrimination follows Chapter 2 in the textbook by Barocas, Hardt, and Narayanan.³

The abalone example was derived from data available at the UCI Machine Learning Repository, which we will discuss in more detail in Chapter 8. We modified the data to ease exposition. The actual data does not have an equal number of male and female instances, and the optimal predictor is not exactly a threshold function.

Bibliography

- ¹ Ruha Benjamin. *Race after Technology*. Polity, 2019.
- ² Ben Hutchinson and Margaret Mitchell. 50 years of test (un) fairness: Lessons for machine learning. In *Conference on Fairness, Accountability, and Transparency*, pages 49–58, 2019.
- ³ Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. <http://www.fairmlbook.org>.
- ⁴ Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Innovations in Theoretical Computer Science*, 2017.
- ⁵ Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.
- ⁶ Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica*, May 2016.
- ⁷ William Dieterich, Christina Mendoza, and Tim Brennan. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. Technical report, 2016.
- ⁸ Jerzy Neyman and Egon S. Pearson. On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika*, pages 175–240, 1928.
- ⁹ Jerzy Neyman 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(694-706):289–337, 1933.
- ¹⁰ Abraham Wald. Contributions to the theory of statistical estimation and testing hypotheses. *The Annals of Mathematical Statistics*, 10(4):299–326, 1939.

- ¹¹ Dimitri P. Bertsekas and John N. Tsitsiklis. *Introduction to Probability*. Athena Scientific, 2nd edition, 2008.
- ¹² W. Wesley Peterson, Theodore G. Birdsall, and William C. Fox. The theory of signal detectability. *Transactions of the IRE*, 4(4):171–212, 1954.
- ¹³ Wilson P. Tanner Jr. and John A. Swets. A decision-making theory of visual detection. *Psychological Review*, 61(6):401, 1954.
- ¹⁴ Chao Kong Chow. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, (4):247–254, 1957.
- ¹⁵ Wilbur H. Highleyman. Linear decision functions, with application to pattern recognition. *Proceedings of the IRE*, 50(6):1501–1514, 1962.
- ¹⁶ Meredith Broussard. *Artificial Unintelligence: How Computers Misunderstand the World*. MIT Press, 2018.
- ¹⁷ Virginia Eubanks. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin’s Press, 2018.
- ¹⁸ Safiya Umoja Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, 2018.
- ¹⁹ Cathy O’Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books, 2016.