Classification

Simply put, the goal of classification is to determine a plausible value for an unknown variable Y given an observed variable X. For example, we might try to *predict* whether a loan applicant will pay back her loan by looking at various characteristics such as credit history, income, and net worth. Classification also applies in situations where the variable Y does not refer to an event that lies in the future. For example, we can try to determine if an image contains a *cat* by looking at the set of pixels encoding the image. This practice is also called *object recognition* or *image classification*. Object recognition might not even seem like a statistical problem, yet statistical methods came to be the method of choice for many important pattern recognition tasks in computer vision.

Supervised learning

A classifier is a mapping from the space of possible values for X to the space of values that the target variable Y can assume. *Supervised learning* is the prevalent method for constructing classifiers from observed data. The essential idea is very simple. Suppose we have labeled data, also called *training examples*, of the form $(x_1, y_1), ..., (x_n, y_n)$, where each *example* is a pair (x_i, y_i) of an *instance* x_i and a *label* y_i.

Instances are usually arranged as vectors of some dimension. You can think of them as arrays with numbers in them. In a classification problem, labels typically come from a discrete set such as $\{-1, 1\}$ in the case of binary classification. We interpret these labels as partitioning the set of instances into positive and negative instances depending on their label.[1] We can interpret such a classifier as a *decision rule* by equating a positive label with *acceptance* and a negative label with *rejection*.

[1] Multi-class prediction is the generalization to label sets with more than two values.

In a *regression* problem, the label y is typically a real number. The goal is no longer to predict the exact value of y but rather to be close to it. The tools to solve classification and regression problems in practice are very similar. In both cases, roughly the same optimization

approach is used to find a classifier f that maps an instance x to a label $\hat{y} = f(x)$ that we hope agrees with the correct label. This optimization process is often called *training*; its specifics are irrelevant for this chapter.

To turn supervised learning into a statistical problem, we assume that there is an underlying distribution from which the data were drawn. The distribution is fixed and each example is drawn independently of the others. We can express this underlying distribution as a pair of random variables (X, Y). For example, our training examples might be responses from a survey. Each survey participant is chosen independently at random from a fixed sampling frame that represents an underlying population. As we discussed in the introduction, the goal of supervised learning is to identify meaningful patterns in the population that aren't just artifacts of the sample.

At the population level, we can interpret our classifier as a random variable by considering $\hat{Y} = f(X)$. In doing so, we overload our terminology slightly by using the word *classifier* for both the random variable $\hat{Y}$ and mapping f. The distinction is mostly irrelevant for this chapter as we will focus on the statistical properties of the joint distribution of the data and the classifier, which we denote as a tuple of three random variables $(X, Y, \hat{Y})$. For now, we ignore how $\hat{Y}$ was learned from a finite sample, what the functional form of the classifier is, and how we estimate various statistical quantities from finite samples. While finite sample considerations are fundamental to machine learning, they are often not specific to the conceptual and technical questions around fairness that we will discuss.

Statistical classification criteria

What makes a classifier *good* for an application and how do we choose one out of many possible classifiers? This question often does not have a fully satisfying answer, but some formal criteria can help highlight different qualities of a classifier that can inform our choice.

Perhaps the most well known property of a classifier $\hat{Y}$ is its *accuracy* defined as $\mathbb{P}\{Y = \hat{Y}\}$, the probability of correctly predicting the target variable. It is common practice to apply the classifier that achieves highest accuracy among those available to us.[2]

[2] We typically don't know the classifier that maximizes accuracy among all possible classifiers, but rather we only have access to those that we can find with effective training procedures.

Accuracy is easy to define, but misses some important aspects. A classifier that always predicts *no traffic fatality in the next year* might have high accuracy, simply because individual accidents are highly unlikely. However, it's a constant function that has no value in assessing the risk that an individual experiences a fatal traffic accident.

Many other formal classification criteria highlight different aspects of a classifier. In a binary classification setting, we can consider the

conditional probability $\mathbb{P}\{\text{event} \mid \text{condition}\}$ for various different settings.

Table 1: Common classification criteria

Event	Condition	Resulting notion ($\mathbb{P}\{\text{event} \mid \text{condition}\}$)
$\hat{Y} = 1$	$Y = 1$	True positive rate, recall
$\hat{Y} = 0$	$Y = 1$	False negative rate
$\hat{Y} = 1$	$Y = 0$	False positive rate
$\hat{Y} = 0$	$Y = 0$	True negative rate

To be clear, the true positive rate corresponds to the frequency with which the classifier correctly assigns a positive label to a positive instance. We call this a *true positive*. The other terms *false positive*, *false negative*, and *true negative* derive analogously from the respective definitions.

It is not important to memorize all these terms. They do, however, come up regularly in the classification setting so the table might come in handy.

Another family of classification criteria arises from swapping event and condition. We'll only highlight two of the four possible notions.

Table 2: Additional classification criteria

Event	Condition	Resulting notion ($\mathbb{P}\{\text{event} \mid \text{condition}\}$)
$Y = 1$	$\hat{Y} = 1$	Positive predictive value, precision
$Y = 0$	$\hat{Y} = 0$	Negative predictive value

We'll return to these criteria later on when we explore some of their properties and relationships.

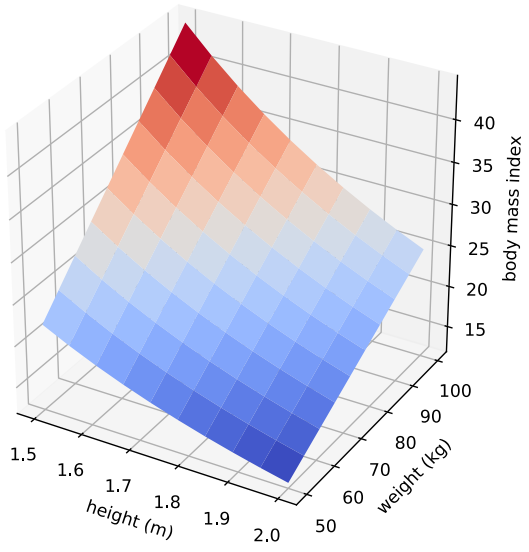
Score functions

Classification is often attacked by first solving a regression problem to summarize the data in a single real-valued variable. We will refer to such a variable as *score*. We can turn a score into a classifier by thresholding it somewhere on the real line.

For an illustrative example consider the well-known body mass index which summarizes *weight* and *height* of a person into a single real number. In our formal notation, the features are $X = (H, W)$ where H denotes height in meters and W denotes weight in kilograms. The body mass index corresponds to the score function $R = W/H^2$.

We could interpret the body mass index as measuring risk of heart disease. Thresholding it at the value 27, we might decide that indi-

Figure 1: Plot of the body mass index.



viduals with a body mass index above this value are at risk of developing heart disease while others are not. It does not take a medical degree to suspect that the resulting classifier may not be very accurate[3]. The body mass index has a number of known issues leading to errors when used for classification. We won't go into detail, but it's worth noting that these classification errors can systematically align with certain demographic groups. For instance, the body mass index tends to be inflated as a risk measure for taller people (due to its scaling issues).

[3] In fact, it seems to be quite poor.

Score functions need not follow simple algebraic formulas such as the body mass index. In most cases, score functions are built by fitting regression models against historical data. Think of a credit score, as is common in some countries, which can be used to accept or deny loan applicants based on the score value. We will revisit this example in detail later.

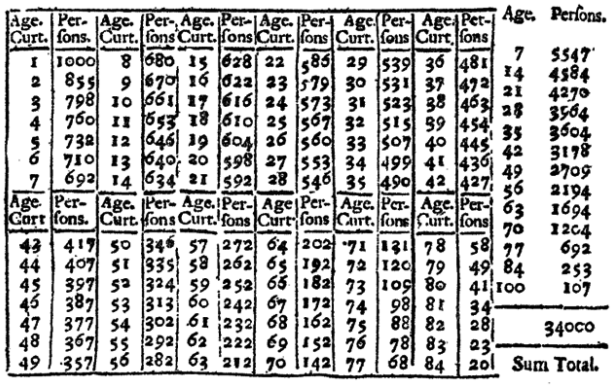
The conditional expectation

A natural score function is the expectation of the target variable Y conditional on the features X we have observed. We can write this score as $R = r(X)$ where $r(x) = \mathbb{E}[Y \mid X = x]$, or more succinctly, $R = \mathbb{E}[Y \mid X]$. In a sense, this score function gives us the *best guess* for the target variable given the observations we have. We can think of the conditional expectation as a *lookup table* that gives us for each setting of features the frequency of positive outcomes given these features.[4]

[4] We can make this statement more precise. This score is sometimes called the *Bayes optimal score* or *Bayes optimal score* as it minimizes the squared error $\mathbb{E}(g(X) - R)^2$ among all functions $g(X)$.

Such lookup tables have a long and fascinating history in applications of risk assessment such as insurance pricing.[5] One of the earliest examples is Halley's *life table* from 1693 that was used to estimate the life expectancy an individual in order to accurately price certain annuities.

[5] Dan Bouk, *How Our Days Became Numbered: Risk and the Rise of the Statistical Individual* (University of Chicago Press, 2015).



Age. Curt.	Per-sons.	Age. Curt.	Per-sons	Age. Curt.	Per-sons	Age. Curt.	Per-sons	Age. Curt.	Per-sons	Age. Curt.	Per-sons
1	1000	8	680	15	628	22	586	29	539	36	481
2	855	9	670	16	622	23	579	30	531	37	472
3	798	10	661	17	616	24	573	31	523	38	463
4	760	11	653	18	610	25	567	32	515	39	454
5	732	12	646	19	604	26	560	33	507	40	445
6	710	13	640	20	598	27	553	34	499	41	436
7	692	14	634	21	592	28	546	35	490	42	427
Age. Curt	Per-sons.	Age. Curt.	Per-sons	Age. Curt.	Per-sons	Age Curt	Per-sons	Age. Curt.	Per-sons	Age. Curt.	Per-sons
43	417	50	346	57	272	64	202	71	131	78	58
44	407	51	335	58	262	65	192	72	120	79	49
45	397	52	324	59	252	66	182	73	109	80	41
46	387	53	313	60	242	67	172	74	98	81	34
47	377	54	302	61	232	68	162	75	88	82	28
48	367	55	292	62	222	69	152	76	78	83	23
49	357	56	282	63	212	70	142	77	68	84	20

Age.	Persons.
7	5547
14	4584
21	4270
28	3564
35	3604
42	3178
49	2709
56	2194
63	1694
70	1204
77	692
84	253
100	107
	34000
	Sum Total.

Figure 2: Halley's life table (1693)

The conditional expectation also makes sense for our example of scoring risk of heart disease. What it would do here is to tell us for every setting of weight (say, rounded to the nearest kg unit) and every physical height (rounded to the nearest cm unit), the incidence rate of heart disease among individuals with these values of weight and height. The target variable in this case is a binary indicator of heart disease. So, $r((176, 68))$ would be the incidence rate of heart disease among individuals who are 1.76m tall and weigh 68kg. Intuitively, we can think of the conditional expectation as a big lookup table of incidence rates given some setting of characteristics.

The conditional expectation is likely more useful as a risk measure of heart disease than the body mass index we saw earlier. After all, the conditional expectation directly reflects the incidence rate of heart disease given the observed characteristics, while the body mass index is a general-purpose summary statistic.

That said, we can still spot a few issues with this score function. First, our definition of target variable was a bit fuzzy, lumping together all sorts of different kinds of heart disease with different characteristics. Second, in order to actually compute the conditional expectation in practice, we would have to collect incidence rate statistics by height and weight. These data points would only tell us about historical incidence rates. The extent to which they can tell us about future cases of heart disease is somewhat unclear. If our data comes from a time where people generally smoked more cigarettes, our statistics might overestimate future incidence rates. There are numerous other features that are relevant for the prediction of heart disease, including age and gender, but they are neglected in our data.

We could include these additional features in our data; but as we increase the number of features, estimating the conditional expectation becomes increasingly difficult. Any feature set partitions the population into demographics. The more features we include, the fewer data points we can collect in each subgroup. As a result, the conditional expectation is generally hard to estimate in *high-dimensional* settings, where we have many attributes.

From scores to classifiers

We just saw how we can turn a score function into a discrete classifier by discretizing its values into buckets. In the case of a binary classifier, this corresponds to choosing a threshold t so that when the score is above t our classifier outputs 1 (*accept*) and otherwise -1 (*reject*).[6] Each choice of the threshold defines one binary classifier. Which threshold should we choose?

[6] The choice of the values 1 and -1 is arbitrary. Any two distinct values will do.

The answer to this question is surprisingly subtle. Roughly speaking, which threshold we choose depends on our notion of utility for the resulting classifier and the problem we're trying to solve. Our notion of utility could be complex and depend on many different considerations.

In classification, it is common to oversimplify the problem quite a bit by summarizing all considerations of utility with just two numbers: a cost for accepting a negative instance (false positive) and a cost for rejecting a positive instance (false negative). If in our problem we face a high cost for false positives, we want to choose a higher threshold than in other applications where false negatives are costly.

The choice of a threshold and its resulting trade-off between true positive rate and false positive rate can be neatly visualized with the help of an *ROC curve*[7]. Note that true positive rate equals 1 - false negative rate.

[7] ROC stands for receiver operating characteristic.

The ROC curve serves another purpose. It can be used to eyeball how predictive our score is of the target variable. A common measure of predictiveness is the area under the curve, which is the probability that a random positive instance gets a score higher than a random negative instance. An area of 1/2 corresponds to random guessing, and an area of 1 corresponds to perfect classification, or more formally, the score equals the target. Known disadvantages[8] make *area under the curve* a tool that must be interpreted with caution.

[8] Steve Halligan, Douglas G. Altman, and Susan Mallett, "Disadvantages of Using the Area Under the Receiver Operating Characteristic Curve to Assess Imaging Tests: A Discussion and Proposal for an Alternative Approach," *European Radiology* 25, no. 4 (April 2015): 932–39.

Sensitive characteristics

In many classification tasks, the features X contain or implicitly encode sensitive characteristics of an individual. We will set aside the

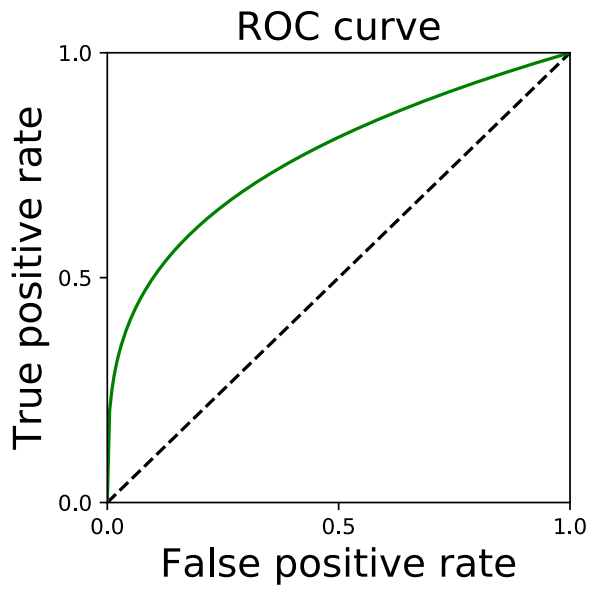


Figure 3: Example of an ROC curve. Each point on the solid curve is realized by thresholding the score function at some value. The dashed line shows the trade-offs achieved by randomly accepting an instance irrespective of its features with some probability $p \in [0, 1]$.

letter A to designate a discrete random variable that captures one or multiple sensitive characteristics[9]. Different settings of A correspond to different groups of the population. This notational choice is not meant to suggest that we can cleanly partition the set of features into two independent categories such as "neutral" and "sensitive". In fact, we will see shortly that sufficiently many seemingly neutral features can often give high accuracy predictions of sensitive characteristics. This should not be surprising. After all, if we think of A as the target variable in a classification problem, there is reason to believe that the remaining features would give a non-trivial classifier for A.

[9] Note that formally we can always represent any number of discrete sensitive attributes as a single discrete attribute whose support corresponds to each of the possible settings of the original attributes.

The choice of sensitive attributes will generally have profound consequences as it decides which groups of the population we highlight, and what conclusions we draw from our investigation. The taxonomy induced by discretization can on its own be a source of harm if it is too coarse, too granular, misleading, or inaccurate. Even the act of introducing a sensitive attribute on its own can be problematic. We will revisit this important discussion in the next chapter.

No fairness through unawareness

Some have hoped that removing or ignoring sensitive attributes would somehow ensure the impartiality of the resulting classifier. Unfortunately, this practice is usually somewhere on the spectrum between ineffective and harmful.

In a typical data set, we have many features that are slightly correlated with the sensitive attribute. Visiting the website `pinterest.com`,

for example, has a small statistical correlation with being female.[10]

The correlation on its own is too small to predict someone's gender with high accuracy. However, if numerous such features are available, as is the case in a typical browsing history, the task of predicting gender becomes feasible at high accuracy levels.

In other words, several features that are slightly predictive of the sensitive attribute can be used to build high accuracy classifiers for that attribute.

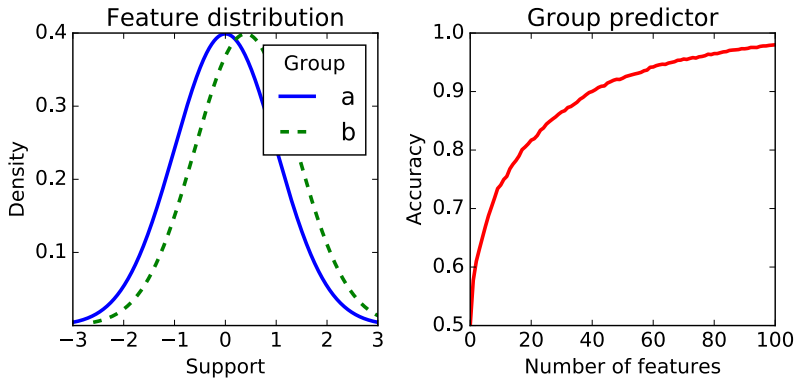


Figure 4: On the left, we see the distribution of a single feature that differs only very slightly between the two groups. In both groups the feature follows a normal distribution. Only the means are slightly different in each group. Multiple features like this can be used to build a high accuracy group membership classifier. On the right, we see how the accuracy grows as more and more features become available.

In large feature spaces sensitive attributes are generally *redundant* given the other features. If a classifier trained on the original data uses the sensitive attribute and we remove the attribute, the classifier will then find a redundant encoding in terms of the other features. This results in an essentially equivalent classifier, in the sense of implementing the same function.

To further illustrate the issue, consider a fictitious start-up that sets out to predict your income from your genome. At first, this task might seem impossible. How could someone's DNA reveal their income? However, we know that DNA encodes information about ancestry, which in turn correlates with income in some countries such as the United States. Hence, DNA can likely be used to predict income better than random guessing. The resulting classifier uses ancestry in an entirely implicit manner. Removing redundant encodings of ancestry from the genome is a difficult task that cannot be accomplished by removing a few individual genetic markers. What we learn from this is that machine learning can wind up building classifiers for sensitive attributes without explicitly being asked to, simply because it is an available route to improving accuracy.

Redundant encodings typically abound in large feature spaces. What about small hand-curated feature spaces? In some studies, features are chosen carefully so as to be roughly statistically indepen-

[10] As of August 2017, 58.9% of Pinterest's users in the United States were female. See here (Retrieved 3-27-2018)

dent of each other. In such cases, the sensitive attribute may not have good redundant encodings. That does not mean that removing it is a good idea. Medication, for example, sometimes depends on race in legitimate ways if these correlate with underlying causal factors.[11] Forcing medications to be uncorrelated with race in such cases can harm the individual.

[11] Vence L Bonham, Shawneequa L Callier, and Charmaine D Royal, "Will Precision Medicine Move Us Beyond Race?" *The New England Journal of Medicine* 374, no. 21 (2016): 2003.

Formal non-discrimination criteria

Many *fairness criteria* have been proposed over the years, each aiming to formalize different desiderata. We'll start by jumping directly into the formal definitions of three representative fairness criteria that relate to many of the proposals that have been made.

Once we have acquired familiarity with the technical matter, we'll have a broader debate around the purpose, scope, and meaning of these fairness criteria in Chapter 3.

Most of the proposed fairness criteria are properties of the joint distribution of the sensitive attribute A, the target variable Y, and the classifier or score R.[12] This means that we can write them as some statement involving properties of these three random variables.

To a first approximation, most of these criteria fall into one of three different categories defined along the lines of different (conditional) independence[13] statements between the involved random variables.

[12] If all variables are binary, then the joint distribution is specified by 8 non-negative parameters that sum to 1. A non-trivial property of the joint distribution would restrict the way in which we can choose these parameters.

[13] Learn more about conditional independence here.

Table 3: Non-discrimination criteria

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

Below we will introduce and discuss each of these conditions in detail. Variants of these criteria arise from different ways of relaxing them.

As an exercise, think about why we omitted the conditional independence statement $R \perp Y \mid A$ from our discussion here.

Independence

Our first formal criterion simply requires the sensitive characteristic to be statistically independent of the score.

Definition 1. *The random variables* (A, R) *satisfy* independence *if* $A \perp R$.

Independence has been explored through many equivalent terms or variants, referred to as *demographic parity*, *statistical parity*, *group*

fairness, *disparate impact* and others. In the case of binary classification, independence simplifies to the condition

$$\mathbb{P}\{R = 1 \mid A = a\} = \mathbb{P}\{R = 1 \mid A = b\}\,,$$

for all groups a, b. Thinking of the event $R = 1$ as "acceptance", the condition requires the acceptance rate to be the same in all groups. A relaxation of the constraint introduces a positive amount of slack $\epsilon > 0$ and requires that

$$\mathbb{P}\{R = 1 \mid A = a\} \geq \mathbb{P}\{R = 1 \mid A = b\} - \epsilon\,.$$

Note that we can swap a and b to get an inequality in the other direction. An alternative relaxation is to consider a ratio condition, such as,

$$\frac{\mathbb{P}\{R = 1 \mid A = a\}}{\mathbb{P}\{R = 1 \mid A = b\}} \geq 1 - \epsilon\,.$$

Some have argued[14] that, for $\epsilon = 0.2$, this condition relates to the *80 percent rule* in disparate impact law.

Yet another way to state the independence condition in full generality is to require that A and R must have zero mutual information[15] $I(A; R) = 0$. The characterization in terms of mutual information leads to useful relaxations of the constraint. For example, we could require $I(A; R) \leq \epsilon$.

[14] Michael Feldman et al., "Certifying and Removing Disparate Impact," in *Proc.* 21*st SIGKDD* (ACM, 2015).

[15] Mutual information is defined as $I(A; R) = H(A) + H(R) - H(A, R)$, where H denotes the entropy.

Limitations of independence

Independence is pursued as a criterion in many papers, for several reasons. For example, it may be an expression of a belief about human nature, namely that traits relevant for a job are independent of certain attributes. It also has convenient technical properties.

However, decisions based on a classifier that satisfies independence can have undesirable properties (and similar arguments apply to other statistical critiera). Here is one way in which this can happen, which is easiest to illustrate if we imagine a callous or ill-intentioned decision maker. Imagine a company that in group a hires diligently selected applicants at some rate $p > 0$. In group b, the company hires carelessly selected applicants at the same rate p. Even though the acceptance rates in both groups are identical, it is far more likely that unqualified applicants are selected in one group than in the other. As a result, it will appear in hindsight that members of group b performed worse than members of group a, thus establishing a negative track record for group b.[16]

This situation might arise without positing malice: the company might have historically hired employees primarily from group a, giving them a better understanding of this group. As a technical matter,

[16] This problem was identified and called *self-fulfilling prophecy* in, Cynthia Dwork et al., "Fairness Through Awareness," in *Proc.* 3*rd ITCS*, 2012, 214–26. One might object that enforcing demographic parity in this scenario might still create valuable additional training data which could then improve predictions in the future after re-training the classifier on these additional data points.

the company might have substantially more training data in group a, thus potentially leading to lower error rates of a learned classifier within that group. The last point is a bit subtle. After all, if both groups were entirely homogenous in all ways relevant to the classification task, more training data in one group would equally benefit both. Then again, the mere fact that we chose to distinguish these two groups indicates that we believe they might be heterogeneous in relevant aspects.

Interlude: How to satisfy fairness criteria

A later chapter devoted to algorithmic interventions will go into detail, but we pause for a moment to think about how we can achieve the independence criterion when we actually build a classifier. We distinguish between three different techniques. While they generally apply to all the criteria and their relaxations that we review in this chapter, our discussion here focuses on independence.

- Pre-processing: Adjust the feature space to be uncorrelated with the sensitive attribute.
- At training time: Work the constraint into the optimization process that constructs a classifier from training data.
- Post-processing: Adjust a learned classifier so as to be uncorrelated with the sensitive attribute.

The three approaches have different strengths and weaknesses.

Pre-processing is a family of techniques to transform a feature space into a representation that as a whole is independent of the sensitive attribute. This approach is generally agnostic to what we do with the new feature space in downstream applications. After the pre-processing transformation ensures independence, any deterministic training process on the new space will also satisfy independence[17].

[17] Formally, this is a consequence of the data processing inequality from information theory.

Achieving independence at training time can lead to the highest utility since we get to optimize the classifier with this criterion in mind. The disadvantage is that we need access to the raw data and training pipeline. We also give up a fair bit of generality as this approach typically applies to specific model classes or optimization problems.

Post-processing refers to the process of taking a trained classifier and adjusting it possibly depending on the sensitive attribute and additional randomness in such a way that independence is achieved. Formally, we say a *derived classifier* $\hat{Y} = F(R, A)$ is a possibly randomized function of a given score R and the sensitive attribute. Given a cost for false negatives and false positives, we can find the derived

classifier that minimizes the expected cost of false positive and false negatives subject to the fairness constraint at hand. Post-processing has the advantage that it works for any *black-box* classifier regardless of its inner workings. There's no need for re-training, which is useful in cases where the training pipeline is complex. It's often also the only available option when we have access only to a trained model with no control over the training process. These advantages of post-processing are simultaneously also a weakness as it often leads to a significant loss in utility.

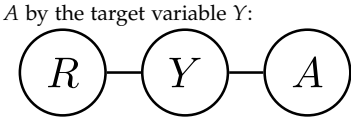
Separation

Our next criterion acknowledges that in many scenarios, the sensitive characteristic may be correlated with the target variable. For example, one group might have a higher default rate on loans than another. A bank might argue that it is a matter of business necessity to therefore have different lending rates for these groups.

Roughly speaking, the separation criterion allows correlation between the score and the sensitive attribute to the extent that it is *justified by the target variable*. This intuition can be made precise with a simple conditional independence statement.

Definition 2. *Random variables* (R, A, Y) *satisfy* separation *if* $R \perp A \mid Y$.[18]

[18] We can display separation as a graphical model in which R is separated from A by the target variable Y:



If you haven't seen graphical models before, don't worry. All this says is that R is conditionally independent of A given Y.

In the case where R is a binary classifier, separation is equivalent to requiring for all groups a, b the two constraints

$$\mathbb{P}\{R = 1 \mid Y = 1, A = a\} = \mathbb{P}\{R = 1 \mid Y = 1, A = b\}$$
$$\mathbb{P}\{R = 1 \mid Y = 0, A = a\} = \mathbb{P}\{R = 1 \mid Y = 0, A = b\}\,.$$

Recall that $\mathbb{P}\{R = 1 \mid Y = 1\}$ is called the *true positive rate* of the classifier. It is the rate at which the classifier correctly recognizes positive instances. The *false positive rate* $\mathbb{P}\{R = 1 \mid Y = 0\}$ highlights the rate at which the classifier mistakenly assigns positive outcomes to negative instances. What separation therefore requires is that all groups experience the same false negative rate and the same false positive rate.

This interpretation in terms of equality of error rates leads to natural relaxations. For example, we could only require equality of false negative rates. A false negative, intuitively speaking, corresponds to denied opportunity in scenarios where acceptance is desirable, such as in hiring.[19]

[19] In contrast, when the task is to identify high-risk individuals, as in the case of recidivism prediction, it is common to denote the undesirable outcome as the "positive" class. This inverts the meaning of false positives and false negatives, and is a frequent source of terminological confusion.

Achieving separation

As was the case with independence, we can achieve separation by post-processing a given score function without the need for retraining.[20]

The post-processing step uses the ROC curve that we saw earlier and it's illustrative to go into a bit more detail. A binary classifier that satisfies separation must achieve the same true positive rates and the same false positive rates in all groups. This condition corresponds to taking the intersection of all group-level ROC curves. Within this constraint region, we can then choose the classifier that minimizes the given cost.

[20] Recall, a derived classifier is a possible randomized mapping $\hat{Y} = F(R, A)$.

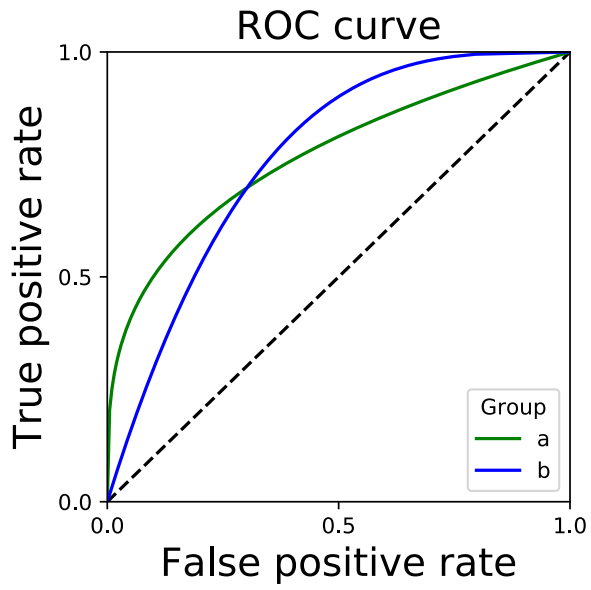


Figure 5: ROC curve by group.

We see the ROC curves of a score displayed for each group separately. The two groups have different curves indicating that not all trade-offs between true and false positive rate are achievable in both groups. The trade-offs that are achievable in both groups are precisely those that lie under both curves, corresponding to the intersection of the regions enclosed by the curves.

The highlighted region is the *feasible region* of trade-offs that we can achieve in all groups. There is a subtlety though. Points that are not exactly on the curves, but rather in the interior of the region, require *randomization*. To understand this point, consider a classifier that accepts everyone corresponding to true and false positive rate 1, the upper right corner of the plot. Consider another classifier that accepts no one, resulting in true and false positive rate 0, the lower left corner of the plot. Now, consider a third classifier that given an

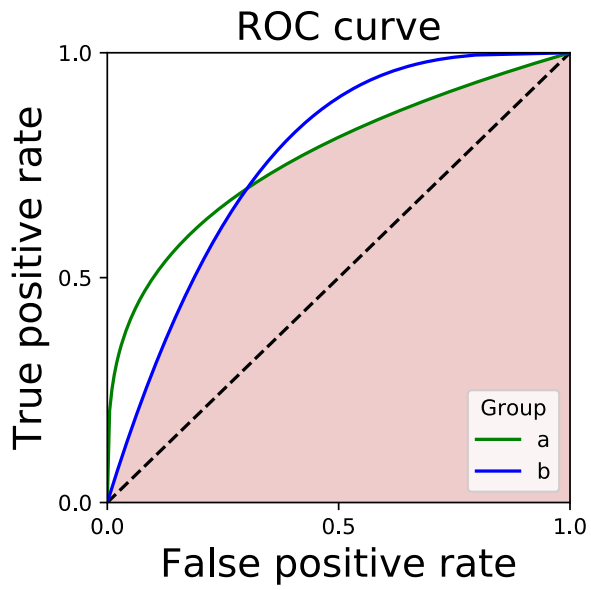


Figure 6: Intersection of area under the curves.

instance randomly picks and applies the first classifier with probability $1 - p$, and the second with probability p. This classifier achieves true and false positive rate p thus giving us one point on the dashed line in the plot. In the same manner, we could have picked any other pair of classifiers and randomized between them. We can fill out the entire shaded region in this way, because it is *convex*, meaning that every point in it lies on a line segment between two classifiers on the boundary.

Sufficiency

Our third criterion formalizes that the score already subsumes the sensitive characteristic for the purpose of predicting the target. This idea again boils down to a conditional independence statement.

Definition 3. *We say the random variables* (R, A, Y) *satisfy* sufficiency *if* $Y \perp A \mid R$.[21]

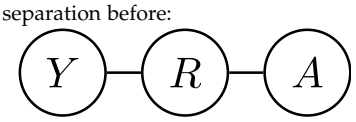
We will often just say that R satisfies *sufficiency* when the sensitive attribute A and target variable Y are clear from the context.

Let us write out the definition more explicitly in the binary case where $Y \in \{0, 1\}$. In this case, a random variable R is sufficient for A if and only if for all groups a, b and all values r in the support of R, we have

$$\mathbb{P}\{Y = 1 \mid R = r, A = a\} = \mathbb{P}\{Y = 1 \mid R = r, A = b\}\,.$$

When R has only two values we recognize this condition as requiring

[21] We can again display sufficiency as a graphical model as we did with separation before:



If you haven't seen graphical models before, feel free to ignore this interpretation.

a parity of positive/negative predictive values across all groups.

While it is often useful to think of sufficiency in terms of positive and negative predictive values, there's a useful alternative. Indeed, sufficiency turns out to be closely related to an important notion called *calibration*, as we will discuss next.

Calibration and sufficiency

In some applications it is desirable to be able to interpret the values of the score functions as probabilities. Formally, we say that a score R is *calibrated* if for all score values r in the support of R, we have

$$\mathbb{P}\{Y = 1 \mid R = r\} = r\,.$$

This condition means that the set of all instances assigned a score value r has an r fraction of positive instances among them. The condition refers to the group of all individuals receiving a particular score value. It does not mean that at the level of a single individual a score of r corresponds to a probability r of a positive outcome. The latter is a much stronger property that is satisfied by the conditional expectation $R = \mathbb{E}[Y \mid X]$.[22]

[22] Formally, we have for every set S, $\mathbb{P}\{Y = 1 \mid R = r, X \in S\} = r$.

In practice, there are various heuristics to achieve calibration. For example, *Platt scaling* is a popular method that works as follows. Platt scaling takes a possibly uncalibrated score, treats it as a single feature, and fits a one variable regression model against the target variable based on this feature. More formally, given an uncalibrated score R, Platt scaling aims to find scalar parameters a, b such that the sigmoid function[23]

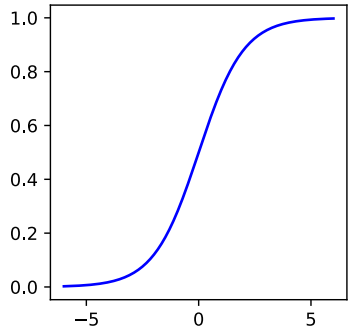
$$S = \frac{1}{1 + \exp(aR + b)}$$

fits the target variable Y with respect to the so-called *log loss*

$$-\mathbb{E}[Y \log S + (1 - Y) \log(1 - S)].$$

This objective can be minimized given labeled examples drawn from (R, Y) as is standard in supervised learning.

[23] A plot of the sigmoid function $1/(1 + \exp(-x))$.



Calibration by group

From the definition, we can see that sufficiency is closely related to the idea of calibration. To formalize the connection we say that the score R satisfies *calibration by group* if it satisfies

$$\mathbb{P}\{Y = 1 \mid R = r, A = a\} = r\,,$$

for all score values r and groups a. Recall that calibration is the same requirement at the population level without the conditioning on A.

Fact 1. *Calibration by group implies sufficiency.*

Conversely, sufficiency is only slightly weaker than calibration by group in the sense that a simple renaming of score values goes from one property to the other.

Proposition 1. *If a score R satisfies sufficiency, then there exists a function $\ell\colon [0,1] \to [0,1]$ so that $\ell(R)$ satisfies calibration by group.*

Proof. Fix any group a and put $\ell(r) = \mathbb{P}\{Y = 1 \mid R = r, A = a\}$. Since R satisfies sufficiency, this probability is the same for all groups a and hence this map ℓ is the same regardless of what value a we chose.

Now, consider any two groups a, b. We have,

$$\begin{aligned} r &= \mathbb{P}\{Y = 1 \mid \ell(R) = r, A = a\} \\ &= \mathbb{P}\{Y = 1 \mid R \in \ell^{-1}(r), A = a\} \\ &= \mathbb{P}\{Y = 1 \mid R \in \ell^{-1}(r), A = b\} \\ &= \mathbb{P}\{Y = 1 \mid \ell(R) = r, A = b\}\,, \end{aligned}$$

thus showing that $\ell(R)$ is calibrated by group.

□

We conclude that sufficiency and calibration by group are essentially equivalent notions. In particular, this gives us a large repertoire of methods for achieving sufficiency. We could, for example, apply Platt scaling for each of the groups defined by the sensitive attribute.

Calibration by group as a consequence of unconstrained learning

Sufficiency is often satisfied by default without the need for any explicit intervention. Indeed, we generally expect a learned score to satisfy sufficiency in cases where the sensitive attribute can be predicted from the other attributes.

To illustrate this point we look at the calibration values of a standard logistic regression model on the standard UCI adult data set.[24]

[24] Source

We fit a logistic regression model using Python's sklearn library on the UCI training data. The model is then applied to the UCI test data[25]. We make no effort to either tune or calibrate the model.

[25] Number of test samples in the UCI data set by group: 1561 Black, 13946 White; 5421 Female, 10860 Male

As we can see from the figure below, the model turns out to be fairly well calibrated by *gender* on its own without any explicit correction.

We see some deviation when we look at calibration by *race*.

The deviation we see in the mid deciles may be due to the scarcity of the test data in the corresponding group and deciles. For example, the 6th decile, corresponding to the score range $(0.5, 0.6]$, on the test data has only 34 instances with the 'Race' attribute set to 'Black'. As a result, the error bars[26] in this region are rather large.

[26] The shaded region in the plot indicates a 95% confidence interval for a binomial model.

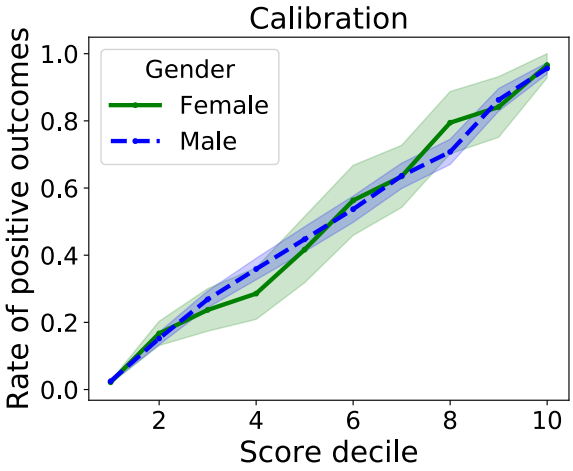


Figure 7: Calibration by gender on UCI adult data. A straight diagonal line would correspond to perfect calibration.

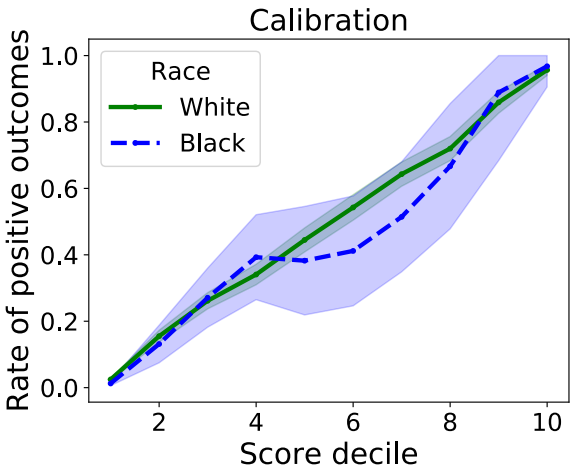


Figure 8: Calibration by race on UCI adult data.

Continue to explore the UCI Adult data in this code example.

The lesson is that sufficiency often comes for free (at least approximately) as a consequence of standard machine learning practices. The flip side is that imposing sufficiency as a constraint on a classification system may not be much of an intervention. In particular, it would not effect a substantial change in current practices.

Relationships between criteria

The criteria we reviewed constrain the joint distribution in non-trivial ways. We should therefore suspect that imposing any two of them simultaneously over-constrains the space to the point where only degenerate solutions remain. We will now see that this intuition is largely correct.

What this shows is that we cannot impose multiple criteria as hard constraints. This leaves open the possibility that meaningful trade-offs between these different criteria exist.

Independence versus Sufficiency

We begin with a simple proposition that shows how in general independence and sufficiency are mutually exclusive. The only assumption needed here is that the sensitive attribute A and the target variable Y are *not* independent. This is a different way of saying that group membership has an effect on the statistics of the target variable. In the binary case, this means one group has a higher rate of positive outcomes than another. Think of this as the typical case.

Proposition 2. *Assume that A and Y are not independent. Then sufficiency and independence cannot both hold.*

Proof. By the contraction rule for conditional independence,

$$A \perp R \quad \text{and} \quad A \perp Y \mid R \quad \Longrightarrow \quad A \perp (Y, R) \quad \Longrightarrow \quad A \perp Y .$$

To be clear, $A \perp (Y, R)$ means that A is independent of the pair of random variables (Y, R). Dropping R cannot introduce a dependence between A and Y.

In the contrapositive,

$$A \not\perp Y \quad \Longrightarrow \quad A \not\perp R \quad \text{or} \quad A \not\perp R \mid Y .$$

□

Independence versus Separation

An analogous result of mutual exclusion holds for independence and separation. The statement in this case is a bit more contrived and requires the additional assumption that the target variable Y is binary. We also additionally need that the score is not independent of the target. This is a rather mild assumption, since any useful score function should have correlation with the target variable.

Proposition 3. *Assume Y is binary, A is not independent of Y, and R is not independent of Y. Then, independence and separation cannot both hold.*

Proof. Assume $Y \in \{0, 1\}$. In its contrapositive form, the statement we need to show is

$$A \perp R \quad \text{and} \quad A \perp R \mid Y \quad \Longrightarrow \quad A \perp Y \quad \text{or} \quad R \perp Y$$

By the law of total probability,

$$\mathbb{P}\{R = r \mid A = a\} = \sum_y \mathbb{P}\{R = r \mid A = a, Y = y\}\mathbb{P}\{Y = y \mid A = a\}$$

Applying the assumption $A \perp R$ and $A \perp R \mid Y$, this equation simplifies to

$$\mathbb{P}\{R = r\} = \sum_y \mathbb{P}\{R = r \mid Y = y\}\mathbb{P}\{Y = y \mid A = a\}$$

Applied differently, the law of total probability also gives

$$\mathbb{P}\{R = r\} = \sum_y \mathbb{P}\{R = r \mid Y = y\}\mathbb{P}\{Y = y\}$$

Combining this with the previous equation, we have

$$\sum_y \mathbb{P}\{R = r \mid Y = y\}\mathbb{P}\{Y = y\} = \sum_y \mathbb{P}\{R = r \mid Y = y\}\mathbb{P}\{Y = y \mid A = a\}$$

Careful inspection reveals that when y ranges over only two values, this equation can only be satisfied if $A \perp Y$ or $R \perp Y$.

Indeed, we can rewrite the equation more compactly using the symbols $p = \mathbb{P}\{Y = 0\}$, $p_a = \mathbb{P}\{Y = 0 \mid A = a\}$, $r_y = \mathbb{P}\{R = r \mid Y = y\}$, as:

$$pr_0 + (1 - p)r_1 = p_a r_0 + (1 - p_a)r_1.$$

Equivalently, $p(r_0 - r_1) = p_a(r_0 - r_1)$.

This equation can only be satisfied if $r_0 = r_1$, in which case $R \perp Y$, or if $p = p_a$ for all a, in which case $Y \perp A$.

□

The claim is not true when the target variable can assume more than two values, which is a natural case to consider.

Exercise 1. *Give a counterexample to the claim in the previous proposition where the target variable Y assumes three distinct values.*

Separation versus Sufficiency

Finally, we turn to the relationship between separation and sufficiency. Both ask for a non-trivial conditional independence relationship between the three variables A, R, Y. Imposing both simultaneously leads to a degenerate solution space, as our next proposition confirms.

Proposition 4. *Assume that all events in the joint distribution of (A, R, Y) have positive probability, and assume $A \not\perp Y$. Then, separation and sufficiency cannot both hold.*

Proof. A standard fact[27] about conditional independence shows

$$A \perp R \mid Y \quad \text{and} \quad A \perp Y \mid R \quad \Longrightarrow \quad A \perp (R, Y)\,.$$

Moreover,

$$A \perp (R, Y) \quad \Longrightarrow \quad A \perp R \quad \text{and} \quad A \perp Y\,.$$

Taking the contrapositive completes the proof. □

[27] See Theorem 17.2 in Larry Wasserman, *All of Statistics: A Concise Course in Statistical Inference* (Springer, 2010)

For a binary target, the non-degeneracy assumption in the previous proposition states that in all groups, at all score values, we have both positive and negative instances. In other words, the score value never fully resolves uncertainty regarding the outcome.

In case the classifier is also binary, we can weaken the assumption to require only that the classifier is imperfect in the sense of making at least one false positive prediction. What's appealing about the resulting claim is that its proof essentially only uses a well-known relationship between true positive rate (recall) and positive predictive value (precision). This trade-off is often called *precision-recall trade-off*.

Proposition 5. *Assume Y is not independent of A and assume $\hat{Y}$ is a binary classifier with nonzero false positive rate. Then, separation and sufficiency cannot both hold.*

Proof. Since Y is not independent of A there must be two groups, call them 0 and 1, such that

$$p_0 = \mathbb{P}\{Y = 1 \mid A = 0\} \neq \mathbb{P}\{Y = 1 \mid A = 1\} = p_1\,.$$

Now suppose that separation holds. Since the classifier is imperfect this means that all groups have the same non-zero false positive rate $\text{FPR} > 0$, and the same positive true positive rate $\text{TPR} > 0$. We will show that sufficiency does not hold.

Recall that in the binary case, sufficiency implies that all groups have the same positive predictive value. The positive predictive value in group a, denoted PPV_a satisfies

$$\text{PPV}_\text{a} = \frac{\text{TPR} p_a}{\text{TPR} p_a + \text{FPR}(1 - p_a)} \,.$$

From the expression we can see that $\text{PPV}_0 = \text{PPV}_1$ only if $\text{TPR} = 0$ or $\text{FPR} = 0$. The latter is ruled out by assumption. So it must be that $\text{TPR} = 0$. However, in this case, we can verify that the negative predictive value NPV_0 in group 0 must be different from the negative predictive value NPV_1 in group 1. This follows from the expression

$$\text{NPV}_\text{a} = \frac{(1 - \text{FPR})(1 - p_a)}{(1 - \text{TPR}) p_a + (1 - \text{FPR})(1 - p_a)} \,.$$

Hence, sufficiency does not hold.

□

A good exercise is to derive variants of these trade-offs such as the following.

Exercise 2. *Prove the following result: Assume Y is not independent of A and assume $\hat{Y}$ is a binary classifier with nonzero false positive rate and nonzero true positive rate. Then, if separation holds, there must be two groups with different positive predictive values.*

Inherent limitations of observational criteria

All criteria we've seen so far have one important aspect in common. They are properties of the joint distribution of the score, sensitive attribute, and the target variable. In other words, if we know the joint distribution of the random variables (R, A, Y), we can without ambiguity determine whether this joint distribution satisfies one of these criteria or not.[28]

We can broaden this notion a bit and also include all other features, not just the sensitive attribute. So, let's call a criterion *observational* if it is a property of the joint distribution of the features X, the sensitive attribute A, a score function R and an outcome variable Y.[29] Informally, a criterion is observational if we can express it using probability statements involving the random variables at hand.

Exercise 3. *Convince yourself that independence, separation, and sufficiency are all observational definitions. Come up with a criterion that is* not *observational.*

[28] For example, if all variables are binary, there are eight numbers specifying the joint distributions. We can verify the property by looking only at these eight numbers.

[29] Formally, this means an observational property is defined by set of joint distributions over a given set of variables.

Observational definitions have many appealing aspects. They're often easy to state and require only a lightweight formalism. They make no reference to the inner workings of the classifier, the decision maker's intent, the impact of the decisions on the population, or any notion of whether and how a feature actually influences the outcome. We can reason about them fairly conveniently as we saw earlier. In principle, observational definitions can always be verified given samples from the joint distribution—subject to statistical sampling error.

At the same time, all observational definitions share inherent limitations that we will explore now. Our starting point are two fictitious worlds with substantively different characteristics. We will see that despite their differences these two worlds can map to identical joint distributions. What follows is that all observational criteria will look the same in either world, thus glossing over whatever differences there are.

To develop these two worlds, we'll use the case of a fictitious advertising campaign that targets a hiring ad to software engineers. A score function estimates the likelihood that an individual is a software engineer given some available features.

Scenario I

Imagine we introduce the following random variables in our classification problem.

- A indicates gender
- X_1 indicates whether the user visited `pinterest.com`
- X_2 indicates whether the user visited `github.com`
- R^* is the optimal unconstrained score
- $\tilde{R}$ is the optimal score satisfying separation
- Y indicates whether the user is a software engineer

We can summarize the conditional independence relationships between the variables in a *directed graphical model*.[30] The main fact we need is that a node is conditionally independent of any node that is not a direct ancestor given its parents.

[30] Learn more about graphical models here.

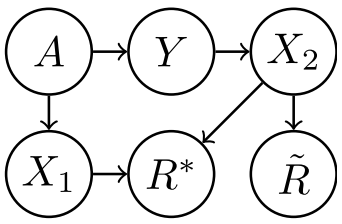


Figure 9: Directed graphical model for the variables in Scenario I

Let's imagine a situation that corresponds to this kind of graphical model. We could argue that gender influences the target variable, since currently software engineers are predominantly male. Gender also influences the first feature, since Pinterest's user base skews female.[31] We assume `github.com` has a male bias. However, this bias is explained by the target variable in the sense that conditional on being a software engineer, all genders are equally likely to visit `github.com`.

[31] As of August 2017, 58.9% of Pinterest's users in the United States were female. See here (Retrieved 3-27-2018)

Once we make these assumptions, we can work out what the optimal unconstrained classifier will do. Both features correlate with the target variable and are therefore useful for prediction. The first feature is predictive since (absent other information) visiting `pinterest.com` suggests female gender, which in turns makes "software engineer" less likely. The second feature is predictive in a more direct sense, as the website is specifically designed for software engineers.

The optimal classifier satisfying separation will refrain from using the first feature (visiting `pinterest.com`). After all, we can see from the graphical model that this feature is not conditionally independent of the sensitive attribute given the target. This score will only use the directly predictive feature `github.com`, which is indeed conditionally independent of gender given the target.

Scenario II

Our two features are different in Scenario II, but all other variables have the same interpretation.

- X_1 indicates whether the user studied computer science
- X_2 indicates whether the user visited the Grace Hopper conference

Although the other variables have the same names and interpretations, we now imagine a very different graphical model.

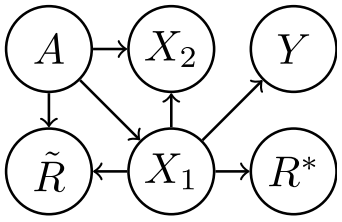


Figure 10: Directed graphical model for the variables in Scenario II

As before, we assume that gender influences the target variable, but now we assume that the target variable is conditionally independent from gender given the first feature. That is, conditional on having studied computer science, all genders are equally likely to go on to become software engineers.[32]

[32] This may not be true in reality. It's an assumption we make in this example.

With these assumptions, we can again work out the optimal unconstrained classifier. This time, the optimal unconstrained classifier will only use one feature, namely the first. The reason is that, given the first feature, all remaining features (including the sensitive attribute) become conditionally independent of the target. Therefore, knowing the second feature does not help in predicting the target, once we have the first.

The optimal classifier under separation turns out to be a bit subtle in Scenario II. The issue is that neither of the two features is conditionally independent from the sensitive attribute given the target. The classifier will therefore actively take the sensitive attribute into account in order to *subtract* its influence on the other features.

Different interpretations

Interpreted in the concrete advertising context, the two scenarios don't seem very similar. In particular, the inner workings of the optimal unconstrained classifier in each scenario are rather different. In the first scenario it uses `pinterest.com` as a weak proxy for being *female*, which it then uses as a proxy for not being a software engineer. Software engineers who visit *pinterest.com* might be concerned about this kind of stereotyping, as they might miss out on seeing the ad, and hence the job opportunity. In the second scenario, unconstrained score leads to a classifier that is natural in the sense that it only considers the directly predictive educational information. Absent other features, this would seem agreeable.

Similarly, the optimal classifier satisfying separation behaves differently in the two scenarios. In the first, it corresponds to the natural classifier that only uses `github.com` when predicting *software engineer*. Since `github.com` is primarily a website for software engineers, this seems reasonable. In the second scenario, however, the optimal constrained score performs a subtle adjustment procedure that explicitly takes the sensitive attribute into account. These score functions are also not equivalent from a legal standpoint. One uses the sensitive attribute explicitly for an adjustment step, while the other does not.

Indistinguishability

Despite all their apparent differences, we can instantiate the random variables in each scenario in such a manner that the two scenarios map to identical joint distributions. This means that no property of the joint distribution will be able to distinguish the two scenarios. Whatever property holds for one scenario, it will inevitably also hold for the other. If by some observational criterion we call one scenario *unfair*, we will also have to call the other *unfair*.

Proposition 6. *The random variables in Scenario I and II admit identical joint distributions. In particular, no observational criterion distinguishes between the two scenarios.*

The indistinguishability result has nothing to do with sample sizes or sampling errors. No matter how many data points we have, the size of our data does not resolve the indistinguishability.

There's another interesting consequence of this result. Observational criteria cannot even determine if the sensitive attribute was fed into the classifier or not. To see this, recall that the optimal constrained score in one scenario directly uses *gender*, in the other it does not.

A forced perspective problem

To understand the indistinguishability result, it's useful to draw an analogy with a *forced perspective* problem. Two different objects can appear identical when looked at from a certain fixed perspective.

A data set always forces a particular perspective on reality. There is a possibility that this perspective makes it difficult to identify certain properties of the real world. Even if we have plenty of data, so long as this data comes from the same distribution, it still represents the same perspective. Having additional data is a bit like increasing the resolution of our camera. It helps with some problems, but it doesn't change the angle or the position of the camera.

The limitations of observational criteria are fundamentally the limitations of a single perspective. When analyzing a data set through the lens of observational criteria we do not evaluate alternatives to the data we have. Observational criteria do not tell us what is missing from our perspective.

What then is *not* observational and how do we go beyond observational criteria? This is a profound question that will be the focus of later chapters. In particular, we will introduce the technical repertoire of measurement and causality to augment the classification paradigm. Both measurement and causality give us mechanisms to interrogate, question, and change the perspective suggested by our data.

Case study: Credit scoring

We now apply some of the notions we saw to credit scoring. Credit scores support lending decisions by giving an estimate of the risk that a loan applicant will default on a loan. Credit scores are widely used in the United States and other countries when allocating credit, ranging from micro loans to jumbo mortgages. In the United States,

there are three major credit-reporting agencies that collect data on various lendees. These agencies are for-profit organizations that each offer risk scores based on the data they collected. FICO scores are a well-known family of proprietary scores developed by FICO and sold by the three credit reporting agencies.

Regulation of credit agencies in the United States started with the Fair Credit Reporting Act, first passed in 1970, that aims to promote the accuracy, fairness, and privacy of consumer of information collected by the reporting agencies. The Equal Credit Opportunity Act, a United States law enacted in 1974, makes it unlawful for any creditor to discriminate against any applicant the basis of race, color, religion, national origin, sex, marital status, or age.

Score distribution

Our analysis relies on data published by the Federal Reserve[33]. The data set provides aggregate statistics from 2003 about a credit score, demographic information (race or ethnicity, gender, marital status), and outcomes (to be defined shortly). We'll focus on the joint statistics of score, race, and outcome, where the race attributes assume four values detailed below.[34]

[33] The Federal Reserve Board, "Report to the Congress on Credit Scoring and Its Effects on the Availability and Affordability of Credit" (https://www.federalreserve.gov/boarddocs/rptcongress/creditscore/, 2007).

[34] These numbers come from the "Estimation sample" column of Table 9 on this web page.

Table 4: Credit score distribution by ethnicity

Race or ethnicity	Samples with both score and outcome
White	133,165
Black	18,274
Hispanic	14,702
Asian	7,906
Total	174,047

The score used in the study is based on the TransUnion TransRisk score. TransUnion is a US credit-reporting agency. The TransRisk score is in turn based on a proprietary model created by FICO, hence often referred to as FICO scores. The Federal Reserve renormalized the scores for the study to vary from 0 to 100, with 0 being *least creditworthy*.

The information on race was provided by the Social Security Administration, thus relying on self-reported values.

The cumulative distribution of these credit scores strongly depends on the group as the next figure reveals.

For an extensive documentation of the data set see the Federal Reserve report.

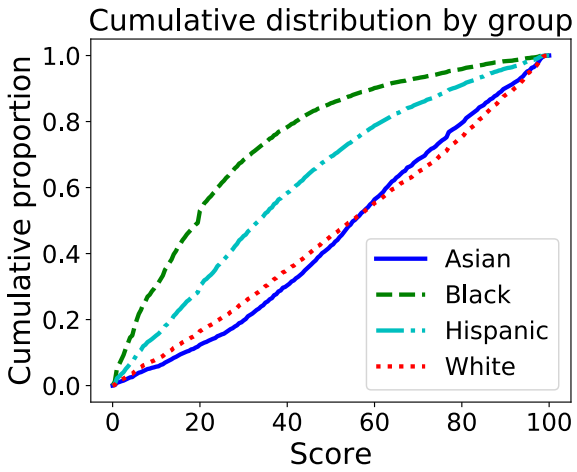


Figure 11: Cumulative density of scores by group.

Performance variables and ROC curves

As is often the case, the outcome variable is a subtle aspect of this data set. Its definition is worth emphasizing. Since the score model is proprietary, it is not clear what target variable was used during the training process. What is it then that the score is trying to predict? In a first reaction, we might say that the goal of a credit score is to predict a *default* outcome. However, that's not a clearly defined notion. Defaults vary in the amount of debt recovered, and the amount of time given for recovery. Any single binary performance indicator is typically an oversimplification.

What is available in the Federal Reserve data is a so-called *performance* variable that measures a *serious delinquency in at least one credit line of a certain time period*. More specifically,

> (the) measure is based on the performance of new or existing accounts and measures whether individuals have been late 90 days or more on one or more of their accounts or had a public record item or a new collection agency account during the performance period.[35]

[35] Quote from the Federal Reserve report.

With this performance variable at hand, we can look at the ROC curve to get a sense of how predictive the score is in different demographics.

The meaning of true positive rate is *the rate of predicted positive performance given positive performance.* Similarly, false positive rate is *the rate of predicted negative performance given a positive performance*.

We see that the shapes appear roughly visually similar in the groups, although the 'White' group encloses a noticeably larger area under the curve than the 'Black' group. Also note that even two ROC curves with the same shape can correspond to very different score

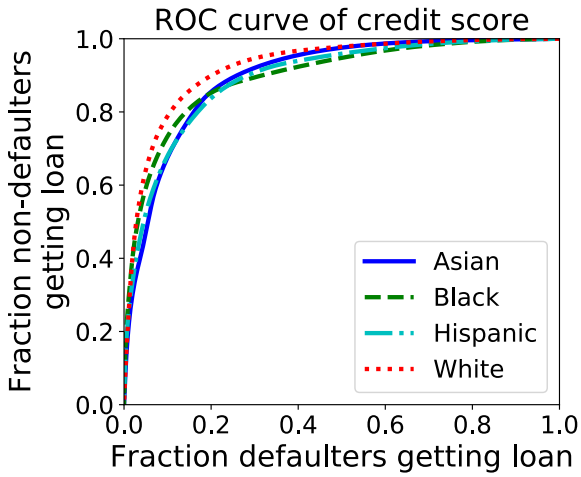


Figure 12: ROC curve of credit score by group.

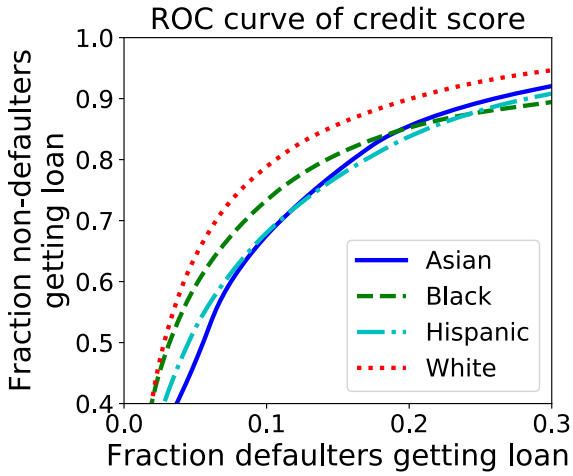


Figure 13: ROC curve of credit score by group zoomed in on region of large differences.

functions. A particular trade-off between true positive rate and false positive rate achieved at a threshold t in one group could require a different threshold t' in the other group.

Comparison of different criteria

With the score data at hand, we compare four different classification strategies:

- *Maximum profit:* Pick possibly group-dependent score thresholds in a way that maximizes profit.
- *Single threshold:* Pick a single uniform score threshold for all groups in a way that maximizes profit.
- *Separation:* Achieve an equal true/false positive rate in all groups. Subject to this constraint, maximize profit.
- *Independence:* Achieve an equal acceptance rate in all groups. Subject to this constraint, maximize profit.

To make sense of maximizing profit, we need to assume a reward for a true positive (correctly predicted positive performance), and a cost for false positives (negative performance predicted as positive). In lending, the cost of a false positive is typically many times greater than the reward for a true positive. In other words, the interest payments resulting from a loan are relatively small compared with the loan amount that could be lost. For illustrative purposes, we imagine that the cost of a false positive is 6 times greater than the return on a true positive. The absolute numbers don't matter. Only the ratio matters. This simple cost structure glosses over a number of details that are likely relevant for the lender such as the terms of the loan.

There is another major caveat to the kind of analysis we're about to do. Since we're only given aggregate statistics, we cannot retrain the score with a particular classification strategy in mind. The only thing we can do is to define a setting of thresholds that achieves a particular criterion. This approach may be overly pessimistic with regards to the profit achieved subject to each constraint. For this reason and the fact that our choice of cost function was rather arbitrary, we do not state the profit numbers. The numbers can be found in the original analysis[36], which reports that 'single threshold' achieves higher profit than 'separation', which in turn achieves higher profit than 'independence'.

[36] Moritz Hardt, Eric Price, and Nati Srebro, "Equality of Opportunity in Supervised Learning," in *Proc. 29th NIPS*, 2016, 3315–23.

What we do instead is to look at the different trade-offs between true and false positive rate that each criterion achieves in each group.

We can see that even though the ROC curves are somewhat similar, the resulting trade-offs can differ widely by group for some of the criteria. The true positive rate achieved by *max profit* for the

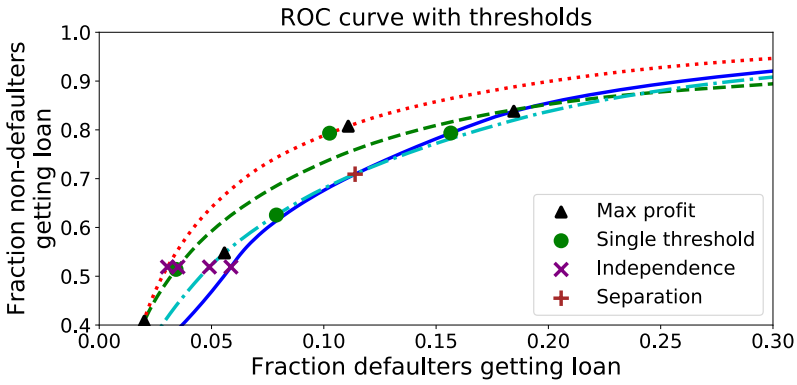


Figure 14: ROC curves with thresholds induced by different criteria.

Asian group is twice of what it is for the Black group. The separation criterion, of course, results in the same trade-off in all groups. Independence equalizes acceptance rate, but leads to widely different trade-offs. For instance, the Asian group has a false positive rate more than three times the false positive rate within the Black group.

Calibration values

Finally, we consider the non-default rate by group. This corresponds to the calibration plot by group.[37]

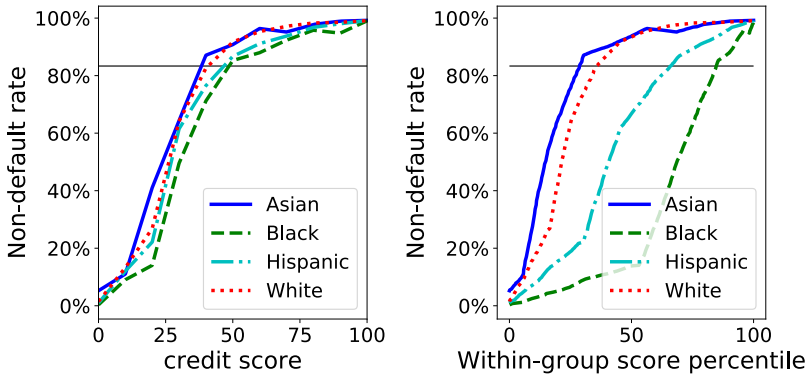


Figure 15: Calibration values of credit score by group.

[37] The error bars on the these plots were omitted as they are generally small except for very low score values (0-5) where few samples are available.

We see that the performance curves by group are reasonably well aligned. This means that a monotonic transformation of the score values would result in a score that is roughly calibrated by group according to our earlier definition. Due to the differences in score distribution by group, it could nonetheless be the case that thresholding the score leads to a classifier with different positive predictive values

in each group.

Feel free to continue exploring the data in this code repository.

Problem set: Criminal justice case study

Risk assessment is an important component of the criminal justice system. In the United States, judges set bail and decide pre-trial detention based on their assessment of the risk that a released defendant would fail to appear at trial or cause harm to the public. While *actuarial risk assessment* is not new in this domain, there is increasing support for the use of learned risk scores to guide human judges in their decisions. Proponents argue that machine learning could lead to greater efficiency and less biased decisions compared with human judgment. Critical voices raise the concern that such scores can perpetuate inequalities found in historical data, and systematically harm historically disadvantaged groups.

In this problem set[38], we'll begin to scratch at the surface of the complex criminal justice domain. Our starting point is an investigation carried out by Propublica[39] of a proprietary risk score, called COMPAS score. These scores are intended to assess the risk that a defendant will re-offend, a task often called *recidivism prediction*. Within the academic community, the ProPublica article drew much attention to the trade-off between separation and sufficiency that we saw earlier.

We'll use data obtained and released by ProPublica as a result of a public records request in Broward Country, Florida, concerning the COMPAS recidivism prediction system. The data is available here. Following ProPublica's analysis, we'll filter out rows where `days_b_screening_arrest` is over 30 or under -30, leaving us with 6,172 rows.

[38] Solutions to these problems are available to course instructors on request.

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

Calibration/sufficiency

- Plot the fraction of defendants recidivating within two years (`two_year_recid == 1`) as a function of risk score (`decile_score`), for black defendants (`race == "African-American"`) and white defendants (`race == "Caucasian"`).
- Based on these plots, does the risk score satisfy sufficiency across racial groups in this dataset? This is somewhat subjective, since we want to allow for approximate equality between groups; justify your answer in a sentence or two.

Error rates/separation

- Plot the distribution of scores received by the positive class (recidivists) and the distribution of scores received by the negative class (non-recidivists) for black defendants and for white defendants.
- Based on these plots, does COMPAS achieve separation between the risk score and race?
- Report the Positive Predictive Value, False Positive Rate, and False Negative Rate for a risk threshold of 4 (i.e., defendants with `decile_score >= 4` are classified as high risk), for black defendants and for white defendants.
- Can we pick two thresholds (one for black defendants, one for white defendants) such that FPR and FNR are roughly equal for the two groups (say, within 1% of each other)? What is the PPV for the two groups in this case? Note: trivial thresholds of 0 or 11 don't count.

Risk factors and interventions

- Report the recidivism rate of defendants aged 25 or lower, and defendants aged 50 or higher. Note the stark difference between the two: younger defendants are far more likely to recidivate.

The following questions are best viewed as prompts for a class discussion.

- Suppose we are interested in taking a data-driven approach to changing the criminal justice system. Under a theory of incarceration as incapacitation (prevention of future crimes by removal of individuals from society), how might we act on the finding that younger defendants are more likely to reoffend?
- How might we act on this finding under a rehabilitative approach to justice, in which we seek to find interventions that minimize a defendant's risk of recidivism?
- Under a retributive theory of justice, punishment is based in part on culpability, or blameworthiness; this in turn depends on how much control the defendant had over their actions. Under such a theory, how might we act on the finding that younger defendants are more likely to reoffend (and, more generally, commit offenses at all)?

Problem set: Data modeling of traffic stops

For this problem we'll use data released by the Stanford Open Policing Project (SOPP) for the state of North Carolina, available here. It

contains records of 9.6 million police stops in the state between 2000 and 2015.

General notes and hints:

- The *stop rates* section of this problem requires linking SOPP data to census data, whereas the rest is based only on SOPP data and no external datasets. So you might want to work on *post-stop outcomes* and the following sections first, so that you can get familiar with the SOPP data before having to also deal with the census data.
- Throughout this problem, report any data cleaning steps (such as dropping some rows) that you took. Also report any ambiguities you encountered and how you resolved them.

Stop rates

Part A

- For each possible group defined by race, age, gender, location, and year, where:
  - race is one of "Asian", "Black", "Hispanic", "White"
  - age is one of the buckets 15–19, 20–29, 30–39, 40–49, and 50+.
  - gender is one of "female", "male"
  - location is a state patrol troop district
  - and year is between 2010 and 2015, inclusive
- report the following:
  - the population of the group from census data, and
  - the number of stops in that group from SOPP data.

The census data is available here and the fields are explained here. Your data should look like the table below.

Table 5: Census data

Race	Age	Gender	Location	Year	Population	Count
Hispanic	30-39	F	B5	2012	434	76
White	40-49	F	C8	2011	2053	213
Asian	15-19	M	A2	2012	2	0
White	20-29	M	A6	2011	8323	1464
Hispanic	20-29	F	D3	2010	393	56
Black	40-49	F	D7	2011	1832	252
Asian	30-39	M	E6	2013	503	34
Asian	15-19	F	B5	2015	12	4
White	20-29	M	A5	2012	12204	1852
Black	15-19	F	H1	2011	1281	55

Notes and hints:

- The table is a small sample of rows from the actual answer. You can use it to check your answers. There should be about 13,000 rows in the table in total.
- The relevant fields in the census data are `AA_[FE]MALE`, `BA_[FE]MALE`, `H_[FE]MALE`, `WA_[FE]MALE`.
- The relevant fields in the SOPP data are `driver_race`, `driver_age`, `driver_gender`, `district`, and `stop_date`.
- The census data is grouped by county, which is more granular than district.The mapping from county to district is available from SOPP here.

Part B

- Fit a negative binomial regression to your data from part (A) as given in page 5 of the SOPP paper. Report the coefficients of race, age, and gender, and the overdispersion parameter ϕ. Based on these coefficients, what is the ratio of stop rates of Hispanic drivers to White drivers, and Black drivers to White drivers, controlling for age, gender, location, and year?

Notes and hints:

- This and the following tasks will be easier using a data modeling framework such as R or statsmodels rather than an algorithmic modeling framework such as scikit-learn.
- The "Population" column in your data corresponds to the "exposure" variable in most frameworks. Equivalently, "offset" is the log of the exposure.
- The coefficients of the different values of each variable (e.g. female and male) are not interpretable individually; only the difference is interpretable.
- Treat `year` as a categorical rather than a continuous variable.

Part C

- Give three distinct potential reasons for the racial disparity in stop rate as measured in part B.

Post-stop outcomes

Part D

- Controlling for age (bucketed as in parts A & B), gender, year, and location, use logistic regression to estimate impact of race on
  - probability of a search (`search_conducted`)

  - probability of arrest (`is_arrested`),
  - probability of a citation (`stop_outcome == "Citation"`)

- For each of the three outcomes, report the coefficients of race, age, and gender along with standard errors of those coefficients. Feel free to sample the data for performance reasons, but if you do, make sure that all standard errors are < 0.1.

Part E

- Interpret the coefficients you reported in part D.
  - What is the ratio of the probability of search of Hispanic drivers to White drivers? Black drivers to White drivers?
  - Repeat the above for the probability of arrest instead of search.
  - What is the difference in citation probability between Hispanic drivers and White drivers? Black drivers and White drivers?
  - Comment on the age and gender coefficients in the regressions.

Notes and hints:

- Interpreting the coefficients is slightly subjective. Since the search and arrest rates are low, in those regressions we can approximate the $1/(1+e^{-\beta x})$ formula in logistic regression as $e^{\beta x}$, and thus we can use differences in β between groups to calculate approximate ratios of search/arrest probabilities.
- This trick doesn't work for citation rates, since those are not low. However, we can pick "typical" values for the control variables, calculate citation rates, and find the difference in citation rate between groups. The results will have little sensitivity to the values of the control variables that we pick.

Part F

Explain in a sentence or two why we control for variables such as gender and location in the regression, and why the results might not be what we want if we don't control for them. (In other words, explain the idea of a confound in this context.)

Part G

However, decisions about what to control are somewhat subjective. What is one reason we might *not* want to control for location in testing for discrimination? In other words, how might we underestimate discrimination if we control for location? (Hint: broaden the idea of discrimination from individual officers to the systemic aspects of policing.)

Data quality

Part H

The SOPP authors provide a README file in which they note the incompleteness, errors, and missing values in the data on a state-by-state level. Pick any two items from this list and briefly explain how each could lead to errors or biases in the analyses you performed (or in the other analyses performed in the paper).

Notes and hints:

- Here is one example: For North Carolina, stop time is not available for a subset of rows. Suppose we throw out the rows with missing stop time (which we might have to if that variable is one of the controls in our regression). These rows might not be a random subset of rows: they could be correlated with location, because officers in some districts don't record the stop time. If so, we might incorrectly estimate race coefficients, because officer behavior might also be correlated with location.

What is the purpose of a fairness criterion?

There is an important question we have neglected so far. Although we have seen several demographic classification criteria and explored their formal properties and the relationships between them, we haven't yet clarified the purpose of these criteria. This is a difficult normative question that will be a central concern of the next chapter. Let us address it briefly here.

Take the independence criterion as an example. Some support this criterion based on the belief that certain intrinsic human traits such as intelligence are independent of, say, race or gender. Others argue for independence based on their desire to live in a society where the sensitive attribute is statistically independent of outcomes such as financial well-being. In one case, independence serves as a proxy for a belief about human nature. In the other case, it represents a long-term societal goal. In either case, does it then make sense to impose independence as a constraint on a classification system?

In a lending setting, for example, independence would result in the same rate of lending in all demographic groups defined by the sensitive attribute, regardless of the fact that individuals' ability to repay might be distributed differently in different groups. This makes it hard to predict the long-term impact of an intervention that imposes independence as a hard classification constraint. It is not clear how to account for the impact of the fact that giving out loans to individuals who cannot repay them impoverishes the individual who defaults (in addition to diminishing profits for the bank).

Without an accurate model of long-term impact it is difficult to foresee the effect that a fairness criterion would have if implemented

as a hard classification constraint. However, if such a model of long-term impact model were available, directly optimizing for long-term benefit may be a more effective intervention than to impose a general and crude demographic criterion.[40]

If demographic criteria are not useful as direct guides to fairness interventions, how should we use them then? An alternative view is that classification criteria have *diagnostic value* in highlighting different social costs of the system. Disparities in true positive rates or false positive rates, for example, indicate that two or more demographic groups experience different costs of classification that are not necessarily reflected in the cost function that the decision maker optimized.

At the same time, the diagnostic value of fairness criteria is subject to the fundamental limitations that we saw. In particular, we cannot base a conclusive argument of fairness or unfairness on the value of any observational criterion alone. Furthermore, Corbett-Davies et al.[41] make the important point that statistics such as positive predictive values or false positive rates can be manipulated through external (and possibly harmful) changes to the real world processes reflected in the data. In the context of recidivism prediction in criminal justice, for example, we could artificially lower the false positive rate in one group by arresting innocent people and correctly classifying them as low risk. This external intervention will decrease the false positive rate at the expense of a clearly objectionable practice.

[40] Lydia T. Liu et al., "Delayed Impact of Fair Machine Learning," in *Proc. 35th ICML*, 2018, 3156–64.

[41] Sam Corbett-Davies et al., "Algorithmic Decision Making and the Cost of Fairness," *arXiv Preprint arXiv:1701.08230*, 2017.

Bibliographic notes and further reading

The fairness criteria reviewed in this chapter were already known in the 1960s and 70s, primarily in the education testing and psychometrics literature.[42] An important fairness criterion is due to Cleary[43] and compares regression lines between the test score and the outcome in different groups. A test is considered *fair* by the Cleary criterion if the slope of these regression lines is the same for each group. This turns out to be equivalent to the sufficiency criterion, since it means that at a given score value all groups have the same rate of positive outcomes.

Einhorn and Bass[44] considered equality of precision values, which is a relaxation of sufficiency as we saw earlier. Thorndike[45] considered a weak variant of calibration by which the frequency of positive predictions must equal the frequency of positive outcomes in each group, and proposed achieving it via a post-processing step that sets different thresholds in different groups. Thorndike's criterion is incomparable to sufficiency in general.

Darlington[46] stated four different criteria in terms of succinct

[42] We are greatful to Ben Hutchinson for bringing these to our attention.

[43] T Anne Cleary, "Test Bias: Validity of the Scholastic Aptitude Test for Negro and White Students in Integrated Colleges," *ETS Research Bulletin Series* 1966, no. 2 (1966): i–23; T Anne Cleary, "Test Bias: Prediction of Grades of Negro and White Students in Integrated Colleges," *Journal of Educational Measurement* 5, no. 2 (1968): 115–24.

[44] Hillel J Einhorn and Alan R Bass, "Methodological Considerations Relevant to Discrimination in Employment Testing." *Psychological Bulletin* 75, no. 4 (1971): 261.

[45] Robert L Thorndike, "Concepts of Culture-Fairness," *Journal of Educational Measurement* 8, no. 2 (1971): 63–70.

[46] Richard B Darlington, "Another Look at 'Cultural Fairness'," *Journal of Educational Measurement* 8, no. 2 (1971): 71–82.

expressions involving the correlation coefficients between various pairs of random variables. These criteria include independence, a relaxation of sufficiency, a relaxation of separation, and Thorndike's criterion. Darlington included an intuitive visual argument showing that the four criteria are incompatible except in degenerate cases.

Lewis[47] reviewed three fairness criteria including equal precision and equal true/false positive rates.

These important early works were re-discovered later in the machine learning and data mining community. Numerous works considered variants of independence as a fairness constraint[48]. Feldman et al.[49] studied a relaxation of demographic parity in the context of disparate impact law. Zemel et al.[50] adopted the mutual information viewpoint and proposed a heuristic pre-processing approach for minimizing mutual information. Dwork et al.[51] argued that the independence criterion was inadequate as a fairness constraint.

The separation criterion appeared under the name *equalized odds*[52], alongside the relaxation to equal false negative rates, called *equality of opportunity*. These criteria also appeared in an independent work[53] under different names. Woodworth et al.[54] studied a relaxation of separation stated in terms of correlation coefficients. This relaxation corresponds to the third criterion studied by Darlington[55].

ProPublica[56] implicitly adopted equality of false positive rates as a fairness criterion in their article on COMPAS scores. Northpointe, the maker of the COMPAS software, emphasized the importance of calibration by group in their rebuttal[57] to ProPublica's article. Similar arguments were made quickly after the publication of ProPublica's article by bloggers including Abe Gong.[58] There has been extensive scholarship on the actuarial risk assessment in criminal justice that long predates the ProPublica debate; Berk et al.[59] provide a survey with commentary.

Variants of the trade-off between separation and sufficiency were shown by Chouldechova[60] and Kleinberg et al.[61] Each of them considered somewhat different criteria to trade off. Chouldechova's argument is very similar to the proof we presented that invokes the relationship between positive predictive value and true positive rate. Subsequent work[62] considers trade-offs between relaxed and approximate criteria. The other trade-off results presented in this chapter are new to this book. The proof of the proposition relating separation and independence for binary classifiers, as well as the counterexample for ternary classifiers, is due to Shira Mitchell and Jackie Shadlen, pointed out to us in personal communication.

The unidentifiability result for observational criteria is due to Hardt, Price, and Srebro[63], except for minor changes in the choice of graphical models and their interpretation.

[47] Mary A Lewis, "A Comparison of Three Models for Determining Test Fairness" (Federal Aviation Administration Washington DC Office of Aviation Medicine, 1978).

[48] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy, "Building Classifiers with Independency Constraints," in *In Proc. IEEE ICDMW*, 2009, 13–18; Faisal Kamiran and Toon Calders, "Classifying Without Discriminating," in *Proc. 2nd International Conference on Computer, Control and Communication*, 2009.

[49] Feldman et al., "Certifying and Removing Disparate Impact."

[50] Richard S. Zemel et al., "Learning Fair Representations," in *Proc. 30th ICML*, 2013.

[51] Dwork et al., "Fairness Through Awareness."

[52] Hardt, Price, and Srebro, "Equality of Opportunity in Supervised Learning."

[53] Muhammad Bilal Zafar et al., "Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification Without Disparate Mistreatment," in *Proc. 26th WWW*, 2017.

[54] Blake E. Woodworth et al., "Learning Non-Discriminatory Predictors," in *Proc. 30th COLT*, 2017, 1920–53.

[55] Darlington, "Another Look at 'Cultural Fairness'."

[56] Angwin et al., "Machine Bias."

[57] William Dieterich, 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.

[58] See this and subsequent posts.

[59] Richard Berk et al., "Fairness in Criminal Justice Risk Assessments: The State of the Art," *ArXiv E-Prints* 1703.09207 (2017).

[60] Alexandra Chouldechova, "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments," in *Proc. 3rd FATML*, 2016.

[61] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," *Proc. 8th ITCS*, 2017.

[62] Geoff Pleiss et al., "On Fairness and Calibration," in *Proc. 30th NIPS*, 2017.

[63] Hardt, Price, and Srebro, "Equality of Opportunity in Supervised Learning."

A dictionary of criteria

For convenience we collect some demographic fairness criteria below that have been proposed in the past (not necessarily including the original reference). We'll match them to their closest relative among the three criteria independence, separation, and sufficiency. This table is meant as a reference only and is not exhaustive. There is no need to memorize these different names.

Table 6: List of demographic fairness criteria

Name	Closest relative	Note	Reference
Statistical parity	Independence	Equivalent	Dwork et al. (2011)
Group fairness	Independence	Equivalent	
Demographic parity	Independence	Equivalent	
Conditional statistical parity	Independence	Relaxation	Corbett-Davies et al. (2017)
Darlington criterion (4)	Independence	Equivalent	Darlington (1971)
Equal opportunity	Separation	Relaxation	Hardt, Price, Srebro (2016)
Equalized odds	Separation	Equivalent	Hardt, Price, Srebro (2016)
Conditional procedure accuracy	Separation	Equivalent	Berk et al. (2017)
Avoiding disparate mistreatment	Separation	Equivalent	Zafar et al. (2017)
Balance for the negative class	Separation	Relaxation	Kleinberg, Mullainathan, Raghavan (2016)
Balance for the positive class	Separation	Relaxation	Kleinberg, Mullainathan, Raghavan (2016)
Predictive equality	Separation	Relaxation	Chouldechova (2016)
Equalized correlations	Separation	Relaxation	Woodworth (2017)
Darlington criterion (3)	Separation	Relaxation	Darlington (1971)
Cleary model	Sufficiency	Equivalent	Cleary (1966)
Conditional use accuracy	Sufficiency	Equivalent	Berk et al. (2017)
Predictive parity	Sufficiency	Relaxation	Chouldechova (2016)
Calibration within groups	Sufficiency	Equivalent	Chouldechova (2016)
Darlington criterion (1), (2)	Sufficiency	Relaxation	Darlington (1971)

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.

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. "Fairness in Criminal Justice Risk Assessments: The State of the Art." *ArXiv E-Prints* 1703.09207 (2017).

Bonham, Vence L, Shawneequa L Callier, and Charmaine D Royal. "Will Precision Medicine Move Us Beyond Race?" *The New England*

Journal of Medicine 374, no. 21 (2016): 2003.

Bouk, Dan. *How Our Days Became Numbered: Risk and the Rise of the Statistical Individual*. University of Chicago Press, 2015.

Calders, Toon, Faisal Kamiran, and Mykola Pechenizkiy. "Building Classifiers with Independency Constraints." In *In Proc. IEEE ICDMW*, 13–18, 2009.

Chouldechova, Alexandra. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments." In *Proc. 3rd FATML*, 2016.

Cleary, T Anne. "Test Bias: Prediction of Grades of Negro and White Students in Integrated Colleges." *Journal of Educational Measurement* 5, no. 2 (1968): 115–24.

———. "Test Bias: Validity of the Scholastic Aptitude Test for Negro and White Students in Integrated Colleges." *ETS Research Bulletin Series* 1966, no. 2 (1966): i–23.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. "Algorithmic Decision Making and the Cost of Fairness." *arXiv Preprint arXiv:1701.08230*, 2017.

Darlington, Richard B. "Another Look at 'Cultural Fairness'." *Journal of Educational Measurement* 8, no. 2 (1971): 71–82.

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.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. "Fairness Through Awareness." In *Proc. 3rd ITCS*, 214–26, 2012.

Einhorn, Hillel J, and Alan R Bass. "Methodological Considerations Relevant to Discrimination in Employment Testing." *Psychological Bulletin* 75, no. 4 (1971): 261.

Feldman, Michael, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. "Certifying and Removing Disparate Impact." In *Proc. 21st SIGKDD*. ACM, 2015.

Halligan, Steve, Douglas G. Altman, and Susan Mallett. "Disadvantages of Using the Area Under the Receiver Operating Characteristic Curve to Assess Imaging Tests: A Discussion and Proposal for an Alternative Approach." *European Radiology* 25, no. 4 (April 2015): 932–39.

Hardt, Moritz, Eric Price, and Nati Srebro. "Equality of Opportunity in Supervised Learning." In *Proc. 29th NIPS*, 3315–23, 2016.

Kamiran, Faisal, and Toon Calders. "Classifying Without Discriminating." In *Proc. 2nd International Conference on Computer, Control and Communication*, 2009.

Kleinberg, Jon M., Sendhil Mullainathan, and Manish Ragha-

van. "Inherent Trade-Offs in the Fair Determination of Risk Scores." *Proc. 8th ITCS*, 2017.

Lewis, Mary A. "A Comparison of Three Models for Determining Test Fairness." Federal Aviation Administration Washington DC Office of Aviation Medicine, 1978.

Liu, Lydia T., Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. "Delayed Impact of Fair Machine Learning." In *Proc. 35th ICML*, 3156–64, 2018.

Pleiss, Geoff, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. "On Fairness and Calibration." In *Proc. 30th NIPS*, 2017.

The Federal Reserve Board. "Report to the Congress on Credit Scoring and Its Effects on the Availability and Affordability of Credit." https://www.federalreserve.gov/boarddocs/rptcongress/creditscore/, 2007.

Thorndike, Robert L. "Concepts of Culture-Fairness." *Journal of Educational Measurement* 8, no. 2 (1971): 63–70.

Wasserman, Larry. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2010.

Woodworth, Blake E., Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. "Learning Non-Discriminatory Predictors." In *Proc. 30th COLT*, 1920–53, 2017.

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gómez Rodriguez, and Krishna P. Gummadi. "Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification Without Disparate Mistreatment." In *Proc. 26th WWW*, 2017.

Zemel, Richard S., Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. "Learning Fair Representations." In *Proc. 30th ICML*, 2013.