6

Detection Theory

Let's continue from Chapter 3, where you are the data scientist building the loan approval model for the (fictional) peer-to-peer lender ThriveGuild. As then, you are in the first stage of the machine learning lifecycle, working with the problem owner to specify the goals and indicators of the system. You have already clarified that safety is important, and that it is composed of two parts: basic performance (minimizing aleatoric uncertainty) and reliability (minimizing epistemic uncertainty). Now you want to go into greater depth in the problem specification for the first part: basic performance. (Reliability comes in Part 4 of the book.)

What are the different quantitative metrics you could use in translating the problem-specific goals (e.g. expected profit for the peer-to-peer lender) to machine learning quantities? Once you've reached the modeling stage of the lifecycle, how would you know you have a good model? Do you have any special considerations when producing a model for risk assessment rather than simply offering an approve/deny output?

Machine learning models are *decision functions*: based on the borrower's features, they decide a response that may lead to an autonomous approval/denial action or be used to support the decision making of the loan officer. The use of decision functions is known as statistical discrimination because we are distinguishing or differentiating one class label from the other. You should contrast the use of the term 'discrimination' here with *unwanted* discrimination that leads to systematic advantages to certain groups in the context of algorithmic fairness in Chapter 10. Discrimination here is simply telling the difference between things. Your favorite wine snob talking about their discriminative palate is a distinct concept from racial discrimination.

This chapter begins Part 3 of the book on basic modeling (see Figure 6.1 to remind yourself of the lay of the land) and uses *detection theory*, the study of *optimal* decision making in the case of categorical output responses,[1] to answer the questions above that you are struggling with.

[1]Estimation theory is the study of optimal decision making in the case of continuous output responses.

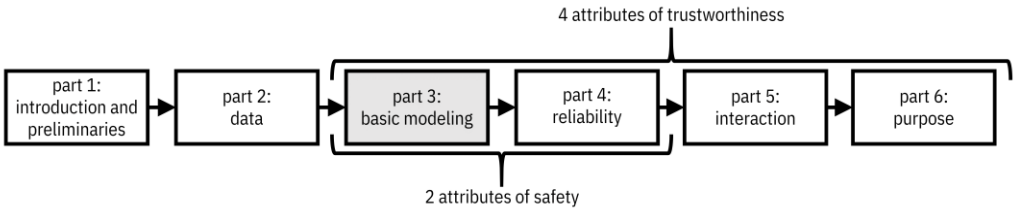

Figure 6.1. *Organization of the book. This third part focuses on the first attribute of trustworthiness, competence and credibility, which maps to machine learning models that are well-performing and accurate.* Accessible caption. A flow diagram from left to right with six boxes: part 1: introduction and preliminaries; part 2: data; part 3: basic modeling; part 4: reliability; part 5: interaction; part 6: purpose. Part 3 is highlighted. Parts 3–4 are labeled as attributes of safety. Parts 3–6 are labeled as attributes of trustworthiness.

Specifically, this chapter focuses on:

- selecting metrics to quantify the basic performance of your decision function (including ones that summarize performance across operating conditions),
- testing whether your decision function is as good as it could ever be, and
- differentiating performance in risk assessment problems from performance in binary decision problems.

6.1 Selecting Decision Function Metrics

You, the ThriveGuild data scientist, are faced with the *binary detection* problem, also known as the *binary hypothesis testing* problem, of predicting which loan applicants will default, and thereby which applications to deny.[2] Let Y be the loan approval decision with label $y = 0$ corresponding to deny and label $y = 1$ corresponding to approve. Feature vector X contains employment status, income, and other attributes. The value $y = 0$ is called a *negative* and the value $y = 1$ is called a *positive*. The random variables for the features and label are governed by the pmfs given the special name *likelihood functions* $p_{X|Y}(x \mid y = 0)$ and $p_{X|Y}(x \mid y = 1)$, as well as by *prior probabilities* $p_0 = P(Y = 0)$ and $p_1 = P(Y = 1) = 1 - p_0$. The basic task is to find a *decision function* $\hat{y}: \mathcal{X} \to \{0,1\}$ that predicts a label from the features.[3]

6.1.1 Quantifying the Possible Events

There are four possible events in the binary detection problem:

1. the decision function predicts 0 and the true label is 0,
2. the decision function predicts 0 and the true label is 1,

[2]For ease of explanation in this chapter and in later parts of the book, we mostly stick with the case of two label values and do not delve much into the case with more than two label values.

[3]This is also the basic task of supervised machine learning. In supervised learning, the decision function is based on data samples from (X, Y) rather than on the distributions; supervised learning is coming up soon enough in the next chapter, Chapter 7.

3. the decision function predicts 1 and the true label is 1, and
4. the decision function predicts 1 and the true label is 0.

These are known as *true negatives* (TN), *false negatives* (FN), *true positives* (TP), and *false positives* (FP), respectively. A true negative is denying an applicant who should be denied according to some ground truth, a false negative is denying an applicant who should be approved, a true positive is approving an applicant who should be approved, and a false positive is approving an applicant who should be denied. Let's organize these events in a table known as the *confusion matrix*:

	$Y = 1$	$Y = 0$
$\hat{y}(X) = 1$	TP	FP
$\hat{y}(X) = 0$	FN	TN

Equation 6.1

The probabilities of these events are:

$p_{\text{TP}} = P(\hat{y}(X) = 1 \mid Y = 1)$	$p_{\text{FP}} = P(\hat{y}(X) = 1 \mid Y = 0)$
$p_{\text{FN}} = P(\hat{y}(X) = 0 \mid Y = 1)$	$p_{\text{TN}} = P(\hat{y}(X) = 0 \mid Y = 0)$

Equation 6.2

These conditional probabilities are nothing more than a direct implementation of the definitions of the events. The probability p_{TN} is known as the *true negative rate* as well as the specificity and the selectivity. The probability p_{FN} is known as the *false negative rate* as well as the probability of missed detection and the miss rate. The probability p_{TP} is known as the *true positive rate* as well as the probability of detection, the recall, the sensitivity, and the power. The probability p_{FP} is known as the *false positive rate* as well as the probability of false alarm and the fall-out. The probabilities can be organized in a slightly different table as well:

$P(\hat{y}(X) \mid Y)$	$Y = 1$	$Y = 0$
$\hat{y}(X) = 1$	p_{TP}	p_{FP}
$\hat{y}(X) = 0$	p_{FN}	p_{TN}

Equation 6.3

These probabilities give you some quantities by which to understand the performance of the decision function $\hat{y}$. Selecting one over the other involves thinking about the events themselves and how they relate to the real-world problem. A false positive, approving an applicant who should be denied, means that a ThriveGuild lender has to bear the cost of a default, so it should be kept small. A false negative, denying an applicant who should be approved, is a lost opportunity for ThriveGuild to make a profit through the interest they charge.

The events above are conditioned on the true label. Conditioning on the predicted label also yields events and probabilities of interest in characterizing performance:

$P\big(Y \mid \hat{y}(X)\big)$	$Y = 1$	$Y = 0$
$\hat{y}(X) = 1$	p_{PPV}	p_{FDR}
$\hat{y}(X) = 0$	p_{FOR}	p_{NPV}

Equation 6.4

These conditional probabilities are reversed from Equation 6.2. The probability p_{NPV} is known as the *negative predictive value*. The probability p_{FOR} is known as the *false omission rate*. The probability p_{PPV} is known as the *positive predictive value* as well as the precision. The probability p_{FDR} is known as the *false discovery rate*. If you care about the quality of the decision function, focus on the first set (p_{TN}, p_{FN}, p_{TP}, p_{FP}). If you care about the quality of the predictions, focus on the second set (p_{NPV}, p_{FOR}, p_{PPV}, p_{FDR}).

When you need to numerically compute these probabilities, apply the decision function to several i.i.d. samples of (X, Y) and denote the number of TN, FN, TP, and FP events as n_{TN}, n_{FN}, n_{TP}, and n_{FP}, respectively. Then use the following estimates of the probabilities:

$p_{\text{TP}} \approx \dfrac{n_{\text{TP}}}{n_{\text{TP}} + n_{\text{FN}}}$	$p_{\text{FP}} \approx \dfrac{n_{\text{FP}}}{n_{\text{FP}} + n_{\text{TN}}}$
$p_{\text{FN}} \approx \dfrac{n_{\text{FN}}}{n_{\text{FN}} + n_{\text{TP}}}$	$p_{\text{TN}} \approx \dfrac{n_{\text{TN}}}{n_{\text{TN}} + n_{\text{FP}}}$

$p_{\text{PPV}} \approx \dfrac{n_{\text{TP}}}{n_{\text{TP}} + n_{\text{FP}}}$	$p_{\text{FDR}} \approx \dfrac{n_{\text{FP}}}{n_{\text{FP}} + n_{\text{TP}}}$
$p_{\text{FOR}} \approx \dfrac{n_{\text{FN}}}{n_{\text{FN}} + n_{\text{TN}}}$	$p_{\text{NPV}} \approx \dfrac{n_{\text{TN}}}{n_{\text{TN}} + n_{\text{FN}}}$

Equation 6.5

As an example, let's say that ThriveGuild makes the following number of decisions: $n_{\text{TN}} = 1234$, $n_{\text{FN}} = 73$, $n_{\text{TP}} = 843$, and $n_{\text{FP}} = 217$. You can estimate the various performance probabilities by plugging these numbers into the respective expressions above. The results are $p_{\text{TN}} \approx 0.85$, $p_{\text{FN}} \approx 0.08$, $p_{\text{TP}} \approx 0.92$, $p_{\text{FP}} \approx$

0.15, $p_{\text{NPV}} \approx 0.94$, $p_{\text{FOR}} \approx 0.06$, $p_{\text{PPV}} \approx 0.80$, and $p_{\text{FDR}} \approx 0.20$. These are all reasonably good values, but must ultimately be judged according to the ThriveGuild problem owner's goals and objectives.

6.1.2 Summary Performance Metrics

Collectively, false negatives and false positives are *errors*. The *probability of error*, also known as the error rate, is the sum of the false negative rate and false positive rate weighted by the prior probabilities:

$$p_{\text{E}} = p_0 p_{\text{FP}} + p_1 p_{\text{FN}}.$$

Equation 6.6

The *balanced* probability of error, also known as the balanced error rate, is the unweighted average of the false negative rate and false positive rate:

$$p_{\text{BE}} = \frac{1}{2} p_{\text{FP}} + \frac{1}{2} p_{\text{FN}}.$$

Equation 6.7

They summarize the basic performance of the decision function. Balancing is useful when there are a lot more data points with one label than the other, and you care about each type of error equally. *Accuracy*, the complement of the probability of error: $1 - p_{\text{E}}$, and *balanced accuracy*, the complement of the balanced probability of error: $1 - p_{\text{BE}}$, are sometimes easier for problem owners to appreciate than error rates.

The F_1-score, the harmonic mean of p_{TP} and p_{PPV}, is an accuracy-like summary measure to characterize the quality of a prediction rather than the decision function:

$$F_1 = 2 \frac{p_{\text{TP}} p_{\text{PPV}}}{p_{\text{TP}} + p_{\text{PPV}}}.$$

Equation 6.8

Continuing the example from before with $p_{\text{TP}} \approx 0.92$ and $p_{\text{PPV}} \approx 0.80$, let ThriveGuild's prior probability of receiving applications to be denied according to some ground truth be $p_0 = 0.65$ and applications to be approved be $p_1 = 0.35$. Then, plugging in to the relevant equations above, you'll find ThriveGuild to have $p_{\text{E}} \approx 0.13$, $p_{\text{BE}} \approx 0.11$, and $F_1 \approx 0.86$. Again, these are reasonable values that may be deemed acceptable to the problem owner.

As the data scientist, you can get pretty far with these abstract TN, FN, TP, and FP events, but they have to be put in the context of the problem owner's goals. ThriveGuild cares about making good bets on borrowers so that they are profitable. More generally across real-world applications, error events yield significant consequences to affected people including loss of life, loss of liberty, loss of livelihood, etc. Therefore, to truly characterize the performance of a decision function, it is important to consider the *costs* associated with the different events. You can capture these costs through a cost function $c\big(Y, \hat{y}(X)\big)$ and denote the costs as $c(0,0) = c_{00}$, $c(1,0) = c_{10}$, $c(1,1) = c_{11}$, and $c(0,1) = c_{01}$, respectively.

Taking costs into account, the characterization of performance for the decision function is known as the Bayes risk R:

$$R = (c_{10} - c_{00})p_0 p_{FP} + (c_{01} - c_{11})p_1 p_{FN} + c_{00}p_0 + c_{11}p_1.$$

Equation 6.9

Breaking the equation down, you'll see that the two error probabilities, p_{FP} and p_{FN} are the main components, multiplied by their relevant prior probabilities and costs. The costs of the non-error events appear just multiplied by their costs. The Bayes risk is the performance metric most often used in finding optimal decision functions. Actually finding the decision function is known as solving the *Bayesian detection* problem. Eliciting the cost function $c(\cdot,\cdot)$ for a given real-world problem from the problem owner is part of value alignment, described in Chapter 14.

A mental model or roadmap, shown in Figure 6.2, to hold throughout the rest of the chapter is that the Bayes risk and the Bayesian detection problem are the central concept, and all other concepts are related to the central concept in various ways and for various purposes. The terms and concepts that have not yet been defined and evaluated are coming up soon.

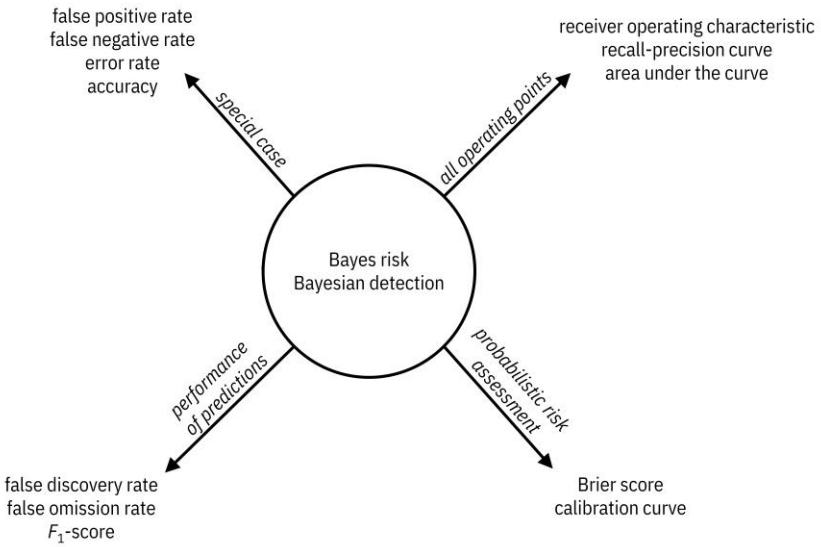

Figure 6.2. *A mental model for different concepts in detection theory surrounding the central concept of Bayes risk and Bayesian detection.* A diagram with Bayes risk and Bayesian detection at the center and four other groups of concepts radiating outwards. False positive rate, false negative rate, error rate, and accuracy are special cases. Receiver operating characteristic, recall-precision curve, and area under the curve arise when examining all operating points. Brier score and calibration curve arise in probabilistic risk assessment. False discover rate, false omission rate, and F_1-score relate to performance of predictions.

Because getting things right is a good thing, it is often assumed that there is no cost to correct decisions, i.e., $c_{00} = 0$ and $c_{11} = 0$, which is also assumed in this book going forward. In this case, the Bayes risk simplifies to:

$$R = c_{10}p_0p_{FP} + c_{01}p_1p_{FN}.$$

Equation 6.10

To arrive at this simplified equation, just insert zeros for c_{00} and c_{11} in Equation 6.9. The Bayes risk with $c_{10} = 1$ and $c_{01} = 1$ is the probability of error.

We are implicitly assuming that $c(\cdot,\cdot)$ does not depend on X except through $\hat{y}(X)$. This assumption is not required, but made for simplicity. You can easily imagine scenarios in which the cost of a decision depends on the feature. For example, if one of the features used in the loan approval decision by ThriveGuild is the value of the loan, the cost of an error (monetary loss) depends on that feature. Nevertheless, for simplicity, we usually make the assumption that the cost function does not explicitly depend on the feature value. For example, under this assumption, the cost of a false negative may be $c_{10} = 100{,}000$ dollars and the cost of a false positive $c_{01} = 50{,}000$ dollars for all applicants.

6.1.3 Accounting for Different Operating Points

The Bayes risk is all well and good if there is a fixed set of prior probabilities and a fixed set of costs, but things change. If the economy improves, potential borrowers might become more reliable in loan repayment. If a different problem owner comes in and has a different interpretation of opportunity cost, then the cost of false negatives c_{10} changes. How should you think about the performance of decision functions across different sets of those values, known as different *operating points*?

Many decision functions are parameterized by a threshold η (including the optimal decision function that will be demonstrated in Section 6.2). You can change the decision function to be more or less forgiving of false positives or false negatives, but not both at the same time. Varying η explores this tradeoff and yields different error probability pairs $(p_{\text{FP}}, p_{\text{FN}})$, i.e. different operating points. Equivalently, different operating points correspond to different false positive rate and true positive rate pairs $(p_{\text{FP}}, p_{\text{TP}})$. The curve traced out on the p_{FP}–p_{TP} plane as the parameter η is varied from zero to infinity is the *receiver operating characteristic* (ROC). The ROC takes values $(p_{\text{FP}} = 0, p_{\text{TP}} = 0)$ when $\eta \to \infty$ and $(p_{\text{FP}} = 1, p_{\text{TP}} = 1)$ when $\eta \to 0$. You can understand this because at one extreme, the decision function always says $\hat{y} = 0$; in this case there are no FPs and no TPs. At the other extreme, the decision function always says $\hat{y} = 1$; in this case all decisions are either FPs or TPs.

The ROC is a concave, nondecreasing function illustrated in Figure 6.3. The closer to the top left corner it goes, the better. The best ROC for discrimination goes straight up to (0,1) and then makes a sharp turn to the right. The worst ROC is the diagonal line connecting (0,0) and (1,1) achieved by random guessing. The area under the ROC, also known as the *area under the curve* (AUC) synthesizes performance across all operating points and should be selected as a metric when it is likely that the same threshold-parameterized decision function will be applied in very different operating conditions. Given the shapes of the worst (diagonal line) and best (straight up and then straight to the right) ROC curves, you can see that the AUC ranges from 0.5 (area of bottom right triangle) to 1 (area of entire square).[4]

[4]The recall-precision curve is an alternative to understand performance across operating points. It is the curve traced out on the p_{PPV}–p_{TP} plane starting at $(p_{PPV} = 0, p_{TP} = 1)$ and ending at $(p_{PPV} = 1, p_{TP} = 0)$. It has a one-to-one mapping with the ROC and is more easily understood by some people. Jesse Davis and Mark Goadrich. "The Relationship Between Precision-Recall

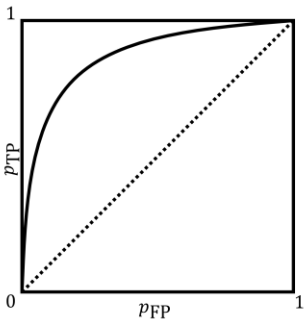

Figure 6.3. *An example receiver operating characteristic (ROC).* Accessible caption. A plot with p_{TP} on the vertical axis and p_{FP} on the horizontal axis. Both axes range from 0 to 1. A dashed diagonal line goes from (0,0) to (1,1) and corresponds to random guessing. A solid concave curve, the ROC, goes from (0,0) to (1,1) staying above and to the left of the diagonal line.

6.2 The Best That You Can Ever Do

As the ThriveGuild data scientist, you have given the problem owner an entire menu of basic performance measures to select from and indicated when different choices are more and less appropriate. The Bayes risk is the most encompassing and most often used performance characterization for a decision function. Let's say that Bayes risk was chosen in the problem specification stage of the machine learning lifecycle, including selecting the costs. Now you are in the modeling stage and need to figure out if the model is performing well. The best way to do that is to optimize the Bayes risk to obtain the best possible decision function with the smallest Bayes risk and compare the current model's Bayes risk to it.

> "The predictability ceiling is often ignored in mainstream ML research. Every prediction problem has an upper bound for prediction—the Bayes-optimal performance. If you don't have a good sense of what it is for your problem, you are in the dark."
>
> —Mert R. Sabuncu, computer scientist at Cornell University

Let us denote the best possible decision function as $\hat{y}^*(\cdot)$ and its corresponding Bayes risk as R^*. They are specified using the minimization of the expected cost:

---

and ROC Curves." In: *Proceedings of the International Conference on Machine Learning*. Pittsburgh, Pennsylvania, USA, Jun. 2006, pp. 233–240.

$$\hat{y}^*(\cdot) = \arg\min_{\hat{y}(\cdot)} E\big[c\big(Y, \hat{y}(X)\big)\big],$$

Equation 6.11

where the expectation is over both X and Y. Because it achieves the minimal cost, the function $\hat{y}^*(\cdot)$ is the best possible $\hat{y}(\cdot)$ by definition. Whatever Bayes risk R^* it has, no other decision function can have a lower Bayes risk R.

We aren't going to work it out here, but the solution to the minimization problem in Equation 6.11 is the Bayes optimal decision function, and takes the following form:

$$\hat{y}^*(\cdot) = \begin{cases} 0, & \Lambda(x) \leq \eta \\ 1, & \Lambda(x) > \eta \end{cases}$$

Equation 6.12

where $\Lambda(x)$, known as the *likelihood ratio*, is defined as:

$$\Lambda(x) = \frac{p_{X|Y}(\,x \mid Y = 1\,)}{p_{X|Y}(\,x \mid Y = 0\,)}$$

Equation 6.13

and η, known as the *threshold*, is defined as:

$$\eta = \frac{c_{10} p_0}{c_{01} p_1}.$$

Equation 6.14

The likelihood ratio is as its name says: it is the ratio of the likelihood functions. It is a scalar value even if the features X are multivariate. As the ratio of two non-negative pdf values, it has the range $[0, \infty)$ and can be viewed as a random variable. The threshold is made up of both costs and prior probabilities. This optimal decision function $\hat{y}^*(\cdot)$ given in Equation 6.12 is known as the *likelihood ratio test*.

6.2.1 Example

As an example, let ThriveGuild's loan approval decision be determined solely by one feature X: the income of the applicant. Recall that we modeled income to be exponentially-distributed in Chapter 3. Specifically, let $p_{X|Y}(\,x \mid Y = 1\,) = 0.5e^{-0.5x}$ and $p_{X|Y}(\,x \mid Y = 0\,) = e^{-x}$, both for $x \geq 0$. Like earlier in this chapter, $p_0 = 0.65$, $p_1 = 0.35$, $c_{10} = 100000$, and $c_{01} = 50000$. Then simply plugging in to Equation 6.13, you'll get:

$$\Lambda(x) = \frac{0.5e^{-0.5x}}{e^{-x}} = 0.5e^{0.5x}, \qquad x \geq 0$$

Equation 6.15

and plugging in to Equation 6.14, you'll get:

$$\eta = \frac{100000}{50000}\frac{0.65}{0.35} = 3.7.$$

Equation 6.16

Plugging these expressions into the Bayes optimal decision function given in Equation 6.12, you'll get:

$$\hat{y}^*(x) = \begin{cases} 0, & 0.5e^{0.5x} \leq 3.7 \\ 1, & 0.5e^{0.5x} > 3.7 \end{cases}$$

Equation 6.17

which can be simplified to:

$$\hat{y}^*(x) = \begin{cases} 0, & x \leq 4 \\ 1, & x > 4 \end{cases}$$

Equation 6.18

by multiplying both sides of the inequalities in both cases by 2, taking the natural logarithm, and then multiplying by 2 again. Applicants with an income less than or equal to 4 are denied and applicants with an income greater than 4 are approved. The expected value of $X \mid Y = 1$ is 2 and the expected value of $X \mid Y = 0$ is 1. Thus in this example, an applicant's income has to be quite a bit higher than the mean to be approved.

You should use the Bayes-optimal risk R^* to lower bound the performance of any machine learning classifier that you might try for a given data distribution.[5] No matter how hard you work or how creative you are, you can never overcome the Bayes limit. So you should be happy if you get close. If the Bayes-optimal risk itself is too high, then the thing to do is to go back to the data understanding and data preparation stages of the machine learning lifecycle and get more informative data.

6.3 Risk Assessment and Calibration

To approve or to deny, that is the question for ThriveGuild. Or is it? Maybe the question is actually: what is the probability that the borrower will default? Maybe the problem is not binary classification, but probabilistic risk assessment. It is certainly an option for you, the data scientist, and the problem owner to consider during problem specification. Thresholding a probabilistic risk assessment yields a classification, but there are a few subtleties for you to weigh.

[5]There are techniques for estimating the Bayes risk of a dataset without having access to its underlying probability distribution. Ryan Theisen, Huan Wang, Lav R. Varshney, Caiming Xiong, and Richard Socher. "Evaluating State-of-the-Art Classification Models Against Bayes Optimality" In: *Advances in Neural Information Processing Systems* 34 (Dec. 2021).

The likelihood ratio ranges from zero to infinity and the threshold value $\eta = 1$ is optimal for equal priors and equal costs. Applying any monotonically increasing function to both the likelihood ratio and the threshold still yields a Bayes optimal decision function with the same risk R^*. That is,

$$\hat{y}^*(\cdot) = \begin{cases} 0, & g\big(\Lambda(x)\big) \leq g(\eta) \\ 1, & g\big(\Lambda(x)\big) > g(\eta) \end{cases}$$

Equation 6.19

for any monotonically increasing function $g(\cdot)$ is still optimal.

It is somewhat more natural to think of a *score* $s(x)$ to be in the range [0,1] because it corresponds to the label values $y \in \{0,1\}$ and could also potentially be interpreted as a probability. The score, a continuous-valued output of the decision function, can then be thought of as a confidence in the prediction and be obtained by applying a suitable g function to the likelihood ratio. In this case, 0.5 is the threshold for equal priors and costs. Intermediate score values are less confident and extreme score values (towards 0 and 1) are more confident. Just as the likelihood ratio may be viewed as a random variable, the score may also be viewed as a random variable S. The *Brier score* is an appropriate performance metric for the continuous-valued output score of the decision function:

$$\text{Brier score} = E[(S - Y)^2].$$

Equation 6.20

It is the mean-squared error of the score S with respect to the true label Y. For a finite number of samples $\{(s_1, y_1), \ldots, (s_n, y_n)\}$, you can compute it as:

$$\text{Brier score} = \frac{1}{n}\sum_{j=1}^{n}\big(s_j - y_j\big)^2.$$

Equation 6.21

The Brier score decomposes into the sum of two separable components: *calibration* and *refinement*.[6] The concept of calibration is that the predicted score corresponds to the proportion of positive true labels. For example, a bunch of data points all having a calibrated score of $s = 0.7$ implies that 70% of them have true label $y = 1$ and 30% of them have true label $y = 0$. Said another way, perfect calibration implies that the probability of the true label Y being 1 given the predicted score S being s is the value s itself: $P(Y = 1 \mid S = s) = s$. Calibration is important for probabilistic risk assessments: a perfectly calibrated score can be interpreted as a probability of predicting one class or the other. It is also an important concept for evaluating causal inference methods, described in Chapter 8, for algorithmic fairness, described in Chapter 10, and for communicating uncertainty, described in Chapter 13.

[6]José Hernández-Orallo, Peter Flach, and Cèsar Ferri. "A Unified View of Performance Metrics: Translating Threshold Choice into Expected Classification Loss." In: *Journal of Machine Learning Research* 13 (Oct. 2012), pp. 2813–2869.

Since any monotonically increasing transformation $g(\cdot)$ can be applied to a decision function without changing its ability to discriminate, you can improve the calibration of a decision function by finding a better $g(\cdot)$. The calibration loss quantitatively captures how close a decision function is to perfect calibration. The refinement loss is a sort of variance of how tightly the true labels distribute around a given score. For $\{(s_1, y_1), \ldots, (s_n, y_n)\}$ that have been sorted by their score values and binned into k groups $\{\mathcal{B}_1, \ldots, \mathcal{B}_k\}$ with average values $\{(\bar{s}_1, \bar{y}_1), \ldots, (\bar{s}_k, \bar{y}_k)\}$ within the bins

$$\text{calibration loss} = \frac{1}{n}\sum_{i=1}^{k} \|\mathcal{B}_i\| \, (\bar{s}_i - \bar{y}_i)^2$$

$$\text{refinement loss} = \frac{1}{n}\sum_{i=1}^{k} \|\mathcal{B}_i\| \, \bar{y}_i(1 - \bar{y}_i).$$

Equation 6.22

As stated earlier, the sum of the calibration loss and refinement loss is the Brier score.

A *calibration curve*, also known as a reliability diagram, shows the $(\bar{s}_k, \bar{y}_k)$ values as a plot. One example is shown in Figure 6.4. The closer to a straight diagonal from (0,0) to (1,1), the better. Plotting this curve is a good diagnostic tool for you to understand the calibration of a decision function.

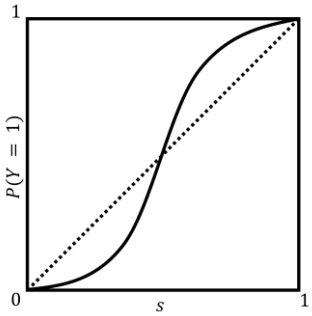

Figure 6.4. *An example calibration curve.* Accessible caption. A plot with $P(Y = 1)$ on the vertical axis and s on the horizontal axis. Both axes range from 0 to 1. A dashed diagonal line goes from (0,0) to (1,1) and corresponds to perfect calibration. A solid S-shaped curve, the calibration curve, goes from (0,0) to (1,1) starting below and to the right of the diagonal line before crossing over to being above and to the left of the diagonal line.

6.4 Summary

- Four possible events result from binary decisions: false negatives, true negatives, false positives, and true positives.
- Different ways to combine the probabilities of these events lead to classifier performance metrics

appropriate for different real-world contexts.

- One important one is Bayes risk: the combination of the false negative probability and false positive probability weighted by both the costs of those errors and the prior probabilities of the labels. It is the basic basic performance measure for the first attribute of safety and trustworthiness.
- Detection theory, the study of optimal decisions, which provides fundamental limits to how well machine learning models may ever perform is a tool for you to assess the basic performance of your models.
- Decision functions may output continuous-valued scores rather than only hard, zero or one, decisions. Scores indicate confidence in a prediction. Calibrated scores are those for which the score value is the probability of a sample belonging to a label class.