— 3 —

Detecting differences

A single statistical problem illuminates much of the mathematical tools necessary for benchmarking. The key lesson is that sample requirements grow quadratically in the inverse of the difference we try to detect.

Source: The Emerging Science of Machine Learning Benchmarks. M. Hardt, 2025. URL: https://mlbenchmarks.org. Compiled on 2025-10-30.

The most basic problem in benchmarking is to decide which of two candidate models is better. Beating the previous best is the essential mechanic of benchmarking and it's fundamentally about comparing two models. Model comparisons, by extension, also support model rankings. By repeatedly comparing models, we can compute or maintain a complete ranking. When a new model enters the leaderboard, we just need to compare it to each other model in order to determine it's place in the ranking.

The previous chapter was about optimal prediction at the population level. Optimal predictors minimize risk, that is, a loss function in expectation over a fixed population. From this perspective, the better of two models is the one that achieves smaller risk. Likewise, a good benchmark should give us a sorting of models in increasing order of risk.

Focusing on risk and nothing else means that we treat models as *black boxes*; we don't look at the wires under the hood, the trained weights of the models. We only compute loss values on input and output pairs. Two representationally different models that compute the same function are equivalent in this view. In practice, there could be other important measures, such as model size, that are not just about input-output behavior. But for now we restrict our attention to comparing models in terms of the risk that they achieve.

3.1 Model comparisons from small samples

The problem of model comparison gets more challenging when we don't know the full population. A population is an idealized construct describing the broad set of all relevant problem instances. In reality, we almost always only have a subset of relevant test cases $S = \{(x_1, y_n), \dots, (x_n, y_n)\}$ consisting of n labeled data points at our disposal. Recall, the empirical risk of a model f on the set S is the sample average of the loss function

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

We previously introduced empirical risk as an objective for model training, but now we're going to use it for model evaluation. Given two models f, g it makes sense to pick whichever model has the lower empirical risk:

$$\text{If } R_S(f) < R_S(g), \text{ pick } f, \text{ else } g\,.$$

This is a simple and natural strategy that's well-defined for any non-empty set of test cases. We don't need to assume anything about the test cases to define this rule. Much of model evaluation and benchmarking reduces to empirical risk comparisons. In this chapter, we'll work out when this selection rule correctly identifies the model of lower risk. This isn't necessarily true and it must depend on the set of test cases that we have.

The problem we face is that our sample is just a small part of the population. If the set is too small, intuitively speaking, there's no reason to believe that it could support reliable comparisons. Even if the set is large enough, perhaps the test cases systematically favor one model over the other. Or the test cases might fail to cover some important parts of the population. For these reasons, we need to make additional assumptions in order to analyze our canonical model comparison rule.

The iid assumption

To analyze model comparisons, we assume that each labeled data point in our sample is a random draw from the population. The draws are independent and each draw follows the data-generating distribution. Machine learning researchers call this the *iid assumption*: labeled data are drawn *independently* and they are *identically distributed* in the sense that they are all drawn from the same joint distribution (X,Y). By extension, for a fixed function f, the loss values $\ell(f(x_1),y_1),\dots,\ell(f(x_n),y_n)$ are also iid with respect to the real-valued random variable $\ell(f(X),Y)$. An implication of the iid assumption is that the ordering of data points doesn't matter. All permutations of the data follow the same distribution.

On the face of it, the iid assumption is a brazenly strong assumption. When things go wrong, researchers therefore often diagnose the problem as a failure of the iid assumption. But this diagnosis misses the more fundamental problem: The world isn't a probability distribution to begin with. The major leap of faith is to assume that the world is a distribution. This is the leap of faith we took in Chapter 2—the heavy lift of the astronomical conception. Once we have already assumed that the population is a probability distribution, to assume that, in addition, we can sample from this distribution repeatedly is the lesser stretch of imagination. If the population is indeed a giant database of instances—e.g., an internet crawl, or the U.S. Census population—we can ensure that the iid assumption holds by randomly sampling from the set of all instances.

Failure of the iid assumption is therefore almost always an indirect way of saying that the world isn't a distribution. Ultimately, though, the assumption again derives its justification from its utility. In return for making the iid assumption we gain useful insights. Primarily, we'll be able to answer the question how large a sample we need to tell apart models of a certain difference in risk. Equivalently, we learn what kind of a difference we can tell apart with the sample size we have. Such *sample size calculations* are the basis for any benchmark design.

From model comparisons to coin tosses

Let's break down the problem of model comparisons into its bare essentials. For now, we focus on classification error as the loss function. The extension to many other loss functions isn't difficult. Recall, taking the classification error as a loss function, the risk of a model f equals

$$R(f) = \mathbb{P}\{f(X) \neq Y\} = \mathbb{E}\left[\mathbb{1}\{f(X) \neq Y\}\right].$$

The indicator $\mathbb{1}\{f(X) \neq Y\}$ is a random variable that follows a *Bernoulli* distribution $B(p)$ with bias p, where $p = R(f)$ is the classification error of the model. Now, suppose a competing model g has risk $q = R(f)$. Without loss of generality, assume $p < q$ and let $\epsilon = q - p > 0$ denote the difference between the two. Our empirical risk based comparison correctly identifies f as the better model provided that

$$R_S(f) < R(f) + \frac{\epsilon}{2} \quad \text{and} \quad R_S(g) > R(g) - \frac{\epsilon}{2}.$$

This condition implies that the differences between risk and empirical risk don't flip our comparison.

From the start, we know that the expected empirical risk equals risk:

$$\mathbb{E}\, R_S(f) = R(f)$$

Here, the expectation is over the random sample. The fact holds so long as our sample follows the data-generating distribution. This needs no independence. What's missing then is a way to argue that the empirical risk is close to its mean. This is where independence comes in.

To further simplify, let's focus on the case where $q = 1/2$ and $p = 1/2 - \epsilon$. This turns out to be the hardest case so that we really don't give up anything here. In this case, the distribution of the indicator $\mathbb{1}\{g(X) \neq Y\}$ now is an unbiased coin toss $B(1/2)$, and the distribution of the indicator $\mathbb{1}\{f(X) \neq Y\}$ is an ϵ-biased coin $B(1/2 - \epsilon)$.

The problem we end up with is that of distinguishing an unbiased coin from an ϵ-biased coin from n samples. Although seemingly simple, this problem illustrates most of the statistical tools we'll need throughout this book. Coin tossing is a surprisingly universal problem in machine learning. Many problems can be reduced to or expressed as an instance of coin tossing.

3.2 *Coin tossing*

Consider the distribution P of a fair coin toss. Recall that this is a Bernoulli distribution with parameter 1/2. For convenience, we'll slightly reparameterize things. A sample from P takes on the value 1 representing *heads* with probability 1/2 and the value -1 representing *tails* also with probability 1/2. The choice of the values $\{-1, 1\}$ is just for mathematical convenience. The choice $\{0, 1\}$ works out similarly.

Contrast the distribution P with the distribution Q of a *biased* coin toss corresponding to a Bernoulli distribution with parameter $(1+\epsilon)/2$. We assume $\epsilon \in (0, 1]$ and think of $\epsilon > 0$ as a small bias term that makes the value 1 just a bit more likely. This way of writing the bias term will make the formulas a bit nicer. That's all.

Suppose you observe n independent random samples from one of the two distributions:

$$+1, -1, +1, +1, -1, -1, +1, +1, +1, +1, +1, -1, 1, -1, -1, -1, +1, +1, -1$$

Can you tell which distribution these samples come from? More quantitatively, we can ask how many samples we need to decide which distribution it is.

Let's start with a basic observation. The distribution P has mean 0, whereas the distribution Q has mean ϵ. This observation motivates a simple test.

Denote the samples by $X_1, \dots, X_n$ and their sum $S = \sum_{i=1}^{n} X_i$. Test if

$$S \overset{?}{>} t \qquad \text{with} \quad t = \frac{\epsilon n}{2}\,.$$

In words, we test if the sum of our samples is closer to 0 or ϵn. If it's closer to 0 we declare the coin unbiased, else biased.

When does this work? What we hope is that for large enough n, coins drawn from P will rarely sum up to a value larger than our decision threshold.

Likewise, coins drawn from Q should rarely sum up to a value smaller than our decision threshold. What can we hope for?

A variance calculation is a good first step. The variance of a coin toss equals 1 in the unbiased case. It is strictly less than 1 in the biased case. The variance of a sum of independent random variables equals the sum of the variances. Therefore,

$$\mathbb{V}\,S = \mathbb{E}(S - \mathbb{E}\,S)^2 = \textstyle\sum_{i=1}^n \mathbb{V}\,X_i \le n\,,$$

with equality in the last step for the fair coin. A *standard deviation* is

$$\sqrt{\mathbb{V}\,S} \le \sqrt{n}\,.$$

Deviations of multiple standard deviations are unlikely. The reason is that for any random variable X with mean μ and standard deviation σ we have:

$$\mathbb{P}\{|X-\mu| > a\sigma\} = \mathbb{P}\left\{\frac{(X-\mu)^2}{a^2\sigma^2} > 1\right\} \le \mathbb{E}\left[\frac{(X-\mu)^2}{a^2\sigma^2}\right] = \frac{\mathbb{E}(X-\mu)^2}{a^2\sigma^2} = \frac{1}{a^2}$$

This basic fact is *Chebyshev's inequality*. We used a clever trick in deriving it: The probability that a nonnegative random variable exceeds 1 is bounded by its expectation. Whatever probability mass the random variable has above the value 1, this mass gets multiplied by a value of at least 1 when computing the expectation. We'll use this fact once more below.

Chebyshev's inequality holds for any random variable with finite mean and variance. In particular, it suggests that the sum S is unlikely to deviate from its mean by significantly more than $\sqrt{n}$. We therefore expect our test to work in the parameter regime:

$$\epsilon n \gg \sqrt{n} \quad \Leftrightarrow \quad \epsilon \gg \frac{1}{\sqrt{n}} \quad \Leftrightarrow \quad n \gg \frac{1}{\epsilon^2}$$

The notation "$a \gg b$" I use here means that $a \ge cb$ for a sufficiently large constant factor $c > 0$.

From here on, we'll work out why this is, in fact, the correct answer. Detecting a bias of order ϵ requires $1/\epsilon^2$ samples. This turns out to be necessary and sufficient up to constant factors. Put differently, from n samples we can generally detect differences of magnitude $1/\sqrt{n}$. This is a fundamental statistical fact. By and large, it's the only statistical fact we'll need. It comes in many variants that go by many different names. But the fundamentals are always the same.

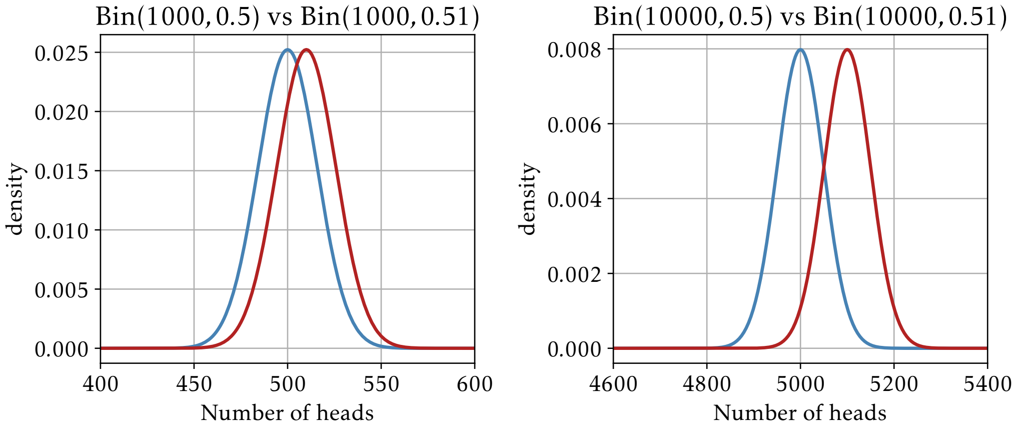

Figure 3.1: Distribution of unbiased and biased coin tosses with 1000 samples (left) and 10000 samples right

Now, this might seem disappointing. Does detecting a 1% bias in a coin really require about 10000 samples?

The answer, in general, is *yes*.

The number of heads in a sequence of n coin tosses with bias p follows a binomial distribution, denoted $\mathrm{Bin}(n, p)$. It's helpful to plot the distribution of coin tosses with biases, say, 0.5 and 0.51, respectively. The result should convince you that after 1000 samples, the corresponding binomial distributions are intuitively quite close. After 10000 samples we have a chance to distinguish them.

There's one exception to this sobering fact. Suppose we want to distinguish a coin of bias 1 from a coin of bias $1-\epsilon$. One coin never comes up tails. For the other coin we expect that after $n \gg 1/\epsilon$ samples we see tails at least once. So, here we only need $n \approx 1/\epsilon$ samples to distinguish the coins. It's tempting to think that this is the correct answer in general, but it really only holds in this special case.

3.3 *Concentration inequalities*

Chebyshev's inequality is an example of a *concentration inequality*. It bounds the probability a random variable deviates from its mean. The inequality holds for *any* random variable with finite mean and variance. There's something a lot stronger we can say for sums of independent random variables.

Coin tosses are Bernoulli trials. The sum of n independent Bernoulli trials follows a Binomial distribution. A standard concentration inequality for sums of independent Bernoulli trials, from the family of *Chernoff bounds*, gives us:

- $\mathbb{P}_{X_i \sim P}\{S \geq \mathbb{E}\,S + a\} \leq \exp(-a^2/2n)$
- $\mathbb{P}_{X_i \sim Q}\{S \leq \mathbb{E}\,S - a\} \leq \exp(-a^2/2n)$

Summing up both *tail bounds* on the right hand side and plugging in $a = \epsilon n/2$, we find that the failure probability of our distinguishing test is bounded by

$$\mathop{\mathbb{P}}_{X_i \sim P}\{S \geq t\} + \mathop{\mathbb{P}}_{X_i \sim Q}\{S \leq t\} \leq 2\exp(-\epsilon^2 n/8).$$

What does this mean? We can see from the bound that for $n > 8/\epsilon^2$, we begin to get a small error probability. For $n \gg 1/\epsilon^2$ our test works with high probability. Whereas Chebyshev gave us a quadratic tail σ^2/a^2, Chernoff gives us an exponentially small probability $\exp(-a^2/2\sigma^2)$. The tail bound is the same tail bound we'd have for a normal distribution of standard deviation σ. As a result, such a tail is called *sub-gaussian*.

To prove the Chernoff bound, we need Markov's inequality. We already saw the proof of Markov's inequality in our derivation of Chebyshev's inequality. It's worth repeating.

Proposition 1. ***(Markov's inequality)*** *Assume X is a nonnegative random variable. Then, for any $a > 0$, we have*

$$\mathbb{P}\{X \geq a\} \leq \frac{\mathbb{E}\,X}{a}$$

Proof.

$$\mathbb{P}\{X \geq a\} = \mathbb{P}\left\{\frac{1}{a}X \geq 1\right\} \leq \mathbb{E}\left[\frac{1}{a}X\right] = \frac{1}{a}\mathbb{E}[X]$$

Here, the inequality uses nonnegativity of X. The last equality follows from the linearity of expectation.

□

We'll prove the most basic Chernoff bound. It gets the key idea across without the more delicate optimizations of other bounds.

Theorem 1. ***(Chernoff's bound)*** *Let $X_1, \dots, X_n$ be independent unbiased coin tosses in $\{-1, 1\}$. Let $S = \sum_i X_i$. Then, for every $a > 0$,*

$$\mathbb{P}\{S \geq a\} \leq \exp(-a^2/2n).$$

Proof. First, exponentiate the random variable to make it nonnegative and apply Markov's inequality. Indeed, for any $a > 0$ and $\lambda > 0$, we have

$$\mathbb{P}\{S \geq a\} = \mathbb{P}\{e^{\lambda S} > e^{\lambda a}\} \leq \mathbb{E}[e^{\lambda S}]e^{-\lambda a}$$

Second, use the independence of the random variables $X_1, \dots, X_n$:

$$\mathbb{E}[e^{\lambda S}] = \mathbb{E}[e^{\lambda \sum_{i=1}^n X_i}] = \mathbb{E}\left[\prod_{i=1}^n e^{\lambda X_i}\right] = \prod_{i=1}^n \mathbb{E}\left[e^{\lambda X_i}\right]$$

Third, apply a clever inequality to each term:

$$\mathbb{E}\left[e^{\lambda X_i}\right] = \frac{e^{\lambda}}{2} + \frac{e^{-\lambda}}{2} \leq e^{\lambda^2/2}$$

To show this inequality, you can compare the Taylor expansion of the left hand side and right hand side.

Finally, put it all together and optimize over the choice of λ :

$$\mathbb{P}\{S \geq a\} \leq \left(e^{\lambda^2/2}\right)^n e^{-\lambda a} = e^{\lambda^2 n/2 - \lambda a} = e^{-a^2/2n} \quad \text{for} \quad \lambda = \frac{a}{n}.$$

□

This gives us the bound we need for unbiased coins. The same bound extends to biased coin tosses. The proof changes slightly. We first center the coin tosses by subtracting the mean. We then repeat the proof and use the magic inequality:

$$(1+\epsilon)e^{(1-\epsilon)\lambda} + (1-\epsilon)e^{-(1+\epsilon)\lambda} \leq e^{\lambda} + e^{-\lambda}$$

The basic proof template behind Chernoff's bound gives a whole slew of related inequalities that hold under various different assumptions. In all cases, we need the random variables to be "independent enough" and "bounded enough".

For many applications in machine learning, a close relative of Chernoff's bound called Hoeffding's inequality comes in handy. It generalizes the bound we proved as well as the bound we need for biased coins.

Theorem 2. ***(Hoeffding's inequality)*** *Let $X_1, \dots, X_n$ be independent random variables with $X_i \in [b_i, c_i]$. Let $S = X_1 + X_2 + \cdots + X_n$ denote their sum. Then, for every positive scalar $a > 0$, we have*

$$\mathbb{P}\{|S - \mathbb{E}\, S| \geq a\} \leq 2\exp\left(-\frac{2a^2}{\sum_{i=1}^n (c_i - b_i)^2}\right).$$

Proof. We start by proving

$$\mathbb{P}\{S - \mathbb{E}\,S \geq a\} \leq \exp\left(-2a^2/\textstyle\sum_{i=1}^n (c_i - b_i)^2\right).$$

Following Chernoff's recipe, apply Markov's inequality and exponentiation:

$$\mathbb{P}\{S - \mathbb{E}\,S \geq a\} = \mathbb{P}\left\{e^{\lambda(S - \mathbb{E}\,S)} \geq e^{\lambda a}\right\} \leq \mathbb{E}\left[e^{\lambda(S - \mathbb{E}\,S)}\right] e^{-\lambda a}$$

By independence,

$$\mathbb{E}\,e^{\lambda(S - \mathbb{E}\,S)} = \textstyle\prod_{i=1}^n \mathbb{E}\left[e^{\lambda(X_i - \mathbb{E}\,X_i)}\right].$$

Now, use the magic inequality

$$\mathbb{E}\left[e^{\lambda(X_i - \mathbb{E}\,X_i)}\right] \leq e^{\lambda^2 (c_i - b_i)^2/8}.$$

The identity follows from some calculus and we omit the proof. So, we conclude

$$\mathbb{P}\{S - \mathbb{E}\,S \geq a\} \leq e^{\lambda^2 \sum_{i=1}^n (c_i - b_i)^2/8} e^{-\lambda a}.$$

Setting $\lambda = 4a/\sum_{i=1}^n (c_i - b_i)^2$, we get

$$\mathbb{P}\{S - \mathbb{E}\,S \geq a\} \leq \exp\left(-2a^2/\textstyle\sum_{i=1}^n (c_i - b_i)^2\right)$$

That's what we wanted. Using the same proof, we can also show

$$\mathbb{P}\{S - \mathbb{E}\,S \leq -a\} \leq \exp\left(-2a^2/\textstyle\sum_{i=1}^n (c_i - b_i)^2\right).$$

Combining the two inequalities via the union bound gives us the statement of the lemma.

□

Hoeffding's inequality generalizes Chernoff's bound. We recover the Chernoff bound above by setting $b_i = -1$ and $c_i = 1$. Note that this gives a constant of $(c_i - b_i)^2 = 4$ in the denominator so that we end up with the same constant as the original Chernoff bound.

Concentration inequalities give us conditions under which our simple mean-based test succeeds with overwhelmingly high probability. But they say nothing about whether there couldn't be some other kind of test that requires far fewer samples.

3.4 Distances between distributions

Comparing the mean of n coin tosses gave us a test to distinguish biased from unbiased coins provided that $n \gg 1/\epsilon^2$. However, we don't know yet that we can't do better. Perhaps our analysis of the mean-based test was far from optimal. Or, perhaps there is an entirely different test that would work much better. Who says that we have to look at the mean?

Formally, it's not clear yet that we can't do better than $1/\epsilon^2$ by way of some more sophisticated test. We'll now rule out any such possibility. To do so, we'll define a distance between distributions that corresponds to the best that a bounded test function can do in distinguishing the two distributions. We'll then show that the distance between n unbiased coins and n biased coins is small so long as $n \ll 1/\epsilon^2$.

Total variation distance

For two discrete distributions P and Q taking on values in a discrete set $\mathcal{X}$, the total variation distance is defined as

$$\mathrm{TV}(P,Q) = \frac{1}{2}\sum_{x\in\mathcal{X}} |P(x) - Q(x)|$$

Total variation distance is directly related to distinguishing probabilities:

$$\mathrm{TV}(P,Q) = \sup_{T\subseteq\mathcal{X}} |P(T) - Q(T)| = \sup_{\theta\colon\mathcal{X}\to\{0,1\}} \left| \mathop{\mathbb{E}}_{x\sim P} \theta(x) - \mathop{\mathbb{E}}_{x\sim Q} \theta(x) \right|$$

The last expression correspond to the best that a bounded test function θ can do in distinguishing P from Q.

If we can show that total variation distance is small, no bounded test function can distinguish between the two distributions with high probability. Think of the test function θ as trying to output 1 if we believe the distribution is P and outputting 0 if the distribution is Q. Our mean-based test is one such test function.

But note that we apply our test function to a sample of n coins. That is the same as a sample from the product distribution P^n that gives n independent draws from P. So, we're generally interested in the total variation distance between the P^n and Q^n. Our goal is to show

$$n \ll \frac{1}{\epsilon^2} \quad \Longrightarrow \quad \mathrm{TV}(P^n, Q^n) \le \frac{1}{2}\,.$$

This means that no test can distinguish P and Q from n samples with high probability. The choice of 1/2 is arbitrary.

Although we're interested in multiple coin tosses, it makes sense to start by analyzing a single coin toss. Returning to our coins P, unbiased, and Q with bias $(1+\epsilon)/2$, a calculation shows $\mathrm{TV}(P,Q)=\epsilon$. But this is a bit of an issue. The total variation distance between a coin with bias 1 and a coin with bias $1-\epsilon$ is also ϵ. From the perspective of total variation, these two pairs of distributions have the same distance. But we argued earlier that we expect these two distinguishing problems to be fundamentally different. One requires $1/\epsilon^2$ samples, the other $1/\epsilon$.

Moreover, trying to calculate the total variation distance between multiple coin tosses from first principles gets tricky. Thankfully, there is another distance measure that comes to the rescue.

Hellinger distance

A different distance measure gets us unstuck. It's called *Hellinger distance*. Define its square as:

$$\mathrm{H}^2(P,Q)=\frac{1}{2}\sum_x\left(\sqrt{P(x)}-\sqrt{Q(x)}\right)^2=1-\sum_x\sqrt{P(x)Q(x)}.$$

Let's see what happens for an unbiased coin P and a coin Q with bias $(1+\epsilon)/2$.

Claim 1. $\mathrm{H}^2(P,Q)\le\epsilon^2/4$

Proof. By Taylor approximation, for $\epsilon\in[-1/2,1/2]$,

$$\sqrt{1+\epsilon}=1+\tfrac{1}{2}\epsilon-\tfrac{1}{8}\epsilon^2+\delta$$

up to error δ with $|\delta|\le\frac{1}{32}\epsilon^2$.

Hence,

$$\begin{aligned}\mathrm{H}^2(P,Q)&=1-\sqrt{\frac{1}{2}\left(\frac{1+\epsilon}{2}\right)}-\sqrt{\frac{1}{2}\left(\frac{1-\epsilon}{2}\right)}\\&=1-\frac{1}{2}\left(\sqrt{1+\epsilon}+\sqrt{1-\epsilon}\right)\\&\le\epsilon^2/4.\end{aligned}$$

□

In contrast, for the bias 1 coin and bias $1-\epsilon$ coin, we get Hellinger squared distance $1-\sqrt{1-\epsilon}\approx\epsilon/2$. We're off to a good start! Hellinger distance is sensitive to the two different kinds of biases. We don't yet know how to analyze multiple coin tosses. But here we got lucky again.

Hellinger squared distance behaves nicely on product distributions; it satisfies a *direct sum theorem*. When we have a sample of n independent coin tosses, its distribution is a product distribution over n-tuples. Denoting by P_1P_2 the product distribution with components P_1 and P_2 we have:

$$\begin{aligned}\mathrm{H}^2(P_1P_2,Q_1Q_2) &= 1-\sum_{x,y}\sqrt{P_1(x)P_2(y)Q_1(x)Q_2(y)}\\ &= 1-\left(\sum_x\sqrt{P_1(x)Q_1(x)}\right)\cdot\left(\sum_y\sqrt{P_2(y)Q_2(y)}\right)\\ &= 1-\left(1-\mathrm{H}^2(P_1,Q_1)\right)\cdot\left(1-\mathrm{H}^2(P_2,Q_2)\right)\end{aligned}$$

In particular,

$$\mathrm{H}^2(P_1P_2,Q_1Q_2)\le\mathrm{H}^2(P_1,Q_1)+\mathrm{H}^2(P_2,Q_2).$$

For n independent identically distributed draws, denoted P^n and Q^n, respectively,

$$\mathrm{H}^2(P^n,Q^n)\le n\cdot\mathrm{H}^2(P,Q)$$

By the direct sum theorem for Hellinger distance, in our coin tossing example, we have

$$\mathrm{H}^2(P^n,Q^n)\le\epsilon^2 n/4.$$

So, for $n\le 1/2\epsilon^2$ the Hellinger distance between n coin tosses from either distribution is at most 1/8. We now relate this back to distinguishing probabilities.

All that's left to do is to go back to total variation distance. This is possible, since the total variation distance is sandwiched by the Hellinger squared distance and the Hellinger distance up to a constant:

$$\mathrm{H}^2(P,Q)\le\mathrm{TV}(P,Q)\le\sqrt{2}H(P,Q)$$

In particular, in our coin tossing example,

$$\mathrm{TV}(P^n,Q^n)\le\sqrt{2\mathrm{H}^2(P^n,Q^n)}\le\sqrt{2/8}=1/2.$$

This means that no test can distinguish the two coins from fewer than $1/2\epsilon^2$ many samples without being wrong a good fraction of the time.

More precisely, if a test θ accepts a sample from P^n with probability p it must also accept a sample from Q^n with probability $p-1/2$. For example, if the former is 0.95, the latter is at least 0.45. If our test accepts the sample, we're left with significant uncertainty.

3.5 From coin tosses back to benchmarking

Let's map what we learned back to benchmarking. Fix any bounded loss function $\ell\colon \mathcal{Y}\times\mathcal{Y}\to[0,1]$. Note that any loss function mapping into $[-B,B]$ can be rescaled to the interval $[0,1]$. Fix a model $f:\mathcal{X}\to\mathcal{Y}$.

Assume the sample $S=\{(x_1,y_1),\dots,(x_n,y_n)\}$ is a random draw from the population. Since each data point is distributed according to the data-generating distribution, we have

$$\mathbb{E}\,R_S(f)=R(f)\,.$$

In addition, the random variables $Z_1,\dots,Z_n$ with $Z_i=\ell(x_i,y_i)$ are independent of each other. Therefore,

$$\mathbb{V}\,R_S(f)=\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\,Z_i\le\frac{1}{n}\,.$$

Here we used that a bounded loss in $[0,1]$ has variance at most 1. Sums of independent random variables concentrate around their mean up to a standard deviation. Therefore, we expect

$$R_S(f)\approx R(f)\pm O(1/\sqrt{n})\,.$$

For bounded losses in $[0,1]$, Hoeffding's bound makes this precise. For a fixed predictor f, we have

$$\mathbb{P}_S\left\{|R_S(f)-R(f)|\ge\frac{t}{\sqrt{n}}\right\}\le 2\exp\left(-2t^2\right).$$

Setting $t=2$, we get $2\exp(-2t^2)\approx 0.00067$. So, risk and empirical risk are within $2/\sqrt{n}$ of each other with high probability.

There are always two equivalent ways to think about this:

1. Estimation errors shrink as the inverse square root of sample size: Given n samples, we can detect differences of about $1/\sqrt{n}$.
2. Sample size requirements grow quadratically in the inverse of the difference we try to detect: Given that the difference is $\epsilon > 0$, we need about $1/\epsilon^2$ samples to be able to detect it.

Coin tossing shows that there is no way around this quadratic dependence. We need at least $1/2\epsilon^2$ samples to distinguish the accuracy of a classifier that does random guessing from one that's ϵ better than random guessing. We can quibble about the precise constants in this trade-off. Although we got it down to a relatively small constant factor between our upper and lower bound already. But there's nothing we can do about the fundamental quadratic relationship. This is a basic lesson for benchmark design. If we want to be able to detect differences of magnitude ϵ on our test set, we better include $1/\epsilon^2$ data points.

There is only that one exception. We can distinguish a perfect predictor f with $R(f) = 0$ from an imperfect predictor g with $R(g) > \epsilon$ using only $O(1/\epsilon)$ samples. Simply draw new data points until we observe a nonzero loss. If after $n \gg 1/\epsilon$ samples we haven't seen a nonzero loss, it's likely that our predictor is perfect.

Multiple model evaluations

What about multiple comparison between classifiers? To reason about a fixed family $\mathcal{F} = \{f_1, \dots, f_k\}$ of k predictors we can couple Hoeffding's bound with the union bound:

$$\mathop{\mathbb{P}}_{S}\left\{\exists f \in \mathcal{F} \colon\ |R_S[f] - R[f]| \geq \frac{t}{\sqrt{n}}\right\} \leq 2k \exp\left(-2t^2\right).$$

In other words, we multiply our tail bound by the number of functions we evaluate. This is called a *Bonferroni* correction in the context of multiple hypothesis testing: If you try out k different things, you need to multiply your p-values by a factor k. It works the same way here. If you evaluate k different models, the probability of a large deviation goes up by a factor k. Thanks to the subgaussian tail bound that Hoeffding provides, this isn't too bad. We can get rid of the factor k by choosing t a factor $\sqrt{\log k}$ larger, since $\exp(-\sqrt{\log k}^2) = 1/k$.

This means the maximum error ϵ that we encounter among k estimates grows as

$$\epsilon = O\left(\sqrt{\frac{\log k}{n}}\right).$$

This shows that if we want to detect differences of magnitude ϵ among k classifiers we need:

$$n \gg \frac{\log k}{\epsilon^2} \quad \Leftrightarrow \quad \epsilon \gg \sqrt{\frac{\log k}{n}}$$

To summarize, sample requirements grow quadratically in the inverse difference, but only logarithmically in the number of functions.

Notes

Wasserman's *All of Statistics*[wasserman2013all] makes for excellent background reading on this topic.

Concentration inequalities. Concentration inequalities are an extremely well studied subject in probability theory. Boucheron, Lugosi, and Bousquet give a primer on concentration inequalities.[boucheron2003concentration] The text by Boucheron, Lugosi, and Massart goes into full detail.[boucheron2013concentration] Vershynin's *High-Dimensional Probability* gives advanced treatment of concentration inequalities.[vershynin2018high]

Chernoff's bound from 1952 and Hoeffding's inequality followed in 1963. Many powerful tools beyond Hoeffding's inequality are available. In particular, other tools come in handy in applications where the loss function is not always bounded. We already saw one example in Chapter 2. Cross entropy can be unbounded if the predictor puts zero weight on the correct class. There are also numerous concentration inequalities in cases where the samples are not fully independent. Often enough, some concentration of measure still works out.

I've often found that when all else fails, *McDiarmid's inequality* might help. It's an instance of a bounded difference method. All you need is that the quantity you're interested in doesn't change much if you toggle any of the random variables. The function computing the quantity may be complicated; it doesn't have to be the mean or anything. The random variables have to be independent, but they don't have to be the sample points. They could be some underlying random seed used to compute the function.

Theorem 3. ***(McDiarmid's inequality)*** *Let $X_1,\dots,X_n$ be independent random variables taking values in a set B and assume that the function $f\colon B^n \to \mathbb{R}$ satisfies the bounded difference condition:*

$$\sup_{x_1,\dots,x_n\in B}\ \sup_{i\in[n],x\in B} |f(x_1,\dots,x_n) - f(x_1,\dots,x_{i-1},x,x_{i+1},\dots,x_n)| \le c_i$$

Then, for all $a > 0$,

$$\mathbb{P}\{|f(X_1,\dots,X_n) - \mathbb{E}f(X_1,\dots,X_n)|\} \le 2\exp\left(-2a^2/\textstyle\sum_{i=1}^n c_i^2\right).$$

Statistical lower bounds. What we proved for coin tosses is called a minimax lower bound on statistical estimation. The method of lower bounding sample requirements of statistical estimation using total variation distance and Hellinger distance is due to LeCam.[lecam1973convergence] For the mathematically inclined, Tsybakov's text, translated from French, is a standard reference on this topic.[tsybakov2009nonparametric]