The learning challenge

Goal

There is some underlying function $f : \mathcal{X} \to \mathcal{Y}$ that captures an input-output relationship which we would like to estimate

Assumption

We do not know f, but we get to observe example input-output pairs which are *generated independently at random*

- we draw $\mathbf{x}_i$ according to some unknown distribution and get to observe the pair $(\mathbf{x}_i, f(\mathbf{x}_i))$
- we draw $\mathbf{x}_i$ according to some unknown distribution and get to observe the pair $(\mathbf{x}_i, f(\mathbf{x}_i) + n_i)$, where n_i represents "noise" with an unknown distribution
- we draw pairs $(\mathbf{x}_i, y_i)$ according to some unknown joint distribution

A first model of learning

Let's restrict our attention to binary classification

- our labels belong to $\mathcal{Y} = \{1, 0\}$ (or $\mathcal{Y} = \{+1, -1\}$)

We observe the data

$$\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$$

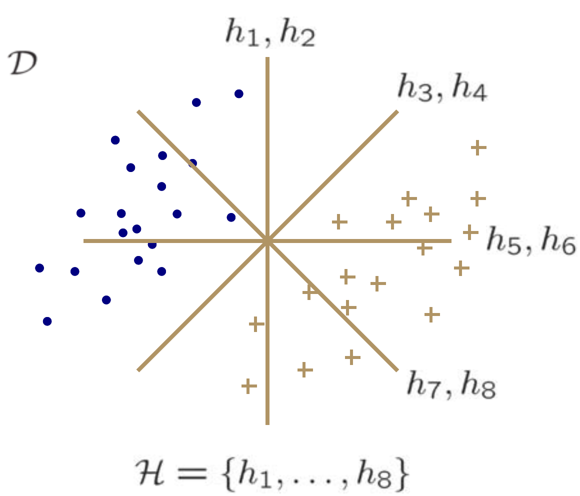
where each $\mathbf{x}_i \in \mathbb{R}^d$

Suppose we are given a list of possible hypotheses

$$\mathcal{H} = \{h_1, \dots, h_m\}$$

From the *training data* $\mathcal{D}$, we would like to select the best possible hypothesis from $\mathcal{H}$

Example



Empirical risk

Recall from last time our definition of *risk* and its empirical counterpart

Risk: $R(h_j) := \mathbb{P}[h_j(X) \neq Y]$

Empirical risk: $\widehat{R}_n(h_j) := \frac{1}{n} \sum_{i=1}^{n} 1_{\{i : h_j(\mathbf{x}_i) \neq y_i\}}(i)$

In our definition of $\widehat{R}_n(h_j)$ we make use of the *indicator function*

$$1_{\{A\}}(t) = \begin{cases} 1 & \text{if } t \in A \\ 0 & \text{if } t \notin A \end{cases}$$

The empirical risk $\widehat{R}_n(h_j)$ gives us an estimate of the true risk $R(h_j)$, and from the law of large numbers we know that $\widehat{R}_n(h_j) \to R(h_j)$ as $n \to \infty$

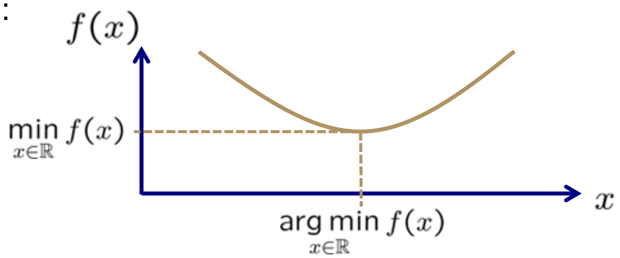
Empirical risk minimization (ERM)

We want to choose a hypothesis from $\mathcal{H}$ that achieves a small risk

Since $\widehat{R}_n(h_j)$ is supposed to be a good estimate of $R(h_j)$, an incredibly natural (and common) strategy is to pick

$$h^* = \arg\min_{h_j \in \mathcal{H}} \widehat{R}_n(h_j)$$

Aside:



The risk in ERM

As we discussed last time, as long as we have enough data, for any particular hypothesis h_j, we expect $\widehat{R}_n(h_j) \approx R(h_j)$

However, if m is very large, then we can also expect that there are some h_k for which $\widehat{R}_n(h_k) \ll R(h_k)$

Thus, what can we say about $R(h^*)$?

- We know that $\widehat{R}_n(h^*)$ is as small as it can be
  - this *could* be because $R(h^*)$ is small
  - *or*, it could be because $\widehat{R}_n(h_k) \ll R(h_k)$ for some h_k
- Which explanation is more likely?
  - *it depends...* just how large is m?

Confidence bounds

We would like to be able to give quantitative answers to questions along the lines of:

If we are deciding between m hypotheses, how much data do we need (how large does n need to be) to ensure that

$$|\widehat{R}_n(h^*) - R(h^*)| \le \epsilon$$

for some $\epsilon \in (0, 1)$ that we define in advance

Asymptotic results like the law of large numbers and the central limit theorem do not give us answers to these questions

Instead, we need *nonasymptotic* results about

$$\mathbb{P}\left[|\widehat{R}_n(h^*) - R(h^*)| \le \epsilon\right]$$

Too much randomness?

Our goal is ultimately to show how to make

$$\mathbb{P}\left[|\widehat{R}_n(h^*) - R(h^*)| \le \epsilon\right] \approx 1$$

by setting n appropriately

What is random here?

- the training data $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$
- $\widehat{R}_n(h_1), \widehat{R}_n(h_2), \dots, \widehat{R}_n(h_m)$, because each depends on $\mathcal{D}$
- h^*, because it depends on $\widehat{R}_n(h_1), \widehat{R}_n(h_2), \dots, \widehat{R}_n(h_m)$

In order to tease all of this apart, let's begin by going back to just a single hypothesis h_j and studying

$$\mathbb{P}\left[|\widehat{R}_n(h_j) - R(h_j)| \le \epsilon\right]$$

Bounding the error

We want to calculate

$$\mathbb{P}\left[|\widehat{R}_n(h_j) - R(h_j)| \le \epsilon\right]$$

Note that $\widehat{R}_n(h_j)$ is a random variable

- we have $\widehat{R}_n(h_j) = \frac{1}{n}\sum_{i=1}^n 1_{\{i:h_j(\mathbf{x}_i)\neq y_i\}}(i) = \frac{1}{n}\sum_{i=1}^n S_i$
  where the S_i are *Bernoulli* random variables
- thus, $n\widehat{R}_n(h_j)$ is a *Binomial* random variable
- since $\mathbb{P}[S_i = 1] = \mathbb{P}[h_j(\mathbf{x}_i) \neq y_i] = R(h_j)$, we have that

$$\begin{aligned}\mathbb{E}\left[n\widehat{R}_n(h_j)\right] = \mathbb{E}\left[\sum_{i=1}^n S_i\right] &= \sum_{i=1}^n \mathbb{E}\left[S_i\right] \\ &= n\mathbb{P}\left[h_j(\mathbf{x}_i) \neq y_i\right] \\ &= nR(h_j)\end{aligned}$$

Deviation from the mean

Thus, an equivalent way to think about our problem is that we would like to calculate

$$\mathbb{P}\left[|n\widehat{R}_n(h_j) - nR(h_j)| \le n\epsilon\right]$$

and this is just asking about the probability that a Binomial random variable will deviate from it's mean by more than $n\epsilon$

If $F(k)$ represents the cumulative distribution function (CDF) of our binomial random variable, then we can write

$$\begin{aligned}&\mathbb{P}\left[|n\widehat{R}_n(h_j) - nR(h_j)| \le n\epsilon\right] \\ &\qquad = F(nR(h_j) + n\epsilon) - F(nR(h_j) - n\epsilon)\end{aligned}$$

Bounding the deviation

Unfortunately, the CDF we are interested in is given by

$$F(k) = \sum_{i=0}^{\lfloor k \rfloor} \binom{n}{i} R(h_j)^i (1 - R(h_j))^{n-i}$$

This has no nice closed form expression, and is rather unwieldy to work with and doesn't give us much intuition

Instead of calculating the probability exactly, it is enough to get a good bound of the form

$$\mathbb{P}\left[|\widehat{R}_n(h_j) - R(h_j)| \le \epsilon\right] \ge 1 - ?$$

or equivalently

$$\mathbb{P}\left[|\widehat{R}_n(h_j) - R(h_j)| \ge \epsilon\right] \le ?$$

Concentration inequalities

An inequality of the form

$$\mathbb{P}\left[|\widehat{R}_n(h_j) - R(h_j)| \ge \epsilon\right] \le ?$$

tell us how a particular random variable (in this case $\widehat{R}_n(h_j)$) *concentrates* around its mean

There are a number of different concentration inequalities that give us various bounds along these lines

We will start with a very simple one, and then build up to a stronger result

Markov's inequality

The simplest of these results is *Markov's inequality*

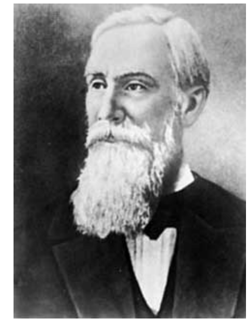
Let X be any nonnegative random variable. Then for any $t \geq 0$,

$$\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}[X]}{t}.$$

This is cool on its own, but can be leveraged to say even more since for any strictly nonotonically increasing (nonnegative-valued) function ϕ

$$\mathbb{P}[X \geq t] = \mathbb{P}[\phi(X) \geq \phi(t)] \leq \frac{\mathbb{E}[\phi(X)]}{\phi(t)}.$$

Chebyshev's inequality



As an example, *Chebyshev's inequality* states that for any random variable X,

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq \epsilon] \leq \frac{\text{var}(X)}{\epsilon^2}$$

Proof.

Note that $|X - \mathbb{E}[X]|$ is a nonnegative random variable. Thus we can apply Markov's inequality to obtain

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq \epsilon] = \mathbb{P}[|X - \mathbb{E}[X]|^2 \geq \epsilon^2]$$

$$\leq \frac{\mathbb{E}\left[|X - \mathbb{E}[X]|^2\right]}{\epsilon^2} = \frac{\text{var}(X)}{\epsilon^2}$$

Proof of Markov (Part 1)

There is a *simple* proof of Markov if you know the (*super useful!*) fact that for any nonnegative random variable X

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}[X \geq x] dx$$

Proof.

Note that we can write $X = \int_0^X dx = \int_0^\infty 1_{\{x \leq X\}}(x) dx$.

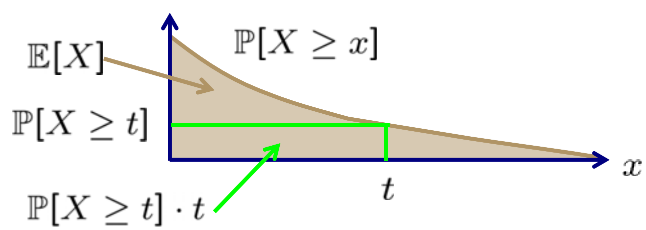
Thus $\mathbb{E}[X] = \mathbb{E}\left[\int_0^\infty 1_{\{x \leq X\}}(x) dx\right]$

$$= \int_0^\infty \mathbb{E}\left[1_{\{x \leq X\}}(x)\right] dx$$

$$= \int_0^\infty \mathbb{P}\left[X \geq x\right] dx$$

Proof of Markov (Part 2)

We can visualize this result as



Thus, we can immediate see that we must have

$$\mathbb{E}[X] \geq \mathbb{P}[X \geq t] \cdot t$$

and hence

$$\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}[X]}{t}.$$

Hoeffding's inequality

Chebyshev's inequality gives us the kind of result we are after, but it is too *loose* to be of practical use

Hoeffding's inequality assumes a bit more about our random variable beyond having finite variance, but gets us a much tighter and more useful result:

Let $X_1, \ldots, X_n$ be independent bounded random variables, i.e., random variables such that $\mathbb{P}[X_i \in [a,b]] = 1$ for all i

Let $S_n = \sum_{i=1}^n X_i$. Then for any $\epsilon > 0$, we have

$$\mathbb{P}\left[|S_n - \mathbb{E}[S_n]| \geq \epsilon\right] \leq 2e^{-2\epsilon^2/n(b-a)^2}$$

Chernoff's bounding method

To prove this result, we will use a similar approach as in Chebyshev's inequality

To begin consider only the upper tail inequality:

$$\begin{aligned}
\mathbb{P}[S_n - \mathbb{E}[S_n] \geq \epsilon] &= \mathbb{P}[\lambda(S_n - \mathbb{E}[S_n]) \geq \lambda\epsilon] && (\lambda > 0) \\
&= \mathbb{P}[e^{\lambda(S_n - \mathbb{E}[S_n])} \geq e^{\lambda\epsilon}] \\
&\leq \frac{\mathbb{E}\left[e^{\lambda(S_n - \mathbb{E}[S_n])}\right]}{e^{\lambda\epsilon}} && \text{(Markov)} \\
&= e^{-\lambda\epsilon}\mathbb{E}\left[e^{\lambda(X_1 - \mathbb{E}[X_1] + \cdots + X_n - \mathbb{E}[X_n])}\right] \\
&= e^{-\lambda\epsilon}\prod_{i=1}^n \mathbb{E}\left[e^{\lambda(X_i - \mathbb{E}[X_i])}\right] && \text{(Independence)}
\end{aligned}$$

Hoeffding's Lemma

It is not obvious, but it is also not too hard to show, that

$$\mathbb{E}\left[e^{\lambda(X_i - \mathbb{E}[X_i])}\right] \leq e^{\lambda^2(b-a)^2/8}$$

The proof uses convexity and then gets a bound using a Taylor series expansion

Plugging this in, we obtain that for any $\lambda > 0$, we have

$$\mathbb{P}[S_n - \mathbb{E}[S_n] \geq \epsilon] \leq e^{-\lambda\epsilon}e^{n\lambda^2(b-a)^2/8}$$

By setting $\lambda = 4\epsilon/n(b-a)^2$, we have

$$\begin{aligned}
\mathbb{P}[S_n - \mathbb{E}[S_n] \geq \epsilon] &\leq e^{-4\epsilon^2/n(b-a)^2}e^{2\epsilon^2/n(b-a)^2} \\
&= e^{-2\epsilon^2/n(b-a)^2}
\end{aligned}$$

Putting it all together

Thus we have proven that

$$\mathbb{P}[S_n - \mathbb{E}[S_n] \geq \epsilon] \leq e^{-2\epsilon^2/n(b-a)^2}$$

An analogous argument proves

$$\mathbb{P}[\mathbb{E}[S_n] - S_n \geq \epsilon] \leq e^{-2\epsilon^2/n(b-a)^2}$$

Combined, these give

$$\mathbb{P}\left[|S_n - \mathbb{E}[S_n]| \geq \epsilon\right] \leq 2e^{-2\epsilon^2/n(b-a)^2}$$

Special case: Binomials

If the X_i are Bernoulli random variables, then S_n is a Binomial random variable and Hoeffding's inequality becomes

$$\mathbb{P}\left[|S_n - \mathbb{E}[S_n]| \geq \epsilon\right] \leq 2e^{-2\epsilon^2/n}$$

Finally going back to our original problem, this means that Hoeffding yields the bound

$$\begin{aligned}\mathbb{P}&\left[|\widehat{R}_n(h_j) - R(h_j)| \geq \epsilon\right] \\ &= \mathbb{P}\left[|n\widehat{R}_n(h_j) - nR(h_j)| \geq n\epsilon\right] \\ &\leq 2e^{-2\epsilon^2 n}\end{aligned}$$

Multiple hypotheses

Thus, after much effort, we have that for a particular hypothesis h_j,

$$\mathbb{P}\left[|\widehat{R}_n(h_j) - R(h_j)| \geq \epsilon\right] \leq 2e^{-2\epsilon^2 n}$$

However, we are ultimately interested in h^*, not just a single hypothesis h_j

One way to argue that $|\widehat{R}_n(h^*) - R(h^*)| \leq \epsilon$ is to ensure that $|\widehat{R}_n(h_j) - R(h_j)| \leq \epsilon$ ***simultaneously*** for all j

Equivalently, we can try to bound the probability that *any* hypothesis h_j has an empirical risk that deviates from its mean by more than ϵ

Formal statement

We can express this mathematically as

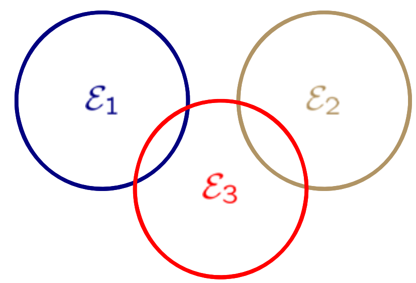
$$\begin{aligned}\mathbb{P}\left[\left|\widehat{R}_n(h^*) - R(h^*)\right| \geq \epsilon\right] \leq \mathbb{P}\Big[& \left|\widehat{R}_n(h_1) - R(h_1)\right| \geq \epsilon \\ \textbf{or} & \left|\widehat{R}_n(h_2) - R(h_2)\right| \geq \epsilon \\ & \vdots \\ \textbf{or} & \left|\widehat{R}_n(h_m) - R(h_m)\right| \geq \epsilon\Big]\end{aligned}$$

We can bound this using something called the *union bound*

Union bound

Union bound For any sequence of events $\mathcal{E}_1, \ldots, \mathcal{E}_m$

$$\mathbb{P}\left[\mathcal{E}_1 \cup \cdots \cup \mathcal{E}_m\right] \leq \mathbb{P}\left[\mathcal{E}_1\right] + \cdots + \mathbb{P}\left[\mathcal{E}_m\right]$$



The events in our case are given by

$$\mathcal{E}_j = \left|\widehat{R}_n(h_j) - R(h_j)\right| \geq \epsilon$$

Final result

$$\mathbb{P}\left[\left|\widehat{R}_n(h^*) - R(h^*)\right| \geq \epsilon\right] \leq \mathbb{P}\Big[\quad \left|\widehat{R}_n(h_1) - R(h_1)\right| \geq \epsilon$$

$$\textbf{or } \left|\widehat{R}_n(h_2) - R(h_2)\right| \geq \epsilon$$

$$\vdots$$

$$\textbf{or } \left|\widehat{R}_n(h_m) - R(h_m)\right| \geq \epsilon \Big]$$

$$\leq \sum_{j=1}^{m} \mathbb{P}\left[\left|\widehat{R}_n(h_j) - R(h_j)\right| > \epsilon\right]$$

$$\leq \sum_{j=1}^{m} 2e^{-2\epsilon^2 n}$$

$$= 2me^{-2\epsilon^2 n}$$

Interpretation

We went through all of this work to show that

$$\mathbb{P}\left[\left|\widehat{R}_n(h^*) - R(h^*)\right| \geq \epsilon\right] \leq 2me^{-2\epsilon^2 n}$$

linearly increasing

exponentially decreasing

When can we be confident that $\widehat{R}_n(h^*) \approx R(h^*)$?

Note that $2me^{-2\epsilon^2 n} = e^{\log(2m) - 2\epsilon^2 n}$

As long as m isn't too big ($m \lessapprox e^n$) then we can be reasonably confident that $\widehat{R}_n(h^*) \approx R(h^*)$

Are we learning yet?

It's not quite enough to have $\widehat{R}_n(h^*) \approx R(h^*)$

Ideally, we would like to have $R(h^*) \approx 0$

Note that if $\widehat{R}_n(h^*) \approx R(h^*)$, then $\widehat{R}_n(h^*) \approx 0$ implies that $R(h^*) \approx 0$

The learning problem

1. Can we ensure that $R(h^*)$ is close to $\widehat{R}_n(h^*)$? ✓
2. Can we make $\widehat{R}_n(h^*)$ small enough?