Beginning of 03/28/2025

... someone didn't show up again :(

Learning Theory

We first go over some result in **sampling**. Suppose given an unknown distribution X we want to estimate $\mu = \mathbb{E}X$. Question: how many samples are sufficient for us to obtain a sufficiently good estimate of μ, via the sample mean $\hat{\mu} = \sum_{i=1}^{m} X_i/m$? A standard Chernoff bound argument gives

$$\mathbb{P}(|\hat{\mu} - \mu| \geqslant \epsilon\mu) \leqslant \delta \qquad \text{for } m = \frac{3}{\mu\epsilon^2}\log\frac{2}{\delta}. \tag{*}$$

For example consider the estimation of π. Given a unit circle we tangentially inscribe it in a square with side length 2. We uniformly sample points from the square and count the fraction of points lying inside the circle, and by multiplying this fraction by 4 we obtain an estimate $\hat{\pi}$. By (*) and some algebra, if $m = 12\epsilon^{-2}\log(1/\delta)$,

$$\mathbb{P}(|\hat{\pi} - \pi| \geqslant \epsilon\pi) \leqslant \delta.$$

This approach is called **rejection sampling** that captures the fraction of desired samples.

What is learning theory? Given some data $(x, y) \sim D$ for some distribution D, we want to learn a model to map x to y. More formally, given **concept class** $\mathcal{C}$ of **hypotheses** (functions), in which each h_i maps x to a prediction $h(x)$, we want to find the most suitable one. For a hypothesis h, define

$$\text{Error}(h) = \mathbb{P}_{(x,y)\sim D}[h(x) \neq y]$$

the probability of a wrong prediction. *We will assume that there exists a perfect hypothesis h^* where* $\text{Error}(h^*) = 0$.

A simple example of $\mathcal{C}$ could be the set of **linear classifiers** $h_{a,b}(x) = \text{sgn}(a^T x + b)$, each of which partitions the space into two halfspaces. Suppose we sample m data points, $(x_1, y_1), \cdots, (x_m, y_m) \sim D$, a distribution we do not have direct access to. Workaround and question: how large should m be so that the best hypothesis for the samples is close to h^*?

Simple. Find large-error hypotheses and exclude them from consideration. By doing so, the final winner h will surely satisfy $\text{Error}(h) \leqslant \epsilon$. [?]

We now define VC-dimension and stuff. A **range space** $(X, \mathcal{R})$ where X is a set of points and $\mathcal{R} = 2^X$, the collection of subsets of X. Given $S \subset X$, let $R_S = \{R \cap S \mid R \in \mathcal{R}\}$. A set S is said to be **shattered** by $\mathcal{R}$ if $|R_S| = 2^{|S|}$, i.e., we recover all subsets of S. The **VC-dimension** is the cardinality of the largest set that can be shattered by $\mathcal{R}$.

The simplest example is the range space $(\mathbb{R}, I)$ where the ranges are the set of intervals. A two-element set $\{a, b\}$ can be shattered by for example intervals $[a-2, a-1]$ (yielding $\varnothing$), $[a-\epsilon, a+\epsilon]$ (yielding $\{a\}$), $[b-\epsilon, b+\epsilon]$ (yielding $\{b\}$), and $[a-\epsilon, b+\epsilon]$ (yielding $\{a, b\}$). However, a set of size 3 $\{a, b, c\}$ cannot be shattered: no element can cover $\{a, c\}$ without covering $\{b\}$. Therefore, $\text{VCdim}(\mathbb{R}) = \text{VCdim}(\mathbb{R}, I) = 2$.

What about $(\mathbb{R}^2, \mathcal{R})$ where $\mathcal{R}$ is the set of all half spaces of $\mathbb{R}^2$? It's easy to see that given three points, we can always find subspaces to split them arbitrarily. However, given 4 points, for example the four corners of a

square, no halfspace can include precisely the two diagonal elements while leaving the two other out. This implies $\mathrm{VCdim}(\mathbb{R}^2) = 3$, and in general, $\mathrm{VCdim}(\mathbb{R}^d) = d + 1$.

Intuition: small VC dimension implies that the concept class has lower complexity, so we need fewer samples to obtain the desired theoretical bounds. In terms of bias-variance tradeoff, halfspaces are simple (low variance) but may fail to fit everything (bias). Now we prove this claim.

Theorem: The ϵ-Net Theorem

Given $(X, \mathcal{R})$ and distribution D, a set $N \subset X$ is an **ϵ-net** if

$$|R \cap N| \neq \varnothing \qquad \text{for all } R \in \mathcal{R} \text{ with } \mathbb{P}_D(R) \geqslant \epsilon.$$

In other words, N intersects with all elements of $\mathcal{R}$ that is not too small. The ϵ-net theorem states the following:

Given ϵ, δ, and $\mathrm{VCdim}(X, \mathcal{R}) = d$, a random set of size $\geqslant m = o(d\epsilon^{-1}\log(d/\epsilon) + \epsilon^{-1}\log(1/\delta))$ is an ϵ-net with probability $\geqslant 1 - \delta$. Importantly, this bound is independent of the size of X.

Proof sketch. STEP 1: Sauer's lemma. For finite X, if $|X| = n$ and $\mathcal{R}$ has VC-dimension d, then $|R| \leqslant n^d$.

STEP 2. Suppose $|X| = n$ and D is uniform. In this case, the ranges that we want to hit are those with $|R| \geqslant \epsilon n$. Choose N, a size m random sample of X. Then by definition and union bound,

$$\mathbb{P}(R \cap N = \varnothing) = (1-\epsilon)^m \Rightarrow \mathbb{P}(\text{some } R \text{ satisfies } R \cap N = \varnothing) \leqslant n^d(1-\epsilon)^m.$$

If we make the RHS < 1 then the converse has nonzero probability (we want this to be $1 - \delta$ by the theorem), i.e., some N will be an ϵ-net. So we really just need $n^d(1-\epsilon)^m < \delta$. Algebra shows $m \leqslant d\epsilon^{-1}\log(n/\delta)$.

STEP 3: generalizing to infinite X.

Beginning of 04/02/2025

We use a technique called **double sampling** — we choose two sets M, T, both of size m from D. We want to bound the probability that M is not an ϵ-net, i.e., the probability $\mathbb{P}(E_1)$ where $E_1 = \{\text{there exists } R \mid \mathbb{P}(R) \geqslant \epsilon, R \cap M = \varnothing\}$.

In addition, we define $E_2 = \{\text{there exists } R \mid \mathbb{P}(R) \geqslant \epsilon, R \cap M = \varnothing, \text{ and } |R \cap T| \geqslant \epsilon m/2\}$. Clearly, $E_2 \subset E_1$ so $\mathbb{P}(E_2) \leqslant \mathbb{P}(E_1)$. We now show that $\mathbb{P}(E_1) \leqslant 2\mathbb{P}(E_2)$, i.e., E_2 is not too unlikely to happen.

We use the definition of conditional probabilities:

$$\frac{\mathbb{P}(E_2)}{\mathbb{P}(E_1)} = \frac{\mathbb{P}(E_2 \cap E_1)}{\mathbb{P}(E_1)} = \mathbb{P}(E_2 \mid E_1).$$

If E_1 has already happened, then there exists some heavy range R' with $R' \cap M = \varnothing$. Looking at this range R',

$$\mathbb{P}(E_2 \mid E_1) \geqslant \mathbb{P}(|R' \cap T| \geqslant \epsilon m/2).$$

Since $\mathbb{P}(R') \geqslant \epsilon$, the expected size of intersection is $\mathbb{E}[|R' \cap T|] = \epsilon m$. By Chernoff bound,

$$\mathbb{P}(|R' \cap T| < \epsilon m/2) \leqslant \exp\left(-\epsilon m/8\right).$$

By choosing a suitably large m this completes the claim that we can make $\mathbb{P}(E_2) \leqslant \mathbb{P}(E_1) \leqslant 2\mathbb{P}(E_2)$.

From now on we effectively only need to upper bound the probability of E_2. Clearly, an upper bound can be obtained by relaxing the first constraint (heavy set):

$$E_2 \subset E_2' = \{\text{there exists } R \mid R \cap M = \varnothing, |R \cap T| \geqslant \epsilon m/2 = k\}.$$

Claim. $\mathbb{P}(E_2') \leqslant (2m)^d \exp(-\epsilon m/2)$. Once this is proven, we immediately get $\mathbb{P}(E_1) \leqslant 2(2m)^d \exp(-\epsilon m/2)$, and the original claim on m follows by bounding this quantity by δ.

Proof. Let $S = M \cup T$, and assume $|S| = 2m$ for we are sampling from an infinite space X.
Fix one such S. The projection $\{R \cap S, R \in \mathcal{R}\}$ of $\mathcal{R}$ onto S has $\leqslant (2m)^d$ ranges. Let E_R be the event that $R \cap M = \varnothing$ and $|R \cap S| \geqslant \epsilon m/2 = k$. Clearly,

$$\mathbb{P}(E_R) \leqslant \frac{\mathbb{P}(R \cap M = \varnothing, |R \cap S| \geqslant k)}{\mathbb{P}(|R \cap S| \geqslant k)} = \mathbb{P}(R \cap M = \varnothing \mid |R \cap S| \geqslant k).$$

as the denominator $\leqslant 1$. Essentially, given S of size $2m$, we want to partition it into M, T, and ask what is the probability that all elements in $|R \cap S|$ are not chosen by M. Algebra shows the above is

$$\leqslant \binom{2m-k}{m}\binom{2m}{m}^{-1} \leqslant 2^{-k} = 2^{-\epsilon m/2}$$

Finally, observe that (for a fixed S), $\mathbb{P}(E_2') = \mathbb{P}(\bigcup_{R \in \mathcal{R}} E_R) \leqslant \sum_R \mathbb{P}(E_R) \leqslant (2m)^d \cdot 2^{-\epsilon m/2}$. □

Going back to our original objective, where $\mathcal{C}$ is the concept class, $h \in \mathcal{C}$ a hypothesis, and $\mathrm{Error}(h) = \mathbb{P}_{(x,y)\sim D}(h(x) \neq y)$ the error rate. We assumed that there exists a perfect hypothesis h^* with $\mathrm{Error}(h^*) = 0$. **How many samples m do we need to ensure that $\mathbf{Error}(\hat{h}) \leqslant \epsilon$?**

Given h, let $\Delta(h) = \{x \in X : h(x) \neq h^*(x)\}$, i.e., the collection of all points from the ambient space where h makes an error. Let $\Delta = \{\Delta(h, h^*), h \in \mathcal{C}\}$. Suppose Δ has VC-dim d, applying the theorem to Δ says by drawing sufficiently large $m = \mathcal{O}(d\epsilon^{-1}\log(d/\epsilon) + \epsilon^{-1}\log(1/\delta))$ samples from D, then for all h with $\mathbb{P}(\Delta(h) \geqslant \epsilon)$, we can find a misclassified sample in $\Delta(h)$.
We were only interested in outputting a hypothesis $\hat{h}$ that perfectly predicts all samples, i.e., $\hat{h}(x) = h^*(x)$ for all $x \in \{x_1, \cdots, x_m\}$. Because our samples would hit anything heavy, i.e., exclude any hypotheses with $\mathbb{P}(h(x) \neq h^*(x)) \geqslant \epsilon$, our final winner $\hat{h}$ must have error rate ϵ.
Just one caveat: we need to show that $\mathrm{VCdim}(\Delta) = \mathrm{VCdim}(\mathcal{C})$. To see this, for a set S, $h_1 \cap S \neq h_2 \cap S$ if and only if $\Delta(h_1) \cap S \neq \Delta(h_2) \cap S$. Better seen by a picture. Thus, whatever shatters the concept class also shatters Δ, and vice versa.