1 Today’s topics
1. Finish No Free Lunch
2. $\epsilon$-nets and Inductive Generalization
3. The “Growth Function”
4. Sauer’s Lemma

2 No Free Lunch Theorem

2.1 WARM–UP PROBLEM
Suppose we have a highly biased coin that only comes up “heads” with probability $\epsilon$. Asymptotically (as a function of $k$ and $\epsilon$), how many times must we toss the coin before it comes up heads $k$ times with constant probability $p \in (0, 1)$?
If $n$ numbers toss...
\[
\sum_{i=k}^{n} \binom{n}{i} (1-\epsilon)^{n-i} \epsilon^i \geq p
\]

2.2 Definitions from last time
The set of dichotomies of a set $S \subseteq X$, realized by the class $C$, $\Pi_C(S) = \{C \cap S : C \in C\}$ (Using $x \in C \mapsto C(x) = 1$). If every subset $S' \subseteq S$ is in $\Pi_C(S)$, we say $S$ is “shattered” by $C$.
The VC-dimension of $C$ is the size of the largest set shattered by $C$.

Note: Chernoff bound
\[
\sum_{i=1}^{n} X_i \ (X_i=1 \leftrightarrow \text{heads})
\]
\[
E[\sum X_i] = \epsilon n
\]
\[
\Pr[\sum X_i \geq (1+\gamma)\epsilon n] \leq e^{-\frac{\gamma^2 \epsilon n}{4}} < e^{-\theta(k)}. \ (\text{if } (1+\gamma)\epsilon n \text{ is } k, \epsilon n \text{ is } \theta(k), \gamma \text{ is a constant})
\]
Consider $\gamma = 2$
\[
\Pr[\sum X_i < (1-\gamma)\epsilon n] \leq e^{-\frac{\gamma^2 \epsilon n}{4}} < e^{-\theta(k)}.
\]

$\epsilon n = k$
$n = \theta(\frac{k}{\epsilon})$

2.3 No Free Lunch Theorem
Any PAC-learning algorithm for a class $C$ of VC-dimension $d$ that produces $h$. $\Pr_{x \in D}[h(x)=c(x)] \geq 1-\epsilon$ for the target $c$ w.p. $1/7 + \delta$ must use $\Omega(\frac{d}{\epsilon^2})$ examples from $D$.

Note: Summery of last time
On a set $\{X^{(1)}, \ldots, X^{(d)}\}=S$ shattered by $C$. If I choose $c \in C$ s.t. a uniform random labeling of $S$ is realized, then if my alg. only sees $\frac{d}{2}$ of the ex’s, then it incorrectly labels $\geq \frac{d}{2}$ with probability $\geq 1/7$ (for some $c \in C$)

To obtain the final bound.
We choose $D$ to force the alg. to take many ex’s just to see a new example. Simultaneously, we need to ensure that the unseen examples can cause error $\geq \epsilon$. 

1
Let’s choose \( D \) to give \( X^{(1)} \) prob. \( 1 - \delta \).
Let’s give the other \( d - 1 \) examples in \( S \) probability \( \frac{8\epsilon}{d-1} \) each.
If the algorithm misses \( \geq \frac{d-1}{d} \) of these, then as before, if we choose \( c \in C \) to induce a uniform, random labeling on \( S \), then it must be wrong on \( \geq \frac{d-1}{2} \) of these unseen points in expectation. By the same calculation as before, the probability that it is wrong on \( \leq \frac{1}{d} \) of these overall satisfies \( p \geq \frac{d-1}{8} + (1 - p)(d-1) \geq \frac{d-1}{4} \).
So \( p \leq \frac{1}{2} \).
So therefore, again, w.p. \( > 1/7 \), it is wrong on \( \geq \frac{d-1}{4} \), then its error is \( \geq \frac{d-1}{8} \).
We now show that \( \Omega(\frac{d}{2}) \) examples from \( D \) are necessary to obtain \( \geq \frac{d-1}{2} \) out of \( \{X^{(2)}, \ldots, X^{(d)}\} \).

Define \( R_i = \begin{cases} 1 & \text{if } i^{th} \text{ example drawn from } D \text{ is among } \{X^{(2)}, \ldots, X^{(d)}\} \\ 0 & \text{if } i^{th} \text{ example is } X^{(i)} \end{cases} \)

If my alg. uses \( m \) examples, and it is obtaining \( \geq \frac{d-1}{2} \) rare examples w.p. \( \delta > 0(\delta \in (0,1)) \).

We then have that \( m = \Omega(\frac{d}{\delta}) \).

\[
\Pr[R_i = 1] = \sum_{k=2}^d \Pr_D[X^{(k)}] \leq (d-1) \frac{8\epsilon}{d-1} = 8\epsilon \text{ by the earlier calculation.}
\]

3 \( \epsilon \)-nets and Inductive Generalization

“\( \epsilon \)-net”: Intuitively, what we need to know for generalization is that if some \( h \in H \) is wrong about the target concept \( c \) on any set \( r \) that has probability \( \geq \epsilon \) under the unknown \( D \), then we have some example \((x, c(x))\) in our training set that catches this error (Then, as before, solving Consistency solves the learning problem).

**Definition** For any target concept \( c \), the class of error regions of \( H \) (w.r.t. \( c \)), denote \( \Delta_H(c) \), is \( \Delta_H(c) = \{[h \neq c](x) : h \in H\} \).
The set of error regions of probability \( \geq \epsilon \) under \( D \) is \( \Delta_{H,\epsilon}(c) = \{r \in \Delta_H(c) : \Pr_{x \in D}[x \in r] \geq \epsilon\} \).

**Definition** For any \( \epsilon > 0 \), we say that a set \( S \subseteq X \) is an \( \epsilon \)-net for \( \Delta_H(c) \) w.r.t. \( D \) if for every \( r \in \Delta_{H,\epsilon}(c) \) we have some \( x \in r \cap S \).

Example: Suppose \( C(X) \equiv 0, H \) is the class of intervals of \( \mathbb{R} \), \( D \) is the uniform distribution on \([0,1]\).

Then \( \Delta_H(c) = \{[h \neq 0](x) : h \in H\} = \mathbb{R}. \) \( \{[h \neq 0](x) = h(x)\} \).

Since for each \( [a, b] \in \Delta_H(c), \Pr_{x \in D}[X \in [a, b]] = \min(1, b) - \max(0, a) \).

\( \Delta_{H,\epsilon}(c) = \{[a, b] : \min(1, b) - \max(0, a) \geq \epsilon\} \).

The set \( S = \{\lfloor \epsilon k \rfloor : k = 1, 2, ..., \lfloor \frac{1}{\epsilon} \rfloor\} \) is an \( \epsilon \)-net for \( \Delta_H(c) \).

Since every \( r \in \Delta_{H,\epsilon}(c) \) has length \( \geq \epsilon \) in \([0,1]\), it must hit this grid of \( \epsilon \)-spaced points.

Revisiting the standard calculation:
If we fix any \( r \in \Delta_{H,\epsilon}(c) \), the probability that we fail to hit \( r \) after \( m \) examples is \( \Pr_{x \in D}[x \notin r]^m = (1 - \Pr_{x \in D}[x \notin r])^m \leq (1 - \epsilon)^m \) since \( r \in \Delta_{H,\epsilon}(c) \).

So the probability that \( m \) examples don’t form an \( \epsilon \)-net for \( \Delta_H(c) \) is at most \( |\Delta_H(c)| (1 - \epsilon)^m \) by a union bound.

\( \Delta_H(c) \): Trouble is usually infinitely for numerical data.

Want: analysis in terms of how \( H \) behaves on finite sample.

4 The Growth Function

**Definition**: The growth function of a representation class \( H \), \( \Pi_H : N \to N \) is given by \( \Pi_H(m) = \max_{s \subseteq x, |s| = m} |\Pi_H(s)| \).

So if \( H \) shatters a set \( S \) of size \( m \), \( \Pi_H(m) = 2^m \).

(Note: there are at most \( 2^m \) possible labelings of any set of \( m \) points, so \( \max_{s \subseteq x, |s| = m} |\Pi_H(S)| \leq 2^m \).

Notice: if \( H \) shatters \( S \), \( H \) also shatters all subsets \( S' \subseteq S \).
So...

**proposition:** If the VC-dimension of $\mathcal{H}$ is $d$, for every $m \leq d$, $\Pi_{\mathcal{H}}(m) = 2^m$.

5 Sauer’s Lemma

If $\mathcal{H}$ has VC-dimension $d$, then for every $m$, $\Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} = \mathcal{O}(m^d)$ for $m > d$.

**Lemma 1** For any $m, d \in \mathbb{N}$, define $\Phi_d(m)$ inductively, by $\Phi_0(m) = 1$ and otherwise $\Phi_d(m) = \Phi_d(m - 1) + \Phi_{d-1}(m - 1)$. Then, if $\text{VCD}(\mathcal{H}) = d$, for every $m$, $\Pi_{\mathcal{H}}(m) \leq \Phi_d(m)$.

**Lemma 2** $\Phi_d(m) = \sum_{i=0}^{d} \binom{m}{i} = \mathcal{O}(m^d)$ for $m > d$.