ECE 0402 - Pattern Recognition

Lecture 8 on 2/9/2022

Review:

- Challenge: Number of hypothesis is $\mid \mathcal{H} \mid$ potentially infinite
- Better: Narrow the scope to the finite training set in order to replace easily infinite m. Dichotomies allow us that.
- $h : \{x_1, ..., x_n\} \mapsto \{-1, 1\} \implies 2^n$ different way of labeling, max! so dichotomy is the way of labeling THAT particular data set
- Hence, in general, $\mid \mathcal{H} \mid > \mid \mathcal{H}(x_1, ..., x_n) \mid$. In English number of hypothesis ¡ number of dichotomies
- So maybe a dichotomies are a better measure of "richness" of the set.
- And then we introduced the idea of "growth function that gets rid of the dependence of dichotomy to a particulars of our training set $x_1, ..x_n$.
- Growth function: $m_{\mathcal{H}}(n) = max_{x1,...,x_n \in \mathcal{X}} \mid \mathcal{H}(x_1, ..., x_n) \mid$.

  We went through some examples of Growth functions:

  - Positive rays: $m_{\mathcal{H}}(n) = n + 1$
  - Positive intervals: $m_{\mathcal{H}}(n) = \frac{1}{2}n^2 + \frac{1}{2}n + 1$
  - Convex sets: $m_{\mathcal{H}}(n) = 2^n$
  - Linear classifiers in $\mathbb{R}^2$:

$$\begin{aligned} m_{\mathcal{H}}(1) &= 2 \\ m_{\mathcal{H}}(2) &= 4 \\ m_{\mathcal{H}}(3) &= 8 \\ m_{\mathcal{H}}(4) &= 14 \\ m_{\mathcal{H}}(n) &= \quad ? \end{aligned}$$

So in the previous lecture we left with linear classifiers example, and we didn't actually calculate the growth function, we just worked out for first four values of n...

Recall

$$\mathbb{P}[\mid \hat{R}_n(h^*) - R(h^*) \mid > \epsilon] \leq 2me^{-2\epsilon^2 n}$$

Another way to write this, by setting $2me^{-2\epsilon^2 n} = \delta$

$$R(h^*) \leq \hat{R}_n(h^*) + \sqrt{\frac{1}{2n} \; log \; \frac{2m}{\delta}}$$

If m is "infinite," it's already exponential

If $m \propto e^n$, we have a problem...

No matter how big n gets $\sqrt{\frac{1}{2n}\, log\, \frac{2m}{\delta}}$ will never be smaller...

What if we replace with m with $m_{\mathcal{H}}(n)$? Suppose that for any $\delta \in (0,1)$, we can guarantee at least $1 - \delta$

$$R(h^*) \leq \hat{R}_n(h^*) + \sqrt{\frac{1}{2n}\, log\, \frac{2m_{\mathcal{H}}(n)}{\delta}}$$

- If $m_{\mathcal{H}}(n) = 2^n$ then $\sqrt{\frac{1}{2n}\, log\, \frac{2m_{\mathcal{H}}(n)}{\delta}}$ is constant   worst case - shattering
- If $m_{\mathcal{H}}(n)$ is a polynomial in n, $\sqrt{\frac{1}{2n}\, log\, \frac{2m_{\mathcal{H}}(n)}{\delta}}$ decays like $\sqrt{\frac{log\ n}{n}}$.

When is learning feasible?   Instead of just memorizing based on sheer quantity of hypotheses

Assuming that we are indeed allowed to substitute $m_{\mathcal{H}}(n)$ for m, we can argue that for a given set of hypothesis $\mathcal{H}$ learning is possible provided that $m_{\mathcal{H}}(n)$ is a polynomial.

Key idea: Break points

def'n: If no data set of size k can be shattered by $\mathcal{H}$, then k is a **break point** for $\mathcal{H}$.

$$m_{\mathcal{H}(k)} < 2^k$$

This also implies that if k is a break point, then so is any $k' > k$.

Examples of Break points

- Positive rays:$m_{\mathcal{H}}(n) = n + 1$
  - break point: $k = 2$
- Positive intervals: $m_{\mathcal{H}}(n) = \frac{1}{2}n^2 + \frac{1}{2}n + 1$
  - break point: $k = 3$
- Convex sets: $m_{\mathcal{H}}(n) = 2^n$
  - break point: $k = \infty$
- Linear classifiers in $\mathbb{R}^2$:
  - break point: $k = 4$

$\boxed{\text{If there exists any break point, then } m_{\mathcal{H}}(n) \text{ is polynomial in } n}$

$\boxed{\text{If no break points, then } m_{\mathcal{H}}(n) = 2^n}$

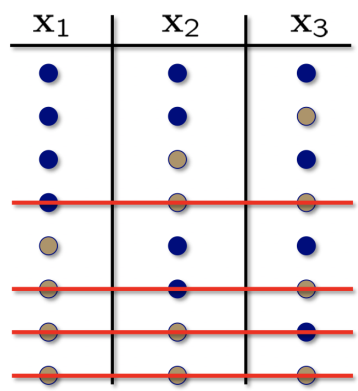
As soon as we have a single break point, this starts eliminating tons of dichotomies.

- We can show that $m_{\mathcal{H}}(n)$ is polynomial in n.
- To show that we worry too much and show that $m_{\mathcal{H}}(n) \leq$ **some** polynomial
- Main approach will center around:
  - $B(n,k) :=$ maximum number of dichotomies on n points such that no subset of size k can be shattered by these dichotomies
  - Notice that this is a purely combinatorial quantity
  - By definition, $m_{\mathcal{H}}(n) \leq B(n,k)$

Example: how many dichotomies?

You are given a hypothesis set which has a break point of 2.

How many dichotomies can you get on 3 data points?



Break point of 2 - can't achieve all combinations of 2 points

Summary: $B(n,k)$ is the combinatorial quantity that's an upper bound on the growth function for any possible set of classifiers.

You can bound $B(n,k)$ recursively which is an algorithmic proof. We will just skip the analytical proof as it is pages and pages math. There is also a "proof by picture" for this which I like, you may review that from "Learning from Data" if you are interested.

$$B(n,k) \leq B(n-1,k) + B(n-1,k-1)$$

Analytical solution: $B(n,k) \leq \sum_{i=0}^{k-1} \binom{n}{i}$ You can prove that it is actually equal,

$$B(n,k) = B(n-1,k) + B(n-1,k-1)$$

but all we really need is an upper bound, so that is all we will prove here.

Proof by induction:

$$B(n,k) \leq B(n-1,k) + B(n-1,k-1)$$

- Base case

  $$B(n,1) = 1$$
  $$B(1,k) = \begin{cases} 1 & \text{if} \quad k = 1 \\ 2 & \text{otherwise} \end{cases}$$

- Inductive step
  - suppose the inequality is true for $B(n-1,k)$ and $B(n-1,k-1)$

$$\begin{aligned} B(n,k) &\leq \sum_{i=0}^{k-1} \binom{n-1}{i} + \sum_{i=0}^{k-2} \binom{n-1}{i} \\ &= 1 + \sum_{i=0}^{k-1} \binom{n-1}{i} + \sum_{i=1}^{k-1} \binom{n-1}{i-1} \\ &= 1 + \sum_{i=1}^{k-1} (\binom{n-1}{i} + \binom{n-1}{i-1}) \\ &= 1 + \sum_{i=1}^{k-1} \binom{n}{i} = \sum_{i=0}^{k-1} \binom{n}{i} \end{aligned}$$