Homework 3

Sara Beery
CS 156A - Learning Systems

October 22, 2018

1 Generalization

The modified Hoeffding Inequality provides a way to characterize the generalization error with a probabilistic bound

$$P[|Ein(g) - Eout(g)| > \epsilon] \leq 2Me^{-2\epsilon^2 N}$$

for any $\epsilon > 0$. If we set $\epsilon = 0.05$ and want the probability bound $2Me^{-2\epsilon^2 N}$ to be at most 0.03, what is the least number of examples N (among the given choices) needed for the case $M = 1$?

To find the smallest N in this case, We first set the upper bound on the Hoeffding Inequality

$$2Me^{-2\epsilon^2 N} \leq 0.03$$

Then we solve for N, given $\epsilon = 0.05$

$$N \geq -\frac{ln(\frac{0.015}{M})}{0.005}$$

1. [**b**] 1000
Plugging in $M = 1$, we get $N \geq 839.94$
2. [**c**] 1500
Plugging in $M = 10$, we get $N \geq 1300.46$
3. [**d**] 2000
Plugging in $M = 100$, we get $N \geq 1760.96$

2 Break Point

As shown in class, the (smallest) break point for the Perceptron Model in the two-dimensional case $\mathbb{R}^2$ is 4 points. What is the smallest break point for the Perceptron Model in $\mathbb{R}^3$? (i.e.,

instead of the hypothesis set consisting of separating lines, it consists of separating planes.)
4. [**b**] 5 points
The break point in $\mathbb{R}^3$ will be 5 points, Fig.1 demonstrates an arrangement of 5 points in $\mathbb{R}^3$ that cannot be shattered by the three-dimensional perceptron.

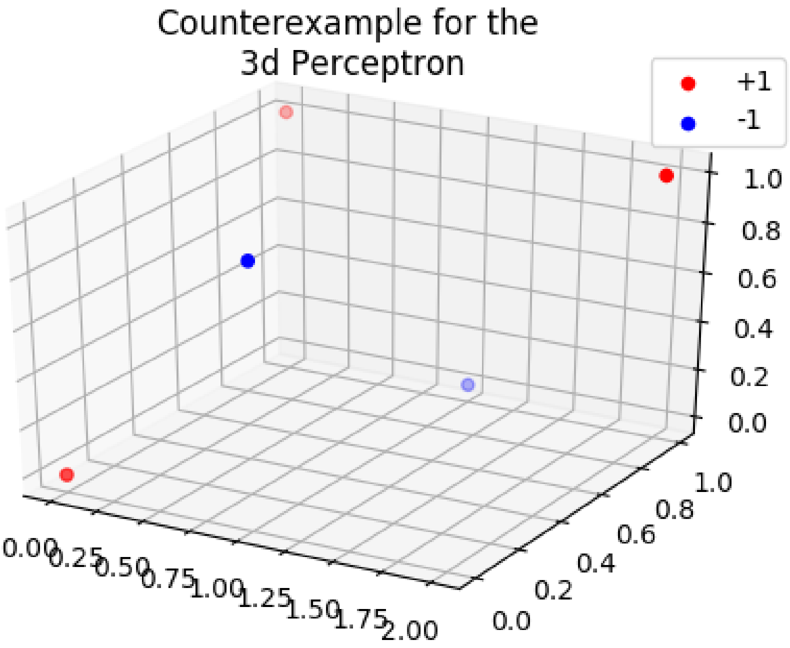

Figure 1: 5 points that break the 3d Perceptron

Rigorously, we can prove that the break point of a perceptron on $\mathbb{R}^d$ will be d+2 (proof originally from homework in CS155).

We will do this in two parts. First, we will show that there exists $d+1$ points that a d-dimensional perceptron can shatter.

Proof. we can write the perceptron as

$$y = f(x) = \begin{cases} 1 & if\ w^T x + b > 0 \\ -1 & otherwise. \end{cases}$$

Let x be the origin and the d orthogonal unit vectors in $\mathbb{R}^d$, a total of $d+1$ data points. Arbitrarily label these $d+1$ points $y = (y_0, ..., y_d)^T \in \{-1, 1\}^{d+1}$. Let $w = (w_1, ..., w_d)$ where $w_i = y_i \forall i \in \{1, ..., d\}$, and let $b = 0.5 \bullet y_0$. By construction, $f(x)$ will label all points correctly for any arbitrary labeling of the d+1 points. Thus the perceptron can shatter these points. □

Next, we will show that d+2 points cannot be shattered.

Proof. For ease of notation, let us rewrite the perceptron as

$$y = f(\tilde{x}) = \begin{cases} 1 & if\ \tilde{w}^T\tilde{x} > 0 \\ -1 & otherwise, \end{cases}$$

where $\tilde{x}^T = (x^T, 1)$, $\tilde{w}^T = (w^T, b)$ (bringing the bias term into the inner product).

Now, let us assume for the sake of contradiction that there exists d+2 points that the d-dimensional perceptron can shatter. Let us denote these points $x^{(1)}, ..., x^{(}d+2) \in \mathbb{R}^d$, corresponding to $\tilde{x}^{(1)}, ..., \tilde{x}^{(d+2)} \in \mathbb{R}^{d+1}$.

Since we have $d+2$ points in $\mathbb{R}^d$, there must be at least one point $\tilde{x}^{(i)}$ that can be written as a nontrivial linear combination of the other $d+1$ points. So there exists i such that

$$\tilde{x}^{(i)} = \sum_{j \neq i} \alpha_j \tilde{x}^{(j)}$$

where α is nonzero. Let $S = \{j | j \neq i, \alpha_j \neq 0\}$.

Now, suppose $y^{(j)} = sign(\alpha_j)\ \forall j \in S$, and suppose $y^{(i)} = -1$. By our assumption, there exists $\tilde{w}$ such that all data is correctly labeled. This implies both

$$\alpha_j \tilde{w}^T \tilde{x}^{(j)} > 0\ \forall j \in S$$

and

$$\tilde{w}^T \tilde{x}^{(i)} \leq 0.$$

But, recall that

$$\begin{aligned} \tilde{w}^T \tilde{x}^{(i)} &= \tilde{w}^T \sum_{j \neq i} \alpha_j \tilde{x}^{(j)} \\ &= \tilde{w}^T \sum_{j \in S} \alpha_j \tilde{x}^{(j)} \\ &> 0, \end{aligned}$$

a contradiction. Thus $d+2$ is the break point of the d-dimensional perceptron. □

3 Growth Function

Which of the following are possible formulas for a growth function $m_H(N)$:

i. $N + 1$

ii. $1 + N + \binom{N}{2}$

iii. $\sum_{i=1}^{\lfloor \sqrt{N} \rfloor} \binom{N}{i}$

iv. $2^{\lfloor N/2 \rfloor}$

v. 2^N

5. [e] i, ii, iii, iv, v

Note: in order for $m_H(N)$ to be a growth function, it must satisfy $m_H(N) <= 2^N$

Any polynomial is less than an exponential, so trivially we see that **i.** $1 + N$ and **ii.** $1 + N + \binom{N}{2} = 1 + \frac{N}{2} + \frac{N^2}{2}$ are both valid growth functions.

Also trivially, we just need $m_H(N) \leq 2^N$, so **iv.** and **v.** are also valid growth functions $(\sqrt{N} < N, N = N)$.

iii. is slightly more challenging to show. For small N it is simple to compute and compare directly.

For $N = 1$:

$$\sum_{i=1}^{1} \binom{1}{i} = 1 \leq 2$$

For $N = 2$:

$$\sum_{i=1}^{1} \binom{2}{i} = 2 \leq 4$$

For $N = 3$:

$$\sum_{i=1}^{1} \binom{3}{i} = 3 \leq 8$$

For $N = 4$:

$$\sum_{i=1}^{2} \binom{4}{i} = 4 + 6 = 10 \leq 16$$

We discussed in class that

$$\sum_{i=1}^{k-1} \binom{N}{i}$$

is polynomial with dominant term N^k but in **iii.** the sum depends on N, not a constant k, so the dominant term as N gets large would be $N^{\sqrt{N}}$ which is non-polynomial.

Essentially, we need to show that

$$N^{\sqrt{N}} \leq 2^N$$

for large N. To examine the behavior as N gets large, note the above is equivalent to

$$N \leq 2^{\sqrt{N}}$$

Which is true for all N. Therefore, all options are valid growth functions.

4 Fun With Intervals

Consider the 2-intervals learning model, where $h : \mathbb{R}1, +1$ and $h(x) = +1$ if the point is within either of two arbitrarily chosen intervals and 1 otherwise. What is the (smallest) break point for this hypothesis set?

6. [**c**] 5

Consider any 5 points on the line, with labels $+1, -1, +1, -1, +1$. No two intervals can correctly label these points, so two intervals cannot shatter 5 points. We can easily see that any 4 points can be shattered by two intervals, as the most difficult case, alternating labels where you get no clustering, can be correctly labeled by putting the intervals around the two positive points. Any other labeling case would only require extending one of the intervals to cover the adjacent points. Therefore the break point is 5.

Which of the following is the growth function $m_H(N)$ for the 2-intervals hypothesis set?

7. [**c**] $\binom{N+1}{4} + \binom{N+1}{2} + 1$

Two intervals can always be defined by 4 points, the end points of the two intervals. When we are considering dichotomies, we need not consider the infinite real-valued possibilities along the real line, but instead only a discrete set of points between the N datapoints, of which there will be N+1. We select our four endpoints out of this set. Using these four points, which we denote $\{p_1, p_2, p_3, p_4\}$ where without loss of generality $p_1 < p_2 < p_3 < p_4$, we can break down all possible dichotomies to find the growth function. First, consider the case where the 4 endpoints are distinct and we set the intervals as between adjacent endpoints in order (e.g. $[p_1, p_2]$ and $[p_3, p_4]$). This provides $\binom{N+1}{4}$ unique dichotomies, as there are $\binom{N+1}{4}$ sets of unique endpoints.

Now, what if we either (a) have chosen the intervals over those 4 unique points in any other way (e.g. $[p_1, p_3]$ and $[p_2, p_4]$, or $[p_1, p_4]$ and $[p_2, p_3]$), or (b) the 4 endpoints are not unique? In both cases, we can show that this will reduce the set of dichotomies to the set of possible dichotomies given a single interval, which we have shown in class to be $\binom{N+1}{2} + 1$. In case (a), note that by overlapping the intervals in any way we see that the coverage of the line, and therefore the dichotomies possible, is identical to a single interval. In case (b), since we only care about the coverage of the N datapoints, if two or more of the endpoints are non-unique we will also have coverage identical to a single interval. Thus, the total number of dichotomies, and thus the growth function, is $\binom{N+1}{4} + \binom{N+1}{2} + 1$.

Now, consider the general case: the M-intervals learning model. Again $h : \mathbb{R}1, +1$, where $h(x) = +1$ if the point falls inside any of M arbitrarily chosen intervals, otherwise $h(x) = 1$. What is the (smallest) break point of this hypothesis set?

8. [**d**] $2M + 1$

Similar to the argument for the ”2-intervals” hypothesis set, we can shatter any set of $2M$ datapoints. The most challenging, alternating-label case can be labeled correctly by placing an interval around each positive label. Any other case can be labeled correctly by extending one or more of the intervals to cover the adjacent positive points. If we have $2M + 1$ points with alternating labels, starting and ending with a poisitive-labeled point, then it will be impossible for us to label all the points correctly. Thus, the smallest break point is $2M + 1$

5 Convex Sets: The Triangle

Consider the triangle learning model, where $h : \mathbb{R}^2 1, +1$ and $h(x) = +1$ if x lies within an arbitrarily chosen triangle in the plane and 1 otherwise. Which is the largest number of points in $\mathbb{R}^2$ (among the given choices) that can be shattered by this hypothesis set?

9. [**b**] 3

For any set of three points on the plane, with any nontrivial alignment (e.g. not considering the case of overlapping points or 3 points in a line, as we take the maximum set of dichotomies over all orientations of points when defining the growth function) a triangle can label those points in $2^3 = 8$ distinct ways. We can draw a triangle around none of the points, around any one of the points, around any set of two of the points, or around all three points. So we can shatter 3 points.

In Fig. 2 we show a labeling of non-trivially aligned points that the triangle cannot label correctly. Thus the triangle cannot shatter 4 points, so 3 is the largest number of points that can be shattered.

6 Non-Convex Sets: Concentric Circles

Compute the growth function $m_H(N)$ for the learning model made up of two concentric circles around the origin in $\mathbb{R}^2$. Specifically, H contains the functions which are +1 for

$$a^2 \leq x_1^2 + x_2^2 \leq b^2$$

and 1 otherwise, where a and b are the model parameters.

10. [**b**] $m_H(N) = \binom{N+1}{2)+1}$

Note that though this problem is defined in two dimensions, the only thing that matters

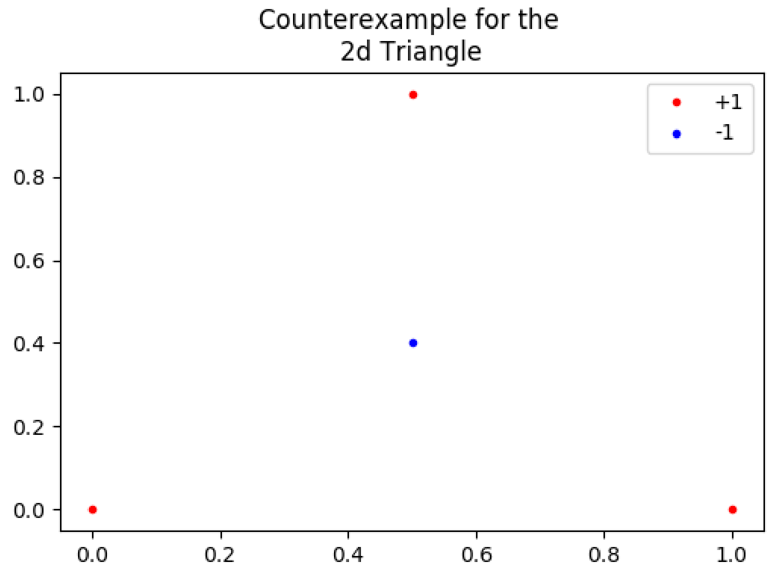

Figure 2: 4 points that break the 2d Triangle

in terms of defining the possible dichotomies is the distance of each datapoint and endpoint from the origin. Recall that for any point in $\mathbb{R}^2$, this distance is defined as the radius r, where $x_1^2 + x_2^2 = r^2$. Note that then

$$a^2 \leq x_1^2 + x_2^2 \leq b^2$$

becomes

$$a^2 \leq r^2 \leq b^2,$$

which is identical to

$$a \leq r \leq b$$

(where without loss of generality due to symmetry, we are considering only the positive root).

Now we can easily see that this $2d$ hypothesis set can be mapped to the $1d$ single-interval hypothesis set along the positive real line, which has growth function $m_H(N) = \binom{N+1}{2)+1}$.