Benjamin Ye
CS/CNS/EE 156a: Learning Systems (Fall 2023)
October 16, 2023

Homework 3

Problem	Answer
1	[b]
2	[c]
3	[d]
4	[b]
5	[b]
6	[c]
7	[c]
8	[d]
9	[d]
10	[b]

Generalization Error

1. The modified Hoeffding's inequality provides a way to characterize the generalization error with a probabilistic bound

$$\mathbb{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] \leq 2Me^{-2\epsilon^2 N}$$

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 $M = 1$?

Answer: [b] 1000

With $M = 1$ and a bound of 0.03, the inequality is

$$\mathbb{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| > 0.05] \leq 2(1)e^{-2(0.05)^2 N} \leq 0.03$$

Solving for N,

$$N \geq -\frac{\ln\big(0.03/\big(2(1)\big)\big)}{2(0.05)^2} \approx 840$$

2. Repeat for the case $M = 10$.

Answer: [c] 1500

With $M = 10$ and a bound of 0.03, the inequality is

$$\mathbb{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| > 0.05] \leq 2(10)e^{-2(0.05)^2 N} \leq 0.03$$

Solving for N,

$$N \geq -\frac{\ln\big(0.03/\big(2(10)\big)\big)}{2(0.05)^2} \approx 1{,}300$$

3. Repeat for the case $M = 100$.

Answer: [d] 2000

With $M = 100$ and a bound of 0.03, the inequality is

$$\mathbb{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| > 0.05] \leq 2(100)e^{-2(0.05)^2 N} \leq 0.03$$

Solving for N,

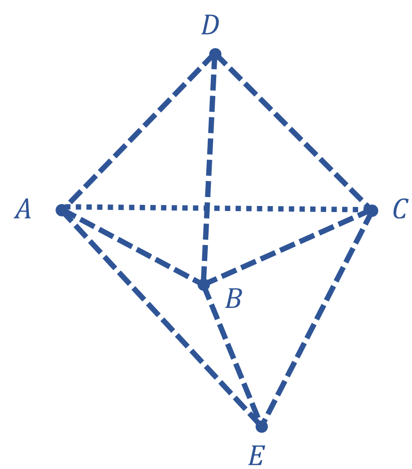
$$N \geq -\frac{\ln\big(0.03/\big(2(100)\big)\big)}{2(0.05)^2} \approx 1{,}761$$

Break Point

4. 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.)

   Answer: [b] 5

   Let us start with $k = 4$, which is the break point for the two-dimensional (2D) case, by considering only points $\mathbf{x}_n = \{A, B, C, D\}$ in the following schematic. Their arrangement in both 2D and three-dimensional (3D) space maximizes the number of possible dichotomies since they are not colinear or coplanar, respectively.

   For a planar arrangement of $\mathbf{x}_n$, where $A = C = +1$ and $B = D = -1$, it is impossible to designate a line using a 2D perceptron that can correctly classify all four points. However, by raising D above the plane containing A, B, and C, a 3D perceptron can now generate a plane in the space between $\{A, C\}$ and $\{B, D\}$ that successfully estimates the target function and shatters $\mathbf{x}_n$. Therefore, the added dimensionality enables a 3D perceptron to generate dichotomies for all possible arrangements of four points, something that is not possible with a 2D perceptron.

   However, even the 3D perceptron can fail to generate a possible dichotomy when a fifth point E is added. When the points are arranged in a triangular bipyramidal fashion with $A = B = C = +1$ and $D = E = -1$, there is no plane that can correctly classify all five points. As such, the break point for a 3D perceptron is $k = 5$.

5. Which of the following are possible formulas for a growth function $m_{\mathcal{H}}(N)$:

i) $1 + N$ 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

where $\lfloor u \rfloor$ is the biggest integer less than or equal to u, and $\binom{M}{m} = 0$ when $m > M$?

Answer: [b] i, ii, v

The growth function $m_{\mathcal{H}}(N)$ can be bounded generally by 2^N, or if it has a break point, by a polynomial in N.

(i) satisfies the criteria since $1 + N$ is less than 2^N, is a polynomial, and is the growth function for positive rays.

(ii) satisfies the criteria since

$$1 + N + \binom{N}{2} = 1 + N + \frac{1}{2}(N-1)N = 1 + \frac{1}{2}N + \frac{1}{2}N^2$$

is less than 2^N, is a polynomial, and is the growth function for positive intervals.

(iii) does not satisfy the criteria since it, despite being bounded by 2^N, is not a polynomial in N. By invoking the binomial theorem, it can be shown that

$$\sum_{j=1}^{\lfloor \sqrt{N} \rfloor} \binom{N}{j} = \sum_{j=0}^{\lfloor \sqrt{N} \rfloor} \binom{N}{j} - \binom{N}{0} = \sum_{j=0}^{\lfloor \sqrt{N} \rfloor} \binom{N}{j} - 1 \geq \sum_{j=0}^{\lfloor \sqrt{N} \rfloor} \binom{\lfloor \sqrt{N} \rfloor}{j} - 1 = 2^{\lfloor \sqrt{N} \rfloor} - 1$$

due to the greater number of elements ($N > \lfloor \sqrt{N} \rfloor$) in the combination. Therefore, as N increases, (iii) grows faster than $2^{\lfloor \sqrt{N} \rfloor} - 1$, which is already greater than a polynomial in N.

(iv) does not satisfy the criteria since $2^{\lfloor N/2 \rfloor}$, despite being bounded by 2^N, is not a polynomial in N.

(v) satisfies the criteria since it is equal to 2^N and is the upper bound of the growth function for convex sets.

Fun with Intervals

6. Consider the "2-intervals" learning model, where $h: \mathbb{R} \to \{-1, +1\}$ 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?

   Answer: [c] 5

   With only two intervals to classify the $+1$ points, this hypothesis set can fail to correctly classify a data set if the sign flips four or more times consecutively, in which there can be three $+1$ intervals, e.g., $\{+1, -1, +1, -1, +1\}$. This arrangement requires at least five points, so the break point is $k = 5$.

7. Which of the following is the growth function $m_{\mathcal{H}}(N)$ for the "2-intervals" hypothesis set?

   Answer: [c] $\binom{N+1}{4} + \binom{N+1}{2} + 1$

   For a data set with N points, there are $N + 1$ spots outside and between the points to start or end an interval.

   - When there are two distinct intervals, there are C_4^{N+1} possible ways to select the four interval bounds.
   - When the two intervals overlap or share a boundary such that they form only one interval, there are C_2^{N+1} possible ways to choose the two interval bounds.
   - When there are no points with a $+1$ classification, there are no intervals. There is only $C_1^{N+1} = 1$ possible way to have this arrangement.

   By summing the three contributions above, the growth function is given by

   $$m_{\mathcal{H}}(N) = \binom{N+1}{4} + \binom{N+1}{2} + 1$$

8. Now, consider the general case: the "M-intervals" learning model. Again $h: \mathbb{R} \to \{-1, +1\}$, 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?

   Answer: [d] $2M + 1$

   By generalizing the pattern in problem 6, there must be at least $M + 1$ nonconsecutive $+1$ points to fail to correctly classify an arrangement of points with M intervals. As such, the smallest set of points that satisfy the criteria has only one $+1$ point in each of the $M + 1$ intervals and one -1 point in the M spaces between the positive intervals (e.g., $\{+1, -1, +1, -1, \dots, +1, -1, +1\}$) for a total of $2M + 1$ points.

Convex Sets: The Triangle

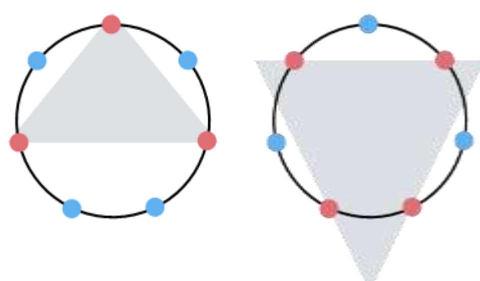
9. Consider the "triangle" learning model, where $h: \mathbb{R}^2 \rightarrow \{-1, 1\}$ and $h(\mathbf{x}) = +1$ if $\mathbf{x}$ lies within an arbitrarily chosen triangle in the plane and -1 otherwise. Which is the largest number of points in $\mathbb{R}^2$ (among the given choices) that can be shattered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

   Answer: [d] 7

   As with a convex set, the points are arranged on the circumference of an arbitrary circle to maximize the number of dichotomies. The triangle can either have its vertices be situated on or be extended past the circumference to enclose $+1$ points. The former is used when there is only one $+1$ point to encompass, while the latter is used when there are consecutive $+1$ points (or a $+1$ interval) to surround.

   From the previous observations, it is obvious that if the points are arranged such that there are three or fewer $+1$ intervals, the triangle hypothesis set will be able to classify all points correctly. With the periodic "boundary conditions" for arranging points on a circle, the largest data set that is guaranteed to have three or fewer $+1$ intervals has seven points.

   (Due to the periodic boundary conditions, it is impossible to have four or more $+1$ intervals with seven points since the first and last points are adjacent to each other on the circle. The only two configurations with three $+1$ intervals are shown below, where a red point is $+1$ and a blue point is -1.)



   With eight points, the break point, it is easy to see that if a $+1$ point is inserted between the two -1 points in the example on the left above, no triangle will be able to enclose all four $+1$ points.

Non-Convex Sets: Concentric Circles

10. Compute the growth function $m_{\mathcal{H}}(N)$ for the learning model made up of two concentric circles in $\mathbb{R}^2$. Specifically, $\mathcal{H}$ contains the functions which are $+1$ for $a^2 \leq x_1^2 + x_2^2 \leq b^2$ and -1 otherwise. The growth function is

    Answer: [b] $\boldsymbol{\binom{N+1}{2} + 1}$

    By applying a nonlinear transformation to the data points, such as $r^2 = x_1^2 + x_2^2$ so that only the radial distance from the origin is relevant, the region between the two concentric circles is reduced to the interval $a \leq r \leq b$ on a line. This is equivalent to the 1-interval hypothesis set and therefore shares its growth function, $m_{\mathcal{H}}(N) = C_2^{N+1} + 1$ (C_2^{N+1} ways to place the two bounds of the $+1$ interval in the $N + 1$ possible spots outside and between the N radii, and 1 way to have no $+1$ interval).