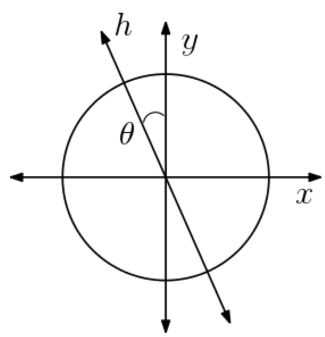
Homework 1

Christopher Mertin
CS6966: Theory of Machine Learning

February 15, 2017

1. Consider the problem of classifying points in the two-dimensional plane, *i.e.*, $\chi = \mathbb{R}^2$. Suppose that the (unknown) true label of a point (x, y) is given by sign(x) (we define sign$(0) = \pm 1$ for convenience). Suppose the input distribution $\mathcal{D}$ is the uniform distribution over the unit circle centered at the origin.

   (a) Consider the hypothesis h as shown in the figure below (h classifies all the points on the right of the line as $+1$ and all the points to the left as -1). Compute the risk $L_{\mathcal{D}}(h)$, as a function of θ (which is, as is standard, given in radians).



Solution:

$$
\begin{aligned}
Area_{circle} &= \pi r^2 \\
Area_{slice} &= \cancel{\pi} r^2 \frac{\theta}{2\cancel{\pi}} = \frac{\theta}{2} r^2
\end{aligned}
$$

For a unit circle, the failure percentage for some uniform random distribution on the unit circle $(r = 1)$ would be for two slices

$$\mathcal{L}_{\mathcal{D}}(h) = \frac{\theta r^2}{\pi r^2} = \frac{\theta}{\pi}$$

   (b) Suppose we obtain $1/\theta$ (which is given to be an integer ≥ 2) training samples (*i.e.*, samples from $\mathcal{D}$, along with their true labels). What is the probability that we find a point whose label is "inconsistent" with h? Can you bound this probability by a constant independent of θ?

Solution:

We can set $k = 1/\theta$, which is the number of points that we're chosing. As each choice is independent, we can sum the independent probabilities such that get get the probability of a point being classified wrong after k points. This gives

$$\sum_{i=1}^{k} \frac{\theta}{\pi} = \left(\frac{\theta}{\pi}\right) k = \frac{1}{\pi}$$

(c) Give an example of a distribution $\mathcal{D}$ under which h has risk zero.

Solution:

The distribution can be a uniform distribution along a line $ax + b$ with a and b being positive constants. This will separate the data such that the classifier would split the data as positive and negative based on the values of x and not allow the classifier to fail, as there would be room to offset the bias.

2. Suppose $A_1, A_2, \ldots, A_n$ are events in a probability space.

(a) Suppose $\Pr[A_i] = \frac{1}{2n}$ for all i. Then, show that the probability that none of the A_i's occur is at least $1/2$.

Solution:

$$P(A_i) = \frac{1}{2n}$$

If independent, the probability of choosing all of them is

$$P(A_{all}) = \left(\frac{1}{2n}\right)\left(\frac{1}{2n}\right)\cdots\left(\frac{1}{2n}\right)$$
$$P(A_{all}) = \prod_{i=1}^{n}\left(\frac{1}{2n}\right) = \left(\frac{1}{2n}\right)^n$$

Therefore, the probability of not choosing any

$$P(A_{none}) = 1 - \left(\frac{1}{2n}\right)^n$$

We can place a lower bound for $n = 1$

$$P(A_1) = 1 - \frac{1}{2} = \frac{1}{2}$$

Therefore, the probability of not being chosen for a given n

$$P(A_i) = 1 - \left(\frac{1}{2n}\right)^n \geq \frac{1}{2}$$

(b) Give a concrete example of events A_i for which $\Pr[A_i] = \frac{1}{n-1}$ for all i, and the probability that none of them occur is zero.

Solution:

Probability of not choosing an integer at random on the infinite domain.

(c) Suppose $n \geq 3$, and $\Pr[A_i] = \frac{1}{n-1}$, but the events are all *independent*. Show that the probability that none of them occur is $\geq 1/8$.

Solution:

Probability of not choosing one

$$P(A_i) = \left(1 - \frac{1}{n-1}\right)$$

probability of not choosing n independent

$$P(A_1, A_2, \ldots, A_n) = \left(1 - \frac{1}{n-1}\right)^n = \left(-\frac{2-n}{n-1}\right)^n$$

We can bound it by using $n = 3$, which gives

$$P(A_1, A_2, A_3) = \frac{1}{8}$$

3. In our proof of the no-free lunch theorem, we assumed the algorithm A to be deterministic. Let us now see how to allow randomized algorithms. Let A be a randomized map from set X to set Y. Formally, this means that for every $x \in X$, $A(x)$ is a random variable, that takes values in Y. Suppose $|X| < c\,|Y|$, for some constant $c < 1$.

(a) Show that there exists $y \in Y$ such that $\max_{x \in X} \Pr[A(x) = y] \leq c$.

Solution:

Assuming that A is a function randomly maps x to some unique value in Y, we can say that the probability of a single mapping is $\frac{1}{|Y|}$. To be unique, as x is assigned, it is lost from the set, so the total probability over the *entire* set is $\frac{1}{|Y|} + \frac{1}{|Y|-1} + \cdots + \frac{1}{|Y|-c|Y|}$, however the probability that they're all the same is $\frac{1}{|Y|}$. Therefore, the probability that all of the x values map to the same value in Y can be represented as

$$\begin{aligned}\max_{x \in X} \Pr[A(x) = y] &\leq \sum_{i=1}^{c|Y|} \frac{1}{|Y|} \\ &\leq c|Y|\frac{1}{|Y|} = c\end{aligned}$$

(b) Show that this implies that for any distribution $\mathcal{D}$ over X, $\Pr_{x \sim \mathcal{D}}[A(x) = y] \leq c$.
Solution:
This problem is asking the probability that for an x in our distribution that our classifier is able to classify it correctly. In other words

$$P(A(x) = y|x) = \frac{p(A(x) = y \cap x)}{p(x)}$$

Where x is bounded by $|Y|$. The entire set of X is bounded by $c|Y|$, but each independent value of x can map to anything in Y. Therefore, we have the domain as being

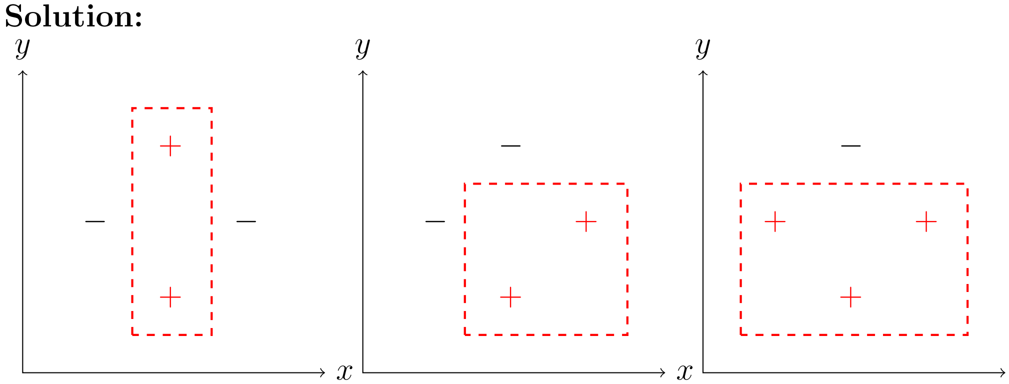
$$P(A(x) = y|x) = \frac{c|Y|}{|Y|} = c$$

This is an upper bound on the probability, which does not account for repetition.

4. Recall that the VC dimension of a hypothesis class $\mathcal{H}$ is the size of the largest set that it can "shatter."

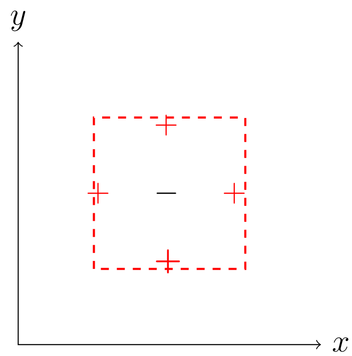
   (a) Consider the task of classifying points on a 2D plane, and let $\mathcal{H}$ be the class of axis parallel rectangles (points inside the rectangle are "+" and the poitns outside are "−"). Prove that the VC dimension of $\mathcal{H}$ is 4.
   In order to prove VC dimension, we need to prove two things

      i. There exists $(d = 4)$ points which can be shattered
      Solution:



      With the last case of all 4 being postive being trivial. These are also trivially applicable when flipping between $+$'s and $-$'s.

      ii. No set of 5 points can be shattered
      Solution:
      The minimum enclosing rectangle is defined where 1 point is each edge



Therefore, the 5^{th} point must lie on the edge or inside of the rectangle.

(b) This time, let $\chi = \mathbb{R}^d \backslash \{0\}$ (origin excluded), and let $\mathcal{H}$ be the set of all hyperplanes through the origin (points on one side are "+" and the other side are "−"). Prove that the VC dimension of $\mathcal{H}$ is $\leq d$.

Hint: Consider *any* set of $d+1$ points. They need to be linearly dependent. Now, could it happen that u, v are "+", but $\alpha u + \beta v$ is "−" for $\alpha,\ \beta \geq 0$? Can you generalize this?

Solution:

Consider d unit base vectors in $\mathbb{R}^d$

$$(1, 0, \dots, 0)(0, 1, 0, \dots, 0), \dots, (0, \dots, 0, 1)$$

It is trivially seen that this set can be defined by d hyper planes through the origin. We can now show that there are no $d+1$ vectors in $\mathbb{R}^d$ that can be shattered by hyperplanes through the origin. We can do this by *proof by contradiction.*

Suppose that $u_1, \dots, u_{d+1}$ can be shattered. This implies that there exists 2^{d+1} vectors $a_i \in \mathbb{R}^d, i = \{1, \dots, 2^{d+1}\}$ such that the matrix of inner products, denoted by $z_{i,j} = u_i^T a_j$ has columns with all possible combination of signs. Therefore, we have the matrix inner products of

$$A = \begin{pmatrix} z_{1,1} & \cdots & z_{1,2^{d+1}} \\ \vdots & \ddots & \vdots \\ z_{(d+1),1} & \cdots & z_{(d+1),2^{d+1}} \end{pmatrix}$$

which has all 2^{d+1} possible combinations of signs.

$$\text{sign}(A) = \begin{pmatrix} - & - & \cdots & - & + \\ - & \cdot & \cdots & \cdot & + \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ - & + & \cdots & \cdot & + \end{pmatrix}$$

Then, the rows of A are linearly independent as there are no constants c such that $\sum_{i=1}^{d+1} c_i z_{i,\forall} = 0$ as for any value of c_i there is a column with the same sign, which makes it always non-zero. This implies that $d+1$ vectors in $\mathbb{R}^d$ are linearly independent but it is a false statement. This contradiction proves there are no

$d+1$ vectors in $\mathbb{R}^d$ that can be shattered by hyperplaces through the origin. Thus, the VC dimension is d

(c) **(BONUS)** Let χ be the points on the real line, and let $\mathcal{H}$ be the class of hypotheses of the form $\text{sign}(p(x))$, where $p(x)$ is a polynomial of degree at most d (for convenience, define $\text{sign}(0) = +1$). Prove that the VC dimension of this class is $d + 1$.

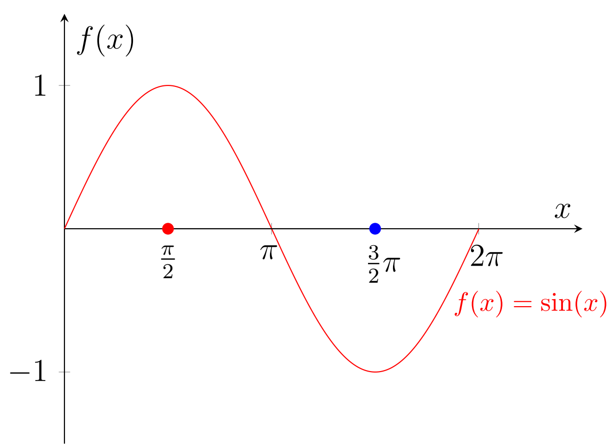
Hint: The tricky part is the uppoer bound. Here, suppose $d = 2$, and suppose we consider any four points $x_1 < x_2 < x_3 < x_4$. Can the sign pattern $+, -, +, -$ arise from a degree 2 polynomial?

5. In the examples above (and in general), a good rule of thumb for VC dimension of a function class is the *number of parameters* involved in defining a function in that class. However, this is not universally true, as illustrated in this problem: Let χ be the points on the real line, and define $\mathcal{H}$ to be the class of functions of the form $h_\theta := \text{sign}(\sin(\theta x))$, for $\theta \in \mathbb{R}$. Note that each hypothesis is defined by the single parameter θ.

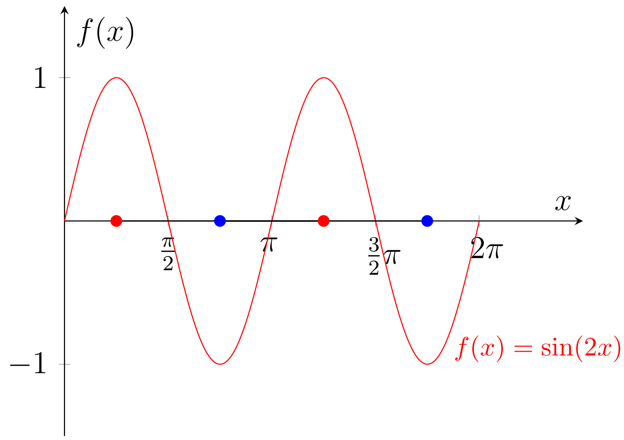
Prove that the VC dimension of $\mathcal{H}$ is infinity.

Solution:

To prove that it's infinite dimensional, we can do so with an example. Suppose we have the following points, where red is the positive and blue is the negative. These points are on the infinite domain and are spaced out every π increments starting at $\pi/2$. This function $\sin(x)$ would correctly classify the infinite number of points on the infinite domain if they keep this structure.



Suppose now we half the distance between the points so that they're separated by $\pi/2$ but start at $\pi/4$. Well, the function $\sin(\theta x)$ still holds in classifying the entire domain of points. This can be seen in the plot below, as it would take $\sin(2x)$ to classify the points.



This can be repeated ad infinitum such that if the points are even closer together, we can simply increase our value of θ such that it will classify the infinite number of points on the enitre domain.