Introduction to Machine Learning

Homework 3

Names: Obaida Khateeb, 201278066

Adham Dahli, 318575834

Question 4 – PAC, VC dimension, Bias vs Variance:

Section 1

A circle (r,c) is defined by its center c and its radius r. Look at the following classifiers family:

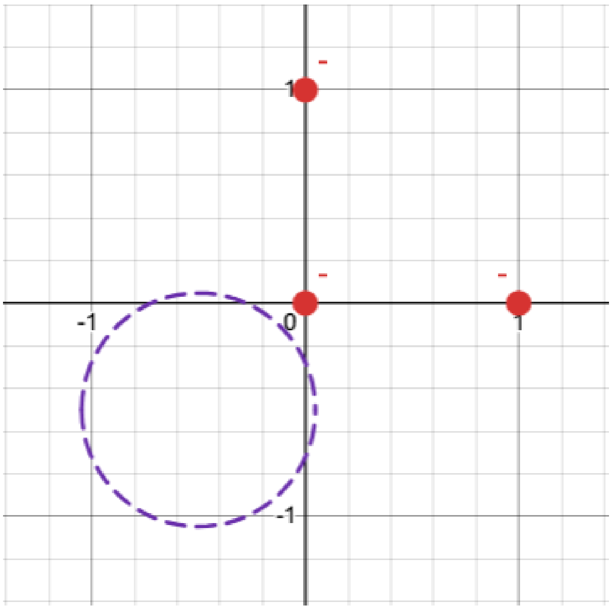
$$\mathcal{H} = \{h_{r,c}:\ r \in \mathbb{R}, c \in \mathbb{R}^2\} \text{ where } h_{r,c}(x) = 1 \text{ iff x inside the circle (r,c)}$$

Find the VCdim of this class with full proof.

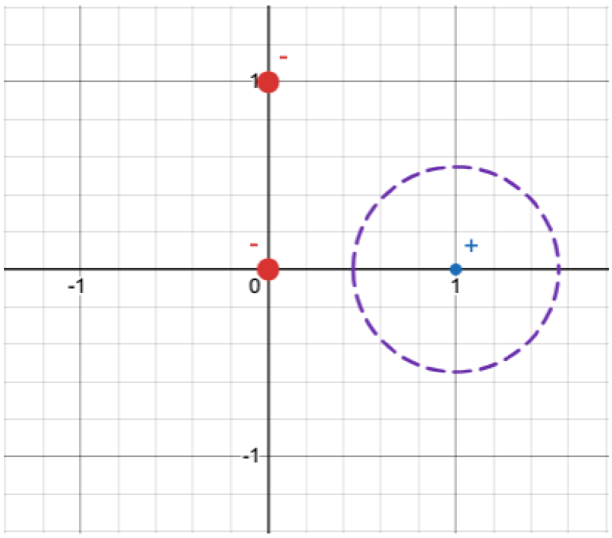
VCdim(H) = 3.

To prove that VCdim(H) = 3, we should show the following:

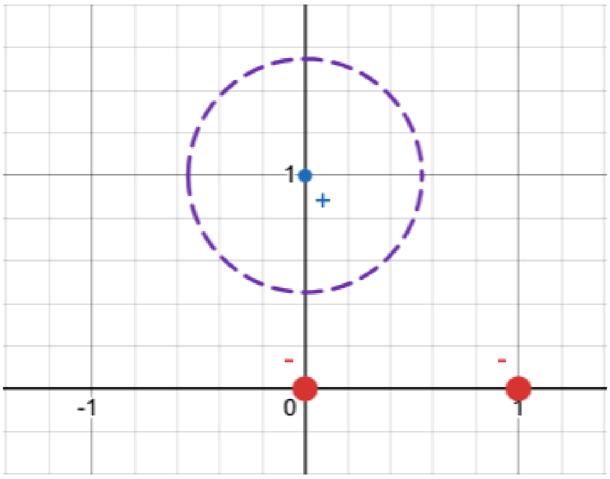
1. The existence of a group of size 3, such that H shatters it:
   Consider the following 3 points in R^2: (0,0), (0,1), (1,0).
   There are $2^3 = 8$ different possible labeling possibilities (binary labels: +1,-1) for them, let's review these possibilities while showing that for each of them there exists a circle that can separate the points according to their labels (so that '+1' inside the circle, '-1' outside the circle):
   $y\big((0,0)\big) = -1\ , y\big((0,1)\big) = -1\ , y\big((1,0)\big) = -1$:



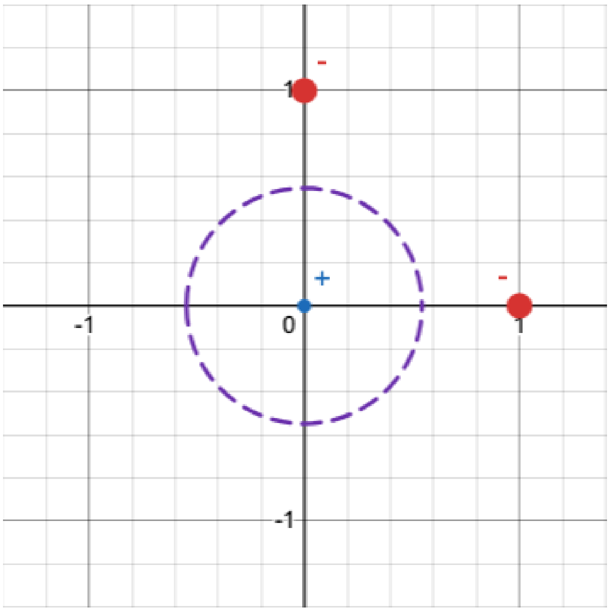
$y\big((0,0)\big) = -1\,, y\big((0,1)\big) = -1\,, y\big((1,0)\big) = +1$:



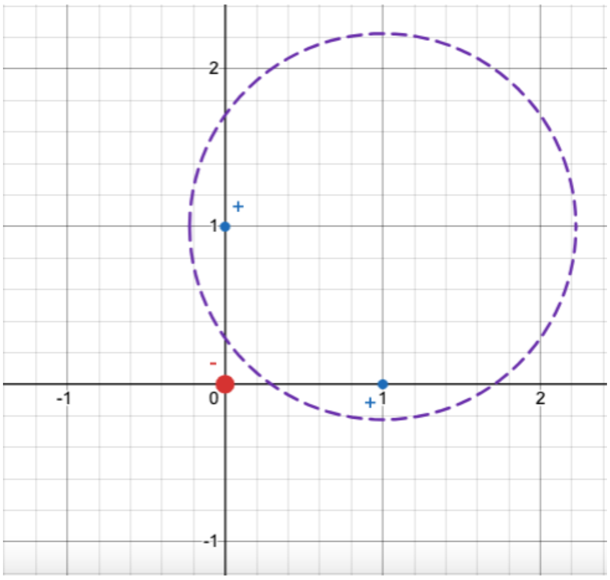
$y\big((0,0)\big) = -1\,, y\big((0,1)\big) = +1\,, y\big((1,0)\big) = -1$:



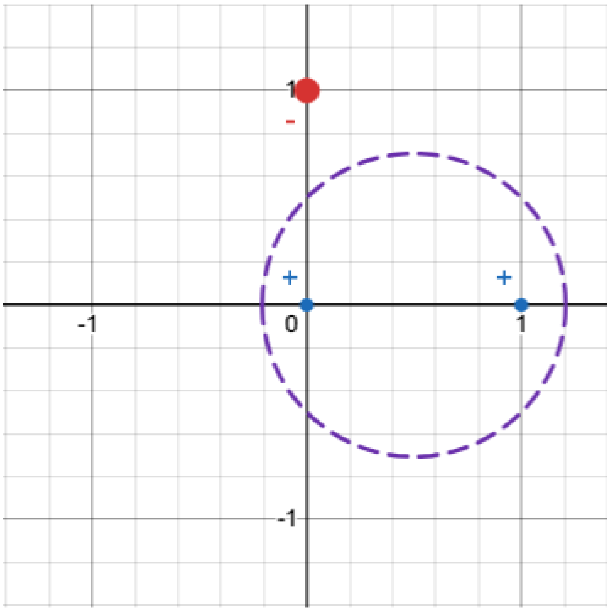
$y\big((0,0)\big) = +1\ , y\big((0,1)\big) = -1\ , y\big((1,0)\big) = -1$:



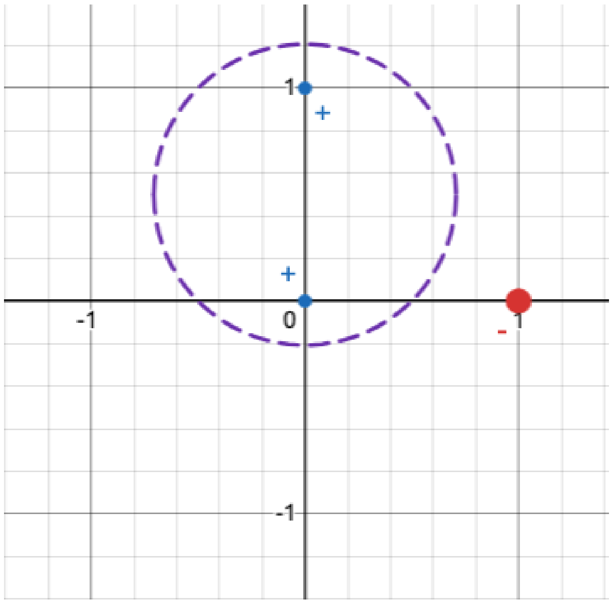
$y\big((0,0)\big) = -1\ , y\big((0,1)\big) = +1\ , y\big((1,0)\big) = +1$:



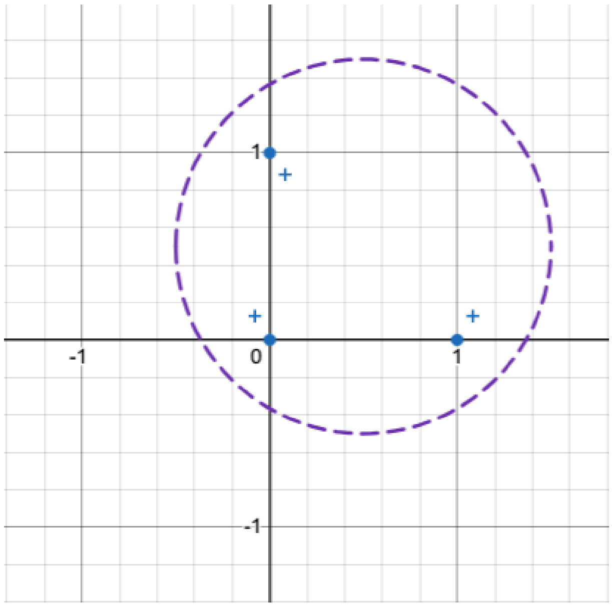
$y\big((0,0)\big) = +1\,, y\big((0,1)\big) = -1\,, y\big((1,0)\big) = +1$:



$y\big((0,0)\big) = +1\,, y\big((0,1)\big) = +1\,, y\big((1,0)\big) = -1$:



$y\big((0,0)\big) = +1\ ,y\big((0,1)\big) = +1\ ,y\big((1,0)\big) = +1$:



Thus, H can shatter 3 points, which means $VCdim(H) \geq 3$.

2. <u>Any group of size 4 can't be shattered by H</u>:

   Consider 4 points in R^2: A, B, C, D. There could be 2 cases:

   - Not all the points are situated in a boundary of a convex hull: In this situation at least one of the points is inner point, and it's impossible to shatter the inner point from the boundary points using a circle.

- All the points are situated in a boundary of a convex hull: this means that they form a four-sided polygon that has four interior angles, lets mark them by $\angle A, \angle B, \angle C, \angle D$. The sum of these interior angles is 360°. Therefore, there are 2 opposite angles out of the four that their sum is less or equal to 180°. Let's assume without loss of generality that these two angles are $\angle A$ and $\angle C$. Therefore, there couldn't be a circle that contains the points A, C but not contain B,D.

  Proof of " there couldn't be a circle that contains the points A,C but don't contain B,D": Let's assume for contradiction that $\angle A + \angle C \leq 180º$ (which means $\angle B + \angle D \geq 180º$), and there's a circle 'C' which is the smaller out of all circles that contain both A and C and excludes B and D. since it's the smaller, A and C resides on its boundaries, and since B and D are out of 'C', their angles have to be less than the points that are inside the circle, therefore $\angle B + \angle D < 180º$. Therefore $\angle A + \angle B + \angle C + \angle D < 360º$, in contradiction to the fact that the polygon's sum of inner angles equals to $360º$

Therefore, H can't shatter 4 points, which means $VCdim(H) < 4$.

Conclusion: $VCdim(H) \geq 3 , VCdim(H) < 4 \rightarrow \boldsymbol{VCdim(H) = 3}$.

Section 2

Consider a training set $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ where $x_i \in \{0,1\}^3$. In other words, each sample has 3 Boolean features $\{X_1, X_2, X_3\}$. You are also given the classification rule $Y = (X_1 \wedge X_2) \vee (\neg X_1 \wedge \neg X_2)$.

We try to learn the function $f: X \to Y$ using a "depth 1 decision trees". A "depth-1 decision tree" is a tree with two leaves, all distance 1 from the root.

Analyze this problem and decide the appropriate sample complexity formula. Justify your answer.

Firstly, the decision tree is of depth 1 which mean it splits only on features which could be X_1 or X_2. Thus we have 2 unique depth-1 decision trees. Therefore $|H| = 2$.

Secondly, since depth 1 decision tree can only base its decision on one feature, it can't perfectly represent the given classification rule which depends on the relation of two features. Thus, we must approach the problem from an agnostic PAC learning perspective, which doesn't assume the existence of a perfect hypothesis in the hypothesis class. Therefore, The sample complexity can be computed using the following formula:

$$m_H(\varepsilon, \delta) \leq \left(\frac{2 * \log\left(\frac{2|\mathrm{H}|}{\delta}\right)}{\varepsilon^2} \right) = \left(\frac{2 * \log(2|\mathrm{H}|) - 2 * \log(\delta)}{\varepsilon^2} \right)$$

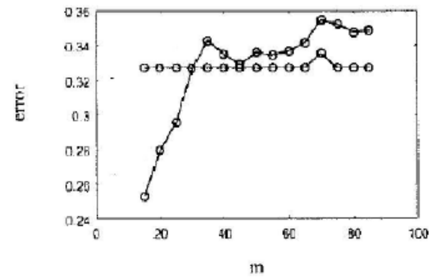
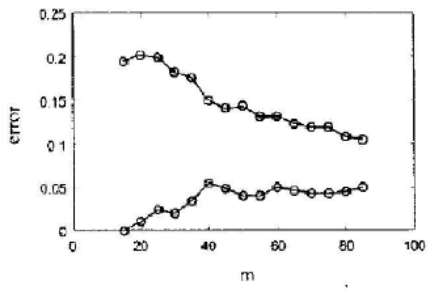
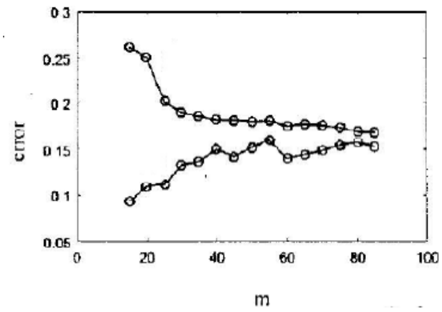
$$= \left(\frac{2 * \log(2 * 2) - 2 * \log(\delta)}{\varepsilon^2} \right) = \boldsymbol{\Theta}\left(\frac{\mathbf{2 - \log(\delta)}}{\boldsymbol{\varepsilon}^2} \right)$$

Section 3

Dana was given a hard classification problem and she decided to use SVM with polynomial kernel with d=2,10,20. For each degree, she tried 15 to 85 training samples, with jumps of 5 (15, 20, ..).

The following graphs describe the train and test error for each d separately. However, she forgot which graph belongs to which d, and for each graph, what line is the train and the test.

Your task is to match each graph to the correct d and mark which lines are the test and the train. No explanation required.

	$d = 2$
	$d = 20$
	$d = 10$