בית ספר ״אפי ארזי״ למדעי המחשב. המרכז הבינתחומי

The Efi Arazi School of Computer Science

The Herzliya Interdisciplinary Center

סמסטר ב׳ תשע״ט

Spring 2019

מבחן מועד ב בלמידה ממוכנת

Machine Learning Exam B

Lecturer: Prof Zohar Yakhini
Time limit: 3 hours

Answer 5 out of 6 from the following question (each one is 20 points)

Good Luck!

מרצה: פרופ זהר יכיני
משך המבחן: 3 שעות

יש לענות על 5 מתוך 6 השאלות הבאות
לכל השאלות משקל שווה (20 נקודות)

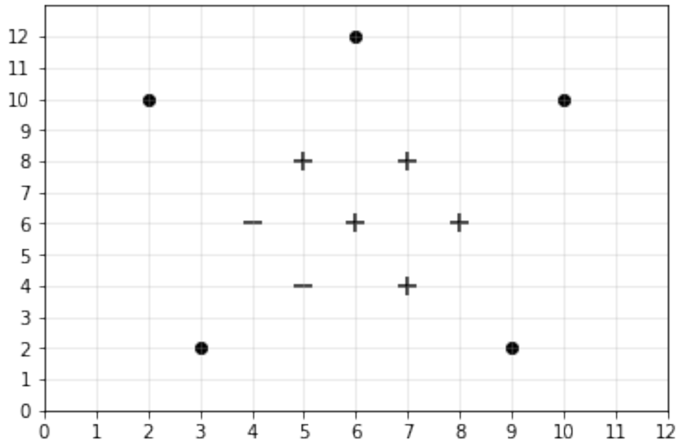
בהצלחה!

Clearly explain all your answers
All answers should be written in exam notebooks

Question 1 (3 parts)

A. Consider the kNN classification algorithm. Specifically assume we are using the L2 (Euclidean) distance metric. Furthermore, assume that we are breaking ties as follows: given k/2 positive instances and k/2 negative instances in the set of k nearest neighbors we will classify the instance according to the k+1st nearest-neighbor.

Given the following dataset:



Where:
The + data points are training points belonging to the positive class.
The – data points are training points belonging to the negative class.
The circles are new instances that you want to classify using kNN with this training dataset.

1. TRUE or FALSE:
   In 7-NN (7 nearest neighbors) the prediction of all the new instances (circle points) will be the same. Explain
2. What is maximum value of k for which not all new instances are assigned to the same class?

B.

1. Consider the following training set in $\mathbb{R}^2$:

x_1	x_2	y
7	2	+1
5	5	-1

You want to classify a new instance $x = (0,0)$ using 1-NN, first with Euclidean distance and then with Manhattan distance (L$_1$).
What will the prediction be in each one of the cases?

2. Can you find a training set in which when using 1NN with Euclidean distance the prediction on $x = (0,0)$ will be +1 and when using 1NN with L_∞ distance the prediction on that same point will be -1?

C. Consider the linear regression task of finding:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \ ||X\theta - y||_2^2$$

Where X is $m \times n$ matrix (m instances and n features including the bias term).

1. Find A, b, expressed in terms of X and y, such that the solution θ^* above is also the solution to the linear equation $A\theta = b$.
   You can't explicitly use θ^* nor the trivial solution ($A = 0, b = 0$).
2. What are the dimensions of A, b?

Question 1 (3 parts) – Solution

A.

1. TRUE
   We have 7 instances in the training set. Therefore, each new instance will classify as the majority of the classes in the training set which is the plus class.
2. The maximum k will be 3. When k>4 the majority of every 5 instances is still the plus class. When k=4 the best we can get is a tie (2 pluses and 2 minuses), and therefore we will break it, according to the definition, with k=5 and once again the prediction will be plus.
   In k=3 we can see the new instance in the bottom left which has 2 neighbors from the minus class and therefore will be classify as minus. All the rest of the new instances will be classify as plus.

B.

1. The predictions will be as follows:
   <u>Euclidean</u>:
   $$\sqrt{7^2 + 2^2} = \sqrt{53} > \sqrt{50} = \sqrt{5^2 + 5^2}$$
   In this case the prediction will be -1.
   <u>Manhattan</u>:
   $$|7 + 2| = 9 < 10 = |5 + 5|$$
   In this case the prediction will be +1.
2. Consider the following training set in $\mathbb{R}^2$:

x_1	x_2	y
6	2	+1
5	5	-1

   <u>Euclidean</u>:
   $$\sqrt{6^2 + 2^2} = \sqrt{40} < \sqrt{50} = \sqrt{5^2 + 5^2}$$
   In this case the prediction will be +1.
   L_∞:
   $$|6| = 6 > 5 = |5|$$
   In this case the prediction will be -1.

C.

1. Recall that the pseudo-inverse is defined as $pinv(X) = (X^T X)^{-1} X^T$.
   Now we can solve the linear regression task and obtain the optimal parameters: $\vec{\theta}^* = pinv(X)\,\vec{y}$.
   From here we can define $A = (X^T X)$ and $b = X^T \vec{y}$ since:
   $$A\vec{\theta}^* = b \rightarrow (X^T X)(X^T X)^{-1} X^T \vec{y} = X^T \vec{y}$$
   Which always holds.

2.

$$A_{\{n\times n\}} = X^T_{\{n\times m\}} X_{\{m\times n\}}$$
$$b_{\{n\times 1\}} = X^T_{\{n\times m\}} \vec{y}_{\{m\times 1\}}$$

Question 2 (2 parts)

A. Consider the following data set:

x_1	x_2	x_3	y
1	0	1	0
1	1	1	1
1	1	0	0
1	0	0	1
1	0	0	0
0	0	0	0
1	0	1	1

1. Given the instance $= [1, 0, 0]$, for each of the following classifiers, state the class to which the above instance would be classified. Show your calculations.
   a. Maximum likelihood Naïve Bayes.
   b. MAP Naïve Bayes classifier.
   c. MAP Full Bayes classifier.
2. Add a row to the table so that the classification in 1.c (MAP Full Bayes classifier) would flip (prediction would change to the other class).

B. Consider the following joint distribution of X and Y under a class C:

$C = 0$

	$X = 0$	$X = 1$
$Y = 0$	$\frac{1}{12}$	$\frac{3}{12}$
$Y = 1$	$\frac{2}{12}$	$\frac{6}{12}$

$C = 1$

	$X = 0$ p	$X = 1$ q
$Y = 0$ p		
$Y = 1$ q		

On the left you see the joint distribution of X and Y when C=0.
On the right you see the **marginals** of the joint distribution of X and Y when C=1, represented by p and q. In addition we are given that:

$$P(C = 0) = P(C = 1) = 0.5$$

1. Find p, q and fill the joint distribution table for C=1 so that X and Y will be conditionally independent given C, but not independent.
2. Find p, q and fill the joint distribution table for C=1 so that X and Y will be both conditionally independent and independent.

Question 2 (2 parts) – Solution

A. Let X = [1, 0, 0]

1. Maximum likelihood Naïve Bayes:
   $P(X|A=0)=0.75\cdot 0.75\cdot 0.75=0.4218$
   $P(X|A=1)=1\cdot 0.66\cdot 0.33=0.2222$
   Maximum likelihood Naïve Bayes will chose 0.
2. MAP Naïve Bayes:
   $P(A=0|X)=P(A=0)\cdot P(X|A=0)=0.57\cdot 0.4218=0.24$
   $P(A=0|X)=P(A=1)\cdot P(X|A=1)=0.43\cdot 0.2222=0.095$
   MAP Naïve Bayes will chose 0.
3. MAP full Bayes:
   X appears once in each class so they both have the same posterior $\frac{1}{7}$, so it will be a random choice.
4. Add the row [1, 0, 0, 1] or [1,0,0,0] and now one of the classes will be chosen deterministically.

B. 1.

$C=1$

	$X=0$ $p=\frac{1}{4}$	$X=1$ $q=\frac{3}{4}$
$Y=0$ $p=\frac{1}{4}$	$\frac{1}{16}$	$\frac{3}{16}$
$Y=1$ $q=\frac{3}{4}$	$\frac{3}{16}$	$\frac{9}{16}$

$C=1$

	$X=0$ $p=\frac{1}{2}$	$X=1$ $q=\frac{1}{2}$
$Y=0$ $p=\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$
$Y=1$ $q=\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$

In both of the tables X and Y Are conditionally independent.
In the table on the left X and Y are independent as well, On the table on the right they are not.
To find p for the independent case you had to solve:

$\frac{1}{2}\left(P(X=0|C=0)+P(X=0|C=1)\right)\cdot\frac{1}{2}\left(P(Y=0|C=0)+P(Y=0|C=1)\right)=\frac{1}{2}(P(X=0,Y=0|C=0)\cdot P(X=0,Y=0|C=1)=$

$\frac{1}{2}\cdot\left(\frac{1}{3}+p\right)\cdot\frac{1}{2}\left(\frac{1}{4}+p\right)=\frac{1}{2}(\frac{1}{12}+p^2)$

Question 3 (3 parts)

A. We want to cluster a set of S instances into k groups. Below is a pseudo code of an algorithm called *k-medoids with L_∞ as the distance measure*:

```
Initialize $c_1, \dots, c_k$ by randomly selecting k elements from S.
Loop:
  Assign all n instances to their closest $c_i$, with $L_\infty$ as the distance
  metric, and create k clusters $S_1, \dots, S_k$
  For each cluster $S_i$ $(1 \le i \le k)$ define a new $c_i$:
    $c_i = \underset{x \in S_i}{\operatorname{argmin}} \sum_{s \in S_i} \|x - s\|_\infty$
Until no change in $c_1, \dots, c_k$
Return $c_1, \dots, c_k$
```

In words: the new center for each cluster is the cluster member that minimizes the sum of L_∞ distances to all cluster members.

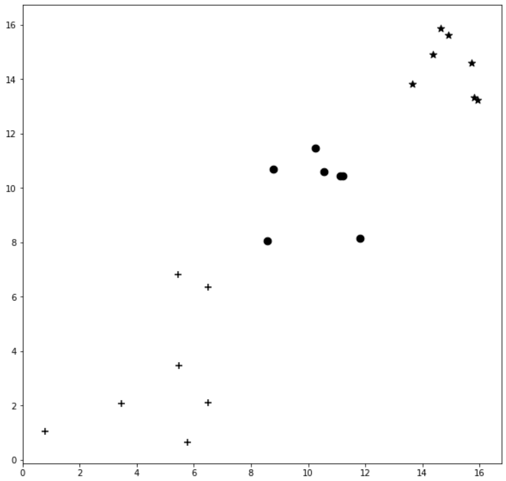
Run 2 iterations of the algorithm, with $k = 2$, $c_1 = p_1$ and $c_2 = p_5$ on the following dataset. Report the new c_1 and c_2 and the cluster assignment of each point (which cluster it belongs to).

Instance	X	Y	Z
p_1	3	9	1
p_2	7	5	9
p_3	1	6	0
p_4	2	0	2
p_5	4	5	4
p_6	9	5	11
p_7	10	1	0

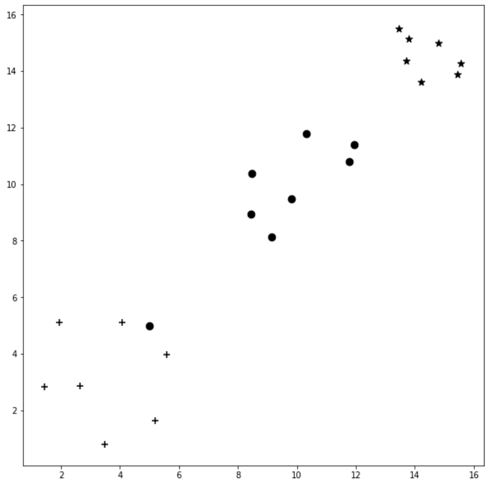
B.

1. For each of the following configurations state whether it could be an output of the k-means algorithm (that is: the algorithm will halt at this configuration).

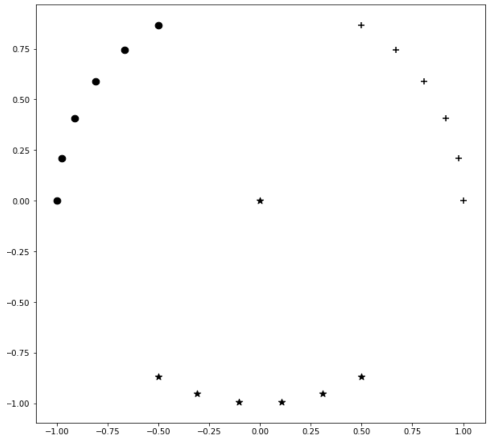
1



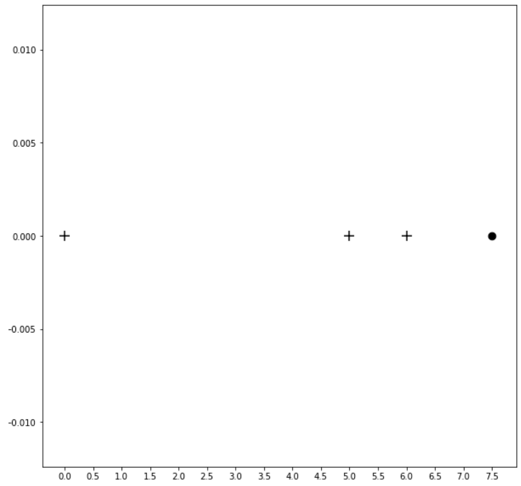
2



3



2. For the following configuration state whether it could be an output of:
   i. The k-means algorithm
   ii. The k-medoids algorithm from part A



C. Recall the naïve cluster growing algorithm:

```
While there are still unclustered elements in the data:
    Pick a seed element s and create the cluster $C_s$
    Mark s as clustered
    While there are unclustered elements e with $d(e, C_s) < T$:
        Insert all elements e with $d(e, C_s) < T$ to $C_s$ and mark them as clustered.
    End
End
```

Here, T is a pre-defined threshold set by the user.

Let $d(e, C_s) = \min_{x \in C_s} \|x - e\|_2$. That is d is defined to be the Euclidean distance to the closest element in C_s.

Given a dataset $X = \{x_1 < x_2 < x_3 \ldots < x_m\}$, on the real line R^1, what is the minimal T one should set in order to get a single cluster?

Question 3 (3 parts)-solution

A. There were a few solutions to this question, all correct ones got full score.
Here is the simplest one (no change after the 1st iteration):
Cluster 1 : c1=[3,9,1], s1=[[3,9,1], [1, 6, 0]]
Cluster 2: c2=[4, 5, 4], s1=[[7, 5, 9], [2, 0, 2], [4, 5, 4], [9, 5, 11], [10, 1, 0,]

B.

1. Yes!
2. No – The dot will clearly be reassigned to the plus instances.
3. Yes! – It can be argued that it won't stop there, but it definitely **could** be an output.
4. K-means: No, the leftmost plus will be reassigned to the dark circles cluster.
   k-mediods: Yes! One center will the dart point and the other will be the middle plus.

C. $T = \max_{1 \le i \le m-1} (x_{i+1} - x_i)$, pick T to be the maximum value between 2 neighboring points.

Question 4 (3 parts)

A. Are decision trees used for classification sensitive to feature transformations of the form $x' = x + a$, where a is some constant?
What about $x' = a \cdot x$, where $a \neq 0$?

B. Consider the training data described below. Decide on which attribute to split using **Entropy** as an impurity measure. Clearly state your calculations and the formulas you are using.

Outlook	Wind	Decision
Sunny	Weak	No
Sunny	Strong	No
Overcast	Weak	Yes
Rain	Weak	Yes
Rain	Weak	Yes
Rain	Strong	No
Overcast	Strong	Yes
Sunny	Weak	No
Sunny	Weak	Yes
Rain	Weak	Yes
Sunny	Strong	Yes
Overcast	Strong	Yes
Overcast	Weak	Yes
Rain	Strong	No

C. How would you change the decision tree learning algorithm to support regression of continuous values? Can you use entropy as a splitting criterion in this case?

Question 4 (3 parts) – Solution

A. Continuous features are tested according to the order of the values. The greater than / less than criteria only cares about the order of the values and not the specific value. This is true for every order preserving transformation. Therefore, decision trees are not to the transformations in the question.

B.

$$\text{entropy}(S) = -\frac{9}{14}\log\frac{9}{14} - \frac{5}{14}\log\frac{5}{14} = 0.94$$

$$\text{entropy}(\text{Wind} = \text{Weak}) = -\frac{2}{8}\log\frac{2}{8} - \frac{6}{8}\log\frac{6}{8} = 0.81$$

$$\text{entropy}(\text{Wind} = \text{Strong}) = -\frac{3}{6}\log\frac{3}{6} - \frac{3}{6}\log\frac{3}{6} = 1.00$$

$$\text{entropy}(\text{Outlook} = \text{Sunny}) = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0.97$$

$$\text{entropy}(\text{Outlook} = \text{Overcast}) = -\frac{4}{4}\log\frac{4}{4} = 0.00$$

$$\text{entropy}(\text{Outlook} = \text{Rain}) = -\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5} = 0.97$$

$$\text{Info} - \text{Gain}(\text{Wind}) = 0.94 - \frac{8}{14}\cdot 0.81 - \frac{6}{14}\cdot 1 = 0.048$$

$$\mathbf{Info} - \mathbf{Gain}(\mathbf{Outlook}) = 0.94 - \frac{5}{14}\cdot 0.97 - \frac{4}{14}\cdot 0 - \frac{5}{14}\cdot 0.97 = 0.24$$

C. Starting from the root, we will calculate the standard deviation of the target value. We will evaluate the standard deviation for each split in the same manner this is performed in classification decision trees. For each discrete feature we will consider splitting all possible attribute values and for continuous features we will consider a single split using the midpoints between the sorted feature values. The gain for each possible split will be the reduction in standard deviation of the target value. We will pick the feature that reduced the standard deviation the most for the split.

Question 5 (4 parts)

A. Find the minimum and the maximum of $3x + 2y$ under the constraint $4x^2 + y^2 = 1$

B. Given the following dataset:

X1	X2	Y
+1	+1	+1
0	+2	-1
-1	+1	+1
0	0	-1

Use the lemma below to show that it is not linearly separable.

Lemma:

Assume that a linear classifier predicts the same $y \in \{-1, +1\}$ for some two points $z, z' \in \mathbb{R}^2$ (that is $h(z) = h(z')$). Then it will produce the same prediction for any intermediate point. That is:

$$\forall \alpha \in [0,1] \quad h\big((1-\alpha)z + \alpha z'\big) = y$$

C. Find a mapping φ into a space of a dimension of your choosing that maps the dataset from Part B into a linearly separable dataset and define the linear classifier.

D. Consider the kernels $K_1, K_2: \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}$. Prove or disprove (by a counter example):

1. For $c > 0$, cK_1 is also a kernel.
2. $K = K_1 + K_2$ is also a kernel.
3. $K = K_1 - K_2$ is also a kernel.

Question 5 (4 parts) – Solution

A.

$$L(x,y,\lambda) = 3x + 2y\ - \lambda(4x^2 + y^2 - 1)$$
$$f'_x = 3 - 8\lambda x = 0 \Rightarrow x = \frac{3}{8\lambda}$$
$$f'_y = 2 - 2\lambda y = 0 \Rightarrow y = \frac{1}{\lambda}$$
$$f'_\lambda = 4x^2 + y^2 - 1 = 0$$
$$\frac{36}{64\lambda^2} + \frac{1}{\lambda^2} = 1 \Rightarrow \lambda^2 = \frac{100}{64} \Rightarrow \lambda = \pm\frac{5}{4}$$

Then we get:

$$x = \pm\frac{3}{10}$$
$$y = \pm\frac{4}{5}$$

And therefore:

$max = \frac{5}{2}$ is attained at $\left(\frac{3}{10},\frac{4}{5}\right)$ $min = -\frac{5}{2}$ is attained at $\left(-\frac{3}{10},-\frac{4}{5}\right)$

B. Choose $\alpha = 0.5$ and the first and third points in the data to see that a linear classifier that perfectly predicts the data must classify the point (0,1) as +1. Likewise, use $\alpha = 0.5$ and the second and forth points in the data, to see that a linear classifier that perfectly predicts the data must classify the point (0,1) as -1.
Contradiction.

C. $\varphi(x_1,x_2) = (x_1,x_2,x_1^2)$.
One possible separator has weights $(w_1,w_2,w_3,b) = (0,0,1,-0.5)$, resulting in the predictor

$$h(x_1,x_2) = sgn(< w,\varphi(x_1,x_2) >) =$$
$$= sgn(< (0,0,1,-0.5),(x_1,x_2,x_1^2,1) >) = sgn(x_1^2 - 0.5)$$

* There are many other possible answers to this question.

D. Consider the kernels $K_1,K_2\colon \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}$. Prove or disprove (by a counter example):

1. TRUE
x, y are two vectors in the lower n-dimension.
Lets d_1 be the higher dimension.
By definition:

$$K_1(x,y) = \varphi_1(x)\cdot\varphi_1(y) = \sum_{i=1}^{d_1}\varphi_1(x)_i\varphi_1(y)_i$$

$$K(x,y) = \mathrm{c}K_1(x,y) = \mathrm{c}\sum_{i=1}^{d_1} \varphi_1(x)_i \varphi_1(y)_i$$
$$= \mathrm{c}\varphi_1(x)_1\varphi_1(y)_1 + \cdots + \mathrm{c}\varphi_1(x)_{d_1}\varphi_1(y)_{d_1} = \varphi(x)\cdot\varphi(y)$$

Where $\varphi(x) = \left(\sqrt{c}\varphi_1(x)_1, \dots, \sqrt{\mathrm{c}}\varphi_1(x)_{d_1}\right)$ Q.E.D

2. TRUE

   x, y are two vectors in the lower n-dimension.

   Lets d_1, d_2 be the higher dimensions respectively.

   By definition:

$$K_1(x,y) = \varphi_1(x)\cdot\varphi_1(y) = \sum_{i=1}^{d_1} \varphi_1(x)_i\varphi_1(y)_i$$
$$K_2(x,y) = \varphi_2(x)\cdot\varphi_2(y) = \sum_{i=1}^{d_2} \varphi_2(x)_i\varphi_2(y)_i$$
$$K(x,y) = K_1(x,y) + K_2(x,y) = \sum_{i=1}^{d_1} \varphi_1(x)_i\varphi_1(y)_i + \sum_{i=1}^{d_2} \varphi_2(x)_i\varphi_2(y)_i$$
$$= \varphi_1(x)_1\varphi_1(y)_1 + \cdots + \varphi_1(x)_{d_1}\varphi_1(y)_{d_1} + \varphi_2(x)_1\varphi_2(y)_1 + \cdots$$
$$+ \varphi_2(x)_{d_2}\varphi_2(y)_{d_2} = \varphi(x)\cdot\varphi(y)$$

   Where $\varphi(x) = \left(\varphi_1(x)_1, \dots, \varphi_1(x)_{d_1}, \varphi_2(x)_1, \dots, \varphi_2(x)_{d_2}\right)$ Q.E.D

3. FALSE

   One possible example is $K_1 = K_2 \rightarrow K = 0$ and therefore, K is not a kernel.

Question 6 (3 parts)

A. Give an example of an instance space X and a binary hypotheses space H on X, such that:

$$VC(H) = 2020$$

B. What is the VC-dimension of 2 concentric rings in $\mathbb{R}^2$, where instances on the rings are classified as positive?

Formally:

$$H = \{h \mid h(\vec{x}) = +1 \Leftrightarrow (r_1 \leq d(\vec{x},(0,0)) \leq r_2) \vee (r_3 \leq d(\vec{x},(0,0)) \leq r_4)\}$$

Where $0 < r_1 < r_2 < r_3 < r_4$

In this formal description:

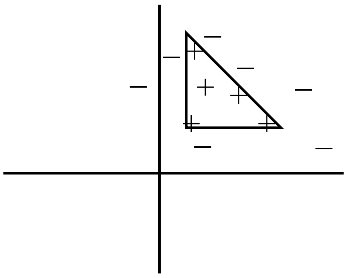
r_1, r_2 – are the radii of the first ring

r_3, r_4 – are the radii of the second ring

$d(\vec{x},\vec{y})$ – is the Euclidean distance between $\vec{x}$ and $\vec{y}$.

Note that the center of both rings is in the origin.

C. Let $X = \mathbb{R}^2$. Let $C = H$ the set of all hypotheses that assign positive values inside isosceles straight triangles with the equal sides parallel to the axes and with their head vertex on the lower left (see picture). Describe a polynomial sample complexity algorithm L that learns C using H. State the time complexity and the sample complexity of your suggested algorithm. Your sample complexity should be linear in $\frac{1}{\epsilon}$. Prove all your steps.



Question 6 (3 parts) – Solution

A. There are several answers to this question. Few of them are:

1. H = all the linear classifier. $X = \mathbb{R}^{2019}$. We saw in class that linear classifiers have VC=d+1.
2. H = n-intervals where $n \leq 1010$. $X = \mathbb{R}$. We saw in class that the hypothesis space of n-intervals in $\mathbb{R}$ has VC=2n. Therefore, for $X = \mathbb{R}$ we have.
   VC(1010 intervals)=2020.

B. $VC \geq 4$:

Consider the following points:

(0,1), (0,2), (0,3), (0,4)

We can find a hypothesis from H that can separate these points for each dichotomy on them:

+ + + + $r_1 = 0.5, r_2 = 4.5$ $r_3, r_4 > 4.5$	+ + + - $r_1 = 0.5, r_2 = 3.5$ $r_3, r_4 > 4.5$	+ + - + $r_1 = 0.5, r_2 = 2.5$ $r_3 = 3.5, r_4 = 4.5$	+ + - - $r_1 = 0.5, r_2 = 2.5$ $r_3, r_4 > 4.5$
+ - + + $r_1 = 0.5, r_2 = 1.5$ $r_3 = 2.5, r_4 = 4.5$	+ - + - $r_1 = 0.5, r_2 = 1.5$ $r_3 = 2.5, r_4 = 3.5$	+ - - + $r_1 = 0.5, r_2 = 1.5$ $r_3 = 3.5, r_4 = 4.5$	+ - - - $r_1 = 0.5, r_2 = 1.5$ $r_3, r_4 > 4.5$
- + + + $r_1 = 1.5, r_2 = 4.5$ $r_3, r_4 > 4.5$	- + + - $r_1 = 1.5, r_2 = 3.5$ $r_3, r_4 > 4.5$	- + - + $r_1 = 1.5, r_2 = 2.5$ $r_3 = 3.5, r_4 = 4.5$	- + - - $r_1 = 1.5, r_2 = 2.5$ $r_3, r_4 > 4.5$
- - + + $r_1 = 2.5, r_2 = 4.5$ $r_3, r_4 > 4.5$	- - + - $r_1 = 2.5, r_2 = 3.5$ $r_3, r_4 > 4.5$	- - - + $r_1 = 3.5, r_2 = 4.5$ $r_3, r_4 > 4.5$	- - - - $r_1, r_2, r_3, r_4 > 4.5$

$VC < 5$:

Sort each 5 points by their distance from the origin:

$$d_1 \leq d_2 \leq d_3 \leq d_4 \leq d_5$$

There is no hypothesis that can separate + - + - +.

QED

C. The algorithm will produce a hypothesis which is the smallest relevant triangle that contains all the positive points. This can be done in O(m) as follows:

Let $\Delta = \Delta^m = (x_i, y_i)_{i=1}^m$ be a set of points in the plane, labeled positive and negative (see Fig1).

Our algorithm seeks to return a hypothesis $h \in H$.

Let $(x_i, y_i)_{i=1}^{m^{(+)}}$ be all positively labeled data points.

Find:

1. $a \coloneqq \min\limits_{1 \leq i \leq m^{(+)}} (x_i)$
2. $b \coloneqq \min\limits_{1 \leq i \leq m^{(+)}} (y_i)$

3. $s := \max_{1\le i\le m^{(+)}}(x_i + y_i)$

The vertices of the hypothesis triangle $h \ = \ L(\Delta)$ will be:
$\gamma = (a, b), \alpha = (a, s-a), \beta = (s-b, b)$.

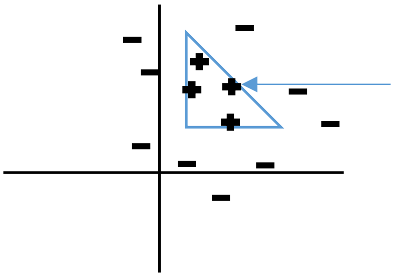

The hypothesis triangle

Fig1

Consider $c \in C$ and let $\Delta^m(c) = (x_i(c), y_i(c))_{i=1}^m$ be training data generated from c without errors and by drawing m independent points according to a probability distribution π on $\mathbb{R}^2$. We will denote the probability distribution thus induced on $(\mathbb{R}^2)^m$ by π^m .

Given $\varepsilon > 0$ and $\delta > 0$ we will now compute a number $m(\varepsilon, \delta)$ so that

(Eq.1) $\quad m \ge m(\varepsilon, \delta) \Longrightarrow e\big(\Delta^m(c)\big) = \pi^m \left(err_\pi\big(L\big(\Delta^m(c)\big), c\big) > \varepsilon\right) \le \delta$

Note that $L\big(\Delta^m(c)\big)$ is the hypothesis h, or the triangle, produced by L when considering data $\Delta^m(c)$ as above. $e\big(\Delta^m(c)\big)$ is a random variable that depends on the stochastic behavior of $\Delta^m(c)$. It is exactly this behavior that we will want to characterize.

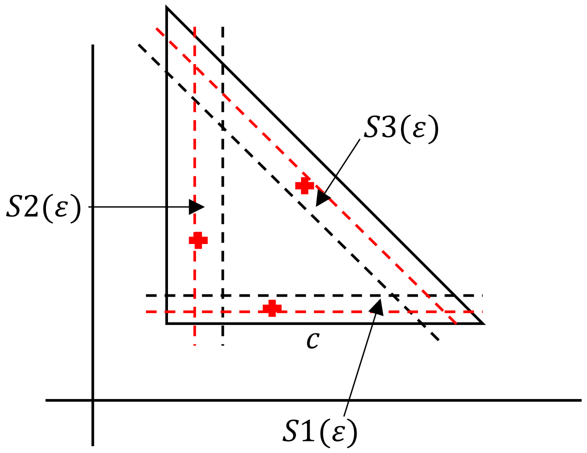


Fig2

Consider the strips parallel to the edges of the triangle c as in Fig2. These are defined to satisfy:

$$\pi\big(S1(\varepsilon)\big) = \pi\big(S2(\varepsilon)\big) = \ \pi\big(S3(\varepsilon)\big) = \frac{\varepsilon}{3}$$

Now, note that

$$\left\{\Delta^m(c) : err_\pi\big(L\big(\Delta^m(c)\big), c\big) > \varepsilon\right\} \subseteq$$

$$\{\Delta^m(c): \Delta^m(c) \cap S1(\varepsilon) = \emptyset\} \cup$$
$$\{\Delta^m(c): \Delta^m(c) \cap S2(\varepsilon) = \emptyset\} \cup$$
$$\{\Delta^m(c): \Delta^m(c) \cap S3(\varepsilon) = \emptyset\}$$

This is because if $\Delta^m(c)$ visits all three strips (note that negative points can not visit these strips as there are no errors) then, according to our construction, the difference between c and $L(\Delta^m(c))$ will have $\pi \leq \pi\big(S1(\varepsilon) \cup S2(\varepsilon) \cup S3(\varepsilon)\big) < \varepsilon$. See the red triangle example in Fig2.

In term of probability we therefore get:

$$\pi^m \left(err_\pi\big(L(\Delta^m(c)), c\big) > \varepsilon\right) \leq$$
$$\pi^m \left(\Delta^m(c) \cap s1(\varepsilon) = \emptyset\right) +$$
$$\pi^m \left(\Delta^m(c) \cap s2(\varepsilon) = \emptyset\right) +$$
$$\pi^m \left(\Delta^m(c) \cap s3(\varepsilon) = \emptyset\right) \leq$$
$$3\left(1 - \frac{\varepsilon}{3}\right)^m$$

We now select $m(\varepsilon, \delta) = \frac{3}{\varepsilon}\left(\ln(3) + \ln\left(\frac{1}{\delta}\right)\right)$ to get (Eq.1) to hold.

Q.E.D

GOOD LUCK!