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)

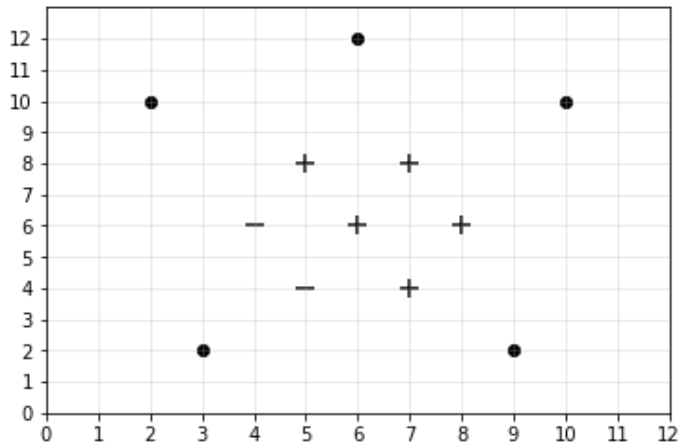
בהצלחה!

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 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 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_∞ 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.

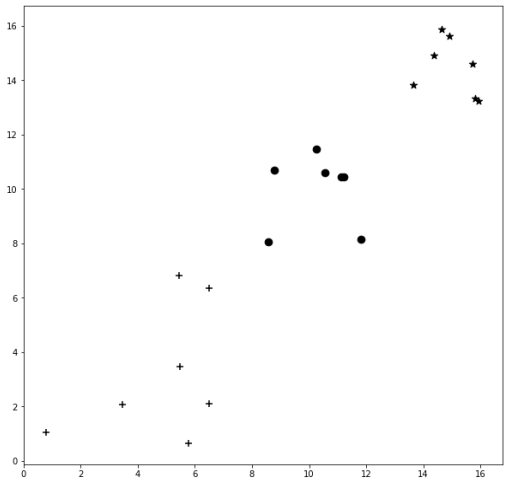
Run 2 iterations of the algorithm, with $k = 2$, $c_1 = p_1$ and $c_2 = p_5$ on the following dataset. Report the new c_1 and c_2 and the cluster assignment of each point (which cluster it belongs to).

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

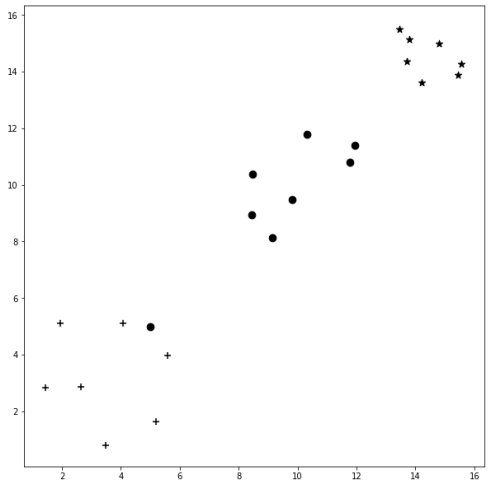
B.

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).

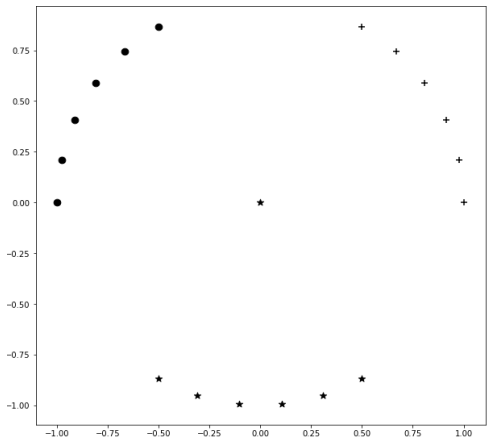
1



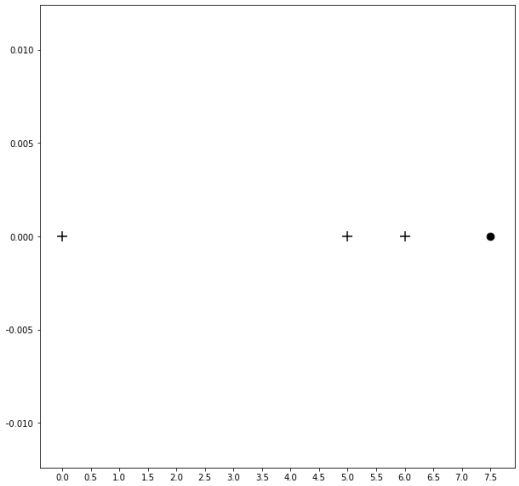
2



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 4 (4 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 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.

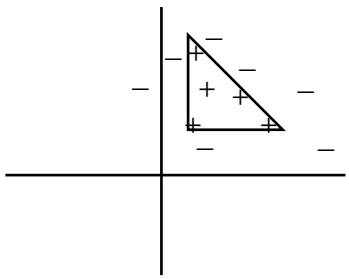
Question 6 (3 parts)

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.



GOOD LUCK!