Notebook No:__________:מס' מחברת **I.D number:______________:מס' ת.ז**

בית ספר "אפי ארזי" למדעי המחשב המרכז הבינתחומי

The Efi Arazi school of computer science

The Interdisciplinary Center

סמסטר ב' תשע"ז
Spring 2017

מבחן מועד א בלמידה ממוכנת

Machine Learning Exam A

Lecturer : Prof Ariel Shamir,
Dr. Zohar Yakhini
Time limit : 3 hours
Additional material or calculators are not allowed in use!

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

Good Luck!

מרצים: פרופ אריאל שמיר,
ד"ר זהר יכיני

משך המבחן: 3 שעות
אין להשתמש בחומר עזר ואין להשתמש במחשבונים!

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

Question 1 (5 parts)

In $\mathbb{R}^2$ consider a hypotheses space of the interiors of "top-right facing right angled isosceles triangles" – all isosceles triangles where the two equal sides are parallel to the x and y axes and the hypotenuse is on the upper right.

Such triangles can be created by intersecting the following lines: x=a, y=b, x+y=r, where r>a and r>b.

Formally:

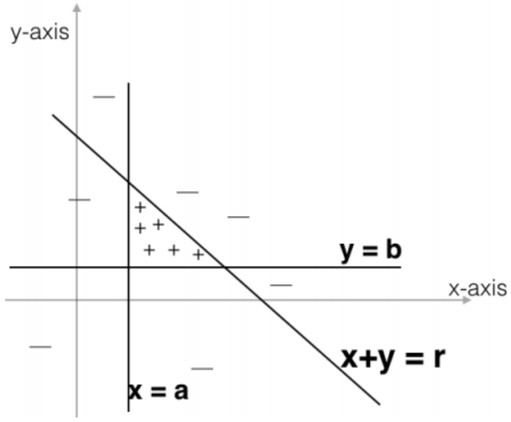
$H_\Delta = \{h_{a,b,r}: a, b, r \in \mathbb{R}, \mathrm{r} > \mathrm{a}, \mathrm{r} > \mathrm{b}\}$

where positives are inside:

$$h_{a,b,r} = \begin{cases} 1 & if\ x \geq a\ and\ y \geq b\ and\ x + y \leq r \\ -1 & otherwise \end{cases}$$

That is: $h_{a,b,r}$ assigns a positive label to any point in the interior or on the edges of the triangle.

This is a picture of a single hypothesis in this space:



Recall the three sample complexity bounds we have learned in class:

$$m \geq \frac{1}{\varepsilon}\left(\ln|H| + \ln\frac{1}{\delta}\right)$$
$$m \geq \frac{1}{2\varepsilon^2}\left(\ln|H| + \ln\frac{1}{\delta}\right)$$
$$m \geq \frac{1}{\varepsilon}\left(8 \cdot VC(H)\log_2\frac{13}{\epsilon} + 4\log_2\frac{2}{\delta}\right)$$

A. Given that we only consider $1 \leq a = b \leq n$ and $1 \leq r \leq 2n$ and for integer valued a, b, r , compute the size of the hypotheses space, as a function of n

B. Let n = 100.
   1. Define a consistent learning algorithm for this case. That is – find $h_c \in H_\Delta$ from any given training data that can be assumed to have been generated from a concept $c \in H_\Delta$
   2. Show that the space H_Δ is PAC learnable by the algorithm you defined above (no need to provide a tight bound)

C. Assume that a=b=0 and that r can be any positive real number. Compute the VC dimension of the resulting hypotheses space H_Δ

D. Consider the learning algorithm you had defined in B1 (recall that n=100 and values are integers). Give a bound on the number of samples required to learn a concept with 90% probability and an error of at most 0.05. Show the required calculation only. (provide a formula, No need to provide final answer)

E. Does your answer to D change if we use the space definition from C? In not – why? If yes – how would you compute a bound on the sample size here (provide a formula, No need to provide final answer)

Question 2 (7 parts)

You are given a set of m instances that are defined by some feature x (i.e. x_i is the value of the feature of instance i), and the values of the target function y defined on each instance as y_i. The formula to calculate the Pearson Correlation Coefficient between the feature x and the target function y is given by:

$$\rho = \frac{\sum_{i=1}^{m}(x_i-\mu_x)(y_i-\mu_y)}{\sqrt{\sum_{i=1}^{m}(x_i-\mu_x)^2\sum_{i=1}^{m}\left(y_i-\mu_y\right)^2}}$$

A. Explain what does the correlation formula measures? what is the meaning of the nominator and what is the meaning of the denominator in the given formula? and what are the possible values of this formula?

B. Explain the connection between correlation and dependency between x and y (as random variables)? Explain the meaning of this connection and the dependency between x and y in the following cases:
- The correlation is 1
- The correlation is -0.8
- The correlation is 0

C. Explain the connection between the correlation measure and linear regression that is trying to explain y using x. In which of the three cases appearing in Section B it is worthwhile to use linear regression? Explain why. Will there be an error in the evaluation of y in each of these cases?

Now assume that the instances have n different features and not just one. We will indicate each feature with a lower index such as x_i is the i-th feature, and to indicate an instance from the set we will use an upper index. For example, x_i^d will mark the i-th feature of the d-th instance from the set of m instances.

D. Explain (including formula) what is the covariance matrix (or scatter matrix) of the features of the set of instances, and explain what is the connection between its elements and the Correlation Coefficient from A.

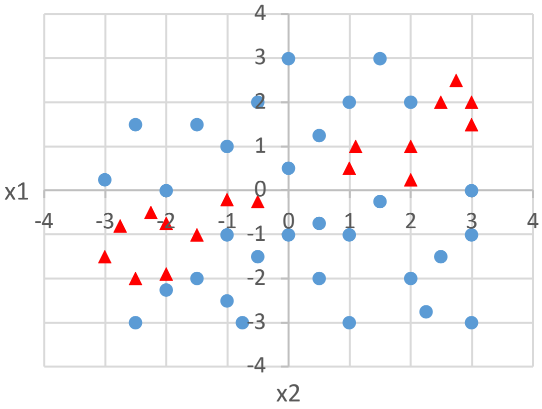
E. Explain what is the purpose of PCA algorithm (principal component analysis)

F. Explain what is the purpose of LDA algorithm (linear discriminant function)

G. Explain what is the different of the scatter matrix in use in PCA algorithm and the one in use in LDA algorithm.

<u>Question 3 (6 parts)</u>

Consider the following training data:



A. Will a perceptron learning algorithm find a good classifier for this training set? If yes – what are the weights for the separator? If not – explain why not.

B. We applied SVM learning to this training data. We have tried different kernels. We obtained the following classification rule (execution algorithm):

$$t(x) = \text{sgn}\big(g(\text{x})\big) = \text{sgn}\left(\sum_{i \in SV} \alpha_i t_i (x_i \cdot x)^2\right)$$

Where

SV is the set of support vectors, α_i is the weight form the i-th support vector, t_i is the class the i-th support.

What kernel $K(x, y)$ was used here?

C. Find $\varphi: \mathbb{R}^2 \to \mathbb{R}^3$, so that the kernel function from B satisfies $K(x, y) = \varphi(x) \cdot \varphi(y)$.

D. Find the weights of the linear separator in $\mathbb{R}^3$ which is equivalent to $x_2(x_2 - x_1) = 0$

E. Assume that an RBF kernel was selected. That is – a kernel of the form:

$$K(x, y) = \varphi(x) \cdot \varphi(y) = exp\left(-\frac{1}{2}\|x - y\|^2\right)$$

Prove that for every two points $x, y \in \mathbb{R}^2$ the following holds:

$$\|\varphi(x) - \varphi(y)\|^2 = 2 - 2exp\left(-\frac{1}{2}\|x - y\|^2\right)$$

F. For each of the following statements decide TRUE or FALSE and explain your answer:

1. After mapping into higher dimensional space using the RBF it is possible that a perceptron achieves better separation than in the original space.
2. After mapping into higher dimensional space using the RBF it is possible that a 1-NN classification algorithm, based on unweighted Euclidean distance, achieves better classification results than in the original space.
   Hint: use the identity proven in E.
3. The answer to F2 doesn't change if we use K-NN with weighted distances, where k>1

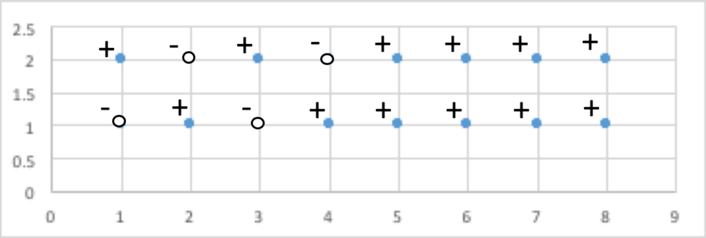
Question 4 (7 parts)

The following table and graph represent a data set containing two features x1 and x2 from two classes that we mark as + and -. This set is used as the training set for a decision tree. We create two decision trees:

TOver Tree is the tree that is extended until the end of the algorithm with no limit and no pruning.

TUnder is the tree that contains only one root node where all the data is included.

instance	x1	x2	Value
1	1	2	+
2	2	1	+
3	3	2	+
4	4	1	+
5	5	1	+
6	5	2	+
7	6	1	+
8	6	2	+
9	7	1	+
10	7	2	+
11	8	1	+
12	8	2	+
13	1	1	-
14	2	2	-
15	3	1	-
16	4	2	-



A. Write the formula for "Goodness of Split" criterion that determines the feature that is chosen to split the decision tree and explain it.

B. To build the TOver tree, assume that we are using the GiniIndex as the function that measures how homogeneous a set is:

$$GiniIndex(S) = 1 - \sum_{i=1}^{c} (p_i)^2 = 1 - \sum_{i=1}^{c} \left(\frac{|S_i|}{|S|}\right)^2$$

Explain (including formula) how this measure is used to determine the Goodness of Split.

C. Explain how the first split in TOver tree is determined? How many Goodness of Split calculations must be done? (there is no need to calculate the numbers or result, just explain how many and which calculation should be done to determine the first split).

D. Assume that the first split in TOver tree is defined by the formula: x1<4.5. How many leaves would the tree have at the end of the learning?

E. Can there be a situation where you stop building a decision tree before all leaves are homogeneous (contain instances of only one class)? Explain why and if so, how is the classification of a new instance that reaches such a leaf determined?

F. What would be the total error of TOver tree using leave-one-out? Explain!

G. What would be the total error of TUnder tree using leave-one-out? Explain!

Question 5 (6 parts)

We want to cluster to k groups a set S of training examples. Below is a pseudo code of an algorithm that is called k-medoids and is similar to k-means:

```
Initialize $c_1, \ldots, c_k$ by randomly selecting k elements from S
Loop:
    Assign all n samples to their closest $c_i$ and create k clusters $S_1, \ldots, S_k$
    For each cluster $S_i$ ($1 \le i \le k$) define a new $c_i$:
            choose $c_i \in S_i$ whose distance to all other members in $S_i$ is the smallest
Until no change in $c_1, \ldots, c_k$
Return $c_1, \ldots, c_k$
```

A. Assume that the set S has 7 instances with 2 features as given in the table. Execute the k-medoids algorithm to cluster the set into two groups (i.e. k=2) when you initialize the execution with p_1 and p_5 as the initial centers. This means that instead of random initialization, $c_1=p_1$ and $c_2=p_5$ (hint: first plot the instances on a 2D Euclidean plane).

B. What is the main difference between k-means and k-medoids?

C. Explain using a formula, what function does k-means minimizes?

D. How would you change the function from C so that it would fit the k-medoids algorithm?

E. Assume that we execute both algorithms on the same set. Should we expect that the value of the function that k-medoid would minimize be smaller, greater or the same as the value of the function that k-means minimizes? Explain why!

F. What serious problem can you find in the k-medoid algorithm as it is presented above in the pseudo code?

instance	x	y
p_1	**2**	**6**
p_2	4	7
p_3	5	8
p_4	6	1
p_5	**6**	**4**
p_6	7	3
p_7	5	6

Question 6 (5 parts)

Consider a diagnostic test that consists of measuring two quantitative features x1 and x2. We know, based on long-term measurement history, that the class-conditional distribution of values for these features are given by (D and H denote the two classed D = disease and H = healthy):

For the first feature

$$f(x_1|D) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$$

$$f(x_1|H) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-3)^2}{2}\right)$$

and, for the second feature

$$f(x_2|D) = \begin{cases} 1 & 0 \le x_2 \le 1 \\ 0 & \text{Otherwise} \end{cases}$$

$$f(x_2|H) = \begin{cases} e & 1 - \frac{1}{e} \le x_2 \le 1 \\ 0 & \text{Otherwise} \end{cases}$$

It is recommended (but not mandatory) that you schematically sketch these distributions to facilitate your answers.

A. Explain (in the general context) the difference between the ML (maximum likelihood) based classification and MAP (maximum aposteriori) based classification approaches.

B. What would the ML prediction be in the following (separate) two cases:
   1. We have measured, for a certain patient, the value $x_2 = 0.25$
   2. We have measured, for a different patient, the value $x_1 = 1$

C. Assuming a prior P(H), clearly state the Naïve Bayes MAP formula for this diagnostic test.
   What is the minimal prior P(H) for which the MAP prediction, in Case B2 above, would be H?

D. Assume that for B2 we also measured $x_2 = 0.95$.
   In this case what is the minimal prior P(H) for which the MAP prediction would be H?

E. For a third case we measured $x_1 = 8$ and $x_2 = 0.95(1-\frac{1}{e})$. Assume that P(H) = 0.9. What is the MAP prediction in this case? Do you think that this represents a bias or a shortcoming of the classification approach? How would you remedy this problem?

Good Luck!