# בית ספר "אפי ארזי" למדעי המחשב המרכז הבינתחומי
# The Efi Arazi school of computer science
# The Interdisciplinary Center

**סמסטר ב' תש"פ**
**Spring 2020**

## מבחן מועד ב בלמידה ממוכנת
## Machine Learning Exam B

**Lecturer**: Prof Zohar Yakhini
**Time limit**: 2 hours

**You should answer Qn 1 (mandatory), for 40 pts and two out of the other three questions (30pts each, 60pts together).**

**In the first page indicate the numbers of the two questions you answered. If there is no indication then the first two solutions will be graded.**

Good Luck!

**מרצה:** פרופ זהר יכיני
**משך המבחן:** 2 שעות

**יש לענות על שאלה מספר 1 (חובה), 40 נקודות, ולבחור שתים מתוך שלוש השאלות הנוספות (30 נקודות לכל אחת, סה"כ 60 נקודות).**

**בעמוד הראשון יש לרשום את מספרי שתי שאלות הבחירה שעליהן בחרת לענות. אם לא יהיה רשום תיבדקנה שתי התשובות הראשונות.**

בהצלחה!

You can use all Moodle course materials as well as student personal notes prepared before the exam. There is no need to print the material. You can use the appropriate files on your PC.
You can use calculators.
Justify all your answers and show your calculations.
All answers should be legibly written and scanned for submission as per IDC's instructions.

ניתן להשתמש בכל החומר הקיים ב Moodle ובנוסף סיכומי סטודנטים שנכתבו לפני תחילת המבחן. אין צורך להדפיס את החומר. אפשר להשתמש במחשב כדי לגשת לקבצים הרלוונטיים.
ניתן להשתמש במחשבון.
יש להצדיק את כל תשובותיך ולהראות את צעדי החישוב.
כל התשובות צריכות להיות כתובות בכתב ברור וקריא ולהיסרק להגשה ע"פ הנחיות IDC.

## Question 1 (3 parts, 40 points) – MANDATORY QUESTION

A. You receive 10,000 samples of labeled data with two classes, split into training and test, to perform classification from 5 real valued features $(x_1, x_2, x_3, x_4, x_5)$.
   1. (4 points) If you preform Logistic Regression with all 5 features, what is the dimension of the inferred vector of coefficients $\vec{\theta}$?
   2. (5 points) The classification results from the previous section were not satisfactory. Your colleague suggests to map the data using the full rational variety of degree 10 and then apply Logistic Regression to the resulting data.
      What is the problem with this approach?
   3. (5 points) Suggest an algorithmic approach, to finding a classifier in your data, that would overcome the problem of the suggestion from the previous section. Provide full details of your proposed approach.

B. Assume that you perform learning for classification on 5,000 samples with two classes A and B. Class A is the positive class. Your training / test split was 4,000 / 1,000. The prior of class A in your data is 0.9. For the test data you observed:

$$\text{Eq. 1} \qquad \frac{TP + TN}{TP + FP + TN + FN} = 0.905$$

   1. (3 points) Provide an example of a confusion matrix for the test data that is consistent with the above information.
   2. (3 points) Is Eq.1 necessarily an indication of good performance? Explain.

C. (10 points) Let $X = \mathbb{R}^2$. Consider $H$ to be a hypotheses space that consists of hypotheses that assign a + (positive) to the inner part at most 5 circles.
Prove that $VC(H) \geq 15$.

Formally:
For any 5 triplets $C = \{(a_i, b_i, r_i)\}_{i=1}^{5}$ where $a_i, b_i \in \mathbb{R}, r_i \in \mathbb{R}_+$
we define

$$h(C) = \bigcup_{i=1}^{5} \{(x, y) \mid (x - a_i)^2 + (y - b_i)^2 \leq r_i^2 \}$$

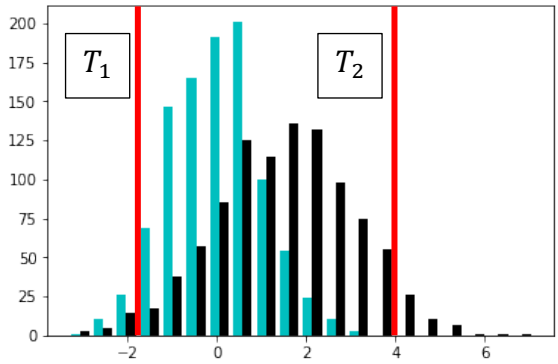
And now we define the hypotheses space as:

$$H = \{h(C) \mid \ a_i, b_i \in \mathbb{R}, r_i \in \mathbb{R}_+\}$$

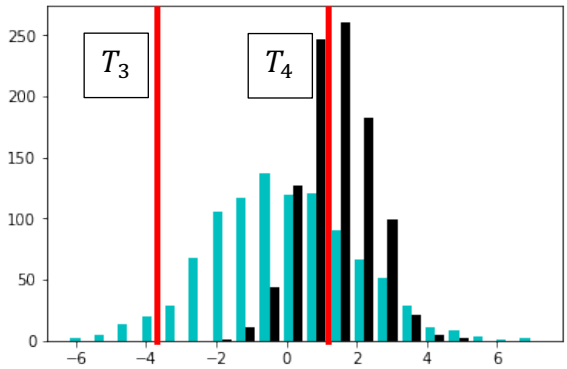
D. Two companies are proposing a Corona test to a hospital. The hospital asks for your advice as to which test is better. The plots below represent the distributions of the tested quantity for the two companies. The black histograms represent the positive cases. We further depict the ROC curves computed to evaluate their performance.

Distributions

A

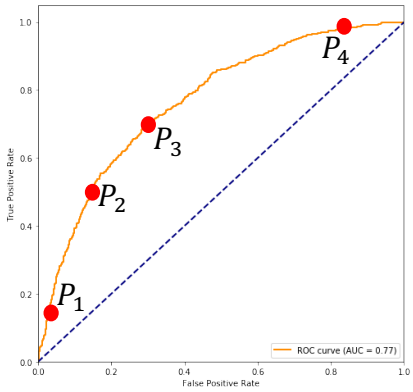



B

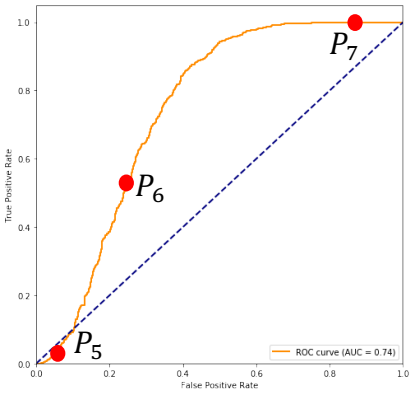



ROC curves

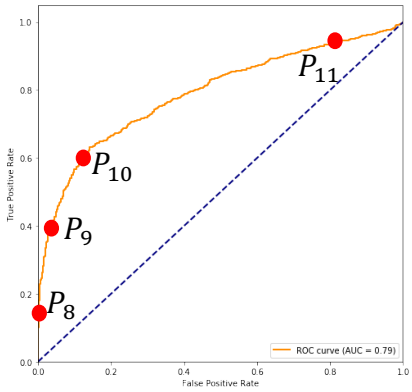
X




Y




Z




1. (2 points) Match two of the three ROC curves above, X, Y and Z, with the two distributions, A and B. Explain.
2. (4 points) Match the indicated thresholds $T_1, T_2, T_3, T_4$ to the appropriate points from the set $P_1, P_2, P_3, P_4, P_5, P_6, P_7, P_8, P_9, P_{10}, P_{11}$. Explain.

Recall that the expected benefit of a test, for the hospital, is measured by

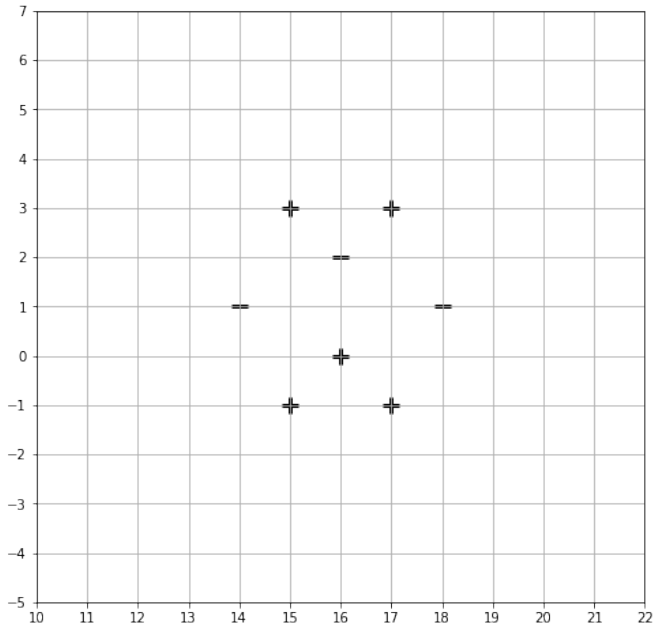
$$\pi = \alpha * TPR - FPR$$

3. (4 points) How would you find a value of $\alpha$ for which both companies can be equivalent in terms of the expected benefit? Explain your answer.

## Question 2 (3 parts 30 points)

A. Consider the kNN classification algorithm. Specifically assume we are using the $L_1$ (Manhattan) 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.

Consider the following training dataset:




Where:
The + data points are training points belonging to the positive class.
The – data points are training points belonging to the negative class.

1. (5 points) What is the leave-one-out cross validation error of 3-NN in the above training dataset?

Now, further consider the following TEST data:

| index | $x_1$ | $x_2$ |
|---|---|---|
| 1 | 13 | -1 |
| 2 | 19 | -1 |
| 3 | 12 | 3 |
| 4 | 16 | 4 |
| 5 | 20 | 3 |

2. (5 points) What will be the prediction for each one of the TEST instances using 5-NN (5 nearest neighbors)? Explain
3. (5 points) What is maximum value of k for which not all new TEST instances are assigned to the same class?

B.

1. (5 points) Consider the following training set in $\mathbb{R}^2$:

| $x_1$ | $x_2$ | $y$ |
|---|---|---|
| 8 | 3 | +1 |
| 6 | 6 | -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. (5 points) Can you find a training set for which:
When using 1NN with Euclidean distance the prediction on $x = (1, -1)$ will be +1 and when using 1NN with $L_\infty$ distance the prediction on that same point will be -1?

C. (5 points) Consider two matrices $M \in \mathbb{R}^{m \times k}$ and $W \in \mathbb{R}^{k \times k}$ and a vector $y \in \mathbb{R}^m$.

Provide a closed form solution for:

$$\theta^* = \underset{\theta \in \mathbb{R}^k}{\operatorname{argmin}} \sum_{i=1}^{m} \left( (M_i W, \theta) - y_i \right)^2$$

Where:
$M_i$ is the $i$-th row of $M$.
$M_i W$ is the vector $M_i$ multiplied by the matrix $W$.
$(u, v)$ represents the standard inner product in Euclidian space.

## Question 3 (4 parts 30 points)

A. (6 points) Are decision trees used for classification sensitive to feature transformations of the form $x' = x^3$?

B. (8 points) Consider a decision tree that only performs binary splits. That is: considering an attribute with discrete (or categorical) values, the split will be into two subsets – cats, dogs and elephants to one side and birds and snakes in the other side.
How many Goodness of Split calculations will a tree construction algorithm perform in the root when there are k discrete attributes and the i-th attribute has V(i) possible values? Explain your answer.

C. (10 points) Consider the training data described below, at a node S. Decide on which attribute to use according to the rules defined in B and using GiniGain. State your calculations.

| X1 | X2 | Y |
|---|---|---|
| Green | Ronaldo | - |
| Green | Ronaldo | + |
| Green | Messi | - |
| Blue | Messi | + |
| Blue | Messi | + |
| Blue | Neymar | - |
| Blue | Neymar | - |
| Blue | Neymar | - |

D. (6 points) Consider the following dataset with features in $\mathbb{R}^2$:

| X1 | X2 | Y |
|---|---|---|
| 1 | 10 | - |
| 1 | 5 | + |
| 2 | 10 | + |
| 2 | 5 | - |
| 3 | 5 | - |

Show that using the standard Goodness of Split with Gini as the impurity function doesn't lead to an optimal decision tree for these data. Optimal here is in the sense of a minimal number of nodes in the resulting decision tree.

## Question 4 (2 parts 30 points)

A. (10 points)
Find the maximum and minimum values of

$$f(x) = 81x^2 + y^2$$

subject to the constraint

$$4x^2 + y^2 = 9$$

B. Consider the instance space $X = \mathbb{R}^3$ with some probability distribution $\pi$. Consider the concept space $C$ that consists of symmetric boxes around the origin. In each of the three axes, the distance of the 2 sides from the origin is equal. Instances inside the box belong to the positive class and instances outside the box belong to the negative class.
Formally, for any $u, v, w > 0$ define
$c(u, v, w) = \{(x, y, z) | \ |x| \leq u, \ |y| \leq v \ , \ |z| \leq w\}$
and then:
$C \ = \{c(u, v, w) | u, v, w > 0 \}$.
Let $H = C$.

1. (6 points) Propose a consistent learning algorithm L that takes as input labeled points $D^m = \{(x_1, y_1, z_1), (x_2, y_2, z_2) \ldots, (x_m, y_m, z_\mathrm{m})\} \subseteq \mathbb{R}^3$ and returns $h \in H$.
2. (7 points) Prove that $C$ is PAC learnable from $H$ by computing a sufficiency bound on the sample complexity.
3. (7 points) For $\varepsilon = 0.1$ and $\delta = 0.05$ compute a sufficiency bound on the size, $m$, of a set $D^m$ of independently drawn training samples, that guarantees that for any $c \in C$ we have:
$Prob\big(Err\big(c, L(D^m)\big) > \ \varepsilon\big) < \delta$

# GOOD LUCK!