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

סמסטר ב׳ תשע"ז
Spring 2018

מבחן מועד א בלמידה ממוכנת
Machine Learning Exam A

Lecturer: Prof Zohar Yakhini
Time limit: 3 hours
Additional material or calculators are not allowed in use!

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

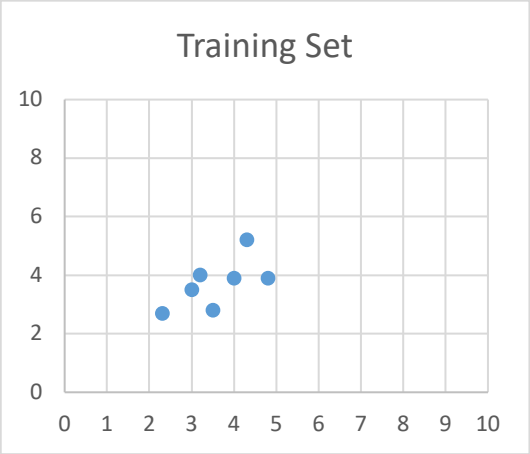
מרצה: פרופ זהר יכיני
משך המבחן: 3 שעות
אין להשתמש בחומר עזר ואין להשתמש במחשבונים!

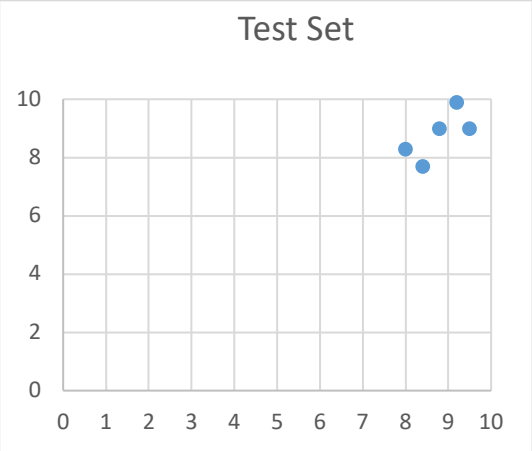
יש לענות על 5 מתוך 6 השאלות הבאות לכל השאלות משקל שווה (20 נקודות)
בהצלחה!

Question 1 (5 parts)

A. Write the formula for MSE (Mean Square Error) in the context of predicting the values of a function $y = f(x)$ (regression)

B. Consider the following training data and test data:





You are using one of the following approaches to predict y=f(x) using the training set:

a. Linear regression
b. 2-NN regression (regression using the closest 2 neighbors)

Which one of the two approaches has a smaller error on the **test** data?

C. How would you change the definition of MSE loss (as in A above) so that every instance can have a different weight, w_i, in computing the loss?

D. Write the pseudo code for linear regression, including the update step, using gradient descent in stochastic mode, as learned in class. How would you modify the algorithm to minimize the loss function you had defined in C?

E. Recall that linear regression seeks to find

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

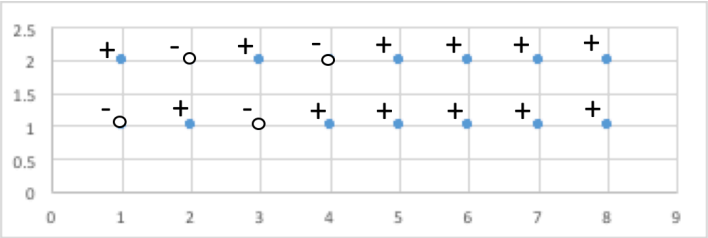
Find a matrix W to help you modify the above equation and define a pseudo-inverse solution for the function you defined in C.

Explain all your steps. In your solution, include the matrix used and the modified minimization task.

Question 2 (4 parts)

Consider the following data table and its graphical representation. The data consists of instances with two numerical attributes, x1 and x2, coming from two classes "+" and "-".
We use these data as training for learning a decision tree.
A tree of type T_N is a binary tree that uses Goodness of Split to perform splits and to grow up to height N or until no further split is possible.
For example:
T_1 will have one split (a root and two children)
T_2 will have one split at the root and then up to one split for each child.
* note: the tree doesn't have to be symmetric.

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. Explain what an impurity function is and how Goodness of Split uses an impurity function φ to determine node splitting in a decision tree. Your explanation should use a clear formula.
B. Is it possible for a tree built using Goodness of Split to be of greater height than that of another tree built by splits that also get to pure leaves? If yes – provide an example. If not – clearly explain why.
C. Will the first split in a T_1 tree learned from training data be different from the first split in a T_3 learned from the same training data? Explain your answer.
D. When constructing trees of type T_1 and of type T_2 using the training data given above, what is the error obtained, evaluated by leave one out?
Which of the two approaches leads to a better result, as evaluated by leave one out?

Question 3 (5 parts)

A. Find the minimum and the maximum of $x + 4y$ under the constraint $x^2 + 9y^2 = 1$
B. Given the following dataset (the XOR function):

X1	X2	Y
+1	+1	-1
+1	-1	+1
-1	+1	+1
-1	-1	-1

use the lemma below to show that it is not linearly separable. You do not need to prove the lemma.

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 choice that maps the dataset from Part B into a linearly separable dataset and define the linear classifier.
D. Consider the mapping $\varphi(x) = (1, x, x^2, x^3, \dots, x^N)$ for $x \in (-1, +1)$ (the open interval between -1 and +1) with some $N \in \mathbb{N}$. Show that a kernel function $K(x, y)$ exists for φ .
E. Show that the function $K(x, y) = \frac{1}{1-xy}$ for $x, y \in (-1, +1)$, is a kernel for some mapping into infinite dimensional space.

Question 4 (4 parts)

Given an instance set S, we want to divide S into k groups (preform clustering).
The k-means-outlier algorithm is a variant of k-means given by the pseudocode below:

```
Initialize c1, …, ck randomly
Loop:
    Assign all n samples to their closest ci and create k clusters S1, …, Sk
    For each cluster Si (1 ≤ i ≤ k) define a new ci:
            bi = the center of the cluster (average point)
            if |Si| > 2 (if the number of samples in Si is larger than 2):
                    x = the sample with the highest distance from bi
                    ci = the center of the cluster without x
            else:
                    ci = bi
Until no change in c1, …, ck
Return c1, …, ck
```

A. What is the function that the standard k-means algorithm seeks to minimize? Provide a formula.
B. Consider an instance set S of size 5 with 2 features as shown in the table. Run both the standard k-means algorithm and the k-manes-outlier algorithm with $k = 2$ and with the starting centers at $c_1 = (0, 0)$ and $c_2 = (2,2)$. You might want to draw the points on a 2-dimensional grid. At each iteration write down the new centers and which center each point is assigned to. No need to show all intermediate calculations.

instance	x_1	x_2
p_1	1	0
p_2	0	1
p_3	2	1
p_4	1	2
p_5	6	6

C. Does the k-means-outlier algorithm converge? If so, prove it. If not, show an example where the algorithm fails to converge.
D. A version of a fuzzy-k-means algorithm goes according to the following pseudocode:

```
Initialize c1, …, ck randomly
Loop:
        Calculate a distance vector for each sample with distances to each cluster
        For each sample j, convert the distance vector to probability vector
```

$$w_j = (w_{j1}, \dots, w_{jk})$$

For each cluster S_i ($1 \le i \le k$) define a new center c_i:

$$c_i = \frac{\sum_{j=1}^{n} w_{ji} x_j}{\sum_{j=1}^{n} w_{ji}}$$

```
Until no change in c1, …, ck
Return c1, …, ck
```

Write a suggested pseudo code for a fuzzy version of the <u>k-means-outlier algorithm</u>.

Question 5 (4 parts)

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

$$VC(H) = 2018$$

B. Recall the 3 sample complexity bounds given in class:

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

Consider an instance space $X = [0,1] \times [0,1]$.

Let $N \in \mathbb{N}$ and $N \geq 2$, we define the set $\mathrm{A} := \{\frac{1}{N}, \frac{2}{N}, \dots, 1\}$ and the following hypothesis spaces:

- $H_1 = \{h: h(x_1, x_2) = +1 \Leftrightarrow x_1 \in [0, a] \wedge x_2 \in [0, a],\ a \in A\}$
- $H_2 = \{h: h(x_1, x_2) = +1 \Leftrightarrow x_1 \in [0, a] \wedge x_2 \in [0, b],\ a, b \in A\}$
- $H_3 = \{h: h(x_1, x_2) = +1 \Leftrightarrow x_1 \in [0, a] \wedge x_2 \in [0, b],\ a, b \in [0,1]\}$

You can view each hypothesis as an axis aligned rectangle with a vertex at the origin. (i.e. an instance is classified as positive iff it's inside the rectangle with vertices [(0,0), (0,a), (a,b), (0,b)])

For each of the following cases, use one of the above bounds to compute the number of instances needed to guarantee an error of less than 0.1 with probability at least 95%:

1. When trying to learn a concept in H_3 using H_2.
2. When trying to learn a concept in H_1 using H_2.
3. When trying to learn a concept in H_3 using H_3.

Question 6 (4 parts)

Consider a quality test that consists of measuring two quantitative features x1 and x2.
We know, based on long-term measurement history, that the class-conditional probability density functions for these features are given below (G and B denote the two classes G =Good and B =Bad).
For the first feature

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

$$f(x_1|B) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-5)^2}{2}\right)$$

and, for the second feature

$$f(x_2|G) = \begin{cases} \frac{1}{3} & 0 \le x_2 \le 1 \\ \frac{2}{3} & 2 \le \mathrm{x}_2 \le 3 \\ 0 & otherwise \end{cases}$$

$$f(x_2|B) = \begin{cases} e^2 & 3 - \frac{1}{e^2} \le x_2 \le 3 \\ 0 & otherwise \end{cases}$$

Sketching out these distributions might help your intuition.

A. What would the ML prediction be in the following (separate) two cases?
   1. We have measured, for a certain product, the value $x_2 = 2.9$
   2. We have measured, for a different product, the value $x_1 = 3$
B. Assuming a prior P(B), clearly state the Naïve Bayes MAP formula for this quality test.
   What is the minimal prior P(B) for which the Naïve Bayes MAP prediction, in Case A2 above, would be B?
C. Assume that for A2 we also measured $x_2 = 2.99$
   In this case what is the minimal prior P(B) for which the Naïve Bayes MAP prediction would be B?
D. For a third case we measured $x_1 = 6$ and $x_2 = 0.975(3 - \frac{1}{e^2})$. Assume that P(B) = 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!