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

סמסטר ב' תשפ"א
Spring 2021

מבחן מועד א בלמידה ממוכנת
Machine Learning Exam A

מרצה: פרופ זהר יכיני
משך המבחן: 3 שעות

יש לענות על 4 מתוך 5 השאלות הבאות.
לכל השאלות משקל שווה (25 נקודות)

בהצלחה!

Lecturer: Prof Zohar Yakhini
Time limit: 3 hours

Answer 4 out of 5 from the following questions. Each question is 25 points.

Good Luck!

ניתן להשתמש בדפי העזר המצורפים, מחשבון ומילון בלבד. כל חומר עזר אחר אסור.

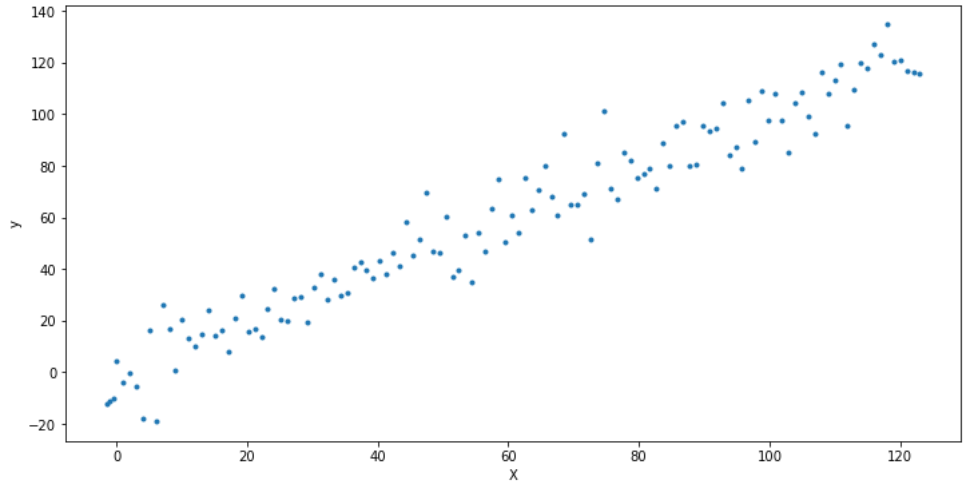
יש להסביר\להוכיח את כל התשובות.

You can use the attached formula sheet, a calculator and a dictionary. All other material should not be used.

Prove/explain all your answers.

Question 1 (25 points) – Linear Regression

1. You are given the following dataset:



You want to check which algorithm will generalize better, in predicting y from X: linear regression or 3-NN (kNN with $k = 3$).
Your colleague suggests to use cross validation as follows:
Divide the data to 3 folds: [-5,40), [40, 80), [80,125].
Use these folds and run cross validation, first with linear regression and second with 3-NN. Then select the one with lower average MSE.

   a) (4 pts) Which algorithm will be selected? Explain.
   b) (4 pts) Is there a problem in your colleague suggested approach? Explain.
   c) (4 pts) Can you think about 3 new data points (in the test data) that will have better results when using 3NN? State the coordinates of these points. Explain your answer.

2. Consider an instance matrix X with dimensions $m \times n$ (m instances and n features).
   a) (3 pts) For standard linear regression, how many coefficients does the model consist of?
   b) (3 pts) For polynomial regression with degree r, how many coefficients does the model consist of? (In linear regression $r = 1$).

3. Recall that linear regression, using MSE loss, seeks to find:

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

   a) (2 pts) How would you change the definition of MSE loss so that every instance can have a different weight, w_i, in computing the loss?
   b) (5 pts) Find a matrix W to help you modify the above equation and define a pseudo-inverse solution for the loss function you defined above.
   Explain all your steps. In your solution, include the matrix used and the modified minimization task.

Question 2 (25 points) – Decision Trees

1. (10 pts) You are trying to discriminate between Humans (H) and Martians (M) using the following characteristics: Green ∈ $\{N, Y\}$, Legs ∈ $\{2,3\}$, Height ∈ $\{S, T\}$ and Smelly ∈ $\{N, Y\}$.
   Using the data in the table, with species representing the label, construct the first two levels (split the root and its two descendants) of a decision tree using entropy and the greedy algorithm learned in class.
   Draw the resulting tree.
   Do you require an additional split in order to obtain a tree with 0 training loss?

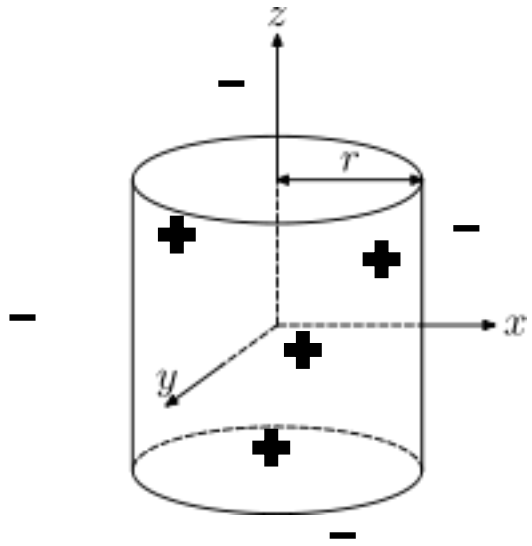
Green	Legs	Height	Smelly	Species
N	3	S	Y	M
Y	2	T	N	M
Y	3	T	N	M
N	2	S	Y	M
Y	3	T	N	M
N	2	T	Y	H
N	2	S	N	H
N	2	T	N	H
Y	2	S	N	H
N	2	T	Y	H

2. (5 pts) Given the same data as (1), provide a decision tree with height = 2 and a training error of 0 (monochromatic leaves)
   (Hint: some features can be omitted).
3. (3 pts) Is it possible to swap the role of chi-square and entropy during the construction and pruning of the decision tree? i.e., use chi-square for best feature selection and entropy for post pruning. Explain your answer.
4. (7 pts) Consider two decision trees as described below.
   a. The first tree was trained using some dataset, $X = X_{m\times n}$ using entropy
   b. Sample indices randomly according to a binomial distribution, $Binom(0.5, n)$. Let S be the set of all sampled indices. Let the matrix A be diagonal matrix of size $n \times n$ where $A_{i,i} = 1$ if and only if the index $i \in S$ and $A_{i,i} = -1$ otherwise. Let $Y = X \cdot A$. The second tree was trained on Y using entropy.

   Do both trees have the same predictions on all training samples?
   Prove/explain your answer.

Question 3 (25 points) – Learning Theory

Let $X = \mathbb{R}^3$ and let $C = H$ be the set of all cylinders threaded on the z axis with center at the origin. Consider data points inside the cylinder as Positives and data points outside of the cylinder as Negatives.
Let π be some probability distribution over X.
Formally:
For any two numbers $l, r \in \mathbb{R}_+$
we define $h(l,r) = \{(x,y,z) \mid \ x^2 + y^2 \leq r \ \wedge \ -l \leq z \leq l\}$.
And now we define the hypotheses space as: $H = \{h(l,r) \mid l, r \in \mathbb{R}_+\}$



1. (5 pts) Calculate the VC-dimension of H. Prove your answer.
2. (6 pts) Propose a consistent learning algorithm L that takes as input labeled points $D^m = \{(x_1, y_1, z_1), (x_2, y_2, z_2), \ldots, (x_m, y_m, z_m)\} \subseteq \mathbb{R}^3$ and returns $L(D^m) = h \in H$.
3. (7 pts) Compute a sufficiency bound on the sample complexity of learning C from H using the algorithm L suggested above. That is – given (ε,δ), directly calculate a 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:
$$Prob\big(Err\big(c, L(D^m)\big) > \ \varepsilon\big) < \delta$$
4. (7 pts) For $\varepsilon = 0.05$ and $\delta = 0.01$ compute two sufficiency bounds, one using the above calculation and one using the VC dimension.
Recall that $m \geq \frac{1}{\varepsilon}\left(4\, log_2\left(\frac{2}{\delta}\right) + 8VC(H)\, log_2\left(\frac{13}{\varepsilon}\right)\right)$ is a sample complexity bound based on the VC dimension of $C = H$.
Which bound is tighter?

Question 4 (25 points) – Bayes & MLE

1. (5 pts) Consider classes $(A_1, A_2, \dots, A_r)$ consisting of elements with properties $(x_1, x_2, \dots, x_n)$. Write the MAP classification formula for:
   a. Naïve Bayes
   b. Full Bayes

In rolling two dice with 6 sides each, we have the following distributions of results for two casino houses, Casino Golden Peacock (CGP) and Casino Silver Turkey (CST):

Casino Silver Turkey

Die 1 / Die 2	1	2	3	4	5	6
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
2	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
3	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
4	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
5	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$

Casino Golden Peacock

Die 1 / Die 2	1	2	3	4	5	6
1	$\frac{1}{12}$	$\frac{1}{12}$	0	0	0	0
2	0	$\frac{1}{12}$	$\frac{1}{12}$	0	0	0
3	0	0	$\frac{1}{12}$	$\frac{1}{12}$	0	0
4	0	0	0	$\frac{1}{12}$	$\frac{1}{12}$	0
5	0	0	0	0	$\frac{1}{12}$	$\frac{1}{12}$
6	$\frac{1}{12}$	0	0	0	0	$\frac{1}{12}$

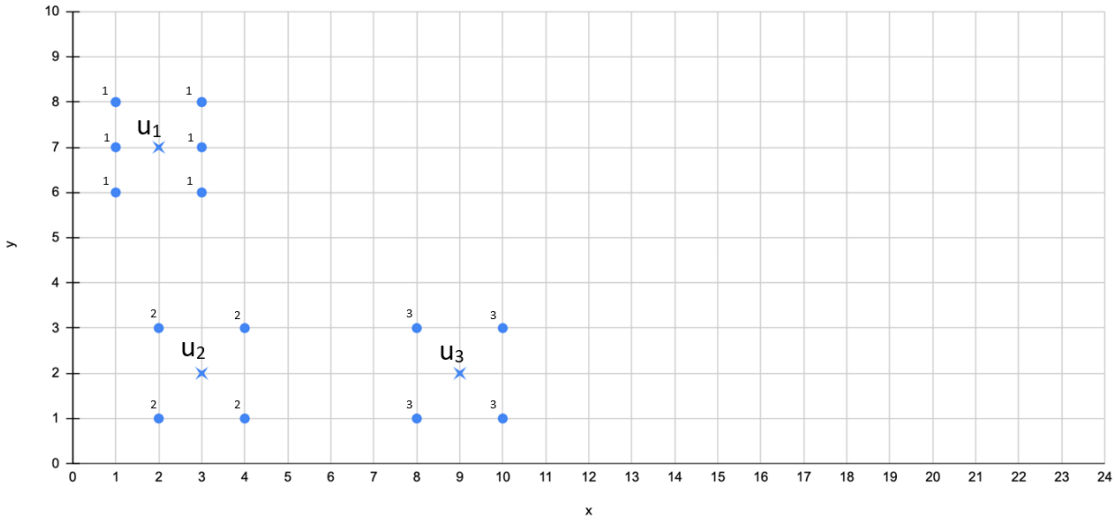
We are given actual dice rolling results and would like to infer which casino they came from. The prior probability for playing in Casino Golden Peacock is 0.6.

2. (5 pts) We observe the following game outcome: 1st die is 1, 2nd die is 6. Which casino will a Full Bayes classifier predict? Show your calculations.
3. (5 pts) Given the same game result as in Part 2, which casino will a Naïve Bayes classifier predict? Show your calculations.
4. (5 pts) What is the minimal prior we need to assign to Casino Silver Turkey for the Full Bayes classifier to predict Silver Turkey regardless of the game outcome? Show/explain your calculations.
5. (5 pts) You can now change two entries in the joint distribution matrix of Casino Golden Peacock.
   Given the same results as in Part 2 (that is: (1,6)), perform a change that will lead the Full Bayes classifier to select Golden Peacock under the prior you had found in Part 4.
   Your newly defined distribution should be an adequate probability distribution.

Question 5 (25 points) – k-Means

Assume the use of Euclidian distance in all parts of this question.

1. Consider the following dataset with 14 data points in $\mathbb{R}^2$, and that this is the current stage of running k-Means with $k = 3$:

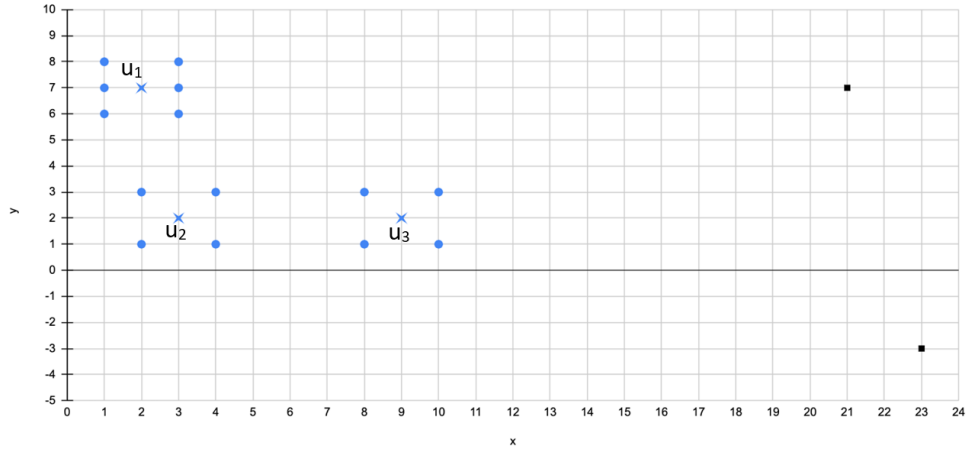


The numbers near each point represent the cluster index to which the point is currently assigned and the X symbols represent the centroids u_1, u_2, u_3.

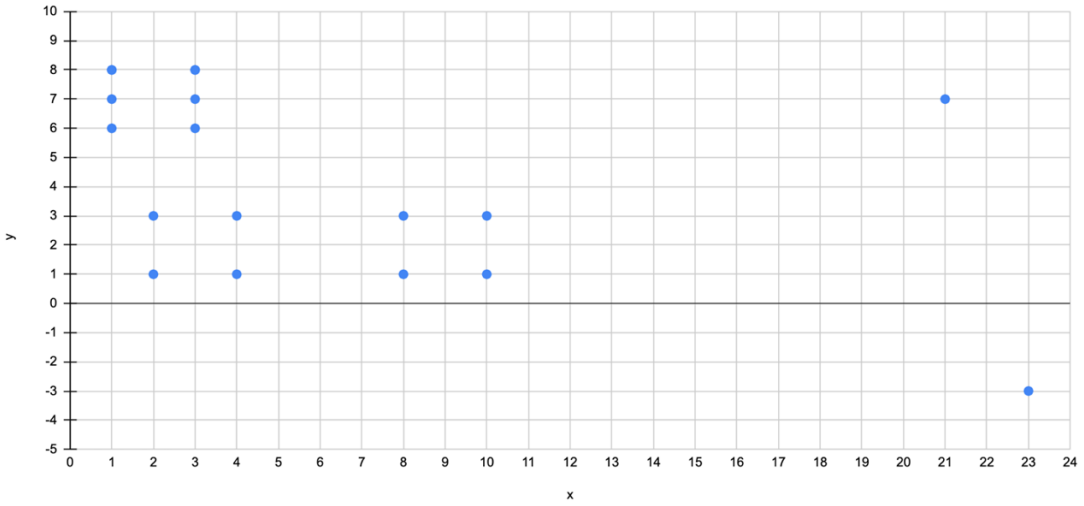
a) (3 pts) State the formula for the loss function of the k-Means algorithm.
b) (3 pts) Is this a stable state for the algorithm? That is – will running another iteration lead to a new assignment? Explain your answer.

After running the algorithm, you found out that there were 2 data points that were accidently omitted from the dataset given to you.
Instead of re-running the algorithm, you decided to add these points and do one more iteration of k-Means, starting with assignments. Running the k-Means single step resulted in a new assignment and in new centroids.

c) (2 pts) Which clusters would the new points be assigned to?
In the plot in the next page, write next to each new point the cluster it would be assigned to.



d) (8 pts) In the plot below, indicate the new cluster centroids using X marks and the new assignment to clusters as a number next to each data point.



2. The loss function for k-Means, which you stated above is also called the *inertia*.
You performed k-Means on data with 16 distinct points and with k = range(16).
As output, you recorded the respective values of the inertia for the cluster structure obtained for each k.
    a. (2 pts) Which column (A-F) of the table below potentially represents the resulting output? (Assume that for each k the algorithm has converged to the global optimum).

k	A Inertia	B Inertia	C Inertia	D Inertia	E Inertia	F Inertia
1	847	847	847	847	847	847
2	290	535	535	290	535	290
3	140	377	377	140	377	140
4	78	180	180	78	180	78
5	26	110	110	26	110	26
6	20	75	75	20	75	20
7	16	20	20	16	20	16
8	12	16	16	12	16	16
9	10	10	10	10	10	10
10	8	8	8	8	8	8
11	6	6	6	6	6	10
12	4	4	4	4	4	4
13	2.5	2.5	2.5	2.5	2.5	2.5
14	1	-1	0.5	1	1	1
15	0.5	-0.1	0	0.5	0.5	0.5
16	0	0	0	0.1	0	0

b. (5 pts) For each of the other columns indicate why it cannot be the output of the process.
c. (2 pts) According to the “knee/elbow” method learnt in class to find the optimal k, which k would you chose to work with?

GOOD LUCK!