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

סמסטר ב' תשע"ט
Spring 2019

מבחן מועד א בלמידה ממוכנת
Machine Learning Exam A

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!

You can use calculators and the formulae sheets provided
Justify all your answers and show your calculations
All answers should be written in exam notebooks

Question 1 (5 parts)

A. In linear regression, why is it common to perform feature normalization prior to the learning phase? Is it common to use feature normalization in the closed-form solution to linear regression ($\vec{\theta} = \text{pinv}\, X \cdot \vec{y}$)? Explain.

B. As part of your job as a machine learning engineer, you are developing a linear regression model using some data your team obtained in laboratory conditions. All of your data (x) comes from a limited range, namely it satisfies $x_i \in [0,10]$. After training the model and getting high accuracy on the training dataset, you start testing your model in the field. You soon realize that the x-values of most of the data points your model encounters in the field are in the range $[100,200]$.
   1. Assume that in reality the function you are modeling is linear. What performance do you expect to see in your field testing? Explain your answer.
   2. One of your colleagues suggests that you use kNN with k=5 for your field predictions. Is this a good suggestion? Explain.

C. A real estate company is interested in predicting house prices based on 10 measurable features. Furthermore, for some houses it is more important that the prediction be accurate than it is for others. How do you change the linear regression framework to address this task?

D. You are performing linear regression on training data (x_i, y_i). What do you expect the MSE to be under the following conditions:
   1. The Pearson correlation of $\vec{x}$ and $\vec{y}$ is 1.
   2. The Spearman correlation of $\vec{x}$ and $\vec{y}$ is 1.

Question 2 (5 parts)

A. Consider classes $(A_1, A_{,2} \dots, A_i)$ consisting of elements with properties $(x_1, x_2 \dots, x_j)$. Write the classification formula for:
   1. Naïve Bayes
   2. Full Bayes

In rolling two dies with 6 sides each we have the following distributions of results for two casino houses – Casino A and Casino B:

Casino A

Dice1 \ Dice2	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 B

Dice1 \ Dice2	1	2	3	4	5	6
1	$\frac{1}{18}$	$\frac{1}{18}$	0	0	0	$\frac{1}{18}$
2	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	0	0	0
3	0	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	0	0
4	0	0	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	0
5	0	0	0	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$
6	$\frac{1}{18}$	0	0	0	$\frac{1}{18}$	$\frac{1}{18}$

The prior probability for playing in Casino A is $\frac{3}{5}$.
Given a game outcome (2 numbers), we want to classify whether the game was played in Casino A or B.

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

Question 3 (4 parts)

A. Let X be a Poisson random variable.

   Recall that then $(\mathrm{X} = k) = e^{-\lambda}\frac{\lambda^k}{k!}$.

   Assume that for a series of n independent samples of X we observed the values $x_1, x_2, \dots, x_n$.

   Write down the expression of the probability of the observed data as a function of the parameter λ.

B. Prove that the value of λ given by MLE is:

$$\lambda = \frac{1}{n}\sum_{i=1}^{n} x_i$$

C. We are given the following observed data which was sampled from two different Poisson distributions with parameters λ_1, λ_2.

2	1	1
5	11	5
2	0	2
2	1	2
6	4	8

   For each row, a coin was tossed. With probability w_1 it led to 3 independent Poisson samples with λ_1 and with probability w_2 it led to 3 independent Poisson samples with λ_2.

   Which algorithm would you use to assess the values of the probability parameters for the given scenario? Explain your answer.

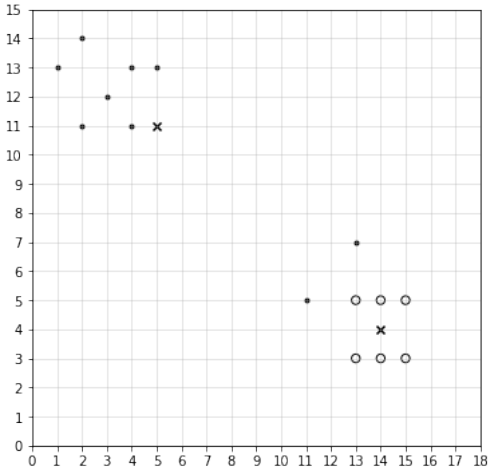
D. Use the algorithm from C to compute the first iteration of the algorithm with the following initial values:

$$\lambda_1 = 1, \qquad \lambda_2 = 5, \qquad w_1 = 0.75$$

Question 4 (5 parts)

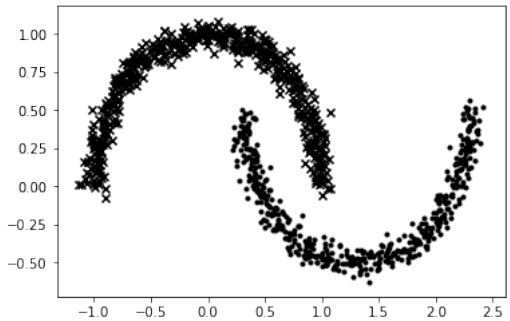
A. What is the function J that the standard k-means algorithm seeks to minimize? Provide a formula.

B. Consider the following status of k-means when applied to data in $\mathbb{R}^2$. The 'dots' belong to Cluster 1 and the 'circles' belong to Cluster 2. The current cluster centers are given by the 'x's.
Compute the exact change in J after updating the assignments (before updating the centers)?



C. Can the function J increase its value during the k-means algorithm? Explain.

D. Consider the following cluster structure. The 'xs' (the upper half circle) belong to Cluster 1 and the 'dots' (the lower half circle) belong to Cluster 2 (see picture).
Can it be a minimum point for the function J?



E. Suggest a clustering algorithm that can produce the above structure as its output, from the given data.

Question 5 (4 parts)

A. Given data in $\mathbb{R}^{10}$ which is not linearly separable, a student suggests to map the data to a full rational variety of degree 10 and then try to find a linear classifier in the higher dimensional space.
What is the problem in this approach and how can you overcome it?

B. Given the following dataset:

X1	X2	Y
+1	0	+1
-1	0	+1
0	+2	+1
0	+1	-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. Let $X = \mathbb{R}^2$. Consider the mapping $\varphi: \mathbb{R}^2 \to \mathbb{R}^3$ given by $\varphi(x_1, x_2) = \left(\sqrt{2}x_1x_2, x_1^2, x_2^2\right)$.

1. Find a kernel for this mapping.
2. Consider the hypotheses space

$$H = \left\{ h: \mathbb{R}^2 \to \{-1, +1\} \;\middle|\; \begin{array}{c} \text{the two sets} \\ \{\varphi(x) \mid h(x) = -1\} \\ \text{and} \\ \{\varphi(x) \mid h(x) = +1\} \\ \text{are linearly separable in } \mathbb{R}^3 \end{array} \right\}$$

What is the VC dimension of H? Justify your answer.

Question 6 (3 parts)

A. Assume that an RBF kernel was selected in applying SVM to a classification task in $\mathbb{R}^2$. That is – a kernel of the form:

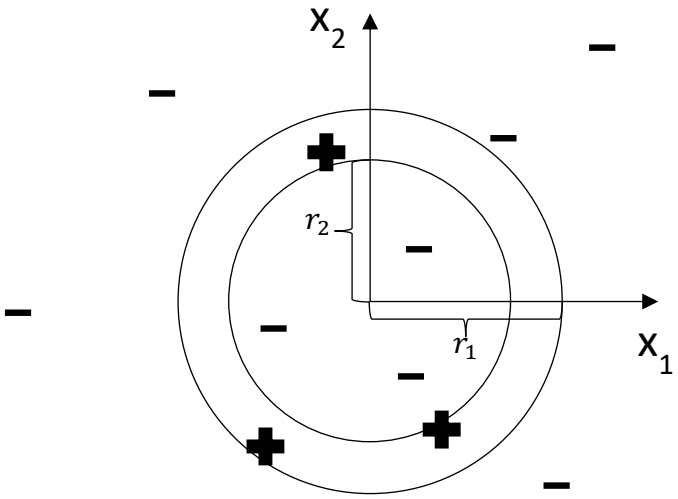
$$K(x,y) = \varphi(x) \cdot \varphi(y) = exp\left(-\frac{1}{2}\|x-y\|^2\right)$$

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

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

B. TRUE or FALSE
After mapping into higher dimensional space using the RBF it is possible that a kNN classification algorithm, based on unweighted Euclidean distance, achieves better classification results than it achieved in the original space.
Hint: use the identity proven in D.

C. Let $X = \mathbb{R}^2$. Let $C = H$ the set of all concentric rings. For a concept $c \in C$ let r_1 and r_2 be the radii of the concept ring where $r_1 \geq r_2$ (see picture). Each instance in the sampled training data is drawn from an unknown distribution π and consists of the position of the instance (x_1, x_2) and the target value ($+1$ if it's inside the ring and -1 otherwise).
Describe a polynomial sample complexity algorithm L that learns C using H. State the time complexity and the sample complexity of your suggested algorithm. Prove all your steps.



GOOD LUCK!