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 ב בלמידה ממוכנת
Machine Learning Exam B

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 (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 Bernoulli distribution with parameters p_1, p_2.

0 0 0 0 1
1 1 1 1 1
0 0 1 1 0
1 1 1 0 1
0 0 0 0 0

   For each row, a coin was tossed. With probability w_1 it led to 5 independent Bernoulli samples with p_1 and with probability w_2 it led to 5 independent Bernoulli samples with p_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:

$$p_1 = 0.5, \qquad p_2 = 1, \qquad w_1 = 0.25$$

Question 2 (5 parts)

A. Explain the difference between Naïve Bayes and Full Bayes.

In a dice game, you roll 2 dice. Where each die has 6 faces.
At Casino A they use fair dice, so on each roll, every face has the same probability.
The probability for each pair of numbers in 2 rolls is 1/36.
At the B casino, the first die is fair, where each face has the same probability, but the second die is skewed so that with probability $\frac{1}{3}$ it will land on the same face as the first die and with probability $\frac{1}{3}$ it will land on each of ± 1 from the result of the first die.
For example, if the first die roll is 3 then the second die, in Casino B, will either be 2,3 or 4, each with probability $\frac{1}{3}$. And if the first die roll is 6 then the second die will either be 5,6, or 1, again, each with probability $\frac{1}{3}$.
The prior probability for playing in Casino A is $\frac{3}{5}$ and for playing Casino B it's $\frac{2}{5}$.
Given a game outcome (2 numbers), we want to classify whether the game was played in Casino A or B.

B. Given the following game outcome: 1st die is 4, 2nd die is 5. Which casino will a Naïve Bayes classifier predict? Explain your answer.
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? And for the Naïve Bayes? Show/explain your calculations.
E. Given the results from 2 pairs of rolls (2 pairs of 2 numbers) played in the same casino, what is the minimal prior we need to assign to Casino A in order for the Full Bayes classifier to predict it regardless of the results? Show your calculations.

Question 3 (4 parts)

We want to cluster to k groups a set S of instances. 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 ($1 \leq i \leq k$) 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 $c_1, \ldots, c_k$

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

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 you start at $c_1=p_1$ and $c_2=p_5$.
   (it is recommended to first plot the instances on a 2D Euclidean plane).
   In every step indicate what the centers are and how instances are assigned. You don't have to show all intermediate calculations.
B. Explain, using a formula, the function, J_S , that k-means seeks to minimize.
C. What is the main difference between k-means and k-medoids? How would you change the function from B so that it would fit the k-medoids algorithm? Call this new target function J_{med}.
D. Assume that we execute both algorithms on the same set and with the same k. Do we expect the value of J_{med} obtained by k-medoid to be smaller, greater or the same as the value of J_S that k-means has obtained? Explain why!

Question 4 (4 parts)

A. State the gain formula for computing the best split of a node in a decision tree. Assume that the impurity function is a generic φ.
B. Prove that when the attributes as well as the class labeling are all binary then GiniGain is always non-negative ($0 \geq$).
C. 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.
D. Consider the training data described below, at a node S. Decide on which attribute to use according to the rules defined in C and using GiniGain. State your calculations. There is no need to get to a final answer.

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

Question 5 (5 parts)

A. Find the minimum and the maximum of $3x + 2y$ under the constraint $4x^2 + y^2 = 1$

B. Given the following dataset:

X1	X2	Y
+1	+1	+1
-1	+1	+1
0	-1	+1
0	0	-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. Find $K: \mathbb{R}^2 \longrightarrow \mathbb{R}$ which is a kernel function for

$$\varphi(x) = \left({x_1}^3, \sqrt{3}{x_1}^2 x_2, \sqrt{3} x_1 {x_2}^2, {x_2}^3\right)$$

E. Show that the function $K(x, y) = e^{xy}$ for $x, y \in \mathbb{R}$ is a kernel for some mapping into infinite dimensional space. (hint: $e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!}$ for $z \in \mathbb{R}$)

Question 6 (4 parts)

A. Consider two hypothesis spaces H and H' such that $H \subseteq H'$. Prove that:

$$VC(H) \leq VC(H')$$

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 $X = [-1,1] \times [-1,1]$.
Let $N \in \mathbb{N}$ and $N \geq 2$, we define the sets $\mathrm{A1} := \{\frac{1}{N}, \frac{2}{N}, \dots, 1\}$, $\mathrm{A}_2 = \{\frac{1}{2N}, \frac{2}{2N}, \dots, 1\}$ and the following hypothesis spaces:

- $H_1 = \{h: h(x_1, x_2) = +1 \Leftrightarrow {x_1}^2 + {x_2}^2 \leq r^2,\ r \in \mathrm{A}_1\}$
- $H_2 = \{h: h(x_1, x_2) = +1 \Leftrightarrow {x_1}^2 + {x_2}^2 \leq r^2,\ r \in \mathrm{A}_2\}$
- $H_3 = \{h: h(x_1, x_2) = +1 \Leftrightarrow {x_1}^2 + {x_2}^2 \leq r^2,\ r \in [0,1]\}$

You can view each hypothesis as a circle centered at the origin. (i.e. an instance is classified as positive iff it's inside the circle with radius r)

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_1 using H_2.
2. When trying to learn a concept in H_3 using H_3.
3. When trying to learn a concept in H_3 using H_2.

GOOD LUCK!