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

סמסטר ב׳ תשע"ז
Spring 2017

מבחן מועד ב בלמידה ממוכנת
Machine Learning Exam B

Lecturer : Prof Ariel Shamir,
Dr. Zohar Yakhini
Time limit : 3 hours
Additional material or calculators are not allowed in use!

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

מרצים: פרופ אריאל שמיר,
ד"ר 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 (7 sections)

a. Explain what is the advantage and disadvantage of using instance-based learning methods such as K-nearest neighbors (KNN).
b. Explain what is the decision rule of KNN in classification.
c. Explain what is the decision rule of KNN in regression.
d. Assume your training set includes instances in Euclidean space. Define a method to accelerate the search for the nearest neighbor of an instance – explain how to find the neighbor in this method!
e. Define a method to choose the parameters K for learning using KNN algorithm. Explain in detail all stages and how the process of choosing is performed.
f. While learning KNN algorithm for a given K, we use a procedure that examines instances in the training set and removes those that are classified correctly (using the training set without the examined instance). For which type of training sets is this procedure useful? Explain your answer.
g. While learning a KNN algorithm for a given K, we use a procedure that examines instances in the training set and removes those that are **not** classified correctly (using the training set without the examined instance). For which type of training sets is this procedure useful? Explain your answer.

Question 2 (5 sections)

Given the three bounds on the number of samples in learning problems we have seen in class:

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

a. Explain why there is a need to have a bound on the number of samples in learning?
b. Explain what are δ, ε and why is the number of samples inversely monotone in these quantities?

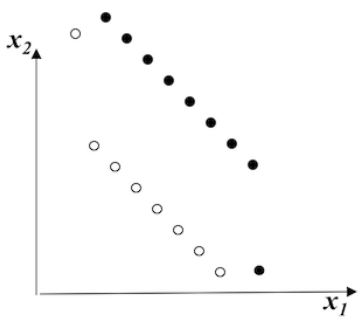
We are interested to know how many training samples we need in the following hypotheses spaces to assure with at least 90% certainty that the error is not more than 0.1. In each of the following sections choose the appropriate formula from the three formulas given above, substitute the appropriate numbers and explain why you chose this formula (show only the calculation, there is no need to reach a final result!)

c. Assume that the instance space X has five binary features $\{X_1, X_2, X_3, X_4, X_5\}$. The real concept we are looking for are all samples that are classified by the following rule $\mathrm{Y} = (X_1 \wedge X_4) \vee (X_2 \wedge X_3)$. The hypothesis space is a simple decision tree (checks one feature in each node) of depth 2 (contains at most 4 leaves).
d. Assume that the instance space X has five binary features $\{X_1, X_2, X_3, X_4, X_5\}$. The real concept we are looking for are all samples that are classified by the following rule $\mathrm{Y} = (X_1 \wedge X_4) \vee (\neg X_1 \wedge X_3$. The hypothesis space is a simple decision tree (checks one feature in each node) of depth 2 (contains at most 4 leaves).
e. Assume that the instance space X is 2D continuous space. The hypothesis space includes axis aligned rectangles where the inside could be either positive or negative (this means each rectangle can represent two different hypotheses – one where the inside is positive and one where the inside is negative).

Question 3 (7 sections)

Consider a training set D of m instances with n numerical features. Assume that each instance is $x_d \in \mathbb{R}^n$

a. Explain what is a discriminant function and how it helps classification problems?
b. Explain (including formula) what is the discriminant function in the Perceptron algorithm and what does it represent geometrically?
c. Explain (including formula) what is the learning rule of the Perceptron algorithm and why does it work?
d. Explain what serious problem is there in using this rule?
e. Explain what is the <u>target</u> function which is minimized in Least Means Square (LMS) learning algorithm.
f. Explain (including formula) what is the learning rule of the LMS algorithm and why does it work?
g. Given the following set of examples as training set (the black examples are positive while the white are negative). Explain what would be the decision boundary of the Perceptron algorithm and what would be its error? Explain what would be the decision boundary of the LMS algorithm and what would be its error? Explain why (you can use a drawing to explain)



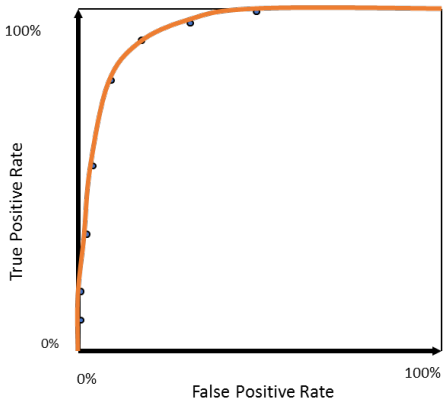
Question 4 (5 sections)

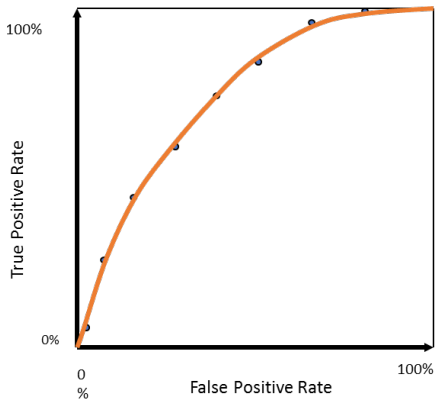
a. Linear Regression
    1. What is the output of the linear regression algorithm at the end of the learning phase?
    2. How is the predicted value of a new instance determined after the end of the learning phase (give a formula)?
    3. Define the loss function that is minimized during learning in the linear regression algorithm (give a formula).
    4. Define what is Gradient Descent and how it is used in linear regression learning algorithm (give a formula)?
b. Explain what are Bias and Variance in the context of machine learning, and what is the Bias-Variance dilemma?
c. Explain how you can affect the Bias and the Variance in a regression problem in learning.
d. Given a training set with one feature that has **Pearson** Correlation value of 1 to the target value, what would be the training error of linear regression? Explain.
e. Can your answer to Section d change when the **Spearman** correlation between the training set feature and the target function is 1? Explain.

Question 5 (5 sections)

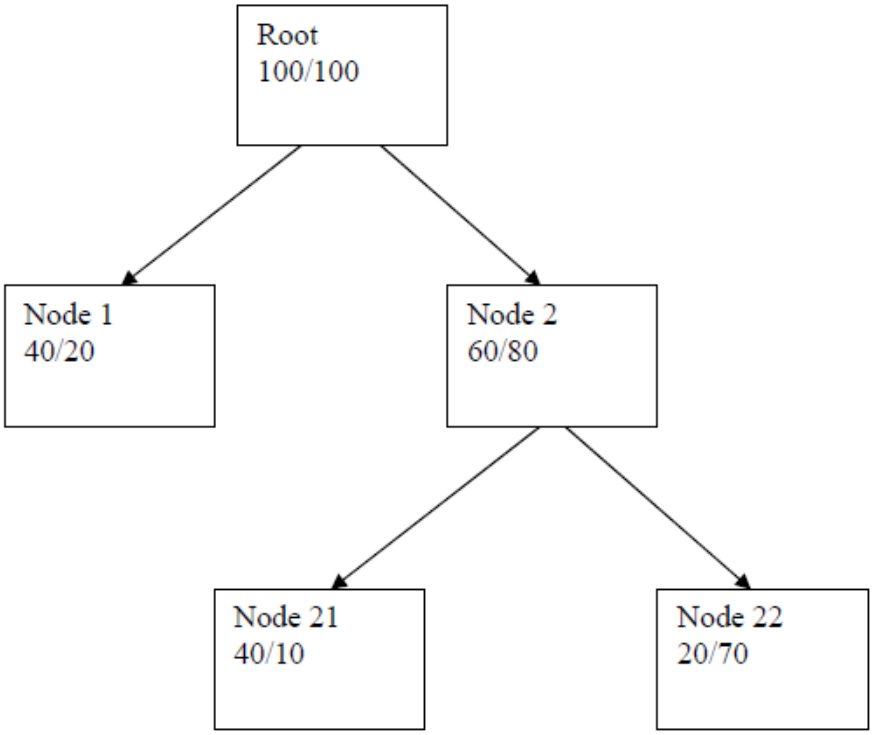
a.

1. Define the confusion matrix of a binary classifier.
2. The following are two ROC curves describing the performance of two classifiers. Which is a better classifier, according to AUC?





Consider the following classification tree. For each node the number of positive (POS) and negative (NEG) training records it contains is indicated, represented as POS/NEG. The training set contains 100 POS samples and 100 NEG samples.



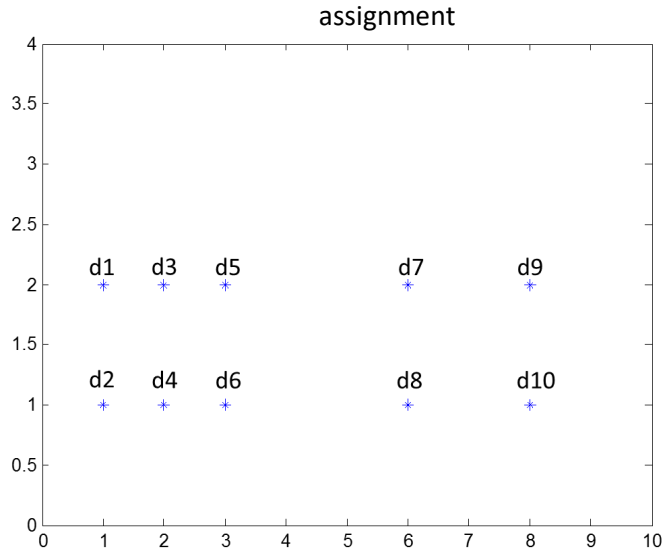
For every fixed $0 < p < 1$ consider classifying records in the leaves of the tree as POS if the proportion of POS records in the leaf node is strictly larger than p and NEG otherwise.
For example, in Node 21 if p > 0.8 the classification will be NEG. If p=0.7 then the classification there will be POS.
As another example, p=0.5 represents a majority vote.

b. For each value of $p \in \Pi = \{0.1, 0.25, 0.75, 0.9\}$ calculate the confusion matrix for this tree.
c. Plot the ROC curve for this classifier, using the points in Π.
d. Suppose that the cost of a FP is C and that the cost of a FN is 0.4 C. Amongst the values in Π, what value of p minimizes the cost of using this classifier as estimated on the training data?
e. Propose two (possibly different) ways of splitting Node 1 into two leaves, Node 11 and Node 12, so that:
    1. Your first suggested split should be such that nothing is changed in the results of (b)
    2. Your second suggested split should yield, at p=0.1, an FPR of 0.9 and a TPR of 1.

Question 6 (6 sections)

a. Explain the difference between learning a classifier on a set of data instances and clustering a set of data instances.
b. What is the k-means optimization criterion? (state as a mathematical formula)
c.
    1. Given 100 data instances, all different from each other, can running a k-means algorithm, with k=5, result in a global minimum wherein one of the clusters is empty (all points are assigned to only 4 of the 5 centroids)?
    2. Can the assumptions in the above question be changed so that the answer is also changed?
d. In the data depicted below – what is an assignment of two centroids for 2-means that will lead to the global optimum? Explain your answer (no need for a formal proof).

    For the selected centroids, compute the value of the target function (from (ii)) to which the algorithm will converge.



e. What is an assignment of initial centroids that will lead to a local minimum that is not a global optimum? Prove your answer.
f. As demonstrated above, one of the shortcomings of k-means is convergence to a local minimum, which isn't a global one. Propose an approach to increasing the likelihood of obtaining a global optimum solution to a k-means problem.

Good luck!