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

סמסטר ב' תש"פ
Spring 2020

מבחן מועד ב בלמידה ממוכנת
Machine Learning Exam B

מרצה: פרופ זהר יכיני
משך המבחן: 2 שעות

יש לענות על שאלה מספר 1 (חובה), 40 נקודות, ולבחור שתים מתוך שלוש השאלות הנוספות (30 נקודות לכל אחת, סה"כ 60 נקודות).

בעמוד הראשון יש לרשום את מספרי שתי שאלות הבחירה שעליהן בחרת לענות. אם לא יהיה רשום תיבדקנה שתי התשובות הראשונות.

בהצלחה!

Lecturer: Prof Zohar Yakhini
Time limit: 2 hours

You should answer Qn 1 (mandatory), for 40 pts and two out of the other three questions (30pts each, 60pts together).

In the first page indicate the numbers of the two questions you answered. If there is no indication then the first two solutions will be graded.

Good Luck!

You can use all Moodle course materials as well as student personal notes prepared before the exam. There is no need to print the material. You can use the appropriate files on your PC.
You can use calculators.
Justify all your answers and show your calculations.
All answers should be legibly written and scanned for submission as per IDC's instructions.

ניתן להשתמש בכל החומר הקיים ב Moodle ובנוסף סיכומי סטודנטים שנכתבו לפני תחילת המבחן. אין צורך להדפיס את החומר. אפשר להשתמש במחשב כדי לגשת לקבצים הרלוונטיים.
ניתן להשתמש במחשבון.
יש להצדיק את כל תשובותיך ולהראות את צעדי החישוב.
כל התשובות צריכות להיות כתובות בכתב ברור וקריא ולהיסרק להגשה ע"פ הנחיות IDC.

Question 1 (3 parts, 40 points) – MANDATORY QUESTION

A. You receive 10,000 samples of labeled data with two classes, split into training and test, to perform classification from 5 real valued features $(x_1, x_2, x_3, x_4, x_5)$.
   1. (4 points) If you preform Logistic Regression with all 5 features, what is the dimension of the inferred vector of coefficients $\vec{\theta}$?
   2. (5 points) The classification results from the previous section were not satisfactory. Your colleague suggests to map the data using the full rational variety of degree 10 and then apply Logistic Regression to the resulting data.
      What is the problem with this approach?
   3. (5 points) Suggest an algorithmic approach, to finding a classifier in your data, that would overcome the problem of the suggestion from the previous section. Provide full details of your proposed approach.

B. Assume that you perform learning for classification on 5,000 samples with two classes A and B. Class A is the positive class. Your training / test split was 4,000 / 1,000. The prior of class A in your data is 0.9. For the test data you observed:

$$\text{Eq. 1} \qquad \frac{TP + TN}{TP + FP + TN + FN} = 0.905$$

   1. (3 points) Provide an example of a confusion matrix for the test data that is consistent with the above information.
   2. (3 points) Is Eq.1 necessarily an indication of good performance? Explain.

C. (10 points) Let $X = \mathbb{R}^2$. Consider H to be a hypotheses space that consists of hypotheses that assign a + (positive) to the inner part at most 5 circles.
Prove that $VC(H) \geq 15$.

Formally:
For any 5 triplets $C = \{(a_i, b_i, r_i)\}_{i=1}^{5}$ where $a_i, b_i \in \mathbb{R}, r_i \in \mathbb{R}_+$
we define

$$h(C) = \bigcup_{i=1}^{5} \{(x, y) \mid (x - a_i)^2 + (y - b_i)^2 \leq r_i^2 \}$$

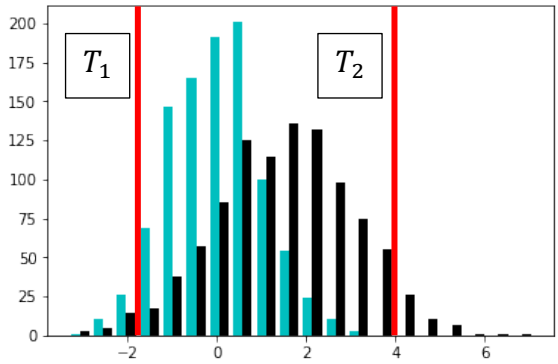
And now we define the hypotheses space as:

$$H = \{h(C) \mid \ a_i, b_i \in \mathbb{R}, r_i \in \mathbb{R}_+\}$$

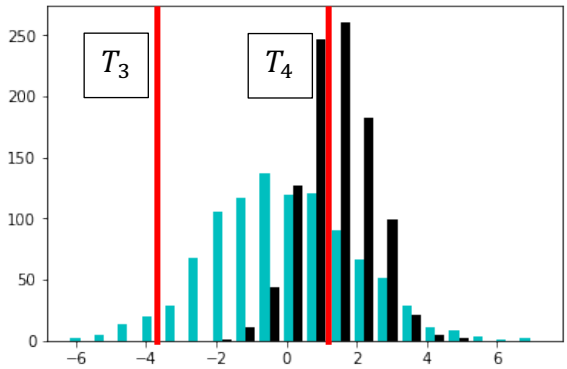
D. Two companies are proposing a Corona test to a hospital. The hospital asks for your advice as to which test is better. The plots below represent the distributions of the tested quantity for the two companies. The black histograms represent the positive cases. We further depict the ROC curves computed to evaluate their performance.

Distributions

A

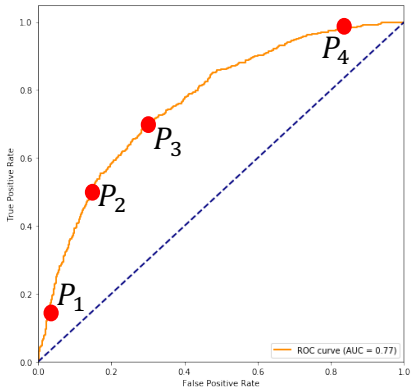


B

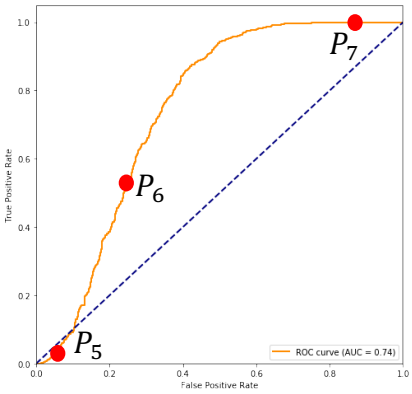


ROC curves

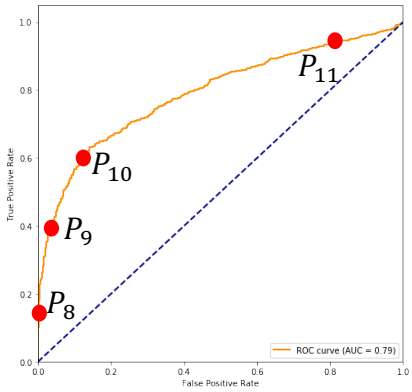
X



Y



Z



1. (2 points) Match two of the three ROC curves above, X, Y and Z, with the two distributions, A and B. Explain.
2. (4 points) Match the indicated thresholds T_1, T_2, T_3, T_4 to the appropriate points from the set $P_1, P_2, P_3, P_4, P_5, P_6, P_7, P_8, P_9, P_{10}, P_{11}$. Explain.

Recall that the expected benefit of a test, for the hospital, is measured by

$$\pi = \alpha * TPR - FPR$$

3. (4 points) How would you find a value of α for which both companies can be equivalent in terms of the expected benefit? Explain your answer.

Question 1 (3 parts, 40 points) – MANDATORY QUESTION - solution

A.

1. The dimension of the inferred vector of coefficients $\vec{\theta}$ will be 6:
$$\vec{\theta} = (\theta_0, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$$
2. Mapping the data using full rational variety of degree 10 and then working in the new space is not practical due to very high time complexity of the mapping and the inner product.
3. One possible algorithmic approach is to use kernel algorithms such as Kernel Perceptron and SVM. We can use these algorithms with the kernel $K = (x \cdot y)^{10}$.

B.

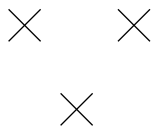
1.

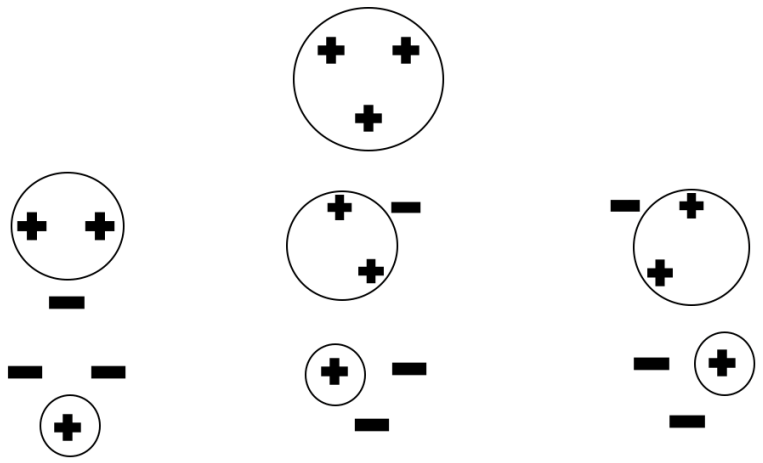
	Predicted Positive	Predicted Negative
Actual Positive	850	50
Actual Negative	45	55

There are many more solutions to this question. All must have Actual Positive of 900 and TP+TN=905.

2. No. The strawman model that predicts all to be positive, will have almost the same accuracy, namely 0.9. But obviously, this is a very naïve model that adds no substantial information. In unbalanced data as described in this question we should use different performance evaluation metric such as ROC curve.

C. In order to prove that $VC(H) \geq 15$ all we need to do is to prove that $VC(one\ circle) \geq 3$. Then we can take 5 separated sets of 3 points and argue that each of the five circles constituting $h \in H$ can separate all dichotomies in each set and therefore, h can separate all dichotomies in the 15 points.
We now show that a single circle shatters the following 3 points in $\mathbb{R}^2$:





Q.E.D

D.

1. A – Z, thresholds in the right side of the distribution lead to TP>0 with FP=0.
   B – Y, thresholds in the left side of the distribution lead to TP=1 with FP<1.
   X is a symmetric ROC curve and therefore doesn't match to any of the two distributions.
2. T1 – P11: High TPR and high FPR.
   T2 – P8: TP>0 with FP=0.
   T3 – P7: TPR=1
   T4 – P6: TPR~0.5 and FPR<0.5.
3. Given that both graphs are convex (minus the small part in the left corner of graph Y), there is only one line that is simultaneously tangential to both graphs. We are seeking α so that:

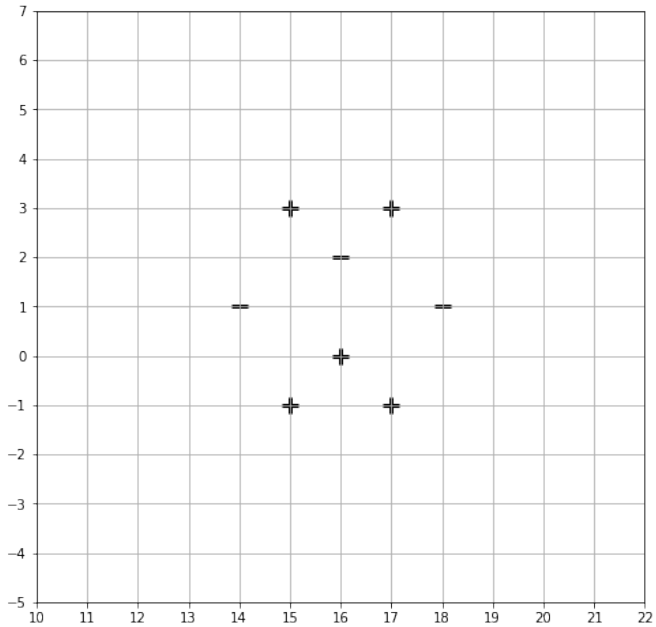
$$\max_{thresholds\ \tau_a}(\alpha * TPR - FPR) = \max_{thresholds\ \tau_B}(\alpha * TPR - FPR)$$

   Graphicly we see that this happens at around $\alpha = 1$.

Question 2 (3 parts 30 points)

A. Consider the kNN classification algorithm. Specifically assume we are using the L_1 (Manhattan) distance metric. Furthermore, assume that we are breaking ties as follows: given k/2 positive instances and k/2 negative instances in the set of k nearest neighbors we will classify the instance according to the k+1st nearest-neighbor.

Consider the following training dataset:



Where:
The + data points are training points belonging to the positive class.
The – data points are training points belonging to the negative class.

1. (5 points) What is the leave-one-out cross validation error of 3-NN in the above training dataset?

Now, further consider the following TEST data:

index	x_1	x_2
1	13	-1
2	19	-1
3	12	3
4	16	4
5	20	3

2. (5 points) What will be the prediction for each one of the TEST instances using 5-NN (5 nearest neighbors)? Explain
3. (5 points) What is maximum value of k for which not all new TEST instances are assigned to the same class?

B.

1. (5 points) Consider the following training set in $\mathbb{R}^2$:

x_1	x_2	y
8	3	+1
6	6	-1

You want to classify a new instance $x = (0,0)$ using 1-NN, first with Euclidean distance and then with Manhattan distance (L_1).
What will the prediction be in each one of the cases?

2. (5 points) Can you find a training set for which:
When using 1NN with Euclidean distance the prediction on $x = (1, -1)$ will be +1 and when using 1NN with L_∞ distance the prediction on that same point will be -1?

C. (5 points) Consider two matrices $M \in \mathbb{R}^{m \times k}$ and $W \in \mathbb{R}^{k \times k}$ and a vector $y \in \mathbb{R}^m$.

Provide a closed form solution for:

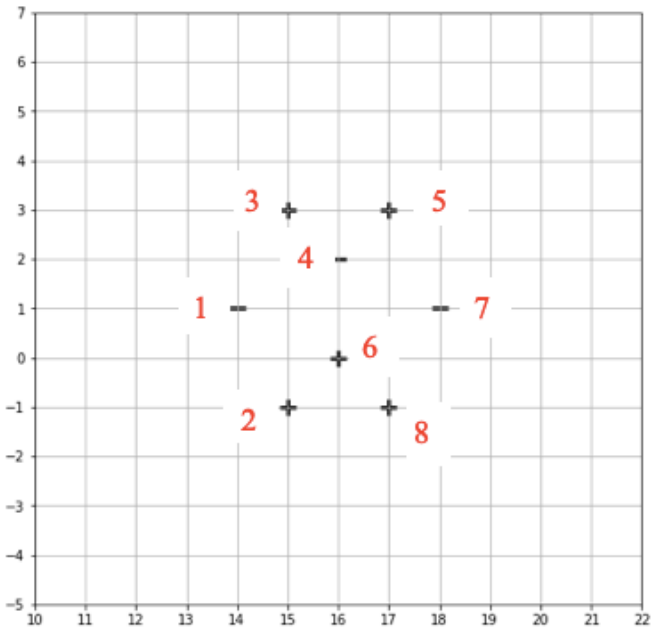
$$\theta^* = \underset{\theta \in \mathbb{R}^k}{\operatorname{argmin}} \sum_{i=1}^{m} \left((M_i W, \theta) - y_i \right)^2$$

Where:
M_i is the i-th row of M.
$M_i W$ is the vector M_i multiplied by the matrix W.
(u, v) represents the standard inner product in Euclidian space.

Question 2 (3 parts 30 points) – solution

A.

1. Using leave-one-out cross validation (LOOCV) means splitting the data into 8 folds (the number of data points in the dataset).
   Let f be the function that maps each data point $\bar{x} = (x_1, x_2)$ to its correct class. For example, $f(15, -1) = +1$.
   Let $\hat{f}$ be the 3-NN (Manhattan distance) function that maps each data point to the majority class of its 3-NNs, according to the Manhattan distance.
   The leave-one-out cross validation error of 3-NN is:
   $$\frac{1}{8}\sum_{i=1}^{8} \mathbb{1}_{f(\bar{x}_i)\neq\hat{f}(\bar{x}_i)}(\bar{x}_i)$$
   $\mathbb{1}_{f(\bar{x}_i)\neq\hat{f}(\bar{x}_i)}$ is an indicator function.
   Let us consider the data points:



   Consider the distance matrix below, where the i, j th index represents the Manhattan distance between points i, j

0	3	3	3	5	3	4	5
3	0	4	4	6	2	5	2
3	4	0	2	2	4	5	6
3	4	2	0	2	2	3	4
5	6	2	2	0	4	3	4
3	2	4	2	4	0	3	2
4	5	5	3	3	3	0	3
5	2	6	4	4	2	3	0

We list for each data point its true class, its neighbors and its assignment.

$\bar{x}_i$	$f(\bar{x}_i)$	$\hat{f}((\bar{x}_i)$	Neighbors
1	-1	+1	2,3,4,6
2	+1	+1	1,6,8
3	+1	-1	1,4,5
4	-1	+1	3,5,6
5	+1	-1	3,4,7
6	+1	+1	2,4,8
7	-1	+1	4,5,6,8
8	+1	+1	2,6,7

The error rate is:

$$\frac{1}{8}\sum_{i=1}^{8} \mathbb{I}_{f(\bar{x}_i)\neq\hat{f}(\bar{x}_i)}(\bar{x}_i) = \frac{1}{8}(1+0+1+1+1+0+1+0) = \frac{5}{8}$$

2. Let us denote these test points as 9-13.
   The distance matrix to the training data is then:

	1 (-)	2 (+)	3 (+)	4 (-)	5 (+)	6 (+)	7 (-)	8 (+)
9	3	2	6	6	8	4	8	4
10	7	4	8	6	6	4	3	2
11	3	6	2	4	5	6	7	7
12	5	6	2	2	2	4	5	6
13	8	9	5	5	3	7	4	7

And we get

$\bar{x}_i$	$\hat{f}((\bar{x}_i)$	Neighbors
9	+1	1,2,3,4,6,8
10	+1	2,4,5,6,7,8
11	+1	1,2,3,4,5,6
12	+1	1,3,4,5,6,7 (2,8 break the tie)
13	+1	3,4,5,6,7,8

3. The maximum will be **k=3**. k=4 will not work. When using k=4 we need a majority of 3:1 to the minus class, that means having points 1,4,7 the closest in order to predict minus. Looking at the distances table from the previous section we can see that there is no set of test points that has all of them in its 4 NNs. Choosing k=3 we need at least 2 points out of 1,4,7 in order to classify a point as -1. Looking at point 11, points 1,3,4 are closest to it and we classify it as -1. Looking at any other point in the test set, it will be classified as +1 and hence we don't classify all points to the same class.

B.

1. Manhattan distance:

$$|8-0|+|3-0|=11$$
$$|6-0|+|6-0|=12$$

We will assign $x=(0,0)$ to class +1.
Euclidean distance:

$$(8-0)^2+(3-0)^2=73$$
$$(6-0)^2+(6-0)^2=72$$

We will assign $x=(0,0)$ to class -1.

2. Consider the points $\left(x_1=(4,2), x_2=(5,0)\right)$, where $x_1$is classified to -1 and x_2 is classified to +1.
Euclidean distance:

$$x_1(-)\rightarrow\sqrt{(1-4)^2+(-1-2)^2}=\sqrt{18}\approx 4.24$$
$$x_2(+)\rightarrow\sqrt{(1-5)^2+(-1-0)^2}=\sqrt{17}\approx 4.12$$

We classify x as +1.
L_∞ distance:

$$x_1(-)\rightarrow\max\{|1-4|,|-1-2|\}=3$$
$$x_2(+)\rightarrow\max\{|1-5|,|-1-0|\}=5$$

We classify x as -1.

C. We can think of the optimization problem in matrix setting:

$$\theta^*=\underset{\theta\in\mathbb{R}^k}{\operatorname{argmin}}\sum_{i=1}^{m}\left((M_iW,\theta)-y_i\right)^2=\underset{\theta\in\mathbb{R}^k}{\operatorname{argmin}}\,\|MW\theta-y\|^2$$

Since $M\in\mathbb{R}^{m\times k}$ and $W\in\mathbb{R}^{k\times k}$, $MW\in\mathbb{R}^{m\times k}$. Denote $MW\coloneqq X$.
So we have:

$$\underset{\theta\in\mathbb{R}^k}{\operatorname{argmin}}\,\|X\theta-y\|^2$$

$$\|X\theta - y\|^2 = < X\theta - y, X\theta - y > = (X\theta - y)^T(X\theta - y) =$$
$$y^Ty - y^TX\theta - \theta^TX^Ty + \theta^TX^TX\theta$$

As the loss is convex the optimum solution lies at gradient zero. The gradient of the loss function is:

$$\frac{\partial(y^Ty - y^TX\theta - \theta^TX^Ty + \theta^TX^TX\theta)}{\partial\theta} = -2y^TX + 2\theta^TX^TX$$

Setting the gradient to zero produces the optimum parameter:

$$= -2y^TX + 2\theta^TX^TX = 0$$
$$\theta^* = (X^TX)^{-1}X^Ty$$

Remembering $X = MW$,

$$\theta^* = (X^TX)^{-1}X^Ty = (W^TM^TMW)^{-1}W^TM^Ty$$

For full credit in the exam there was no need to present all the above derivation process. We can also directly use the pinv result learned in class to obtain the same solution.

Question 3 (4 parts 30 points)

A. (6 points) Are decision trees used for classification sensitive to feature transformations of the form $x' = x^3$?

B. (8 points) 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.

C. (10 points) Consider the training data described below, at a node S. Decide on which attribute to use according to the rules defined in B and using GiniGain. State your calculations.

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

D. (6 points) Consider the following dataset with features in $\mathbb{R}^2$:

X1	X2	Y
1	10	-
1	5	+
2	10	+
2	5	-
3	5	-

Show that using the standard Goodness of Split with Gini as the impurity function doesn't lead to an optimal decision tree for these data. Optimal here is in the sense of a minimal number of nodes in the resulting decision tree.

Question 3 (4 parts 30 points) – solution

A. Since $f(x) = x^3$ is a bijective function, the values of the feature $x' = f(x)$ will maintain their original order. Thus, the tree will perform the same classification on the data (however, it might have different node threshold values).

B. For each of the k attributes, we need to calculate the best possible split out of $V(i)$ values. Since all splits are binary, and since the number of possible ways to split a set of $V(i)$ values are $2^{V(i)-1}$ (The -1 is because order is not relevant), the number of goodness of split calculation in the root is $\sum_k 2^{V(i)-1}$.

C.

Gini impurity on S:

$$Gini(S) = 1 - \sum_{i=1}^{2} \left(\frac{|S_i|}{|S|}\right)^2 = 1 - \left(\frac{3}{8}\right)^2 - \left(\frac{5}{8}\right)^2 = 0.46875$$

Splitting according to X_1:

$$Gini(S) - \sum_v \frac{|S_v|}{|S|} Gini(S_v) = 0.46875 - \frac{3}{8} Gini(S_{green}) - \frac{5}{8} Gini(S_{blue}) =$$

$$= 0.46875 - \frac{3}{8} \cdot \left(1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2\right) - \frac{5}{8} \cdot \left(1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2\right) = 0.002$$

Splitting according to X_2:

$$Gini(S) - \sum_v \frac{|S_v|}{|S|} Gini(S_v) =$$

$$= 0.46875 - \frac{3}{8} Gini(S_{Messi}) - \frac{5}{8} Gini(S_{Ronaldo+Neymar}) =$$

$$= 0.46875 - \frac{3}{8} \cdot \left(1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2\right) - \frac{5}{8} \cdot \left(1 - \left(\frac{1}{5}\right)^2 - \left(\frac{4}{5}\right)^2\right) = 0.102$$

$$Gini(S) - \sum_v \frac{|S_v|}{|S|} Gini(S_v) =$$

$$= 0.46875 - \frac{3}{8} Gini(S_{Ronaldo}) - \frac{5}{8} Gini(S_{Messi+Neymar}) =$$

$$= 0.46875 - \frac{2}{8} \cdot \left(1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2\right) - \frac{6}{8} \cdot \left(1 - \left(\frac{2}{6}\right)^2 - \left(\frac{4}{6}\right)^2\right) = 0.01$$

$$Gini(S) - \sum_v \frac{|S_v|}{|S|} Gini(S_v)$$

$$= 0.46875 - \frac{3}{8} Gini(S_{Neymar}) - \frac{5}{8} Gini(S_{Ronaldo+Messi}) =$$

$$= 0.46875 - \frac{3}{8} \cdot \left(1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2\right) - \frac{5}{8} \cdot \left(1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2\right) = 0.168$$

X_2 will used for the split since the GiniGain value is greater.

D.

Gini impurity on S:

$$Gini(S) = 1 - \sum_{i=1}^{2} \left(\frac{|S_i|}{|S|}\right)^2 = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48$$

Splitting according to X_1 with a threshold of 1.5:

$$0.48 - \frac{2}{5} \cdot \left(1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2\right) - \frac{3}{5} \cdot \left(1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2\right) = 0.013$$

Splitting according to X_1 with a threshold of 2.5:

$$0.48 - \frac{1}{5} \cdot \left(1 - \left(\frac{1}{1}\right)^2 - \left(\frac{0}{1}\right)^2\right) - \frac{4}{5} \cdot \left(1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right) = 0.079$$

Splitting according to X_2:

$$0.48 - \frac{3}{5} \cdot \left(1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2\right) - \frac{2}{5} \cdot \left(1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2\right) = 0.013$$

After this split on the root there must be two additional splits for the tree to be completely pure. Thus, the height of the tree is 3 and the tree has 4 internal nodes. However, if we split the root using X_2, we will obtain the following tree with height 2 and 3 internal nodes:

Root: $X_2 > 7.5$

- Right: $X_1 > 1.5$
  - Leaf
  - Leaf
- Left: $X_1 > 1.5$
  - Leaf
  - Leaf

Question 4 (2 parts 30 points)

A. (10 points)
Find the maximum and minimum values of

$$f(x) = 81x^2 + y^2$$

subject to the constraint

$$4x^2 + y^2 = 9$$

B. Consider the instance space $X = \mathbb{R}^3$ with some probability distribution π. Consider the concept space C that consists of symmetric boxes around the origin. In each of the three axes, the distance of the 2 sides from the origin is equal. Instances inside the box belong to the positive class and instances outside the box belong to the negative class.
Formally, for any $u, v, w > 0$ define
$c(u, v, w) = \{(x, y, z) | \ |x| \leq u, \ |y| \leq v \ , \ |z| \leq w\}$
and then:
$C \ = \{c(u, v, w) | u, v, w > 0 \}$.
Let $H = C$.

1. (6 points) Propose a consistent learning algorithm L that takes as input labeled points $D^m = \{(x_1, y_1, z_1), (x_2, y_2, z_2) \ldots, (x_m, y_m, z_\mathrm{m})\} \subseteq \mathbb{R}^3$ and returns $h \in H$.
2. (7 points) Prove that C is PAC learnable from H by computing a sufficiency bound on the sample complexity.
3. (7 points) For $\varepsilon = 0.1$ and $\delta = 0.05$ compute a sufficiency 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:
$Prob\big(Err\big(c, L(D^m)\big) > \ \varepsilon\big) < \delta$

Question 4 (2 parts 30 points) – solution

A.

$$L = 81x^2 + y^2 - \lambda(4x^2 + y^2 - 9)$$
$$\frac{\partial L}{\partial x} = 162x - 8\lambda x = 0$$
$$\frac{\partial L}{\partial y} = 2y - 2\lambda y = 0$$
$$\frac{\partial L}{\partial \lambda} = 4x^2 + y^2 - 9 = 0$$

We have two options:
First

$$y = 0$$
$$x = \pm\frac{3}{2}$$
$$f\left(\frac{3}{2}, 0\right) = 81\frac{9}{4} + 0 = \frac{729}{4}$$
$$f\left(-\frac{3}{2}, 0\right) = 81\frac{9}{4} + 0 = \frac{729}{4}$$

Second

$$\lambda = 1$$
$$162x = 8x$$
$$x = 0$$
$$x = \pm 3$$
$$f(0,3) = 0 + 9 = 9$$
$$f(0,-3) = 0 + 9 = 9$$

We get maximum of $\frac{729}{4}$ at $(\pm\frac{3}{2}, 0)$ and minimum of 9 at $(0, \pm 3)$

B.

1. The algorithm will produce a hypothesis which is the smallest relevant box that contains all the positive points. This can be done in O(m) as follows:
   Let $\Delta = \Delta^m = (x_i, y_i, z_i)_{i=1}^m$ be a set of points in $\mathbb{R}^3$, labeled positive and negative.
   Our algorithm seeks to return a hypothesis $h \in H$.
   Let $(x_i, y_i, z_i)_{i=1}^{m^{(+)}}$ be all positively labeled data points.
   Find:
   1) $l := \max_{1 \le i \le m^{(+)}}(|x_i|)$
   2) $m := \max_{1 \le i \le m^{(+)}}(|y_i|)$
   3) $n := \max_{1 \le i \le m^{(+)}}(|z_i|)$

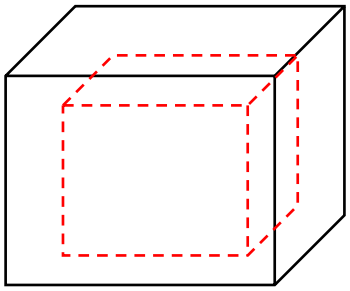
   l, m, n will be the distances of the sides from the origin in the axes x, y, z respectively.

2. Consider $c \in C$ and let $\Delta^m(c) = \left(x_i(c), y_i(c), z_i(c)\right)_{i=1}^m$ be training data generated from c without errors and by drawing m independent points according to a probability distribution π on $\mathbb{R}^3$. We will denote the probability distribution thus induced on $(\mathbb{R}^3)^m$ by π^m.

   Given $\varepsilon > 0$ and $\delta > 0$ we will now compute a number $m(\varepsilon, \delta)$ so that (Eq.1)

   $$m \geq m(\varepsilon, \delta) \Longrightarrow e\left(\Delta^m(c)\right) = \pi^m\left(err_\pi\left(L\left(\Delta^m(c)\right), c\right) > \varepsilon\right) \leq \delta$$

   Note that $L\left(\Delta^m(c)\right)$ is the hypothesis h, or the box, produced by L when considering data $\Delta^m(c)$ (as we describe in the previous section). $e\left(\Delta^m(c)\right)$ is a random variable that depends on the stochastic behavior of $\Delta^m(c)$. It is exactly this behavior that we will want to characterize.



   Consider the margins parallel to the sides of the box c.
   Note that there are 6 margins one for each side. Further note that the size of each two parallel margins are determined by only one parameter from l, m, n.
   Let $S1$ be the area defined by both margins relevant to the x-axis.
   Let $S2$ be the area defined by both margins relevant to the y-axis.
   Let $S3$ be the area defined by both margins relevant to the z-axis.
   These are defined to satisfy:

   $$\pi\left(S1(\varepsilon)\right) = \pi\left(S2(\varepsilon)\right) = \pi\left(S3(\varepsilon)\right) = \frac{\varepsilon}{3}$$

   Now, note that

   $$\begin{aligned}&\left\{\Delta^m(c): err_\pi\left(L\left(\Delta^m(c)\right), c\right) > \varepsilon\right\} \subseteq \\ &\{\Delta^m(c): \Delta^m(c) \cap S1(\varepsilon) = \emptyset\} \cup \\ &\{\Delta^m(c): \Delta^m(c) \cap S2(\varepsilon) = \emptyset\} \cup \\ &\{\Delta^m(c): \Delta^m(c) \cap S3(\varepsilon) = \emptyset\}\end{aligned}$$

   This is because if $\Delta^m(c)$ visits the three strips (note that negative points can not visit these strips as there are no errors in the training labels) then, according to our construction, the difference between c and $L\left(\Delta^m(c)\right)$ will have $\pi \leq \pi\left(S1(\varepsilon) \cup S2(\varepsilon) \cup S3(\varepsilon)\right) < \varepsilon$.

   In terms of probability we therefore get:

$$\pi^m\left(err_\pi\left(L\left(\Delta^m(c)\right),c\right)>\varepsilon\right)\leq$$
$$\pi^m\left(\Delta^m(c)\cap s1(\varepsilon)=\emptyset\right)+\pi^m\left(\Delta^m(c)\cap s2(\varepsilon)=\emptyset\right)+$$
$$\pi^m\left(\Delta^m(c)\cap s3(\varepsilon)=\emptyset\right)\leq 3\left(1-\frac{\varepsilon}{3}\right)^m$$

We now select $m(\varepsilon,\delta)=\frac{3}{\varepsilon}\left(\ln(3)+\ln\left(\frac{1}{\delta}\right)\right)$ to get (Eq.1) to hold.

Q.E.D

3. We will substitute $\varepsilon=0.1$ and $\delta=0.05$ in the equation from the previous section:

$$m>\frac{3}{\varepsilon}\left(\ln(3)+\ln\left(\frac{1}{\delta}\right)\right)$$
$$m>\frac{3}{0.1}\left(\ln(3)+\ln\left(\frac{1}{0.05}\right)\right)$$

$$m>122.96$$

GOOD LUCK!