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 ב בלמידה ממוכנת
Machine Learning Exam B

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) – Theory

Let $X = \mathbb{R}^2$. Let $C = H$ be the set of all rectangles with the bottom left vertex on the origin and the top right in the first quadrant. Consider data points inside the rectangle as Positives and data points outside of the rectangle to be Negatives.

Formally:
For any two numbers $t, r \in \mathbb{R}_+$
we define $h(t,r) = \{(x,y) \mid \ 0 \leq x \leq t \wedge 0 \leq y \leq r\}$.
And now we define the hypotheses space as:

$$H = \{h(t,r) \mid t, r \in \mathbb{R}_+\}$$

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 $D^m = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\} \subseteq \mathbb{R}^2$ and returns $h \in H$.

3. (7 pts) Compute a sufficiency bound on the sample complexity of learning C from H using the algorithm L you suggested.

4. (7 pts) The bound from the previous section is not tight. Explain why.

Question 1 (25 points) – Theory – Solution

1. $VC \geq 2$:

   Let $p_1 = (1,0), p_2 = (0,1)$

$$p_1 = -, p_2 = - \Longrightarrow h(0,0)$$
$$p_1 = -, p_2 = + \Longrightarrow h(0,2)$$
$$p_1 = +, p_2 = - \Longrightarrow h(2,0)$$
$$p_1 = +, p_2 = + \Longrightarrow h(0,0)$$

   $VC < 3$:

   Given any three points p_1, p_2, p_3. WLOG the $p_1(y) \geq p_2(y), p_3(y)$ and $p_2(x) \geq p_1(x), p_3(x)$. The following labeling is not possible for separation $p_1 = +, p_2 = +, p_3 = -$.

2. The algorithm will produce a hypothesis which is the smallest relevant rectangle that contains all the positive points. This can be done in O(m) as follows:

   Let $\Delta = \Delta^m = (x_i, y_i)_{i=1}^m$ be a set of points in the plane, labeled positive and negative.

   Our algorithm seeks to return a hypothesis $h \in H$.

   Let $(x_i, y_i)_{i=1}^{m^{(+)}}$ be all positively labeled data points.

   Find:

   1) $t := \max_{1 \leq i \leq m^{(+)}} (|x_i|)$

   2) $r := \max_{1 \leq i \leq m^{(+)}} (y_i)$

   The top right vertex hypothesis rectangle will be (t, r), $\big(L(\Delta) = h(t,r)\big)$

3.

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

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

   (Eq.1) $\quad m \geq m(\varepsilon, \delta) \Longrightarrow e\big(\Delta^m(c)\big) = \pi^m \left(err_\pi\big(L\big(\Delta^m(c)\big), c\big) > \varepsilon\right) \leq \delta$

Note that $L\left(\Delta^m(c)\right)$ is the hypothesis h, or the rectangle, produced by L when considering data $\Delta^m(c)$ as above. $e\left(\Delta^m(c)\right)$ 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.

Let $c_0 = h(t,r) \in C$ be the concept we are trying to estimate.
Let $S1(\varepsilon) = \{(x,y): \alpha \leq x \leq t, 0 \leq \mathrm{y} \leq r\}$.
Let $S2(\varepsilon) = \{(x,y): 0 \leq x \leq t, \beta \leq \mathrm{y} \leq r\}$.
α, β where chosen so that $\pi\left(S1(\varepsilon)\right) = \pi\left(S2(\varepsilon)\right) = \frac{\varepsilon}{2}$.
Let $I(\varepsilon) = h(t-\alpha, r-\beta)$.
If $\exists p_1, p_2 \in \Delta = \Delta^m(c_0)$, such that $p_1 \in S1(\varepsilon),\ p_2 \in S2(\varepsilon)$, it follows that $I(\varepsilon) \subseteq h = L(\Delta) \subseteq c$.
Hence, $err_\pi\left(L\left(\Delta^m(c)\right), c\right) = err_\pi(h,c) \ \leq\ err_\pi(I(\varepsilon), c) = \pi\left(c - I(\varepsilon)\right) = \pi\left(S1(\varepsilon) \cup S2(\varepsilon)\right) \leq \pi\left(S1(\varepsilon)\right) + \pi\left(S2(\varepsilon)\right) \leq \varepsilon$.

If $err_\pi\left(L\left(\Delta^m(c)\right), c\right) > \varepsilon$ then the above argumentation suggests that the training set does not visit one of the two sets $S1(\varepsilon)$ and $S2(\varepsilon)$, and therefore:

$$\pi^m\left(err_\pi\left(L\left(\Delta^m(c)\right), c\right) > \varepsilon\right) \leq$$
$$\pi^m\left(\Delta^m(c) \cap S1(\varepsilon) = \emptyset\right) + \pi^m\left(\Delta^m(c) \cap S2(\varepsilon) = \emptyset\right) \leq$$
$$2\left(1 - \frac{\varepsilon}{2}\right)^m$$

We now select $m(\varepsilon, \delta) = \frac{2}{\varepsilon}\left(\ln(2) + \ln\left(\frac{1}{\delta}\right)\right)$ to get (Eq.1) to hold.

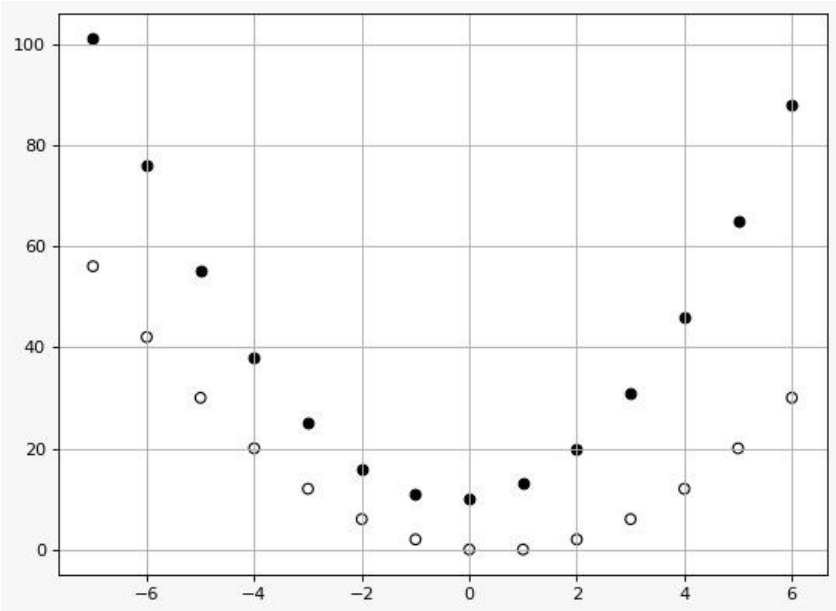
Q.E.D

4. The main reason that the bound is not tight is due to the fact that we consider only hypotheses that have points in both strips to be epsilon good. In fact, there are several more hypotheses that are NOT epsilon bad but, at the same time, visit at most $n-1$ strips (in this case one strip). These hypotheses are not considered when we calculate the bound and therefore the bound is not tight. Note that, if we have only one strip (the circle example) there are no such hypotheses.

Question 2 (25 points) – SVM

You are given the following dataset D in the figure where the empty circles are the negative class and the full circles are the positive class. The data includes the following points:

x	y	$class$
-6	76	+
2	20	+
5	65	+
-2	6	-
-4	20	-
5	20	-



1. Recall that a parabola is defined using the following function:

$$y = ax^2 + bx + c$$

   a. (3 pts) Find the equation for the top (full circles) parabola.
   b. (3 pts) Find the equation for the bottom (empty circles) parabola.

2. Observe the following claim:

Assume that a linear classifier predicts the same $y \in \{-1, +1\}$ for some two points $z, z' \in \mathbb{R}^n$ (that is $h(z) = h(z')$). Then it will produce the same prediction for any intermediate point. That is:

$$\forall \alpha \in [0,1] \quad h\big((1-\alpha)z + \alpha z'\big) = h(z) = h(z')$$

a. (8 pts) Prove the claim.
b. (5 pts) Use the above claim to prove that the depicted data is not linearly separable.

3. (6 pts) Find a mapping $\Phi(x, y)$ to $\mathbb{R}^4$ such that $\Phi(D)$ is linearly separable and prove the validity of this mapping by showing a vector w such that the point (x, y) is positive (full circle) iff $w \cdot \Phi(x, y) \geq 0$.

<u>Question 2 – Solution</u>

1. We solve the parabola equation by using the 3 points of each parabola from the table:
   a. We use the following equations:
      i. $a_1 \cdot (-6)^2 + b_1 \cdot -6 + c_1 = 76$
      ii. $a_1 \cdot (2)^2 + b_1 \cdot 2 + c_1 = 20$
      iii. $a_1 \cdot (5)^2 + b_1 \cdot 5 + c_1 = 65$

   And get that: $a_1 = 2, b_1 = 1, c_1 = 10$

   b. We use the following equations:
   c. $a_2 \cdot (-2)^2 + b_2 \cdot -2 + c_2 = 6$
   d. $a_2 \cdot (-4)^2 + b_2 \cdot -4 + c_2 = 20$
   e. $a_2 \cdot (5)^2 + b_2 \cdot 5 + c_2 = 20$
   And get that: $a_2 = 1, b_2 = -1, c_2 = 0$

2.
   a. Let there be $x = (1 - \alpha)z + \alpha z'$ where $\alpha \in [0,1]$.
   Let w be the weight vector of the linear classifier h.
   Observe that:
   $$w \cdot x = w \cdot \big((1 - \alpha)z + \alpha z'\big) =$$
   $$(1 - \alpha) \cdot (w \cdot z) + \alpha \cdot (w \cdot z')$$
   Divide to cases:
   $$h(z) = h(z') = 1$$
   Therefore $w \cdot z > 0$ and $w \cdot z' > 0$.
   α and $1 - \alpha$ are non-negative and therefore do not change the sign of the multiplication.
   Hence:
   $$(1 - \alpha) \cdot (w \cdot z) + \alpha \cdot (w \cdot z') > 0 \rightarrow h(x) = 1$$
   $$h(z) = h(z') = 1$$
   Therefore $w \cdot z \leq 0$ and $w \cdot z' \leq 0$.
   α and $1 - \alpha$ are non-negative and therefore do not change the sign of the multiplication.
   Hence:
   $$(1 - \alpha) \cdot (w \cdot z) + \alpha \cdot (w \cdot z') \leq 0 \rightarrow h(x) = 0$$
   $$Q.E.D$$

b. Take $z = (-4,20)$ and $z' = (5,20)$. Both Points Belong to the black circle class (denote it as +1).
Any linear classifier according to the lemma will classify any point $x = (1-\alpha)z + \alpha z'$ as $+1$.
Find α for $x' = (2,20)$:

$$2 = (1-\alpha)\cdot -4 + 5\cdot\alpha \rightarrow$$
$$2 = -4 + 4\alpha + 5\alpha \rightarrow$$
$$6 = 9\alpha$$
$$\alpha = \frac{2}{3}$$

Therefore, according to the lemma any linear classifier will classify it as $+1$ but it belongs to the white circle class -1.
Hence the data is not linearly separable.

3. Define a mapping: $\phi = (1, x, x^2, y)$ And a Linear classifier: $w = (0, -1, 1, -1)$.

Question 3 (25 points) – Clustering

1. (5 pts) Consider the following dataset: $\{0, 4, 5, 20, 25, 39, 43, 44\}$ and the hierarchical clustering algorithm learned in class. Using two distance methods – single linkage and complete linkage what will be the elements in the last two clusters you combine?

   Recall that **single linkage** uses the minimum of the distances between all observation of the two sets and **complete linkage** uses the maximum distance between all observation of the two sets. Also recall that these two methods represent different ways for combining clusters.

2. (5 pts) Consider the Naïve Cluster Growing algorithm as shown in class with a predetermined threshold, $T = 1$. Provide two example datasets as follows:
   a. Dataset A contains 5 points, and after running Naïve Cluster Growing 5 clusters are created.
   b. Dataset B contains 5 points, and after running Naïve Cluster Growing a single cluster is created.

   Use Euclidean distance and any data dimensionality. For every dataset, you should list all point coordinates.

3. (6 pts) For each option, determine if it is true or false. Justify your answers.
   a. The number of clusters must be pre-specified as a parameter for both k-means and hierarchical clustering.
   b. The k-means clustering algorithm requires random assignments while the hierarchical clustering algorithm does not.

4. (9 pts) The k-means-outlier-L1 algorithm is a variant of the k-means algorithm given by the following pseudocode:

   1. Initialize k centers $c_1, \dots, c_k$ randomly unless centers are given.
   2. Loop until there is no change in $c_1, \dots, c_k$:
      a. Assign all n samples to their closest center c_i
         and create k clusters, $S_1, \dots, S_k$.
      b. For each cluster define a new center:
         If $|S_i| > 2$:
            let x_i be the point in S_i with the largest L_1 distance from c_i.
            Calculate the new center c_i using $S_i \backslash \{x_i\}$.
         otherwise:
            Calculate the new center c_i using S_i.

#	x_1	x_2
1	1	0
2	0	1
3	1	5
4	3	1
5	6	5
6	6	6

   Run The k-means-outlier-L1 algorithm on the following data for $k = 2$ and the following starting centers: $c_1 = (0,0),\ c_2 = (4,1)$.

Question 3 – Solution

1. For both methods: $\{0, 4, 5\}, \{20, 25, 39, 43, 44\}$
2.
   a. 5 points such that in every clustering step, the distance between the clusters is > 1. i.e. $\{0, 2, 4, 6, 8\}$
   b. 5 points such that in every clustering step, the distance between the clusters is < 1. i.e. $\{0, 0.5, 1, 1.5, 2\}$
3.
   a. False. K-means requires the number of clusters as a parameter, however hierarchical clustering does not require such argument to run. The number of clusters can be determined after the algorithm execution.
   b. True. Hierarchical clustering does not require any random assignment and is completely deterministic. K-means uses random centers in order to make the initial assignment.
4.

#	x_1	x_2	c
1	1	0	1
2	0	1	1
3	1	5	1
4	3	1	2
5	6	5	2
6	6	6	2
$c_1 = (0,0),\ c_2 = (4,1)$			
$c_1 = (0.5,0.5),\ c_2 = (4.5,3)$			

x_1	x_2	c
1	0	1
0	1	1
1	5	1
3	1	1
6	5	2
6	6	2
$c_1 = (0.5,0.5),\ c_2 = (4.5,3)$		
$c_1 = (1.33,0.66),\ c_2 = (6,5.5)$		

x_1	x_2	c
1	0	1
0	1	1
1	5	1
3	1	1
6	5	2
6	6	2
$c_1 = (1.33,0.66),\ c_2 = (6,5.5)$		
$c_1 = (1.33,0.66),\ c_2 = (6,5.5)$		

Question 4 (25 points) – Logistic Regression

1. (6 pts) Recall the logistic regression loss function:

$$J = -\sum_{d=1}^{m} y^{(d)} \ln\left(h_\theta\left(x^{(d)}\right)\right) + \left(1 - y^{(d)}\right) \ln\left(1 - h_\theta\left(x^{(d)}\right)\right)$$

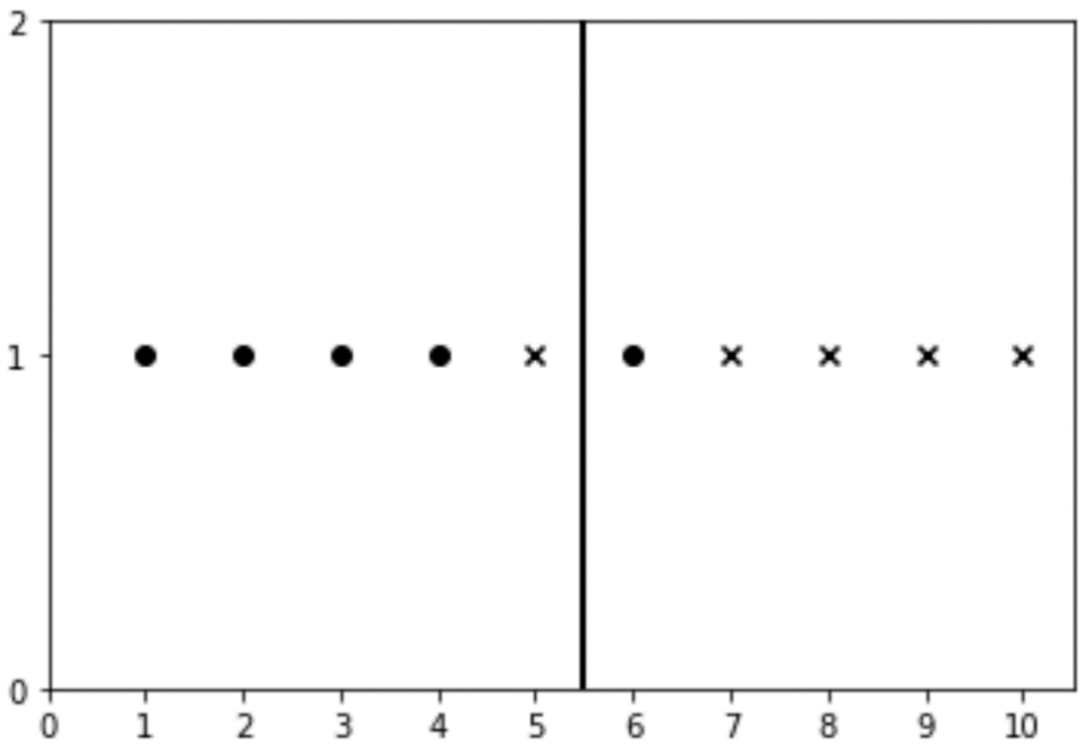
$$h_\theta\left(x^{(d)}\right) = \frac{1}{1 + e^{-\theta^T x^{(d)}}}$$

   Given a linearly separable dataset, does logistic regression always find a perfect linear separator (training error = 0)?

2. (9 pts) Answer the following questions:

   a) What method is used to optimize the cost function of logistic regression? (i.e., to find that the optimal θ)?
   b) Can pseudo inverse be used to find the optimal θ in logistic regression?
   c) Can logistic regression be applied to non-binary classification? If so, how? If not, why not?

3. (10 pts) Consider the following data set of two classes (negative class: o, positive class: x) and the classifier found by using logistic regression (straight line in the middle). Construct the ROC curve for this classifier by drawing 11 points from the curve (draw in your notebook – **not in the exam itself**).



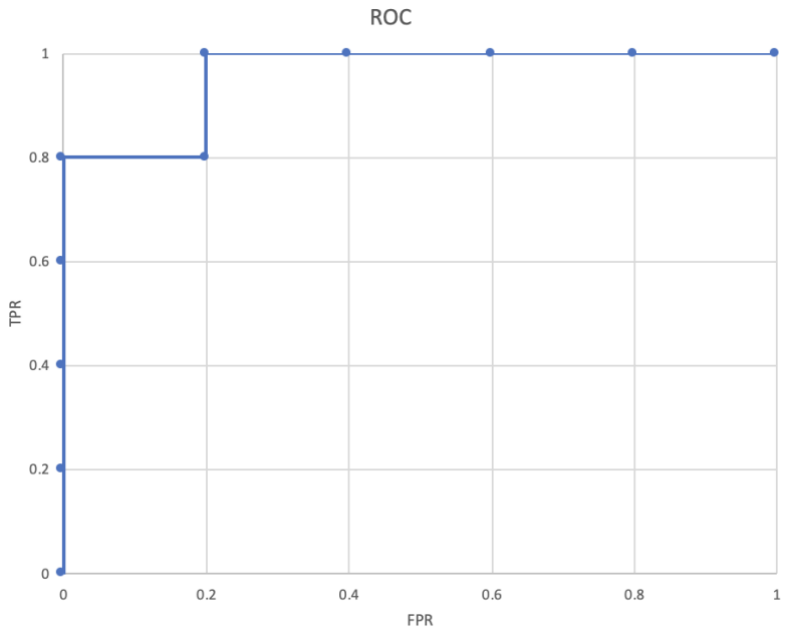
Question 4 (25 points) – Logistic Regression – Solution

1. No. The Logistic Regression trying to find a minimum for a function that penalize probabilities. The fact that it doesn't penalize errors can result in learning hypothesis that does not perfectly separate the separable data.

2. (9 pts) Answer the following questions:

   a) In Logistic Regression we use Gradient Descent in order to find optimal θ
   b) No. There is no close form solution for the cost function of Logistic Regression and therefore we can't use pseudo inverse solution.
   c) Yes. We can use One vs All. The prediction will be the class with the highest score.

3.



Question 5 (25 points) – Decision Trees

The following data consists of instances with two numerical attributes, x_1 and x_2, coming from two classes "+" and "-". We learn a decision tree using this data.
A tree of type T_N is a binary tree that uses Goodness of split and grows up to height N or until no further split is possible.
For example:
T_1 will have one split (a root and two children)
T_2 will have one split at the root and then up to one split for each child.
Note: the tree doesn't have to be symmetric.

Instance	X1	X2	Value
1	2	3	+
2	1	3	-
3	1	4.5	-
4	1	2	+
5	1	5	-
6	2	5	-
7	1	6	-
8	2	6	-
9	1	7	-
10	2	7	-
11	1	8	-
12	2	8	-
13	2	2	+

1. (5 pts) Explain what an impurity function is and how Goodness of Split uses an impurity function φ to determine node splitting in a decision tree. Your explanation should use a clear formula.

2. (10 pts) Is it possible for a tree built using Goodness of Split to be of greater height than that of another tree built by splits that also get to pure leaves? If yes – provide an example. If not – clearly explain why.

3. (5 pts) Would the second splits in a T_2 tree learned from training data be different from the second splits in a T_3 learned from the same training data? Explain your answer.

4. (5 pts) When constructing trees of type T_1 and of type T_2 using the training data given above, what is the error obtained, evaluated by leave one out? Which of the two trees leads to a better result, as evaluated by leave one out?

Question 3 – Solution

1. Impurity measures the distance from perfect classification. The maximum value will be attained when the distribution is uniform, and the minimum value will be attained when there is perfect classification (all the instances in the node have the same class). Goodness of Split checks the reduction in the impurity given the split feature. It calculates the impurity in the current node and then subtract the weighted average of the impurity in the children given the split feature.

$$Goodness\_of\_Split = \varphi(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} \varphi(S_v)$$

2. Yes. Consider the following data:

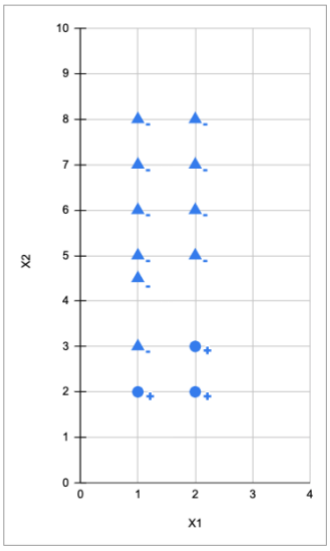
Instance No.	X1	X2	X3	Y
1	1	1	0	+
2	1	0	1	+
3	1	0	0	-
4	0	0	0	-
5	0	1	1	-

Splitting according to X2 first and X3 afterwards yields perfect classification with a smaller tree than the goodness of split method, which will split according to X1 first, and cannot use a single additional split to reach perfect classification.

3. Since the trees are built in a greedy manner, they both are identical in the overlapping depths.

4. Those are the prediction made by each tree for the sample that is left out:

Instance	X1	X2	Value	T_1 Prediction	T_2 prediction
1	2	3	+	- (wrong)	- (wrong)
2	1	3	-	+ (wrong)	+ (wrong)
3	1	4.5	-	-	-
4	1	2	+	+	- (wrong)
5	1	5	-	-	-
6	2	5	-	-	-
7	1	6	-	-	-
8	2	6	-	-	-
9	1	7	-	-	-
10	2	7	-	-	-
11	1	8	-	-	-
12	2	8	-	-	-
13	2	2	+	+	+



In this case T_1 has 2 errors and T_2 has 3 errors, so T_1 is better.