10-601: Homework 1

Due: 18 September 2014 11:59pm (Autolab)

TAs: Abhinav Maurya, Jingwei Shen

Name: ARAVIND SELVAN

Andrew ID: ASELVAN

Please answer to the point, and do not spend time/space giving irrelevant details. You should not require more space than is provided for each question. If you do, please think whether you can make your argument more pithy, an exercise that can often lead to more insight into the problem. Please state any additional assumptions you make while answering the questions. You need to submit a single PDF file on autolab. Please make sure you write legibly for grading.

You can work in groups. However, no written notes can be shared, or taken during group discussions. You may ask clarifying questions on Piazza. However, under no circumstances should you reveal any part of the answer publicly on Piazza or any other public website. The intention of this policy is to facilitate learning, not circumvent it. Any incidents of plagiarism will be handled in accordance with CMU's Policy on Academic Integrity.

⋆: Code of Conduct Declaration

- Did you receive any help whatsoever from anyone in solving this assignment? (Yes) / No.
- If you answered *yes*, give full details: From 'Allan Wang' on understanding the concepts of decision tree & KNN for problem 5&6. (e.g. *Jane explained to me what is asked in Question 3.4*)
- Did you give any help whatsoever to anyone in solving this assignment? (Yes) / No.
- If you answered *yes*, give full details: Pradheep Shanmugam' on question 5(b) to help him with the diagrammatic understanding of what 'leave-one out' accuracy is. (e.g. *I pointed Joe to section 2.3 to help him with Question 2*)

1: The truth will set you free. (TA:- Abhinav Maurya)

State whether true (with a brief reason) or false (with a contradictory example). Credit will be granted only if your reasons/counterexamples are correct.

(a) During decision tree construction, if you reach a node where the maximum information gain for a node split using any attribute is zero, then all training examples at that node have the same label.

False, it is not necessary that all the labels to be the same when the information gain is zero. Ex:

x_1	x_2	x_3	class (label)
a	K	P	c_1
a	K	P	c_2

[2 points] Though both have all 3 features the same, label can still be different.

(b) Whenever a set S of labeled instances is split into two sets S_1 and S_2, the average entropy will not increase, irrespective of the split attribute or the split point.

– Next page –

True, because, entropy will always decrease / remain the same, when the instances are split into two sets. [will never increase]. [2 points]

(c) A decision tree can be represented as a decision list and vice versa. (Hint: A decision list is a sequentially applied list of decision rules of the form: *If condition$_1$ and condition$_2$ and ... condition$_n$, then output is y_i*. Each condition is a test on a single feature similar to the nodes of a decison tree.)

True, a decision tree can be represented by a list of decision rules. i.e) a decision rule for each instance (root-leaf) in a decision tree. [2 points]

(d) If X_1 and X_2 are independent gaussian random variables, $X = \frac{1}{4}(X_1 - X_2)$ is a gaussian random variable.

True, from definition, $X_1 \sim N(\mu_1, \sigma_1^2)$ [2 points]

$X_2 \sim N(\mu_2, \sigma_2^2)$

$X = \frac{1}{4}X_1 - \frac{1}{4}X_2 = N\left(\frac{1}{4}\mu_1 - \frac{1}{4}\mu_2, \frac{1}{16}\sigma_1^2 - \frac{1}{16}\sigma_2^2\right) = N\left(\frac{1}{4}(\mu_1 - \mu_2), \frac{1}{16}(\sigma_1^2 - \sigma_2^2)\right)$

(e) If $f_{X_1}(x_1)$ and $f_{X_2}(x_2)$ are the probability density functions of independent gaussian random variables, $f(X) = \frac{1}{2}\{f_X(x) + f_Y(y)\}$ is a probability density function corresponding to a gaussian random variable.

False, because this can be true only if $f_{X_1}(x_1)$ and $f_{X_2}(x_2)$ are equal and this is a conditional false. [2 points]

2: Maximum Likelihood Estimation. (TA:- Jingwei Shen)

(a) $X_1, X_2, \cdots, X_n$ are random variables that are uniformly distributed between $[-\theta/2, \theta/2]$, $\theta \in \mathbb{R}$. Write down the MLE for the parameter θ and explain it. (You do not have to derive it.)

As the random variables are uniformly distributed, $P(X) = \begin{cases} \frac{1}{b-a} & -\theta/2 \le X \le \theta/2 \\ 0 & \text{otherwise} \end{cases}$ [3 points]

$\therefore P(X_1) = \frac{1}{(\theta_2 + \theta_2)} = \frac{1}{\theta}$

$P(X_n) = \left(\frac{1}{\theta}\right)^n = \frac{1}{\theta^n}$.

MLE for $\theta \Rightarrow \frac{\partial}{\partial\theta}\left(\frac{1}{\theta^n}\right) = \frac{\partial}{\partial\theta}(\theta^{-n}) = -n\theta^{-(n+1)} = \left(\frac{-n}{\theta^{n+1}}\right)$

Equating to zero. $\left(\frac{-n}{\theta^{n+1}}\right) = 0$. ; As the '$\theta$' is in the denominator, the MLE cannot be estimated. //

(b) We have two coins - an unbiased one with probability $p_1 = 1/2$ of showing heads on a toss, and a biased one with probability $p_2 = 1/3$ for showing heads. We choose one of the two coins. With an unknown probability p, we choose the biased coin, and with probability $1 - p$, we choose the unbiased one. We toss the chosen coin 100 times and observe 40 heads during the 100 tosses. Write down the MLE estimate of parameter p and explain it. (You do not have to derive it.)

[3 points]

MLE estimate of parameter p:

Consider p_1 = probability of heads on an unbiased coin
p_2 = probability of heads on an biased coin
H = # of heads; T = # of tails
$p_3 = 1 - p_1$ = probability of tails on an unbiased coin
$p_4 = 1 - p_2$ = probability of tails on an biased coin.

$\Rightarrow [p_1 \cdot p + p_2(1-p)]^H \times [p_3 \cdot p + p_4(1-p)]^T$

$\Rightarrow [(1/3)p + 1/2(1-p)]^{40} \times [(2/3)p + 1/2(1-p)]^{60}$ //.

3: Three Prisoners and a Warden (TA:- Jingwei Shen)

Three prisoners - A, B, and C - are on death row. The governor decides to pardon one of the three and chooses the prisoner to pardon at random. He informs the warden of his choice but requests the name to be kept as a secret.

Having heard of the pardon rumor through grapevine, A tries to get the warden to tell him his fate. The warden refuses. Then A asks which of B or C will be executed. The warden thinks a while and tells A that B is to be executed. (Assume that the warden picks a random legal answer for A's question).

(a) Let A, B, C denote the event that A, B, C will be pardoned respectively. Let $!B$ denote the event that the warden says B will die. Compute $P(A\,|!B)$. Does the chance of A's survival increase with the additional information about B's death? (Hint: compare $P(A|!B)$ and $P(A)$).

[3 points]

$$P(A/!B) = \frac{P(!B/A)P(A)}{P(!B)} = \frac{P(!B/A)P(A)}{P(!B/C)P(C) + P(!B/B)P(B) + P(!B/A)P(A)}$$

$$= \frac{1/2 \times 1/3}{1(1/3) + 0 \times 1/3 + (1/2)(1/3)} = \frac{1/6}{1/3 + 0 + 1/6} = \frac{1/6}{1/2} = 1/3 \text{ //.}$$

$\boxed{P(A|!B) = 1/3}$

(b) Suppose A reveals all of the above to C. Show the probability of C surviving at this time is 2/3. (Hint: Prove $P(C|!B) = 2/3$).

[3 points]

$$P(C|!B) = \frac{P(C \cap !B)}{P(!B)} =$$

Using total probability calculation from 3)a).

$$= \frac{1/3}{1/2} = 2/3 \text{ //}$$

$\boxed{P(C|!B) = 2/3}$

4: Probability Theory (TA:- Jingwei Shen)

(a) Let A, B, C be three discrete random variables. Show that

1. $P(A \mid B, C) = \frac{P(A,B|C)}{P(B|C)}$
2. $P(A \mid C) = \sum_B P(A, B \mid C)$
3. $P(A \mid C) = \sum_B P(A \mid B, C) \cdot P(B \mid C)$

[3 points]

L.H.S.

(1) $P(A|B,C) = \frac{P(A,B,C)}{P(B,C)} = \frac{P(A,B|C).P(C)}{P(B|C).P(C)} = \frac{P(A,B|C)}{P(B|C)} = R.H.S.$ //

(2) We know that,
$P(A|B,C) \cdot P(B|C) = P(A,B|C) \Leftrightarrow P(B|A,C).P(A|C) = P(A,B|C)$;
Take Σ_B on both sides
$\Sigma_B P(B|A,C) \cdot P(A|C) = \Sigma_B P(A,B|C) \Rightarrow$ Since the probability of $\Sigma_B P(B|A,C) = 1$
$\therefore P(A|C) = \Sigma_B P(A,B|C)$

(3) $R.H.S = \Sigma_B P(A|B,C) . P(B|C) = \Sigma_B \frac{P(A \cap B \cap C)}{P(B \cap C)} \cdot \frac{P(B \cap C)}{P(C)}$ $\Rightarrow$ Since probability of sum of all B = 1.
$= \frac{P(A \cap C)}{P(C)} \cdot \frac{P(C)}{P(C)} = \frac{P(A \cap C)}{P(C)} = P(A|C) = L.H.S.$ //

(b) Suppose that 0.5% men and 0.25% women are color-blind. A person is chosen randomly at the university where the number of men is twice of that of women. The chosen person is color-blind. What is the probability that the person is male?

[2 points]

$P(\text{Male}|\text{colorblind}) = \frac{P(\text{color-blind}|\text{Male})\,P(\text{Male})}{P(\text{colorblind blind}|\text{Male})\,P(\text{Male}) + P(\text{colorblind}|\text{Female})\,P(\text{Female})}$

$= \frac{(0.005) \times 2/3}{(0.005 \times 2/3 + 0.0025 \times 1/3)} = \frac{1/3\,(0.010)}{1/3\,(0.0125)} = 0.80$ //

Therefore, the probability that the person is Male = 80%.

(c) Consider the probability density function $f_{X,Y}(x, y)$ over a 2-dimensional random variable $[X, Y]$.

$$f_{X,Y}(x,y) = \begin{cases} c(x+y^2) & 0 \le x \le 1 \text{ and } 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Here, c is a constant appropriate for $f_{X,Y}(x, y)$ to be a density function. Find $P(X < \frac{1}{2} | Y = \frac{1}{2})$

[3 points]

$f_{X,Y}(x,y) = c(x+1/4)$, for $x < 1/2$.

Pdf for $f_{X,Y}(x,y) = \int_{-\infty}^{\infty} c(x+1/4)\,dx = \int_{0}^{1/2} c(x+1/4)\,dx = \left[c\left(x^2/2 + x/4\right)\right]_0^{1/2}$

$= c(1/8 + 1/8) = \boxed{c/4}$

5: Nearest neighbors to the rescue. (TA:- Jingwei Shen)

(a) Consider two classes C_1, C_2 in the two-dimensional space. The data from class C_1 are uniformly distributed in a circle of radius r. The data from class C_2 are uniformly distributed in another circle of radius r. The centers of two circles are at a distance greater than $4r$. Show that the accuracy of 1-NN is greater than or equal to the accuracy of k-NN, where k is an odd integer and $k \geq 3$. *[3 points]*

The accuracy of 1NN ≥ accuracy of KNN because, we've given the data points are uniformly distributed but there is no relation defined between C_1 and C_2 datapoints. There are 2 cases:

Case(1): When the new datapoint lies at P_1 (exact mid point of the distance); then 1-NN assigns to the nearest class and the accuracy remains same.

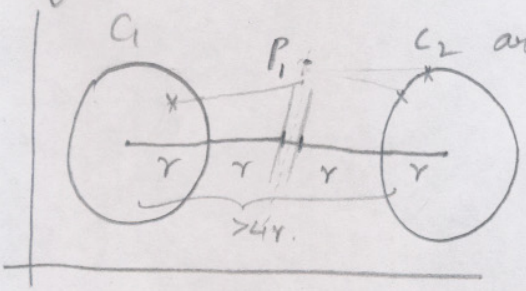
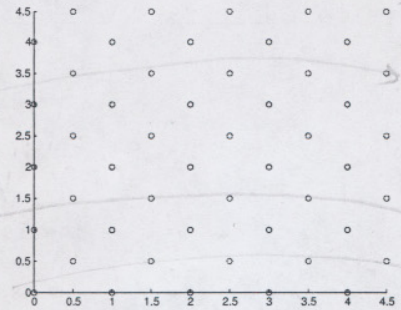


Figure 1: Q4 Dataset

Case(2): New data point at P_1, say C_1 (and is the desired one), if there are 2 datapoints closer to P_1 than one datapoint in C_1 and would classify as C_2.



K-NN assigns to the same class as that of 1-NN. This may not be accurate classification, but the accuracy remains same.

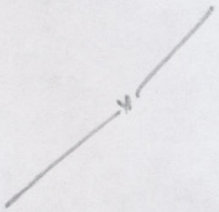
1NN assigns to nearest classifier

KNN might classify wrongly

(b) In the dataset shown in figure 1, what is the leave-one-out accuracy of the k-NN method when $k = 2$? Remember that a data point cannot be considered its own neighbor since it is left out. *[2 points]*

∴ Accuracy of 1NN ≥ KNN

Leave-out accuracy of any data point in the above figure will be 0 % because the nearest two data points will always belong to the other (wrong) class. Thus, the data point would be classified wrongly.



(c) In this problem, explain briefly why you think k-NN performs worse than randomly guessing, which has an accuracy near 50%? [2 points]

For any K, the number of data points of the wrong (other) class will be always greater than that of the other class and accuracy will be less than the accuracy of random guessing (ie., less than 50%.)

6: A tree about the important things in life. (TA:- Abhinav Maurya)

The following dataset will be used to learn a decision tree for predicting whether a person is Happy (H) or Sad (S) based on the color of their shoes, whether they wear a wig and the number of ears they have.

Color	Wig	Num. Ears	Emotion
G	Y	2	S
G	N	2	S
G	Y	2	S
B	N	2	S
B	N	2	H
R	N	2	H
R	N	2	H
R	N	2	H
R	Y	3	H

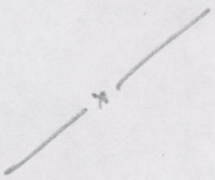
(a) What is Entropy(Emotion | Wig=Y)? [1 points]

$\text{Entropy}(\text{Emotion} \mid \text{Wig}=Y) = H(2/3, 1/3) = -2/3 \log_2(2/3) - 1/3 \log_2(1/3)$

$= 0.91493 \simeq 0.92$

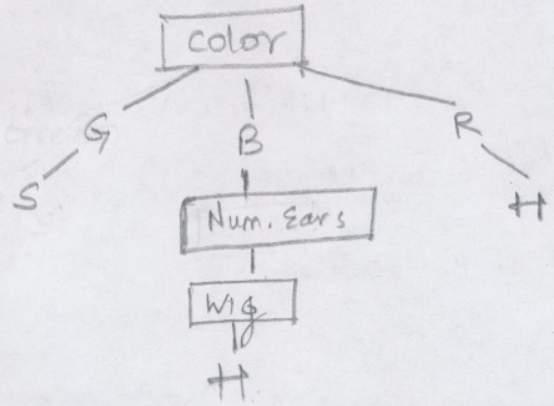
(b) Which attribute would the decision-tree building algorithm choose to use for the root of the tree (assume no pruning)? [2 points]

The 'color' attribute will be chosen as the root by the decision tree algorithm.



(c) Draw the full decision tree that would be learned for this data (assume no pruning).

[3 points]



(d) What would be the training set error for this dataset? Express your answer as the percentage of records that would be misclassified.

[2 points]

11.11%

7: Digging up the dense binary tree. (TA:- Abhinav Maurya)

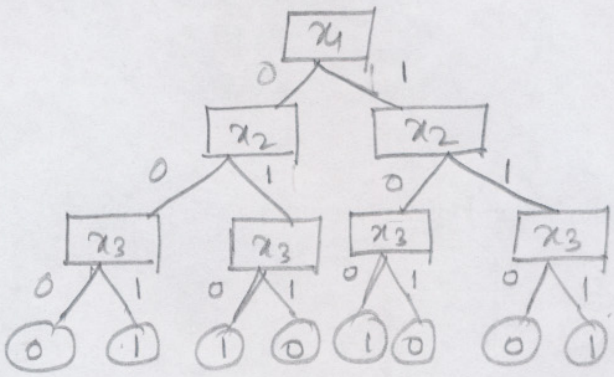
Consider the following data with three binary attributes, where x^i denotes the i^{th} datapoint, x_j denotes the j^{th} feature of the datapoint, and y denotes the class label:-

	x_1	x_2	x_3	y
x^0	0	0	0	0
x^1	0	0	1	1
x^2	0	1	0	1
x^3	0	1	1	0
x^4	1	0	0	1
x^5	1	0	1	0
x^6	1	1	0	0
x^7	1	1	1	1

(a) Draw the decision tree for the above dataset using the entropy criterion to decide node splits (assume no pruning).

- next page -

[3 points]



(b) Decision trees are often pruned so that they can better generalize for prediction on the test set. Do you think you could prune any of the lower levels of the above decision tree used to predict the XOR of 3 binary digits? Give reasons for your decision.

No, the above decision tree cannot be pruned at the lower level, because pruning of any lower level nodes will reduce of the decision tree for predicting the XOR of 3 binary digits. [2 points]

(c) Considering a generalization of the above problem, let's say that we train a decision tree without any pruning to output the XOR function using *all* possible binary strings of length n. Out of the decision tree and KNN classifier (using l_1 distance and $k = 1$), which one would be more accurate when the test samples are also binary strings of length n?

Both KNN and decision tree will be equally accurate. Because, KNN needs to go through the same distance (l_1) as that of the distance taken by the decisiontree. [2 points]

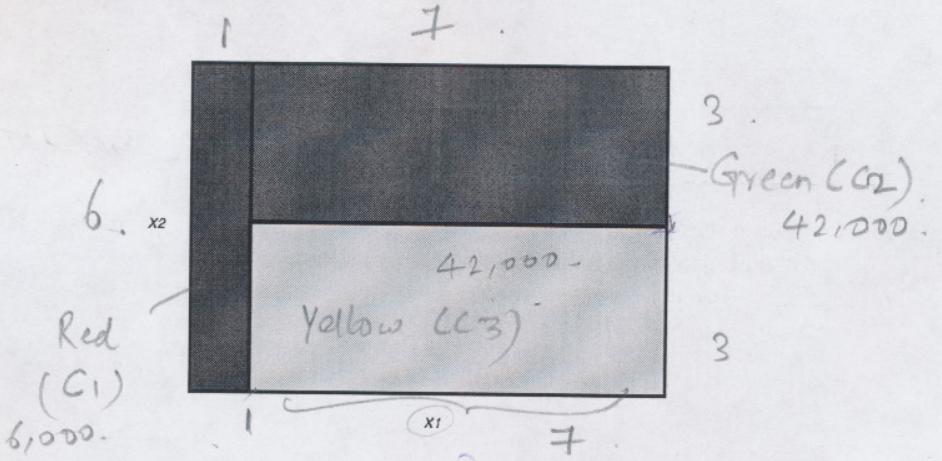
(d) Out of the decision tree and KNN classifiers considered in the previous question, which one will take lesser time to predict the output label of a new test datapoint? Why? (Hint: Note that there are 2^n possible datapoints due to n binary input features. Consider the number of nodes traversed by the decision tree and the number of distance computations performed by the KNN classifier to predict the label of a test datapoint with n binary input features.)

The decision tree will take lesser time to predict the output label of a new test datapoint when compared to the KNN classifier. As, KNN classifier requires to calculate all the nearest data points in the space (2^n), therefore it takes $O(2^n)$; while decision tree takes only $\log 2^n$ time, and its order would $O(n)$ i.e., $O(n) < O(2^n)$ // [2 points]

8: On the hardness of learning optimal binary decision trees (TA:- Abhinav Maurya)

In figure 2, assume that the rectangular region consisting of two features x_1 and x_2 is densely packed with points. The red, green, and yellow subrectangles represent the three classes C_1, C_2, and C_3 of datapoints. The x_1 X x_2 dimensions of the red, green, and yellow rectangles are 1 X 6, 7 X 3, and 7 X 3 respectively. The red rectangle is uniformly populated with 6,000 datapoints of class C_1. The green rectangle is uniformly populated with 42,000 datapoints of class C_2. The yellow rectangle is uniformly populated with 42,000 datapoints of class C_3.

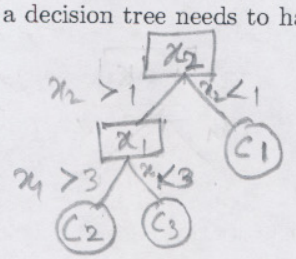
Figure 2: A 2D dataset with three classes



(a) What is the minimum number of nodes that a decision tree needs to have in order to classify the above dataset correctly?

Two //

Tree:



[2 points]

(b) What is the number of nodes in the decision tree trained on the above dataset using the entropy criterion?

x_1	x_2	class
>3	<1	C1
>3	>1	C3
<3	<1	C1
<3	>1	C2

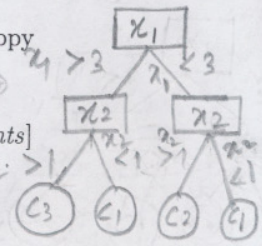
Three //

$H(class) = 1.2867$

$H(class|x_1) = 0.17668$ | $H(class|x_2) = 0.092900$

$I(class; x_1) = 0.9333$ | $I(class; x_2) = 0.086708$

Sample tree ⇒

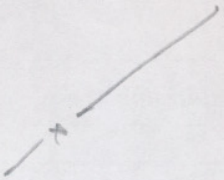


[2 points]

(c) Are the number of nodes in the two cases identical or different? Why do you think that is?

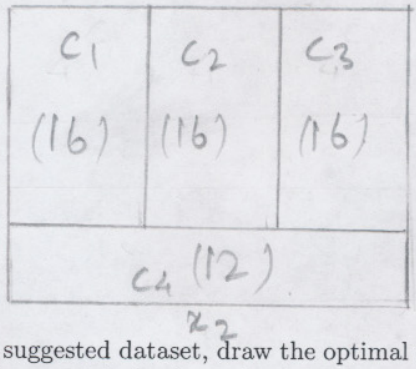
The number of nodes are different in both these cases, because the first uses an optimal strategy but whereas the second follows the entropy criterion and it chooses x_1 as the first node split (due to higher info. gain) and continues to build in a greedy fashion.

[3 points]



(d) Construct another toy dataset where the entropy gain criterion leads to a suboptimal decision tree i.e. one with more nodes than another tree of comparable accuracy. Your dataset should have at least four labels and be sufficiently different from the given toy dataset.

Custom toy dataset:



Two features = x_1, x_2

Total datapoints = 60

Four classes:

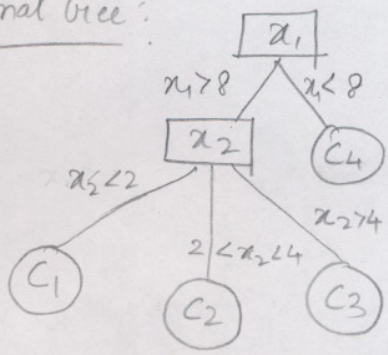
$C_1 = C_2 = C_3 = 8 \times 2$ (each)
= 16 datapoints (each)

$C_4 = 2 \times 6$
= 12 datapoints

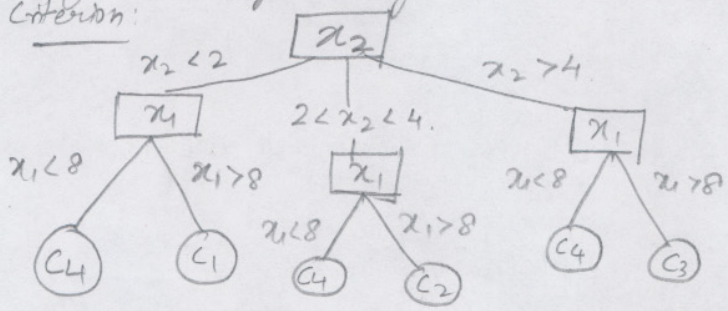
[3 points]

(e) For your suggested dataset, draw the optimal decision tree as well as the decision tree obtained using the entropy minimization criterion.

Optimal tree:



Tree from Entropy Minimization Criterion: [3 points]



(f) A decision tree can classify the dataset in figure 2 with 100% test accuracy (assuming that there is no label noise). What are the general conditions on a dataset under which a decision tree can provide 100% test accuracy? (Hint: Each internal node of a decision tree performs a split based on a single feature. Think about the class of separation functions such a decision tree entails.)

[3 points]

For 100% test accuracy, the values in the datasets shouldn't overlap over two (or) more classes. for the same set of given features.

Total: 70