Homework 4   
Answer questions 8, 9, and 11 in the Exercises section from Chapter 7 of the Larose   
textbook.   
Also, answer questions 1, 2, 3 and 4 in the Exercises section from Chapter 8 of   
the Larose textbook.   
In your submission, you must:   
• Clearly answer each question in a Word document and submit the Word   
document to the Module 5 Homework assignment.   
• Write your answers completely and elaborate on your ideas as much as   
possible.

Chapter 7

1. **The example in the text regarding using the *k*-nearest neighbour algorithm for estimation has the closest record, overwhelming the other records in influencing the estimation. Suggest two creative ways that we could dilute this strong influence of the closest record.**

**Ans. Ways that we could dilute strong influence of the closest record—**

1. We can allow data itself to follow a cross validation procedure to find optimal values z1, z2, z…zm for axis stretching. i.e. by various randomly choosing training data sets to solve the problem
2. We can try various values of k with different randomly selected training sets to train the model. This will help to choose value of k that minimize classification or estimation error.
3. Or if we want to dilute the influence of closes record we should increase the k, or else we end up leading the algorithm towards overfitting the model, that will tend to memorize the training dataset and choosing nearest neighbour values. Hence it would be good to increase the value of K that is greater that will smooth out any behaviour learnt from the training dataset. But if we again take this value of k too large, locally interesting value behaviour will be overlooked. Hence there need to be balance these considerations when choosing value of k.
4. **Discuss the advantages and drawbacks of using a small value versus a large value for *k*.**

**Ans.   
Advantages of using small value of K –**It will help in eliminating the effect of classification or estimation by outliers or unusual observations or noise. The algorithm will simply return target value of nearest observation.  **Drawback –**Since using a small value of k will return the closest target value of observation, it will lead to overfitting ot the algorithm tending to memorize the training data set instead of generalizing the algorithm to work with other dataset values. **Advantage of using large value of k –**Since small value of k will choose nearest neighbour value and overfit the model, choosing a larger value of k will help in smoothing out this effect.   
**Drawback -**Also if we take a value of k that is too large , some of the locally weighted average behaviours will be overlooked.

1. **What is locally weighted averaging, and how does it help in estimation?**

**Ans.**Locally weighted averaging is a method used for estimation and prediction and for continuous valued target variables for classification. Locally weighted averaging estimates using the inverse square of the distances for weights.

The estimated target value of y^ is calculated as **–**

Text

Description automatically generatedwhere Wi = 1/ d(new\_record\_to\_find, *xi*)2 for existing records *x*1*, x*2*,* ... *, xk*.

**Chapter 8**

1. **Describe the possible situations when no further splits can be made at a decision node.**

**Ans.** Situations when no further splits can be made at a decision node   
1. Savings with medium category have good credit risks and 100% accuracy with the data set hence does not require further classification or further splits to be made at a decision node.   
2.Education - Graduated, not-graduated  
3. Marital Status – Married , Unmarried  
In above scenarios there is no need to further splits since a single sdecision can be made based on the classification of dataset category. Hence, the decision tree reaches to its best Purity stage

1. **Suppose that our target variable is continuous numeric. Can we apply decision trees directly to classify it? How can we work around this?**

**Ans. We cannot apply decision trees directly to target variable that is continuous numeric because--**- Decision tree algorithm represent supervised learning and require pre-classified target variables.   
- if the target variable is continuous numeric there is no varied dataset that will provide with.a healthy scross section of types of records for which classification can be applied in the future.   
- Decision trees learn by example, if they are lacking with definable subset of records, classification and prediction of this subset will be difficult or impossible  
- The target attribute classes must be discrete. That means for our target varoiable that is continuous numerc wont be a good dataset to apply for the decision tree analysis.  
- Target variable should be taken on values that are clearly distinguished into definitive classes.

Hence, in order to work around such data set we need to have group a category or set a specific range for which the values of continuous numeric target variable will be applicable for to categorise.

1. **True or false: Decision trees seek to form leaf nodes to maximize heterogeneity in each node.**

**Ans. False**

1. **Discuss the benefits and drawbacks of a binary tree versus a bushier tree.  
   Table

   Description automatically generated**

**Ans. Benefits of binary tree vs a bushier tree—**

* **CART – Classification and Regression Tree always produce a binary decision trees where as c4.5 produces a tree of more variable shape or bushier tree.**
* **For categorical attributes, C4.5 produces a separate branch for each value of categorical attribute. It creates more bushiness in the tree than desired.**
* **The Decision trees are easier to construct and interpret the given data.**
* **Decision rules are constructed from a decision tree by traversing any path from root node to any leaf.**
* **Complete set of decision rules generated by a decision tree is equivalent to decision tree itself**

In given dataset following are types of variables -  
**Numerical variable** – Age, Salary  
**Categorical variable** – Occupation, Gender

Since we are discretizing based on salary following will be **our C5.0 decision tree based** on the dataset given–

* Salary < $35,000
  + Age range 30<=25
    - Gender – Male
      * Age: 25 Occupation – Service [Sal $25, 000]
      * Age: 25 Occupation— Staff [Sal $25,000]
* Salary > = $35,000
  + Salary in the range $35,000<= $45,000
    - Age range >= 45
      * Gender – Female
        + Age: 50 Occupation— Staff [Sal $40,000]
    - Age range 45<=30
      * Gender – Male
        + Age: 33 Occupation – Service [Sal $35, 000]
        + Age: 30 Occupation— Sales [Sal $40,000]
    - Age range 30<=25
      * Gender – Male
        + Age: 25 Occupation— Management [Sal $45,000]
        + Age: 26 Occupation— Management [Sal $45,000]
  + Salary in the range $45,000<= $55,000
    - Age range >= 45
      * Gender – Female
        + Age: 45 Occupation – Service [Sal $48, 000]
    - Age range 45<=30
      * Gender – Female
        + Age: 40 Occupation— Sales [Sal $50,000]
  + Salary > $55,000
    - Age range> =45
      * Gender – Female
        + Age: 65 Occupation— Management [Sal $65,000]
        + Age: 45 Occupation— Management [Sal $70,000]

The above is C5.0 representation of given dataset

Lets, do it with **Decision Tree based on CART method**

As compared to both of above implementation of claasiification methods it is clearly observed that if we further classify data according to CART method , it is visually appealing, understandable, and easy to construct as well as Interpretable with the given dataset and boundaries set.   
- CART uses strictly binary splits. As seen above we have to further divide in some category as Age groups in C5 method, where as here we can efficiently dicelty classify the gven dataset to fit in the decision tree.   
- Drawback of C5.0 Strategy is that it has lead to bushy tree as seen above with many leaf node containing few or no records.   
- The records with low capital gains or loss for the next slpit creates a separate branch for each categorical field values.   
- As compared to the decision Tree – using CART method the data was restricted to binary splits only and helps in better understanding the data structure without more details in to the depths.