Homework 3   
Answer questions 16 and 17 in the Exercises section from Chapter 2 of the Larose   
textbook.   
Also, answer questions 1 through 7 in the Exercises section from Chapter 7 of the Larose   
textbook.   
In your submission, you must:   
• Clearly answer each question in a Word document and submit the Word   
document to the Homework 3 assignment.   
• Write your answers completely and elaborate on your ideas as much as   
possible.

**Chapter 2**

Use the following stock price data (in dollars) for Exercises 13–18.

X = 10 7 20 12 75 15 9 18 4 12 8 14

**16.** Calculate the midrange stock price.

**Ans.**

Midrange stock price = (max(x)+min(x))/2 = 20+4/2 = 12

**Midrange stock price for X is 12.**

1. Compute the *Z*-score standardized stock price for the stock worth $20.

**Ans.**

Z-score standardized stock price for $20—  
  
Mean(X) = 17  
SD(X) = 18.0277563

Z-score = *X* − mean(*X*) = (20 – 17) / 18.0277563 = 3/ 18.0277563 **= 0.166410059581291** SD(*X*)

**Z-score standardized price for stock $20 is - 0.166410059581291**

**Chapter 7**

1. Clearly describe what is meant by classification.

**Ans** – Classification –

* There are target categorical variables that are partitioned in predetermined classes or categories.
* Like income bracket which can be classified into high , medium and low income bracket with specific range defined.
* This gives a clear distinction between the values and their respective categories they fall into.
* If for example based on the characteristics associated to a person, such as age, gender, occupation it is easy to classify the persons income to classified categories.
* This task of classifying the person to the defined income category set is called as data mining method and technique. Hence, classification is a better data mining technique.
* For example – banking, education, medicine,

1. What is meant by the term *instance-based learning*?

**Ans.**

Instance based learning is meant for k-nearest neighbour algorithm, it is important to have access to database having set of various attributes and values with examples, in K-nearest neighbor algorithm we use a training data set that is stored so that the classification for other unclassified records can be done easily based on predefined dataset. Rare classification to be represented sufficiently so algorithm not only predict common classification. Data needs to be balanced with large percentage of less common classification. So as to reduce the proportion of records with more common classifications.   
Maintaining this rich database for easy access could be problematic if there is restriction on main memory space. Main memory may fill up and access to auxiliary storage is slow. Hence k-nearest neighbour method may be helpful to store/retain only those points which are at the boundary.

For example if all records are greater than 19 could be omitted from database without loss of classification accuracy because all records in this region will be classified differently knowing their boundary values.

1. Make up a set of three records, each with two numeric predictor variables and one categorical target variable, so that the classification would not change regardless of the value of *k*.

**Ans** -

|  |  |  |
| --- | --- | --- |
| Weight | Age | Overweight |
| 99 | 22 | TRUE |
| 58 | 21 | FALSE |
| 120 | 23 | TRUE |

Consider a set of three records having two numeric predictor variables and one categorical target variable – Overweight.

1. Refer to Exercise 3. Alter your data set so that the classification changes for different values of *k*.  
   **Ans.**

|  |  |  |
| --- | --- | --- |
| Weight | Age | Overweight |
| 99 | 22 | TRUE |
| 58 | 21 | FALSE |
| 120 | 23 | TRUE |

This classification changes according to K-NN

K=1 ,

When a new record is added it will be classified as True or false based on the variable closest neighbor

K-2

There can be a tie since the nearest neighbor are a,b and b,c hence additional parameter required

If K=3, New record will be classified as Ture o or false based on confidence.

1. Refer to Exercise 4. Find the Euclidean distance between each pair of points. Using these points, verify that Euclidean distance is a true distance metric.

**Ans.**  
Eucledian distance formula = d =√[(x2– x1)2 + (y2– y1)2]

Given the data we need to standardize the data, I have used Min-max standardization to normalize the data.

Minmax = 

Using Excel to get the new values of Age and Weight with Min Max calculation –

|  |  |  |  |
| --- | --- | --- | --- |
| Person | Weight | Age | Overweight |
| 1 | 99 | 22 | TRUE |
| 2 | 58 | 21 | FALSE |
| 3 | 120 | 23 | TRUE |

New normalized values –

|  |  |
| --- | --- |
| Weight | Age |
| 0.66129032 | 0.5 |
| 0 | 0 |
| 1 | 1 |

= √(0.66129032-0) 2 + (0-1) 2 +(1-0.66129032) 2 +(0.5-0) 2 +(0-1) 2 +(1-0.5) 2

= √ 0.437304887325702+1 +0.114724247325702 +0.25 +1 +0.25

= √3.052034887325702

**= 1.747007409064341**

1. Compare the advantages and drawbacks of unweighted versus weighted voting.

These are both Combination function – unweighted and weighted function.

|  |  |
| --- | --- |
| Unweighted Voting | Weighted Voting |
| Unweighted voting is based on no votes for the sample neighbours | Weighted classification is based on more coser similar records with their weights of having possibility of the specified category |
| For simple unweighted voting, their distance from new record no longer matters, it is simple one record and one vote to classify further based on k-records that have been chosen. | For weighted voting the neighbours that’s are closer or more similar to new record will be weighted more heavily than distant neighbours. |
| Unweighted Voting has number of k-records based on which the voice for classifying the new record | In Weighted Voting closer neighbors have a larger voice in the classification decision than do more distant neighbors. |
| In Unweighted Voting there could be a tie among the three classifications represented by the records like that for k=3 (light, medium dark points in dataset having no weighted bias) | Weighted voting also makes it much less likely for ties to arise. |
| Decide on value of k, how many records will have a voice in classifying the new record | In weighted voting, the influence of a particular record is inversely proportional to the distance of the record from the new record to be classified. |
| Compare the new record to the k nearest neighbors that is, to the k records that are of minimum distance from the new record in terms of the Euclidean distance or whichever metric the user prefers. | the analyst may choose to apply weighted voting, where closer neighbors have a larger voice in the classification decision than do more distant neighbors. |
|  |  |

1. Why does the database need to be balanced?

* Access to a rich database full of as many different combination of attribute values as possible
* It is important that rare classifications be represented sufficiently, so that the algorithm does not only [predict common classification but something that are trick to classify as well.
* Therefore, data set would need to be balanced with sufficiently large percentage of less common classification
* If there is more diverse datasets it will help train a better model and bright a balanced solution to reduce the proportion of records with more common classifications.