1. Consider the different classification methods we have discussed in class. Name one classification method for each of the following scenarios and briefly explain why.

(a) A classification method that supports incremental data.
(b) A classification method whose classification decisions are easy to interpret.
(c) A classification method whose classification decisions are not easy to interpret and has many parameters to train.

a) Sol.
A classification method that supports incremental data is Bayesian Classification.
In Bayesian Classification, in case of new data we can just update present model instead of reconstructing the whole model. It is based on Baye's theorem on probability. When new data or features are added, it determines bayesian classifier for each additional feature in terms of existing features and update probabilities. Bayesian Classification offers results comparable to decision tree.

b) Sol.
Decision tree classification method involves classification decisions which are easy to interpret.
Decision tree creates visual representation of all possible outcomes including decisions. This visual representation makes it easier to understand and is easy to interpret. Decision tree is one of the best classification method when it comes to performance.

c) Sol.
Neural Network as Classifier are not easy to interpret and has many parameters to train.
The weakness of this classification method is long training time and its parameters are determined empirically. This results in poor interpretability.

References:
1.
https://moodle.cs.colorado.edu/pluginfile.php/46085/mod_resource/content/2/dm-lec12-mar01-s16.pdf
2.
https://moodle.cs.colorado.edu/pluginfile.php/46087/mod_resource/content/5/dm-lec13-mar03-s16.pdf

2. Given a set of points A1 (2, 10), A2 (2, 5), A3 (8, 4), B1 (5, 8), B2 (7, 5), B3 (6, 4), C1 (1, 2), C2 (4, 9), where (x, y) represents location and
the distance function is Euclidean distance. Our goal is to cluster these points into three clusters using k-means clustering, and the initial three centroids
are A1, B1, and C1. Show the key steps of the first round of clustering.

Sol.
We are given that initial centroids are at the points A1, B1, C1.
Lets say Init_Cent_A = A1, Init_Cent_B = B1 and Init_Cent_C = C1
Let's say dist(x,y) denotes Euclidean distance between x and y.

First Round of Clustering:

From A1:

First, determine euclidean distances from A1 to all initial centroids.
dist(A1, Init_Cent_A) = 0 since A1 is Init_Cent_A
dist(A1, Init_Cent_B)= $\sqrt{13}$
dist(A1, Init_Cent_C)= $\sqrt{65}$

dist(A1, Init_Cent_A) is smaller than remaining values and closer to first centroid so A1 belongs to cluster 1.

From A2:

First, determine euclidean distances from A2 to all initial centroids.
dist(A2, Init_Cent_A) = 5
dist(A2, Init_Cent_B) = $\sqrt{18}$
dist(A2, Init_Cent_C) = $\sqrt{10}$

dist(A2, Init_Cent_C) is smaller than remaining values and closer to third centroid so A2 belongs to cluster 3.

From A3:

First, determine euclidean distances from A3 to all initial centroids.
dist(A3, Init_Cent_A) = $\sqrt{72}$
dist(A3, Init_Cent_B) = 5
dist(A3, Init_Cent_C) = $\sqrt{53}$

dist(A3, Init_Cent_B) is smaller than remaining values and closer to second centroid so A3 belongs to cluster 2.

From B1:

First, determine euclidean distances from B1 to all initial centroids.
dist(B1, Init_Cent_A) = $\sqrt{13}$
dist(B1, Init_Cent_B) = 0 since B1 is Init_Cent_B
dist(B1, Init_Cent_C) = $\sqrt{52}$

dist(B1, Init_Cent_B) is smaller than remaining values and closer to second centroid so B1 belongs to cluster 2.

From B2:

First, determine euclidean distances from B2 to all initial centroids.
dist(B2, Init_Cent_A) = $\sqrt{50}$
dist(B2, Init_Cent_B) = $\sqrt{13}$
dist(B2, Init_Cent_C) = $\sqrt{45}$

dist(B2, Init_Cent_B) is smaller than remaining values and closer to second centroid so B2 belongs to cluster 2.

From B3:
First, determine euclidean distances from B3 to all initial centroids.
dist(B3, Init_Cent_A) = $\sqrt{52}$
dist(B3, Init_Cent_B) = $\sqrt{17}$
dist(B3, Init_Cent_C) = $\sqrt{29}$

dist(B3, Init_Cent_B) is smaller than remaining values and closer to second centroid so B3 belongs to cluster 2.

From C1:
First, determine euclidean distances from C1 to all initial centroids.
dist(C1, Init_Cent_A) = $\sqrt{65}$
dist(C1, Init_Cent_B) = $\sqrt{52}$
dist(C1, Init_Cent_C) = 0 since C1 is Init_Cent_C

dist(C1, Init_Cent_C) is smaller than remaining values and closer to third centroid so C1 belongs to cluster 3.

From C2:
First, determine euclidean distances from C2 to all initial centroids.
dist(C2, Init_Cent_A) = $\sqrt{5}$
dist(C2, Init_Cent_B) = $\sqrt{2}$
dist(C2, Init_Cent_C) = $\sqrt{58}$

dist(C2, Init_Cent_B) is smaller than remaining values and closer to second centroid so C2 belongs to cluster 2.

Given points are clustered into required 3 clusters (mentioned below) using k-means clustering.
Cluster 1: {A1}, Cluster 2: {A3, B1, B2, B3, C2}, Cluster 3: {A2, C1}

3. Provide an application example for each of the three types of outliers: global outliers, contextual outliers, and collective outliers.
Briefly describe the related attributes and how the specific type of outliers may be defined and detected.

Sol.
Outlier represents an object or instance that significantly deviates from normal behavior.
Three types of outliers are global outliers, contextual outliers, and collective outliers.

Global Outliers:

If an object deviates from the rest of the data objects then object is called as global outlier. Real time examples are Intrusion detection in computer networks and credit card fraud detection where abnormal buying patterns can be observed. To detect these type of outliers, consider values of data objects to determine any deviation if exists.

Contextual outliers:
In contextual outliers, object significantly deviates based on selected context.
Example is whether condition or temperature in specific city because it is based on seasons (summer, winter, spring, fall). To detect these type of outliers, consider context or behavior of an object.

Each object can be divided into below two sets.
Contextual attributes:
The contextual attributes determine or defines the context of the object. For example, contextual attributes are the longitude and latitude of a particular location in any data sets (related to geo spatial or global maps). Time is also a contextual attribute which determines the position of an object.

Behavioral attributes:
It defines the non-contextual characteristics or behavior of an object. For example, describing the amount of rainfall or humidity at any location is a behavioral attribute.

Collective outliers:
In collective outliers, collection of data objects deviates from the rest of the whole data set even though individual object may not be considered as outlier.
Real time example is, human electrocardiogram (ECG) report where same low value is observed for long time continuously. This behavior represents contraction and here low value is not outlier but collection of low values occurring repeatedly is an outlier.

To detect these type of outliers, consider overall behavior of group of objects not just individual object and explore relationship among data objects.

4. Briefly describe one data mining tool that you have used either in this course or in other settings. What did you use this tool for? What are the key strengths and possible limitations of this tool?

Sol. Data mining tool I have used is R language. I have used R to implement and generate reports for my final project. My project topic was Movie Recommender System and used R to process large movie data set. Generated plots and reports using R for movie data set. Association rules are mined along with logistic regression using R on this data set. Used inbuilt package “arules” to apply apriori algorithm to determine interesting patterns in the project. There are lot of packages available to generate plots. Used “Amelia” package to determine logistic regression on movie data set.

R language is environment for statistical graphics and computing. It is mainly used for data analysis. It even supports statistical techniques like classification, clustering, linear and nonlinear modeling, statistical tests etc.

Strengths of it are listed below.

- Ease of use
- Extensibility
- Widely available packages from various repositories
- Compatible and works well with many other tools
- Runs on multiple operating systems with different hardware

Limitations of it are listed below.

- Lack of documentation makes it harder to work with few packages.
- Speed
- Efficiency
- Memory management
- Lack of security restrict the usage.