CSCI 5502 - Data Mining - Homework 5
Name: Krishna Chaitanya Sripada
Student ID: 104375417
Honor Pledge: On my honor, as a University of Colorado at Boulder student, I have neither given nor received unauthorized assistance on this work.

Ans 1

(a) Given that the P(S|UG) = 0.15, P(S|G) = 0.23, P(G) = 0.2 and thus P(UG)=0.8. To compute P(G|S), according to Bayes' theorem:

$$P(G|S) = \frac{P(S|G) \times P(G)}{P(S|UG) \times P(UG) + P(S|G) \times P(G)}$$

$$P(G|S) = \frac{0.23 \times 0.2}{0.15 \times 0.8 + 0.23 \times 0.2} = 0.2771$$

(b) From the information given, a randomly chosen college student is more likely an undergraduate than a graduate because P(UG) > P(G).

(c) From (a), we have P(G|S) = 0.2771 and therefore P(UG|S) = 1- P(G|S)= 0.7229. Thus, a smoker student is more likely to be an undergraduate student because P(UG|S) > P(G|S).

(d) Firstly, we need to find all the probabilities needed.
P(D|UG) = 0.1 and P(D|G) = 0.3.
By considering independence between students who live in a dorm and those who smoke,
P(D,S|G) = P(D|G) × P(S|G) = 0.3 × 0.23 = 0.069.
P(D,S|UG) = P(D|UG) × P(S|UG) = 0.1 × 0.15 = 0.015.

We need to compute P(UG|D,S) and P(G|D,S) which are calculated below:

$$P(UG|D,S) = \frac{P(D,S|UG)P(UG)}{P(D,S)} = \frac{0.015 \times 0.8}{P(D,S)} = \frac{0.012}{P(D,S)}$$

$$P(G|D,S) = \frac{P(D,S|G)P(G)}{P(D,S)} = \frac{0.069 \times 0.2}{P(D,S)} = \frac{0.0138}{P(D,S)}$$

Since P(G|D,S) > P(UG|D,S), he or she is more likely to be a graduate student.

Ans 2

Given the dataset containing 6 objects = {93, 137, 48, 71, 162, 119} and the means/initial centroids are 100 and 150.
We take the 6 objects in the dataset and add them based on the Euclidean distance the object and the centroid. Let centroid of C1 be G1 and centroid of C2 be G2. Therefore, 93 is closer to G1 and thus is added to C1. 137 is closer to G2 and is added to C2. 48 is closer to G1 and is added to C1. 71 is closer to G1 and is thus added to C1. 162 is closer to G2 and is thus added to C2. 119 is closer to G1 and is added to C1.

Thus after Iteration-1, C1 = {93, 48, 71, 119} and C2 = {137, 162}.

By recalculating the centroids, we get G1 = $\frac{93+48+71+119}{4}$ = 82.75 and G2 = $\frac{137+162}{2}$ = 149.5

By reassigning the objects, we get, 93 is closer to G1 and it remains in C1. 48 is closer to G1 and it remains in C1. 71 is closer to G1 and it remains in C1. 119 is closer to G2 and is removed from C1 and added to C2. 137 is closer to G2 and it remains in C2. 162 is closer to G2 and it remains in C2.

Thus after Iteration-2, C1 = {93, 48, 71} and C2 = {119, 137, 162}.

By recalculating the centroids, we get G1 = $\frac{93+48+71}{3}$ = 70.7 and G2 = $\frac{119+137+162}{3}$ = 139.3

By reassigning the objects, we get, 93 is closer to G1 and it remains in C1. 48 is closer to G1 and it remains in C1. 71 is closer to G1 and it remains in C1. 119 is closer to G2 and it remains in C2. 137 is closer to G2 and it remains in C2. 162 is closer to G2 and it remains in C2.

Thus after Iteration-3, C1 = {93, 48, 71} and C2 = {119, 137, 162}.

Since there is no change in C1 and C2 in Iteration-2 and Iteration-3, we stop here.

Ans 3

1. **Global Outlier:** An example can be the intrusion detection in computer networks where if the communication behavior of a computer is very different from the normal patterns (i.e., a large number of packages is broadcast in a

short time), this behavior may be considered as a global outlier and the corresponding computer is a suspected victim of hacking.

2. **Contextual Outlier:** An example would be in credit card fraud detection, in addition to global outliers, an analyst may consider outliers in different contexts. Consider customers who use more than 90% of their credit limit. If one such customer is viewed as belonging to a group of customers with low credit limit, then such behavior may not be considered as an outlier. However, similar behavior of customers from a high-income group may be considered outliers if their balance often exceeds their credit limit. Such outliers may lead to business opportunities i.e., raising credit limits for such customers can bring in new revenue.

3. **Collective Outlier:** An example is in intrusion detection, a denial-of-service package from one computer to another is considered normal, and not an outlier at all. However, if several computers keep sending denial-of-service packages to each other, they as a whole should be considered as a collective outlier. The computers involved mat be suspected of being compromised by an attack.

4. **Related Attributes:** Attributes of the data objects are divided into two groups:
Contextual Attributes: The contextual attributes of a data object define the object's context.
Behavioral Attributes: These define the object's characteristics and are used to evaluate whether the object os an outlier in the context to which it belongs.

5. **Global Outlier Definition and Detection:** In a given dataset, a data object is a global outlier if it deviates significantly from the rest of the dataset. To detect global outliers, a critical issue is to find an appropriate measurement of deviation with respect to the application in question. In global outlier detection, whether a data object is an outlier depends on the behavioral attributes.

6. **Contextual Outlier Definition and Detection:** In a given dataset, a data object is a contextual outlier if it deviates significantly with respect to a specific context of the object. In contextual outlier detection, the context has to be specified as part of the problem definition. In contextual outlier detection, whether a data object is an outlier depends on not only the behavioral attributes but also the contextual attributes.

7. **Collective Outlier Definition and Detection:** In a given dataset, a subset of data objects form a collective outlier of the objects as a whole deviate significantly from the entire dataset. In collective outlier detection, we have to consider not only the behavioral attributes of individual objects but also that of the groups of objects. To detect collective outliers, we need background knowledge of the relationship among the data objects such as distance or similarity measurements between objects.

Ans 4

One of the data mining tools that I have used is "RapidMiner". It is an extremely rich tool built on Java that performs all the basic data mining operations that are required for data processing. RapidMiner offers an integrating environment with visually appealing and user-friendly GUI. Everything in RapidMiner is focused on processes that may contain subprocesses. Processes contain operators in the form of visual components. Operators are implementations of Data Mining algorithms, data sources, and data sinks. The data flow is constructed by drag-and-drop of operators and by connecting the inputs and outputs of corresponding operators. RapidMiner also offers the option of application wizards that construct the process automatically based on the required project goals (e.g. direct marketing, churn analysis, sentiment analysis). There are tutorials available for many specific tasks so the tool has a stable learning curve.

I have used this tool because of its ability to import csv files that are extremely heavy i.e., almost 1GB. Although many other tools are available for importing csv content, this tool provides support for heavy file sizes. The key strengths of this tool are with its basic set of operators and it is the extensions that make it even more useful. Popular extensions include sets of operators for text mining, web mining, time series analysis, etc. Most of the operators from Weka are also available through extension, which increases the number of implemented Data Mining methods. The tool has very few limitations. The most important one is the transition to a novel model of business. It remains to be seen whether the transition to proprietary license will limit the number of its users, but it may not be helpful. The support for deep learning methods and some of the more advanced specific machine learning algorithms (e.g. extremely randomized trees, various inductive logic programming algorithms) is currently limited. However, big data analysis via Hadoop cluster (Radoop) is supported.

Few concepts learnt in class that are not supported in "RapidMiner" are:

- BIRCH Hierarchial clustering
- Partition Around Medoids as part of Centroid based clustering.
- Fuzzy clustering