**Problem 1 (10 points). Answer the following questions. Your answer must be clear, unambiguous, and accurate.**

**(1). Describe the difference between classification and clustering. Show one real-world example of classification and one real-world example of clustering.**

Classification is the process of predicting the predefined label of a given data point based on its features. Essentially, we have labeled training data, and we utilize this data to determine which category a new data point should belong to. It's supervised learning.

eg: Email services use classification to categorize incoming emails as either "spam" or "not spam" based on various features such as sender information, content of the email, or patterns observed in other labeled emails.

Clustering, on the other hand, is the process of grouping data points together based on their similarity, without prior knowledge of labels. It’s unsupervised learning.

eg: Online shopping platforms might use clustering to segment their customers into different groups based on purchasing behavior. These clusters can then be targeted with more personalized marketing campaigns, even when the platform didn’t initially know what these groups would be.

In summary, while classification allocates predefined labels to data points, clustering groups data points based on intrinsic similarities without prior label knowledge.

**(2). We illustrated association rule mining using a market-basket example. Show another realworld example of association rule mining.**

It’s common to see medical facilities recruiting data type professionals. This is a great example.

Scenario: Let's consider a database of patient records in a hospital. Each patient record is a list of medical conditions, symptoms, and medications for a particular patient.

Objective: We want to find out if there are any interesting associations between medical conditions, symptoms, and the medications prescribed.

Association Rule Mining Results: After processing the data, we might discover rules such as:

If a patient has Diabetes AND Hypertension, then they are 80% likely to be prescribed Medication X.

Patients with Symptom A are 75% likely to also report Symptom B.

If a patient is prescribed Medication Y, they are 90% likely to undergo Test Z.

**(3). What are two basic principles of clustering (you must describe these two principles)?**

It's compactness and separability.

Compactness:

This principle emphasizes that the data points within a single cluster should be as close or similar to each other as possible. The idea is to ensure that the intra-cluster distance is minimized.

Separability:

While compactness stresses the closeness of data points within a cluster, separability emphasizes that the different clusters themselves should be as distinct and far from each other as possible. This principle aims to minimize inter-cluster distances.

**(4). What is the most important challenge of mining from a very large amount of data?**

It is the quality of the data. Data quality refers to the accuracy, completeness, consistency, timeliness, and relevance of the data. Poor data quality can lead to erroneous outcomes, misleading conclusions, and a waste of resources.

**Problem 2 (10 points). Consider the following decision tree, which we discussed in the class,**

**Classify the following three persons using the above decision tree:**

Personal Loan: 0

0

1

First person income > 99, but education < 2, so it is 0.

Second person income < 99, so it is 0.

Third person income > 99, education > 2, income < 117, CCAVG > 2.7, so it is 1.

**Problem 3 (10 points). Consider the following transactional database:**

**(1). Determine the supports of the following itemsets:**

**{beer, egg}**

**{chip, egg, milk}**

{beer, egg}:

Transactions containing both beer and egg are C2, C4, and C6.

Support({beer, egg}) = 3/9 = 1/3

{chip, egg, milk}:

Transactions containing chip, egg, and milk are C1 and C8.

Support({chip, egg, milk}) = 2/9

**(2). Calculate the confidences of the following rules:**

**{butter} => {milk}**

**{bread, egg} => {milk}**

{butter} => {milk}:

Transactions containing just butter:

C8 and C9.

Transactions containing both butter and milk:

C8 and C9.

Confidence({butter}=>{milk}) = (2/9)/(2/9) = 1

{bread, egg} => {milk}:

Transactions containing both bread and egg:

C1, C4, C6, and C8.

Transactions containing bread, egg, and milk:

C1, C4, and C8.

Confidence ({bread, egg}=>{milk}) = (3/9)/(4/9) = 3/4