CIND 110 Data Organization for Data Analysts
Assignment 3

Note: This assignment is pen and paper hand work. Thereby, except a calculator, nothing else is expected to be required. So, please do not do any coding in R or Python or Weka and upload that in your submission. You are also expected to show the details of all steps in your calculation in order to score full marks.

1. Association rules: Marks: 30

One of the major techniques in data mining involves the discovery of association rules. These rules correlate the presence of a set of items with another range of values for another set of variables. The database in this context is regarded as a collection of transactions, each involving a set of items, as shown below.

Trans ID	Items Purchased
2001	Meat, Potato, Onion
2002	Meat, Noodle
2003	Noodle, Spinach
2004	Meat, Potato, Onion
2005	Onion, Potato, Noodle
2006	Eggs, Spinach
2007	Eggs, Noodle
2008	Meat, Potato, Salt, Onion
2009	Salt, Spinach
2010	Meat, Potato

1.1 Apply the **Apriori** algorithm on this dataset.

Note that, the set of items is {Meat, Potato, Onion, Noodle, Spinach, Eggs, Salt}.
You may use **0.3 for the minimum support** value.

1.2. Show the rules that have a confidence of 0.8 or greater for an itemset containing three items.

2. Classification: Marks: 40

Classification is the process of learning a model that describes different classes of data and the classes should be pre-determined. Consider the following set of data records:

ID	Age	City	Gender	Education	Profile
101	20-30	NY	F	College	Employed
102	31-40	NY	F	College	Employed
103	51-60	NY	F	College	Unemployed
104	20-30	LA	M	High School	Unemployed
105	41-50	NY	F	College	Employed
106	41-50	NY	F	Graduate	Employed
107	20-30	LA	M	College	Employed
108	20-30	NY	F	High School	Unemployed
109	20-30	NY	F	College	Employed
110	51-60	SF	M	College	Unemployed

Assuming, that the class attribute is Profile, apply a classification algorithm to this dataset.

3. Clustering: Consider the following set of two-dimensional records:

RID	Age	Years of Service
101	30	5
102	50	25
103	50	15
104	25	5
105	30	10
106	55	25

3.1 **Marks: 20**

Use the K-means algorithm to cluster this dataset. You can use a value of 2 for K and can assume that the records with RIDs 103, and 104 are used for the initial cluster centroids.

3.2 **Marks: 10**

What is the difference between describing discovered knowledge using clustering and describing it using classification?