Homework 1

CIS787 – Analytical Data Mining

**Problem 1**

Question 2:

1. Time in AM or PM – Binary, Qualitative, Nominal
2. Brightness as measured by light meter – Continuous, Quantitative, Ratio
3. Brightness measured by people – Discrete, Qualitative, Ratio
4. Angles between 0 and 360 – Continuous, Quantitative, Ratio
5. Bronze, Silver and Gold Medal – Discrete, Quantitative, Ordinal
6. Height above sea level – Continuous, Quantitative, Ratio
7. Number of patients in hospital – Discrete, Quantitative, Ratio
8. ISBN number of books – Discrete, Qualitative, Nominal
9. Ability to pass light – Discrete, Qualitative, Ordinal
10. Military rank – Discrete, Qualitative, Ordinal
11. Distance from the center of campus – Continuous, Quantitative, Interval
12. Density of a substance (g/cm3) – Continuous, Quantitative, Ratio
13. Coat check number – Discrete, Qualitative, Ordinal

Question 5:

If a student id has information, then it can be used as a good predictor. For example – if a student id has the year when the student has joined, then we can use it as a predictor for graduation year. But if the student id is just a random sequence of digits then we cannot use it as a predictor because no information can be derived from it.

Question 13:

1. The algorithm works well for the cases when there are no duplicates for the objects present in the data. If we have enough duplicates, then the output list of the algorithm would be only a list of duplicates. Also, the object that we are using as a reference might not be its own nearest neighbor.
2. To fix the problem as stated above what we can do is only one object should represent the duplicates or a group of duplicates. Also, what we can do is only feed the unique objects to the algorithm without any duplicates.

Question 15:

The first part where there is a formula involved has the number of elements proportional to the size of the group which is the main concept behind stratified sampling which is, we draw equal number of elements irrespective of the group size. The second part is where any element can be drawn out from any group and there is no proportionality, this is the idea behind random sampling.

Question 16:

tf 'ij = tfij  \* log m/dfi

1. If we look at the formula, the number of documents (m) does not vary. The only thing that varies is dfi. So, the words occurring more frequently would make the value very small and because of the log it goes close to 0, while the words occurring rarely would make the overall value greater. If there are more unique words in the document, it would have a greater value of inverse document frequency.
2. This transformation helps to identify or distinguish different documents based on the unique words used in them irrespective of any common words occurring in the documents.

**Problem 2**:

1. Sampling with replacement is more appropriate when we are not interested in frequency of events, rather . If we train a model, we need to ensure that it does not miss key patterns so we use sampling with replacement as we want the data to be replaced and picked up again with some other data points which we might have missed. Sampling without replacement is more appropriate when we are more concerned with the frequency of data rather than observing patterns. For sampling without replacement, the probability for a data point to be picked again decreases as we are not replacing it back in the population.

Sampling with replacement – To periodically check workers if they are practicing their labor properly. In this case the worker if selected once has the same probability to be selected again.

Sampling without replacement – To make teams while playing games by putting everyone’s name in a jar and selecting at random.

To find out three random lottery winners from a pack of 100 people.

1. Sampling is all about probability with and without replacement but when we talk about PCA it involves covariance, variance calculation and a reduction in dimensionality which captures the maximum variance of the data. Also, the principal components are included in such a way that we can get maximum variance which is not the case in sampling. Sampling is just picking out random data points without any covariance or variance calculation and it should be a near accurate representation of the dataset.

**Problem 4:**

• center data (having zero mean):

weka.filters.unsupervised.attribute.Center

• Removing attribute 2 to 4:

weka.filters.unsupervised.attribute.Remove

parameter attributeIndices was given the value 2-4

• Removing all attributes but the last:

weka.filters.unsupervised.attribute.Remove

parameter attributeIndices was given the value ‘last’

invertSelection was set to True

• Reordering attributes 1,2,3,4,5 as 5,4,1,2,3:

weka.filters.unsupervised.attribute.Reorder

parameter Indices was given the value last-4,first-3

• Removing instances with missing values:

weka.filters.unsupervised.instance.RemoveWithValues

attributeIndex was given the value for each attribute

matchMissingValues was set to True

• How is a missing value denoted in an ARFF file?

A missing value is denoted by a question mark (?)

• What does “visualize all” do?

It opens a window showing all possible distributions at once.

• Sampling 20% of your data:

weka.filters.unsupervised.instance.Resample

sampleSizePercent was given the value 20.0

• Removing all instances where the 3rd feature value is equal to ’x’:

weka.filters.unsupervised.instance.RemoveWithValues

attributeIndices was set to 3

nominalIndices was set to the instance number of ‘x’

**Problem 5:**

The setosa category of iris dataset is distinct and can be easily differentiated from others. This can be helpful in classifying flowers from measurements as a boundary can be drawn to separate the setosa category from the other two categories.