INFO411/911: Data mining and Knowledge Discovery
Assignment One (15%)

General Instructions: *Please Read Carefully*

- Submit a single PDF document which contains your answers to the questions of both tasks. All questions are to be answered. A clear and complete explanation and analysis needs to be provided with each answer.
- The PDF must contain typed text of your answers (do not submit a scan of a handwritten document. Any handwritten document will be ignored). The document can include computer generated graphics and illustrations (hand-drawn graphics and illustrations will be ignored).
- The PDF document of your answers should be no more than 12 pages including all graphs and illustrations. If it is over 12 pages, only the first 12 pages will be marked. The size limit for this PDF document is 20MB.
- Late submission will not be accepted without academic consideration being granted.

Overview:

This assignment consists of **two tasks**. There are several questions to be answered for each of the two tasks. You may need to do some research on background information for this assignment. For example, you may need to develop a deeper understanding of writing code in R, or study the general characteristics of GPS, obtain general geographic information about Rome, and study other topics that are related to the tasks in this assignment.

What you need:

- The R software package (RStudio is optional), the file **a1.zip** from the Moodle site.
- You need to install the package `kohonen` and other packages. Please follow Week 3 Lab.
- Task One
  - **`taxi.csv.zip`** that can be downloaded from this link *https://cloudstor.aarnet.edu.au/plus/s/2V9DbYfmcUbCgKf (Caution:* The file `taxi.csv.zip` is about 260MB in size; uncompressed the file is 1.2GB in size!)
- Task Two
  - **`creditworthiness.csv`** inside file `a1.zip`
  - Successful completion of Week 4 Lab and Week 5 Lab. You may use the R-script from the labs as a basis for attempting this question.

Task1

Preface: The analysis of results from urban mobility simulations can provide valuable information for the identification and addressing of problems in an urban road network. Public transport vehicles such as busses and taxis are often equipped with GPS location devices and the location data is submitted to a central server for analysis.

The metropolitan city of Rome, Italy collected location data from 320 taxi drivers that work in the centre of Rome. Data was collected during the period from 01/Feb/2014 until 02/March/2014. An extract of the dataset is found in `taxi.csv`. The dataset contains 4 attributes:

1. ID of a taxi driver. This is a unique numeric ID.
2. Date and time in the format `Y:m:dH:m:s.msec+tz`, where msec is micro-seconds, and tz is a time-zone adjustment. (You may have to change the format of the date into one that R can

understand).

3. Latitude
4. Longitude

For a further description of this dataset: http://crawdad.org/roma/taxi/20140717/

Purpose of this task:
Perform a general analysis of this dataset. Learn to work with large datasets. Obtain general information of the behaviour of some taxi drivers. Analyse and interpret results.

Questions: (7 marks)

By using the data in `taxi.csv`, perform the following tasks:

a) Plot the location points (2D plot using all of the latitude, longitude value pairs in the dataset). Clearly indicate points that are invalid, outliers or noise points. The plot should be informative! Clearly explain the rationale that you used when identifying invalid points, noise points, and outliers.

Remove invalid points, outliers and noise points before answering the subsequent questions.

b) Compute the minimum, maximum, and mean location values (with respect to the longitude and latitude, respectively.)

c) Obtain the most active, least active, and averagely active taxi drivers (i.e., most time driven, least time driven, and mean time driven). Explain the rationale of your approach and explain your results.

d) Look up the file `Student_Taxi_Mapping.txt`. Use the taxi ID that is listed next to your Student Number to answer the following questions:

i. Plot the location points for taxi=ID

ii. Compare the mean, min, and max location values of taxi=ID with the global mean, min, and max location values.

iii. Compare the total time driven by taxi=ID with the global mean, min, and max driven time values.

iv. Compute the distance travelled by taxi=ID. To compute the distance between two points on the surface of the earth use the following method:

dlon = longitude2 -longitude1

dlat = latitude2 -latitude1

*a = (sin(dlat/2))^2 + cos(lat1) * cos(lat2) * (sin(dlon/2))^2*

*c = 2 * atan2(sqrt(a), sqrt(1-a))*

*distance = R * c (where R is the radius of the Earth)*

Assume that R=6,371,000 meters.

Note that angles need to be in radians to pass to trig functions! More information on this calculation can be found at https://www.movable-type.co.uk/scripts/latlong.html

With each of your answers: Explain what knowledge can be derived from your answer.

(Task 2 starts from the next page)

Task 2

Preface: Banks are often posed with a problem to determine whether or not a client is credit worthy. Banks commonly employ data mining techniques to classify a customer into risk categories such as category A (highest rating) or category C (lowest rating).

A bank collects data from **past** credit assessments. The file `"creditworthiness_data_with_labels.csv"` contains 1962 of such assessments. Each assessment lists 46 attributes of a customer. **The last attribute "credit rating" is the result of the assessment**. Open the file and study its contents. You will notice that the columns are coded by numeric values. The meaning of these values is defined in the file `"definitions.txt"`. For example, a value "3" in the last column means that the customer credit worthiness is rated "C". Any value of attributes not listed in `definitions.txt` is "as is".

This poses a "prediction" problem. A machine is to learn from the outcomes of past assessments and, once the machine has been trained, to assess any customer who has not yet been assessed.

Purpose of this task:
You are to start with an analysis of the general properties of this dataset by using suitable visualization and clustering techniques (i.e., such as those introduced during the lectures and labs), and you are to obtain an insight into the degree of difficulty of this prediction task. Then you are to design and deploy an appropriate supervised prediction model (i.e., MLP as introduced in the lab of Week 5) to obtain a prediction of customer ratings.

Question 1: Statistical analysis and visualization (4 marks)
Analyse the general properties of the dataset. Create a statistical analysis of the attributes and their values, then list *5 most valuable attributes for predicting "credit rating".* Explain the reasons that make these attributes valuable.

A set of R-script files are provided with this assignment (included in the zip-file). **These are similar to the scripts used in Week 3 Lab**. The scripts provided will allow you to produce some first results. However, virtually none of the parameters used in these scripts are suitable for obtaining a good insight into the general properties of the given dataset. Hence your task is to modify the scripts such that informative results can be obtained from which conclusions about the learning problem can be made. Note that finding a good set of parameters is often very time consuming in data mining. An additional challenge is to make a correct interpretation of the results.

This is what you need to do: Find a good set of parameters (i.e., through a trial-and-error approach) for the steps of SOM training, visualisation, and clustering. Summarize your approach to conducting the experiments, explain your results, and offer a meaningful interpretation of the results. Do not forget that you are also to provide an insight into the degree of difficulty of this learning problem (i.e., from the results that you obtained, can it be expected that a prediction model will be able to achieve a 100% prediction accuracy?). Always explain your answers succinctly.

Note that when you use SOM (an unsupervised learning method) to analyse the data, the last column of the data (i.e., "*credit rating*") shall not be included because it is class label, which will be used in the following supervised task.

Question 2: MLP (4 marks)
Deploy a prediction model to predict the credit worthiness of customers. The prediction capabilities of the MLP in Week 5 Lab was not good enough. Your task is to:

a) Randomly split the data such that 80% is used for training and 20% is reserved for test.
b) Train MLP(s) and report your prediction result on the test set. Give an interpretation of your results. What is the best classification accuracy (expressed in % of correctly classified

data) that you can achieve on the test set?

c) Describe any strategies that you have tried in order to increase the accuracy of predicting the credit rating. Explain why your strategy can be expected to increase the prediction capability.
d) You will find that 100% accuracy cannot be obtained on the test data. Explain reasons to why a 100% accuracy could not be obtained. What can you do to further improve the prediction accuracy for this prediction task?

Note that in this assignment the term "prediction capability" refers to a model's ability to predict the credit rating of samples that were not used to train the model (i.e., samples in a test set).

--- END ---