INFO411/INFO911
Data Mining and Knowledge Discovery

Project 1

Instructions:

This tasks is a real-world data mining problem. You are required to prepare a set of presentation slides which must include (1) the full name and student number of each student in the group, the contribution (in percent) of each group member, (2) a description of the task, (3) your proposed data mining approach and methodology; (4) the strengths and weaknesses of your proposed approach; (5) the performance measures that can evaluate your data mining results; (6) the results a brief discussion and a conclusion.

Below is the recommended structure of your slides:

- Introduction (define the problem and the goal)
- Methods (propose approaches, and discuss their strengths and weaknesses)
- Results (Figures and tables of data analysis)
- Discussion (discovered knowledge from data mining)

Task: Vehicular Data Mining (activity analysis)

Background:

The mining of vehicular data has numerous practical applications such as in traffic control, planning of road networks, flow prediction, intelligent driver navigation, and many more. Vehicular Data Mining is one of the fundamental data mining problems which focuses on discovering patterns by analysing vehicular data (i.e. location data). One popularly used dataset in this context is the DRAWDAD roma taxi driver dataset. The dataset is available from: http://crawdad.org/roma/taxi/20140717/

Several studies have been conducted on this dataset. A good starting point to understanding Vehicular Data Mining can be found in:
Cristian Chilipirea, Andreea Petre, Ciprian Dobre, Florin Pop, Fatos Xhafa. Enabling Vehicular Data with Distributed Machine Learning. In *Transactions on Computational Collective Intelligence XIX*, Vol. 9380, pp. 89-102, 2015.

Definition of the task:

Use the CRAWDAD roma taxi driver dataset for activity analysis of taxi drivers in Rome. Find how the rate of activity varies over time. Questions to be answered are:

1. Are there regions in Rome which are more (or less) frequented than typically by taxi drivers at certain periods of time (or days)?
2. Discover interesting cases of work ethics among the drivers. Describe how the activity pattern of those drivers deviate from the norm.

Requirements:

1. Present a general description of the dataset and present a the general properties of the dataset.
2. Propose two different approaches to answering the afore mentioned questions. Discuss the strengths and weaknesses of the two approaches.
3. Present and explain the answers to the two questions.
4. Offer a qualitative comparison between the two approaches. How did the result of the two approaches differ?
5. Summarize: What new and interesting things did you discover about the activity of roman taxi drivers in Rome?