INFO411/INFO911
Data Mining and Knowledge Discovery

Project 7

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: Atmospheric Chemistry Mining

Background:

The Dataset: Lauder Hourly Atmospheric Green House Dataset
This dataset contains a range of environmental measured atmospheric conditions, together with a set of green-house gas recordings, including: N20, CO, CH4 and CO2. This data covers conditions from 2006 to 2008 arranged as some 16,996 hourly observations. Lauder is one of several data monitoring stations distributed in New Zealand as part of the National Institute of Water and Atmospheric Research (NIWA).

Definition of the task:

As this given set of data set has no initially specified targets to mine for; part of the task is search for plausible patterns of interest, or “modes”. Such modes which may occur frequently, or rarely, and could be determined by unsupervised learning or clustering techniques applied to various sections of this hourly average data set.

Modes of interest might be formed from mixtures of various attributes such as certain green house gases, prevailing weather conditions, time of day and season. Once you have determined a range of modes, you will need to subsequently determine what the characteristic properties might be that particularly differentiate one mode from another.

One method for approaching this is to label (or score) the data with these predicted modes and subsequently utilise a supervised learning or predictive modelling approach to express the essential (minimal) divisive differences between modes. This can often be used to identify the essential factors that determine a mode of interest.

Subsequent analysis of the outcomes may lead to further insights as to how one mode may gradually change to become another mode type – as an example, it should also be noted that there is a natural daily sequence of gas mixtures. These will obviously be effected by the associated seasonal weather conditions and prevailing winds from various directions. These for example will mask, or weaken certain modes, or indeed transport higher concentrations of gases from other sources such industrial, farming or nearby townships.

Requirements:

You are required to carry out a range of exploratory data analysis on the allocated historical data:

1. Present a general description of the dataset and present a the general properties of the dataset.
2. Data Cleaning: identification and adjustments or re-mediation for missing values and selection of a target dataset -- followed by a comprehensive survey of all distributions and relationships it entails.
3. Transformation: any appropriate conversion, remapping or construction of existing or new attributes.
4. Data mining:
   a) For unspecified targets: the identification of interesting clusters or modes, followed by explanatory models that support their existence, performance analysis of the three best models.
5. Interpretation and analysis, if no clear insights, some revision and re-start at stage 3?
6. Summarize: What new and interesting things did you discover about the activity of roman taxi drivers in Rome?