INFO411: Data Mining and Knowledge Discovery, Au2020

Project 12

Instructions:

This task 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) your proposed data mining approach and methodology; (3) the strengths and weaknesses of your proposed approach; (4) the performance measures that can evaluate your data mining results; (5) the results and a brief discussion. 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: Air pollution prediction in Beijing

Background:

Beijing has been struggling with air pollution for years. As a result, several monitoring stations have been set up in and around the city in order to study the temporal and spatial behaviour of particulate matter (PM).

The data set for this task contains hourly PM2.5 data for Beijing, and associated meteorological data. The data are available from

`https://archive.ics.uci.edu/ml/machine-learning-databases/00394/`

Please consider only the data associated with Beijing (`BeijingPM20100101_20151231.csv`) and, in particular, with the station labelled *Dongsihuan*.

Requirements:

1. Load the data into `R` and create two new variables, one identifying whether the gas concentration was measured on a weekday or a weekend, and the other identifying whether the concentration was measured during work hours (8:00 – 17:59) or not.

2. Explore the relationships between air pollution at *Dongsihuan*, the meteorological variables, the time/type of day, season and year.

3. Present relevant visualisations of the data, which help to illustrate the relationships, trends and differences found in the previous items.

4. Develop models to predict air pollution (PM2.5 concentrations) using the meteorological data, the time/type of day, season and year. Two of these that you develop should be the standard linear model and the random forest.

5. Provide the performance evaluation of any fitted models, including details of cross-validation or splitting into training, validation and/or testing sets.

6. Present your interpretations and conclusions.