Intelligent systems 2015-2016

Assignment 1

The assignment is to be done in **pairs**. The presentations will be held during the lab practices in the week of **December 7–13.**

The file "pollution.txt", which can be found on the course web page, contains data about air pollution during the period from 1996 to 2014. The dataset consists of 6804 measurements of 10 attributes.

Independent variables:

DATE	Measurement date in YYYY-MM-DD format
TRAJ	A discrete attribute, nominal description of 7 day air mass movement
SHORT_TRAJ	A discrete attribute, nominal description of 1-2 day air mass movement
AMP_TMP2M_mean	A continuous attribute, day's mean air temperature
AMP_RH_mean	A continuous attribute, day's mean relative humidity
AMP_WS_mean	A continuous attribute, day's mean wind speed
AMP_PREC_sum	A continuous attribute, day's total precipitation

Dependant variables:

O3_max	A continuous attribute, day's max ozone level
PM10	A continuous attribute, day's large pollution particles concentration
PM2.5	A continuous attribute, day's small pollution particles concentration

Assignment

The main objective is to apply machine learning methods to forecast air pollution.

Specific tasks:

1. Data preparation
   Dataset contains missing values (marked as NA). You can ignore them (completely remove them from the dataset) or try to reconstruct a good approximation to missing values.

2. Data summary and visualisation
   Summarize the data with summary statistics and plots. Such exploratory analysis may help you identify potential characteristics that you can use when constructing new attributes.

3. Attribute evaluation
   Evaluate existing attributes and possibly construct new attributes that will improve the quality of trained models.

4. Prediction: Classification
   Train at least three different types of classification models for predicting:

   a. max ozone level – there are four classes: LOW (under 60.0), MODERATE (between 60.0 and 120.0), HIGH (between 120.0 and180.0), and EXTREME (above 180.0).

   b. the concentration of large pollution particles – there are three classes: LOW (under 35.0), MODERATE (between 35.0 and 50.0), and HIGH (above 50.0).

5. Prediction: Regression
   Train at least three different types of regression models for predicting:

   a. the max ozone level (O_3),

   b. the concentration of small pollution particles (PM2.5)

6. Model evaluation
   Compare the chosen models in terms of predictive accuracy and comprehensibility of results. Present the best classification and regression model.

7. Prepare a report (doc or pdf document)
   Describe your approach, present obtained results, and summarize your conclusions based on the experimental evaluation.

Grading

The final score will be based on the predictive accuracy of selected models and the quality of attributes, your exposition and justification of the chosen approach, and your interpretation of the final results.