Statistical Modelling & Machine Learning HW2
(Due: 11/14/2020, Sunday)

Instruction:

- There is no correct or unique answer in this homework.
- I will give your HW score based on your results and model building procedure.
- Submit your report for HW including a `R` code.
- Your report should describe data pre-processing and predictive modelling processes. Also, you should report test error rate for your final model.
- Your `R` code should show the procedure that you obtain the final model (**Do NOT include all models that you have tried**).

1. Consider the training data in '`train.csv`' file and the test data in '`test.csv`'. Our goal is to build the best model predicting `Y` variable based on `X1 - X4` variables.

Power plant generates electric power and it is always set to work with full load. Suppose that we want to predict electric power that the power plant produces for an hour. There are four factors (`X1 - X4`) that affect the generation of electric power. The description of the variables in the dataset is as follows:

`X1`: Temperature.
`X2`: Amount of Gas.
`X3`: Pressure.
`X4`: Humidity.
`Y`: Amount of electric power that the power plant produces for an hour.

Note that there are missing values (`NA`) for the input variables in the training dataset (No missing values in the test dataset). You should build your own predictive model using the training data and apply the model to the test dataset. Report the test MSE value. You can use any predictive models including both data modelling and algorithmic modelling.

2. Consider the training data in '`pm25_tr.csv`' and the test data in '`pm25_te.csv`'. Suppose that our interest is to predict pm 2.5 concentration based on some meteorological factors. In the dataset, the output variable is `pm25`. The training set has data measured from March 1st to May 20th, 2010 and the test set has data measured from May 21st to May 25th, 2010 (next 5 days). The descriptions of the variables in the dataset are as follows:

`year`: Year of data.
`month`: Month of data.

`day`: Day of data.
`hour`: Hour of data.
`pm25`: PM2.5 concentration.
`DEWP`: Dew Point.
`TEMP`: Temperature.
`PRES`: Pressure.
`cbwd`: Combined wind direction.
`Iws`: Cumulated wind speed.

You should build your own predictive model using the training data and apply the model to the test dataset. Report the test MSE value. You can use any predictive models including both data modelling and algorithmic modelling.

Instruction for prediction of Q2:

1. To predict pm2.5 at time t, you may use input variable values at time t.
2. That is, to predict pm2.5 of next 5 days, assume that we know meteorological factor values for next 5 days from the meteorology center.
3. Do not use any pm2.5 values of next 5 days to predict them. Those values should be used only for calculating the test MSE.