Practical 1

Dr Simon Denman
CAB420: Machine Learning

This weeks practical will focus on linear regression. This practical uses the same data used in the week 0 activity on data wrangling.

This data contains:

1. Data from the Bureau of Meteorology data for Brisbane City showing daily rainfall, maximum daily temperature data, daily solar exposure data.
2. Data for Brisbane City Council (BCC) cycleways, showing pedestrian and cyclist counts for various points throughout the BCC network.

Data is provided for 2014-18. `CAB420_Prac1.zip` contains the raw files containing the data, while `combined.csv` is a pre-merged version. Those that have completed the week 0 activity may wish to use the raw data and their own code to merge the data, otherwise you are welcome to use the pre-merged data.

Problem 1. Linear Regression. Using the provided dataset (either the individual files and your own approach to merge the data, or the `combined.csv` merged dataset), split the data into training, validation and testing as follows:

- Training: All data from the years 2014-2016
- Validation: All data from 2017
- Testing: All data from 2018

Develop a regression model to predict one of the cycleway data series (select whichever one takes your fancy) in your dataset. In developing this model you should:

- Initially, use all weather data (temperature, rainfall and solar exposure) and all other data series for a particular counter type (i.e. if you're predicting cyclists inbound for a counter, use all other cyclist inbound counters).
- Use p-values, qqplots, correlation between predictors and response, correlation between pairs of predictor, sand performance on the validation set to remove terms and improve the model.

When you have finished refining the model, evaluate it on test set, and compare the Root Mean Squared Error (RMSE) for the training, validation and test sets.

In training the model, you will need to ensure that you have no samples (i.e. rows) with missing data. As such, you should remove samples with missing data from the dataset before training and evaluating the model. This may also mean that have to remove some columns that contain large amounts of missing data (i.e. determine how many samples are missing in each column, remove columns with lots of missing data, remove any other rows where data is missing).

We recommend the `statsmodels` and `pandas` packages for this problem. In particular, you may wish to use:

- `isna`: A member function of a pandas dataframe that indicates if a variable is missing.
- `dropna`: A member function of a pandas dataframe that drops missing values.
- `statsmodels.api.OLS`: Ordinary Least Squares regression function within `statsmodels`.

You may also wish to explore the `sklearn` package which also contains methods for linear regression, and data splitting.