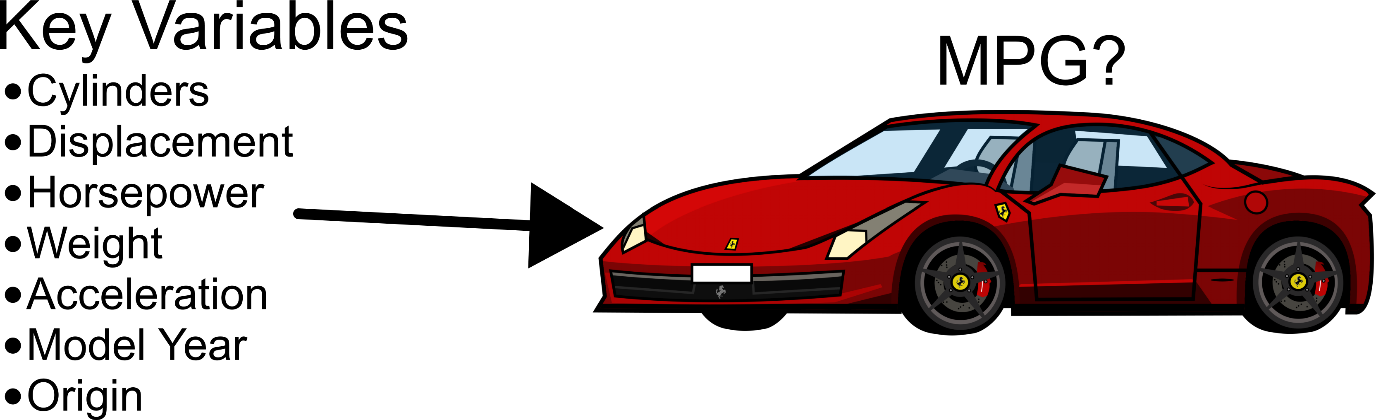
Gordon Chen z5161163 COMP9417, 2019 S2  
Chengbin Feng  
Alisa Yang

# Topic 2.1: Perceptron / Linear Unit

## Topic Introduction

This topic is a good introduction to AI as this is about predicting the car MPG (Miles per Gallon) based on continuous (e.g. weight) and multi-valued discrete (e.g. model year) variables. The key objective of this topic is to find the best perceptron equation that has minimal cost in both training and testing phase. This involves the concept of correlation, deciding which variables to use or not and implementing key part of algorithms to aid perceptron training algorithm and learning performance.



## Planning out

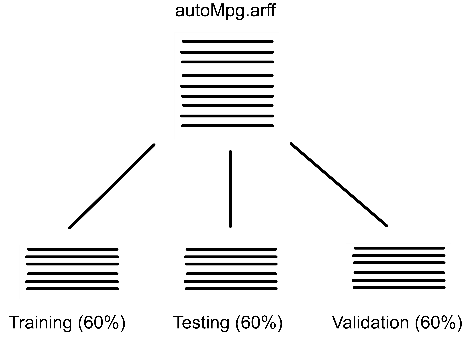
On Week 6, we all met each other and discussed about what we need to do in this topic. After some discussion, we agreed who does which task.

* Gordon’s task is to modify the original data files for our own benefits, implement a method that reads from our modified data files and return it as an array for our linear regression model and implement pocket algorithm to aid the learning performance.
* Chengbin’s task is to implement our linear regression model for both the training and predicting phase. This involves a lot of reading of both lecture notes and other websites on linear regression concept and equation before coding it.
* Alisa’s task is to find key relationship between each key variable, display it on graph and discuss the results with Gordon and Chengbin on key observations and adjustment we need to make on our linear regression model better.

## Data set architecture

We initially made the decision to load our arff data as a numpy array before processing it for training and predicting MPG. However, with a lack of experience in numpy array and its core function to modify data files and array, we decided to use bash since Gordon is currently doing COMP2041 and has some current experience in modifying file and its content via bash.

Using bash, we first modify our arff data to **transformedData.data** which makes Alisa job easier to plot relationship between each variable. We then separate our arff files into 3 different data sets which are training, testing and validation sets. We split the whole dataset in 60:20:20 ratio such that the test data has 60% of data while testing and validation sets each has 20% of data.

With interest in testing specific cases, we included line id since our bash files are sorting in random before allocated to our 3 sub data sets. The training, testing and validation sets also has some key variables removed which will be explained later.

## Key observation of variables relationship

Alisa made correlation.py to find any relationship or correlation between any 2 variables. The key objective to ensure that:

* Avoid getting same information more than once, effectively avoiding over fitting. A key example is horsepower and cylinder which is explained more below.
* Any variables that has no clear correlation with MPG are not evaluated. A key example is acceleration.

We then use matplotlib.pyplot as a graphical representation of her observation. From that, we found that:

* While horsepower is correlated with MPG, it’s also correlated with displacement and cylinders. Along with the issue of missing values for horsepower in the original dataset, it’s best to not include horsepower in our linear regression model.
* Acceleration doesn’t correlate with MPG at all, so we don’t include acceleration in our linear regression model.
* From previous testing, we found that model origin impacts the model too much as it’s sometime has a very low cost or a very high cost. After some determination, we feel that it’s best to not include model year into our model.

Overall, we have the following features to determine our car MPG:

* Cylinders
* Displacement
* Weight
* Origin
* Year

## Independent Continuous Variable vs Dependent Variable (Mpg) – Pearson Coefficient

|  |  |
| --- | --- |
| Key Variable | Pearson Coefficient |
| Cylinders | -0.83 |
| Displacement | -0.81 |
| Horsepower | -0.78 |
| Weight | -0.83 |
| Acceleration | 0.42 |

## Independent Variable vs Independent Variable

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Pearson Coefficient | Cylinders | Displacement | Horsepower | Weight | Acceleration |
| Cylinders | 1 |  |  |  |  |
| Displacement |  | 1 |  |  |  |
| Horsepower |  |  | 1 |  |  |
| Weight |  |  |  | 1 |  |
| Acceleration |  |  |  |  | 1 |

## Implementation of Linear Regression Model

## Results

## Conclusion

## Implementation of Pocket Algorithm