**Data Mining Project Proposal- Predicting MPG For Cars**

**Jiannan Zhang**

[jiannanzhang1992@gmail.com](mailto:jiannanzhang1992@gmail.com)

**I. Introduction:**

Nowadays, the fuel economy is catching more and more attentions. Therefore, it is essential to have good estimates of fuel consumption for different types of cars. MPG, which stands for mile per gallon, is one of most important measures for that. In this project, I will use the ridge regression algorithm to predict the MPG from the data set from the StatLib library, which is maintained at Carnegie Mellon University (The dataset was used in the 1983 American Statistical Association Exposition). I will also use some methods like cross-validation holdout method to find the best model.

**II. Motivation:**

Predicting MPG can help people to make decision when they buy cars, for someone who drives often, a more fuel-efficient vehicle can translate into major monthly savings for the driver. On the other hand, it can also be used to estimate the car’s fuel economy before they come out, which is good for car industry.

**III. Project Summary:**

In this project, I am trying to find the interesting correlations/associations, visualizations, or analyses on this dataset. I am trying to build up a model to predict MPG by given eight features, which are:

1. Cylinders: multi-valued discrete
2. Displacement: continuous
3. Horsepower: continuous
4. Weight: continuous
5. Acceleration: continuous
6. Model year: multi-valued discrete
7. Origin: multi-valued discrete
8. Car name: string (unique for each instance)

**IV. Project Details:**

a. Architecture and Environment

In this project, the mainly used software is MATLAB. The main method is to do the ridge regression and use cross-validation to find the best model. The holdout method can be used to train and test data.

b.Implementation Issues and Challenges

The challenges of this project are that the model may not be a linear model and there are some missing values. Also, some features maybe have almost no effect on MPG. So I need to consider all these issues to build and test the model.

c. Timeline and schedule

First week (Apr.2– Apr.8): Learning and reviewing the machine learning algorithm about regression and ridge regression and other essential knowledge to finish this project.

Second week (Apr.9– Apr.15): input all the data to MATLAB and use MATLAB to build up the model.

Third week (Apr.16 – Apr.22): Test the model and fix model, find interesting correlations/associations, visualizations, or analyses.

Last week (Apr.23 – Apr.29): Review the model and do the final modification and report out the final project.

Notes: 1. You should use ridge regression and use cross validation to find the best lamda. (+ lamda.w’w)

2. You could use two -third to train and one­-third to test.

3. You can multiply two features to make a new feature to see what is going on. That would be beyond the linear model.

4 .test the model when you take out a weight. If the different is very small or see R suqrese , just delete it.