**Assignment 1 - Save the Planet: Predicting Car Fuel Efficiency**

In the code package, you are given a dataset including information about different cars. The final goal is to predict the gas mileage of each car (miles per gallon), given its information. The original data came from the StatLib library from CMU. It was modified by Ross Quinlan to remove entries with unknown miles per gallon (mpg). We have modified it further by removing six entries with unknown horsepower.

We want to inspect these features, extracted from raw collected data, and consider the implications of those feature choices on the learning algorithm. In class, we often considered examples only with 2 features. In this assignment, you will need to extrapolate your knowledge to work with multidimensional feature sets (and multivariate regression.)

**Dataset**

There are 392 entries in the dataset. The 9 field names should be self-explanatory except possibly acceleration, displacement and origin. Acceleration indicates the time to accelerate from 0 to 60 mph (sec.) You can read about displacement here; in this data set displacement is in cubic inches. Origin is 1=USA, 2=Europe, 3=Asia. Weight is in lbs, and the number of cylinders is between 4 and 8.

# Part I - Short Answers

**1.a Defining the problem**. Let's assume for our first approach, we do not necessarily need to know the exact value of mpg; instead, we are trying to simplify the problem to know whether it is a fuel efficient car or not. Discuss how you will try to solve the problem. Select a value you could consider as a threshold for an "efficient" car according to the data and why?

Instruction: **write your full response in your report.pdf**

**1.b** - **Preprocessing** is a vital step in doing any machine learning task. Here you will discuss, given the data, what considerations should be taken and why. This includes selecting features and feature engineering.

Concretely, for each of the 8 features in the data, indicate how you could represent it to make classification easier and obtain good generalization on unseen data, by choosing one of:

- 'drop' - leave the feature out,
- 'raw' - use values as they are,
- 'standard' - standardize values by subtracting out average value and dividing by standard deviation,
- 'one-hot' - use a one-hot encoding, i.e. a vector of 0's and 1's where the presence of the feature is indicated by 1, and nonexistence by 0

There could be multiple answers that make sense for each feature; please mention the tradeoffs between each answer. Write down your choices.

Instruction: **write your full response in your report.pdf**

**1.c. General** Provide a short answer to the following questions:

- 3.c.1. Can we use our machine learning model to predict whether cars in 2019 will have good gas mileage? Why?
- 3.c.2. Since weight values and cylinders values are on different scales, is it hard for a model to predict whether a car is efficient or not? Discuss your answer for each of KNN, logistic regression and decision trees. If so, how can it be mitigated?

Instruction: **write the full responses to your report.pdf.**

# Part II - Coding and Report

**2 - Classification** Implement KNN classification discussed in class based on the features and decisions made in the previous questions. Use Python and 10-fold cross validation, with at least 3 values of K. Evaluate them using 2 metrics based on what has been discussed in the class. Discuss why you chose those metrics, and report the results in a table similar to that which was shown in class.

Instruction: **write your full response in your report.pdf**
*It is not necessary to submit the code for 2a.*

**3 - Regression** Now assume we want to predict the actual value of the mpg for each car. For convenience, we will choose to not include model year and car name as features.

Two choices of feature set:

- [cylinders=standard, displacement=standard, horsepower=standard, weight=standard, acceleration=standard, origin=one_hot]

- [cylinders=one_hot, displacement=standard, horsepower=standard, weight=standard, acceleration=standard, origin=one_hot]

We will use 10-fold cross-validation to try all possible combinations of these feature choices and test which is best.

The file q3.py contains functions, some of which will need to be filled in with your definitions from this homework. You will be implementing a **regularized** version of linear regression, called **ridge regression** which adds an L2 penalty term to the loss function. The effect of the penalty term is controlled by a hyperparameter λ. See page 680 in the textbook for the exact formula for **L2 regularization**.

The functions you define will be called by ridge_min, defined for you, which takes a dataset (*X*,*y*) and a hyperparameter λ as input and returns θ and $\theta_0$ (in our class notes, these correspond to our **w** vector $w_2$ and $w_1$) minimizing the ridge regression objective using stochastic gradient descent (SGD). The code will run a grid search over a large range of λ, with choices $\lambda=\{0.0,0.01,0.02,\cdots,0.1\}$ for polynomial features of orders 1 and 2, and the choices $\lambda=\{0,20,40,\cdots,200\}$ for polynomial features of order 3 (as this is approximately where we found the optimal λ to lie).

The learning rate and number of iterations are fixed in this function and should not be modified for the purpose of answering the below questions (although you should feel free to experiment with these if you are interested!).

The ridge_min function will then further be called by xval_learning_alg, which returns the average RMSE across all splits of your data when performing cross-validation. (Note that this RMSE is reported in standardized y units).

The file auto.py will be used to implement the auto data regression. The file contains code for creating the two feature sets that you are asked to work with here. Load the data and run the cross-validation function, based on your implementations in q3.py, you should be able to answer the following questions:

Instruction: **upload your filled q3.py to canvas. You are only allowed to change the block bellow** `#Your code here [n].` **There are in total 10 blocks that need to be completed.**

Instruction: **write the responses below in your report.pdf**

- 3.e.1 What combination of feature set and lambda minimizes the average cross-validation RMSE?

- 3.e.2 What is the average cross-validation RMSE value in mpg that you obtain using the best combination in the previous question (with 3 decimal precision)?