Linear and Polynomial Regression

J.R. Powers-Luhn

October 8th, 2018

1 Objective

2 Preprocessing the data

The data set analyzed consisted of continuous (weight, power, displacement, and acceleration), quasi-continuous (year), and discrete (origin) properties for a variety of car models along with their fuel efficiency (expressed in miles per gallon). One other property (car name) was ignored as outside of the scope of this analysis[1].

The data were loaded and examined visually. Six rows with missing horsepower values were removed. Imputation was not attempted since these rows consisted of less than 2% of the data. The continuous variables were plotted against each other to visually identify correlations and non-linear relationships. It was determined that weight, horsepower, and displacement had an apparently inverse relationship to fuel efficiency. As such, the inverse of the values was added to the data set.

The data were split into training (50% of the data remaining), testing (25%), and validation (25%) sets by random selection of samples. The random seed used to select the rows was specified for future reproduction of the analysis. This method was selected for its simplicity and based on the expectation that the samples (car models) were independent of each other–no property of any car depended on any other car.

The data were examined for potential outliers or bad measurements. Some problematic measurements were identified, but ultimately determined to be benign. A few of the cars in the set had an odd number of cylinders–one was a diesel engine with five cylinders while others (listed as three cylinders) had rotary engines. If a more accurate model is desired for conventional, gas powered cars, these should be omitted, but they were included in this model in an attempt to extend the predictive range of the model.

[1] It might prove interesting to see if a semantic analysis of the model name could predict other properties of the car, e.g. model year or acceleration

3 Linear modeling

Models of the data were generated using the ordinary least squares method. It was assumed that the model would follow a linear (or linear in parameters) form, as in equation 1.

$$\mathbf{X}\vec{b} = \vec{y} \tag{1}$$

Since there is no expectation that $\mathbf{X}$ is square, each side of the equation is left-multiplied by $\mathbf{X}^T$. We must then find the parameters $\vec{b}$ that minimize the sum of squared errors.

$$0 = \frac{\mathrm{d}}{\mathrm{d}\vec{b}}(\mathbf{X}^T\mathbf{X}\vec{b} - \mathbf{X}^T\vec{y})^T(\mathbf{X}^T\mathbf{X}\vec{b} - \mathbf{X}^T\vec{y}) \tag{2}$$

$$\mathbf{X}^T\vec{y} = \mathbf{X}^T\mathbf{X}\vec{b} \tag{3}$$

$$\vec{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{y} \tag{4}$$

Equation 4 is the one used to solve for the coefficients, which can then be used to predict future cases. Since these equations are linear, they are not in general expected to have a zero intercept. As such, a column of fixed values (ones) are added to the measured input values. This sets the last element in $\vec{b}$ to the y-intercept value.

3.1 Model selection

Models were trained on the training data, then evaluated using the testing data. Final results are reported using the validation data, which is only used to evaluate the selected "best" model.

3.2 Scaling

Linear models, unlike others, do not require that the input data be scaled to unit variance or mean-centered. In order to verify this, the same linear