Theoretical 1

Steven Raaijmakers

April 2017

1. **What is the number of features?**

   11

2. **What is the target variable?**

   The car price

3. **Is this a regression or classification task? Explain your choice.**

   This is a regression task, since the outcome should be a number (from a continuous set)

4. **Write down the linear function that express your hypothesis - define a set of input variables xi and output variable y to represent your features and the value you wish to predict (the car price), respectively.**

   $Y_i = O_1X_1 + O_2X_2 + ... + O_nX_n$

5. **Write down the mean squared error cost function you wish to minimize. What does the cost function express. Why use the squared error instead of the raw error?**

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(Y_i - \hat{Y}_i)^2 \tag{1}$$

   It is the difference between the hypothesis and the real value (for your test set). By squaring the difference we always obtain a positive value, and it emphasis larger differences.

6. **[Research] Come up with an alternative cost function and explain your choice (i.e. why is this a good cost function? How is it different from the squared error one?**

   Mean Absolute Error (MAE) uses the absolute value to measure the average magnitude of errors. However it doesn't emphasis big error as much as RMSE does.

7. **[Research] Features 5, 6, 9, 10, and 11, are real-valued continuous features. So they can be used by the linear regression algorithm. The rest of the features however are categorical (nominal-scale) features. How can you represent these feature so that the linear regression algorithm can use them? For instance, "num-of-doors", clearly cannot be used as it is - it does not make sense to multiply a weight with the words "four" or "two". Explain your choice of representation of these categorical features.**

   One could use dummy variables to represent these categorical features (which is often used in MLR) as numbers. These numbers actually are one-to-one mapping from categorial features to number.