Introduction to Machine Learning
ECE 580

Spring 2020
Homework #2: Data Exploration and Regression

To the extent possible, submit your homework using this template, as that will facilitate (expedite) scoring.

This homework assignment is worth **180 points**.

Submit a **single, self-contained PDF file** to the assignment in Sakai. Do not include a print-out of your code in the PDF file you submit. Be prepared to provide your source code if we ask to see it.

Each problem is worth some multiple of 10 points, and will be scored on the below letter scale.
The letter grades B through D may be modified by + (+3%) and A through D may be modified by a - (-3%).

A+ = 100%: Exceeds expectations, and no issues identified
A = 95%: Meets expectations, and (perhaps) minor/subtle issues
B = 85%: Issues that need to be addressed
C = 75%: Significant issues that must be addressed
D = 65%: Major issues, but with noticeable perceived effort
F = 50%: Major issues, and insufficient perceived effort
Z = 30%: Minimal perceived effort
N = 0%: Missing, or no (or virtually no) perceived effort

Structuring and Organizing Your Code

I am language agnostic; you may use your preferred computing platform. You may also shoose to use packages/toolboxes authored by others. If you use packages/toolboxes authored by others, you are expected to reference the packages/toolboxes so we know what external code supported your completion of the assignment. You are also responsible for knowing how to use the packages/toolboxes to achieve what we want. In past semesters, many students have commented toward the end of the semester that, in hindsight, they spent more time looking for functions and figuring out how to make them do what they wanted them to do than they would have spent writing their own functions. You may wish to keep this in mind as you decide when to write your own functions and when to leverage existing packages/toolboxes.

The majority of homework assignments this semester will involve coding. I suggest you think about how you can structure and organize your code so it can easily be extended to additional use cases. For example, you can use data structures to support extensible code so input/output argument lists for your functions do not become unwieldy.

I also suggest thinking about how to modularize your code. For example, if a quantity could be calculated in more than one context, consider making the calculation of that quantity a separate function rather than embedding (and replicating!) the code to calculate that quantity within several functions. When a code block exists once, extending it or correcting it can be achieved by revising that single code block. When a code block is replicated within several functions, it must be revised every place it exists in order to extend it or correct it across all instances.

Generating PDF files

Your homework assignments this semester will be submitted electronically as a **single, self-contained PDF file** through Sakai. The TAs and I will provide feedback by inking on the PDF file and will return the annotated file to you via Sakai. Since we will be working with the PDF file just as we would be working with a paper-based submission (the only difference being it exists electronically rather than physically), we are asking for the submitted PDF file to be identical in form to a paper-based submission.

Duke provides Adobe Creative Cloud accounts for all Duke students at no charge. These accounts are automatically created upon enrollment, so you should already have an account. If, for some reason, you do not have an account, you can request one through OIT[1].

Adobe Acrobat is a component of Adobe Creative Cloud that provides PDF editing capabilities, and is a resource you can use to generate the single, self-contained PDF file that you submit for your lab assignments this semester.

What you should be able to do with PDF files...

- Scan paper-based answers to a PDF file.
- Insert a new page into an existing PDF file.
- Insert a blank page into a PDF file (helpful if you need more room to ink your answer).
- Delete a page from a PDF file.
- Insert an image into a PDF file.
- Annotate (add text) to a PDF file, either by inking or typing.
- Merge two PDF files.

[1] `https://software.duke.edu/node/272`

Data Preparation and Exploration

Download the Automobile Data Set from the UCI Machine Learning Repository[2] Although there are several potential uses of this data set, we are going to use this data to predict a car's price from its characteristics (including horsepower!). For the time being, we are going to restrict ourselves to the 13 continuous predictor variables (in the order in which they appear in the data base): wheel-base, length, width, height, curb-weight, engine-size, bore, stroke, compression-ratio, horsepower, peak-rpm, city-mpg, and highway-mpg.

(10) 1. (a) Document (list) the steps you take to clean this data set, including removing the (non-continuous) features that are not of interest, and removing any data points for which the target variable (price) is unknown. The documentation should include the number of data points that are removed at each step (do not list the removed data points themselves), and conclude with a statement as to the number of data points that remain after cleaning the data set.

[2] `https://archive.ics.uci.edu/ml/datasets/Automobile`

(10) (b) In a future assignment, you will explore systematic feature (predictor variable) selection. For now, you are going to select features you believe are most promising via data exploration.

For each of the 13 continuous predictor variables (features), plot the target variable (price) as a function of the predictor variable (feature) using a scatter plot.[3] (This will produce 13 scatter plots, one for each feature.)

[3] I recommend writing, or leveraging, a function that accepts a matrix of features and a vector of target variables, and produces a feature vs. target scatter plot for each feature in the data matrix.

(10) (c) For each of the 13 features, explain, based on your scatter plots, why you believe that feature holds promise for predicting a car's price, or why you believe it does not hold promise for predicting a car's price. You may notice that for some features there appears to be a nonlinear relationship between the feature and the car's price. For example, $price = feature^2$, or $price = 1/feature$... keep this in mind when you propose candidate models for predicting a car's price. (There should be 13 explanations, one for each feature.)

(10) (d) When performing regression, it is preferable to have features that are as independent as possible, as strongly related (correlated) features do not provide much additional information and may lead to computational challenges. For example, if there were two additional continuous features, "km per gallon city" and "km per gallon highway", these features would be highly (perfectly?) correlated with the existing features "city-mpg" and "highway-mpg," respectively, because 1 km = 0.6241 miles. For this reason, we would want to include only one of "km per gallon city" and "city-mpg" in our model, and only one of "km per gallon highway" and "highway-mpg" in our model.

Plot each pair-wise combination of features using scatter plots to aid in (visually) identifying features that are related (correlated).[4] (There will be **a lot** of plots. Write code (or leverage a package) to do the heavy, repetitive, lifing for you!) Since you are using these plots to identify correlation trends the plots do not need to be high-resolution, so it is ok if the plots are "small".

[4] I recommend writing, or leveraging, a function that accepts a matrix of features (and, optionally, a vector of target variables), and produces a set of pair-wise feature scatter plots (with symbols that may be color-coded by target variable).

(10) (e) Identify, based on your pair-wise scatter plots, variables that are related and preferably would not both (or all) be used in the model simultaneously.

Explain how you arrived at your conclusions.

Regression

Continuing with the 13 continuous predictor variables from the Automobile Data Set from the UCI Machine Learning Repository to predict a car's price from its characteristics...

(10) 2. (a) Based on your data exploration, propose 3 unique linear models (3 unique subsets of the 13 continuous features) for predicting a car's price from its characteristics. If you noticed that for some features there appears to be a nonlinear relationship between the feature and the car's price, such as $price = feature^2$ or $price = 1/feature$, you may propose a model that uses a transformed feature that you expect to capture the nonlinear relationship you observed. Limit your proposed models to use no more than three features.

Explain why you proposed each of these models as a candidate model to test. (There should be 3 explanations, one for each model.)

(30) (b) For your proposed model #1,

i. Perform linear regression to find the model parameters,[5] and provide the specific model (*i.e.*, write down the equation $\hat{price} = f(features, \hat{\mathbf{w}})$ with the values for each element of $\hat{\mathbf{w}}$ specified.

ii. What is R^2 for this model?

iii. Scatter plot the predicted price as a function of the true price. Also plot the line $\hat{price} = price$ (the line representing perfect prediction) as a reference.

iv. What is your impression of this model? How do the predicted prices compare to the true prices? Are there price ranges where the model is particularly good? Are there price ranges where the model is particularly bad?

[5] I recommend writing, or leveraging, a function that accepts a matrix of features (or transformed features) and a vector of target variables, and performs linear regression. This function may (optionally) return R^2 and/or other diagnostic information.

(30) (c) For your proposed model #2,

i. Perform linear regression to find the model parameters, and provide the specific model (*i.e.*, write down the equation $\hat{price} = f(\text{features}, \hat{\mathbf{w}})$ with the values for each element of $\hat{\mathbf{w}}$ specified.

ii. What is R^2 for this model?

iii. Scatter plot the predicted price as a function of the true price. Also plot the line $\hat{price} = price$ (the line representing perfect prediction) as a reference.

iv. What is your impression of this model? How do the predicted prices compare to the true prices? Are there price ranges where the model is particularly good? Are there price ranges where the model is particularly bad?

(30) (d) For your proposed model #3,

i. Perform linear regression to find the model parameters, and provide the specific model (*i.e.*, write down the equation $\hat{price} = f(\text{features}, \hat{\mathbf{w}})$ with the values for each element of $\hat{\mathbf{w}}$ specified.

ii. What is R^2 for this model?

iii. Scatter plot the predicted price as a function of the true price. Also plot the line $\hat{price} = price$ (the line representing perfect prediction) as a reference.

iv. What is your impression of this model? How do the predicted prices compare to the true prices? Are there price ranges where the model is particularly good? Are there price ranges where the model is particularly bad?

(10) (e) Which of your three proposed models would you select? Why?

3. In the case of simple linear regression of y onto x, both the coefficient of determination, R^2, and the sample correlation between x and y, r, are measures of the linear relationship between x and y. It can be shown in this case (simple linear regression) the coefficient of determination is equal to the square of the sample correlation between y and y, $R^2 = r^2$.

(10) (a) Does the equivalence between the coefficient of determination and the square of the sample correlation extend to multiple linear regression (regression with more than one predictor, or feature, variable)?

(10) (b) What does R^2 provide that r (or r^2) does not?