Regression Homework

* Datasets are available as XLS or CSV files.
* For simplicity you don’t have to normalize the data for Problem 1 and 2 if you use Excel.
* You can submit a Word file with pictures/notes pasted with the file name “HW1-yourname”. Python users can submit a Jupyter (.ipynb) file.
* If you write Python or Matlab codes for Problem 3 please attach your program as part of your submission. A Jupyter notebook is good for Python users.
* There is no standard answer but there are better models. Learn to go through the model building process and try to improve your models. The grading is not necessarily based on the model accuracy; it’s more to do with your effort.

1. (3 points) Use the Advertising data (Advertising)
   1. Explore the data and provide a sample of the plots you created.

Answer is on ‘Advertising\_fnish.xlsx'

* 1. Build a linear regression model of Sales using all 3 media data as features using Excel
     1. Is this a good model?

NO, since the R squared value is only 0.89, fitting through these existed features is not so well.

What does the coefficient and t-test’s p-value tell you?

The coefficient tell us how important of the feature and whether it is positive related or negative related; p-valued reflect the significance of the t-test results

, when p-value is too high, we can say that the feature is not important to the result.

* + 1. Rebuild the model without newspaper as a feature.

Answer is on ‘Advertising\_fnish.xlsx’

* + 1. Add TV\*Radio as a new feature, rebuild the model and compare with the previous models.

Answer is on ‘Advertising\_ifish.xlsx’, and I found that the model is better.



1. (3 points) The fish dataset (Fish-3Only-Homework) contains the weight, height, and width data of 3 different kinds of fish: Bream, Roach, and Perch. Use Excel or software of youir choice to do the following:
   1. Explore the dataset and watch for possible missing data.
   2. Since fish species is a categorical variable please do one-hot encoding (name your dummy variables Bream-D, Roach-D, and Perch-D if your software allows). Note: if you use Excel you can use only 2. The third one would then be represented by (0,0).
   3. Split the data into Train (80%) and test (20%) for each kind of fish. How do you make the split? Would it be appropriate to just take the last 20% of the data as the test set?
   4. Build a linear regression model of fish’s Weight with height, width and the one-hot variables as features. Write the formula of your model.
   5. Report the R-sqaure and RMSE of the training set. Check other key regression performance criteria.
   6. Calculate the R-sqaure and RMSE of the test set. Are they close to the training set?
2. (6 points) Concrete is the most important material in civil engineering. The concrete compressive strength could be modeled as a highly nonlinear function of age and ingredients. The relationships could be nonlinear.
   1. Explore the dataset (concrete-Homework). Watch for missing data. Provide a sample of the plots you created.
   2. Split the dataset into 80% training and 20% test sets.
   3. Build the best Linear Regression model you can.
      1. Report any feature engineering (feature choice, transformation etc) you did.
      2. Report the RMSE of your Train and Test datasets.
   4. Build the best Decision Tree model you can. Report the RMSE of your Train and Test datasets. Report the hyperparameters you use,
   5. Build the best Random Forest model you can. Report the RMSE of your Train and Test datasets. Report the hyperparameters you use.

After Boost Tree and Neural Network (NN) are covered you can come back and build boost tree and NN models of this dataset.

For your reference the following are additional info about the variables:

