Intro to Data Science
Homework 7: Due Friday November 8 at 2:00pm

Exercises:

1. For each part, indicate whether we would generally expect the performance of a flexible model (*i.e.* one with more degrees of freedom) to be better or worse than an inflexible method. Justify your answer.

   (a) The sample size (number of observations) n is extremely large, and the number of predictors p is small.

   (b) The number of predictors p is extremely large, and the number of observations n is small.

   (c) The relationship between the predictors and responses is highly non-linear.

   (d) The variance of the error terms ϵ, is extremely high.

2. Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide the number of observations n and the number of predictors p.

   (a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

   (b) We are considering launching a new product and wish to know whether it will be a *success* or a *failure*. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

3. The following problems relate to linear regression and are taken from *An Introduction to Statistical Learning*, `http://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf`.

   (a) Do Conceptual Exercise 2 from Chapter 3 of *An Introduction to Statistical Learning.* **Hint:** Read section 3.5 first.

   (b) Do Conceptual Exercise 3 from Chapter 3 of *An Introduction to Statistical Learning.*

   (c) Do Applied Exercise 9 from Chapter 3 of *An Introduction to Statistical Learning.*

4. In this problem you are asked to write R code to do simple quadratic regression "by hand." Recall that in the lecture R notebook for simple linear regression, we computed simple linear regression "by hand" by carrying out the following steps:

   (a) We wrote a function to represent the linear expression $ax + b$, where x is the predictor variable and a and b are the parameters.

   (b) We wrote a function to compute the MSE for the predictor values x and for input values of the parameters a and b. This will be a function of the parameters which we made sure was vectorized.

(c) We used the `optim()` function to minimize the MSE function as a function of the parameter values.

Your task is to carry out the same steps now using a quadratic expression $ax^2 + bx + c$ instead of a linear expression. In order to test your code, simulate some data (make sure to add a small amount of noise) that is well-approximated by a quadratic function. Then, simulate some data that is should not be well-approximated by a quadratic function.

5. If possible, try using both simple linear regression and multiple linear regression on your project dataset. Alternatively, you can use the `Carseats` data from the `ISLR` package. Describe what you do and assess the results. Specifically, what if anything does (either simple or multiple) linear regression tell you about (some or all of) the variables in your data? Do you have any evidence that linear regression helps to accurately explain at least part of your dataset? Be sure to make plots such as scatter plots and residual versus fitted plots. Use some form of cross-validation to estimate test error.

6. Solve the expression

$$B = \frac{A}{1+A},$$

for A. Apply your result to show that

$$p(X) = \frac{e^{mX+b}}{1+e^{mX+b}},$$

is equivalent to

$$\frac{p(X)}{1-p(X)} = e^{mX+b}.$$

Now use the previous result to show that

$$\ln\left(\frac{p(X)}{1-p(X)}\right) = mX + b.$$

7. The following code produces the confusion matrix resulting from a logistic regression model with 0.5 threshold value decision criterion applied to classify responses in the `Default` data set from the `ISLR` package.

```r
log_fit <- glm(default~balance,data=Default,family = "binomial")
pred_vals <- predict(log_fit,newdata = Default,type="response")
threshold_val <- 0.5
Default$pred_class <- factor(ifelse(pred_vals > threshold_val,"Yes","No"))
conf_mat <- confusionMatrix(Default$pred_class,Default$default,positive="Yes")
conf_mat$table
```

```
##           Reference
## Prediction   No  Yes
##        No  9625  233
##        Yes   42  100
```

Modify the code to recompute the confusion matrix for different threshold values. Use each resulting confusion matrix to compute the true positive rate and the false positive rate. Is 0.5 the optimal threshold value? Explain why or why not.

8. This question should be answered using the `Weekly` data set, which is part of the `ISLR` package. This data is similar in nature to the `Smarket` data set which is examined in section 4.6 of the book *An Introduction to Statistical Learning.*

   (a) Produce some numerical and graphical summaries of the `Weekly` data. Do there appear to be patterns?

   (b) Use the full data set to perform a logistic regression with `Direction` as the response variable and the five lag variables plus `Volume` as predictors. Use the tidy() function in the `broom` package to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?

   (c) Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.

   (d) Construct an ROC curve for the logistic regression and compute the area under the ROC curve. Remember that this can be done with the functions in the `ROCR` package used in class.

   (e) Now fit the logistic regression model using as training data period from 1990 to 2008, with `Lag2` as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 to 2010).

   (f) Repeat (e) using KNN with $K = 1$.

   (g) Experiment with different combinations of predictors, including possible transformations and interactions, for each of the methods. Report the variables, method, and associated confusion matrix that appears to provide the best results on the held out data. Also experiment with values of K when using KNN.