Homework 2

STAT 430, Spring 2017

Due: Friday, February 10 by 11:59 PM

Please see the homework instructions document for detailed instructions and some grading notes. Failure to follow instructions will result in point reductions.

Exercise 1

[**14 points**] In this exercise, you will investigate the bias-variance tradeoff when estimating the function f defined below.

```
f = function(x1, x2) {
  x1 ^ 3 + x2 ^ 3
}
```

The following code defines the data generating process and should we used to simulate data.

```
get_sim_data = function(f, sample_size = 100) {
  x1 = runif(n = sample_size, min = -1, max = 1)
  x2 = runif(n = sample_size, min = -1, max = 1)
  y = f(x1, x2) + rnorm(n = sample_size, mean = 0, sd = 0.5)
  data.frame(x1, x2, y)
}
```

Use simulation to investigate the bias and variance of *five* models at the point $\mathbf{x} = (x_1, x_2) = (0.75, 0.95)$. The five models are of the form

- `y ~ poly(x1, degree = k) + poly(x2, degree = k)`

for $k = 1, 2, 3, 4, 5$. Use 500 simulated samples each of size 200. Before performing the simulations, you should set a seed equal to your UIN. For example,

```
uin = 123456789
set.seed(uin)
```

Summarize your results as a *single* plot which compares both **squared bias** and **variance** of the estimates to the **degree** of the polynomials used. That is, the x-axis should be **degree** and you should have a line for both **squared bias** and **variance**. Comment on the plot. Are the results what you expected? Explain. (A few points may not strictly follow the general pattern as a result of the randomness of the simulation.)

Exercise 2

[**8 points**] For this exercise use the data found in `hw02-train.csv` and `hw02-test.csv` which contain train and test data respectively.

Find a model by fitting to the training data which achieves:

- Train RMSE less than 1.08
- Test RMSE less than 1.01

Report the model you found (you may use `R` formula notation), as well as the two metrics.

Exercise 3

[**8 points**] For this exercise use the data found in `auto-train.csv` and `auto-test.csv` which contain train and test data respectively. `auto.csv` is provided but not used. It is a modification of the `Auto` data from the `ISLR` package. For information on the original data:

```r
library(ISLR)
#?Auto
```

Use the training data to train a classifier which achieves:

- Train Accuracy greater than 0.89
- Test Accuracy greater than 0.89

Report these metrics, as well as the confusion matrix, sensitivity, and specificity for the test data.