Homework 5

STAT 430, Spring 2017

Due: Friday, March 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

[**15 points**] For this homework we will use data found in `wisc-train.csv` and `wisc-test.csv` which contain train and test data respectively. `wisc.csv` is provided but not used. This is a modification of the Breast Cancer Wisconsin (Diagnostic) dataset from the UCI Machine Learning Repository. Only the first 10 feature variables have been provided. (And these are all you should use.)

- UCI Page
- Data Detail

You should consider coercing the response to be a factor variable. Do not use cross-validation for this exercise.

Use KNN. Consider $k = 1, 2, \ldots, 50$. Find the best k using both scaled and unscaled predictors. For both, plot train and test accuracy vs k on a single plot, report the best k, and report the associated test accuracy.

So, your answer will be two plots (both with two lines), two values of k, and two test accuracies. Was the scaling helpful?

Exercise 2

[**15 points**] Calculate *train*, *test*, and *5-fold cross-validated* **accuracy** for both an **additive logistic regression** and **LDA**. You may use the `createFolds()` function from `caret`, but you may not use the `train()` function from `caret`.

Use your UIN in place of `uin`.

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