Homework 06

STAT 430, Fall 2017

Due: Friday, October 27, 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.

For this homework we will use data found in `wisc-trn.csv` and `wisc-tst.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.

You should use the `caret` package and training pipeline to complete this homework. Any time you use the `train()` function, first run `set.seed(1337)`.

Exercise 1 (Tuning KNN with `caret`)

[6 points] Train a KNN model using all available predictors, **no data preprocessing**, 5-fold cross-validation, and a well chosen value of the tuning parameter. Consider $k = 1, 3, 5, 7, \ldots, 101$. Store the tuned model fit to the training data for later use. Plot the cross-validated accuracies as a function of the tuning parameter.

```
library(caret)
```

```
## Loading required package: lattice
```

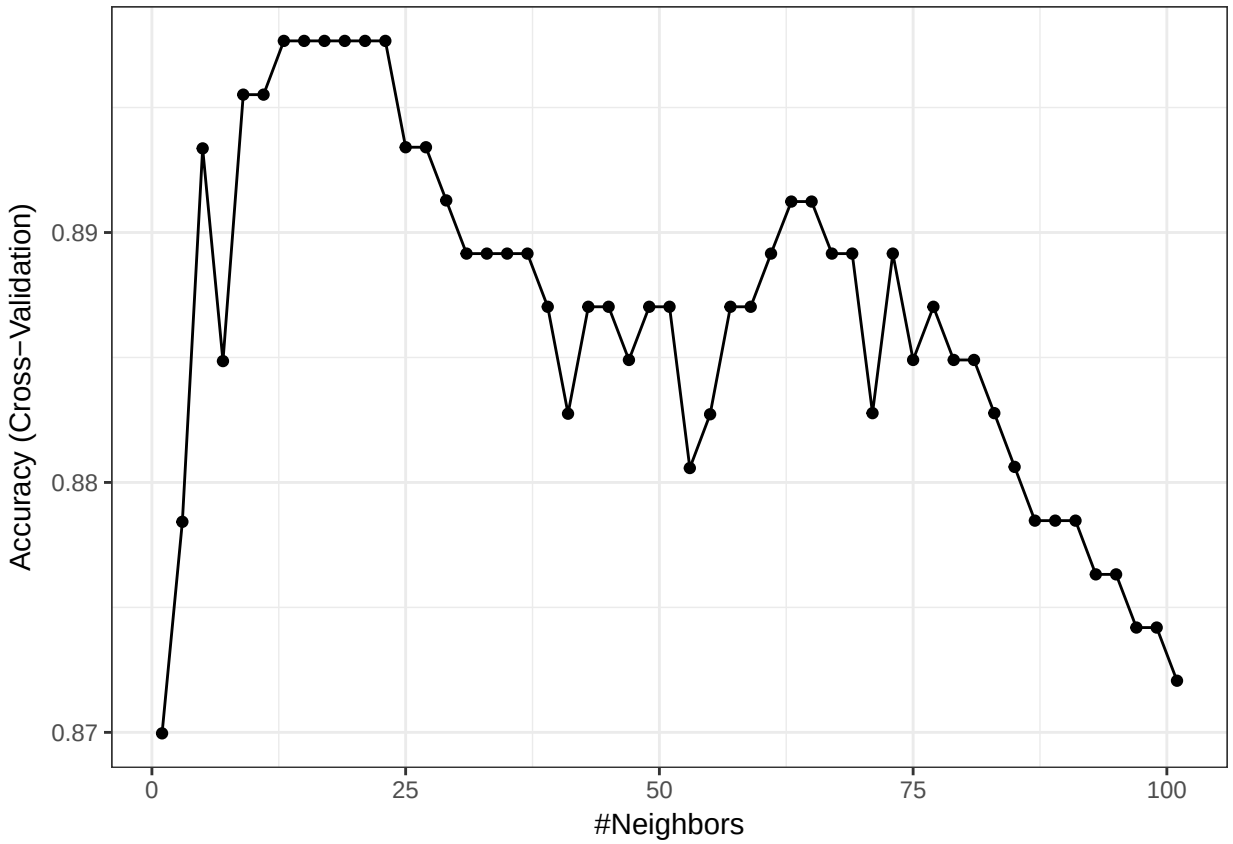
```
## Loading required package: ggplot2
```

```
trn = read.csv('wisc-trn.csv')
tst = read.csv('wisc-tst.csv')
```

```
set.seed(1337)
```

```
knn_mod = train(class ~ .,
  data = trn,
  method = "knn",
  trControl = trainControl(method = "cv", number = 5),
  tuneGrid = expand.grid(k = seq(1, 101, by = 2))
)
```

```
ggplot(knn_mod) + theme_bw()
```



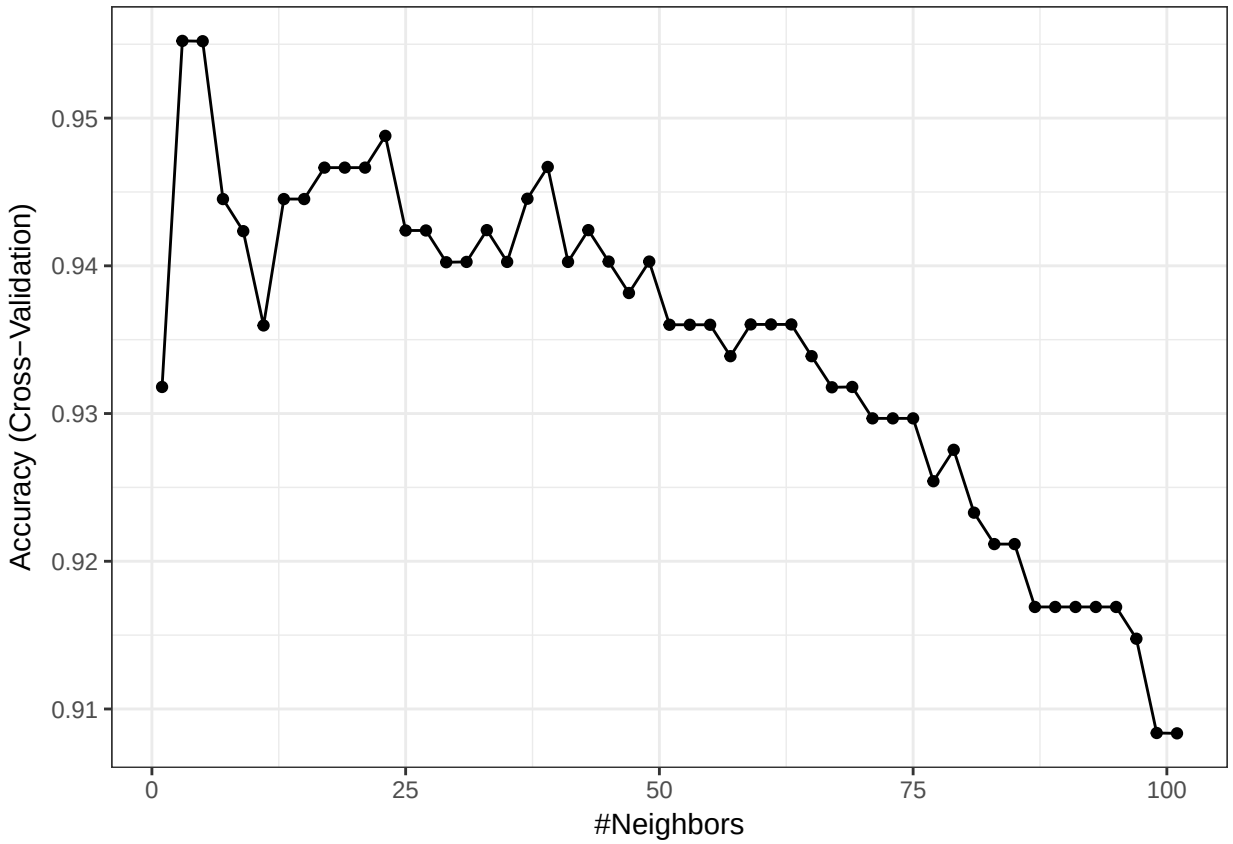
Exercise 2 (More Tuning KNN with `caret`)

[**6 points**] Train a KNN model using all available predictors, predictors scaled to have mean 0 and variance 1, 5-fold cross-validation, and a well chosen value of the tuning parameter. Consider $k = 1, 3, 5, 7, \ldots, 101$. Store the tuned model fit to the training data for later use. Plot the cross-validated accuracies as a function of the tuning parameter.

```
set.seed(1337)
```

```
knn_scale = train(class ~ .,
  data = trn,
  method = "knn",
  trControl = trainControl(method = "cv", number = 5),
  preProcess = c("center", "scale"),
  tuneGrid = expand.grid(k = seq(1, 101, by = 2))
)
```

```
ggplot(knn_scale) + theme_bw()
```



***

Exercise 3 (Random Forest?)

[**6 points**] Now that we've introduced `caret`, it becomes extremely easy to try different statistical learning methods. Train a random forest using all available predictors, **no data preprocessing**, 5-fold cross-validation, and well a chosen value of the tuning parameter. Using `caret` to perform the tuning, there is only a single tuning parameter, `mtry`. Consider `mtry` values between 1 and 10. Store the tuned model fit to the training data for later use. Report the cross-validated accuracies as a function of the tuning parameter using a well formatted table.

```
library(randomForest)
```

```
## randomForest 4.6-12
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
set.seed(1337)
```

```
mtry = seq(1,10)
metric = 'Accuracy'
tunegrid = expand.grid(.mtry=mtry)
```

```
rf_default = train(class~., data=trn, method="rf", metric=metric,
                   trControl = trainControl(method = "cv", number = 5), tuneGrid=tunegrid)
```

```
df = rf_default$results
knitr::kable(df)
```

mtry	Accuracy	Kappa	AccuracySD	KappaSD
1	0.9402654	0.8707109	0.0208842	0.0456183
2	0.9466484	0.8849867	0.0262238	0.0575091
3	0.9402654	0.8716088	0.0246152	0.0535592
4	0.9487989	0.8898150	0.0231601	0.0505391
5	0.9403111	0.8718995	0.0244552	0.0516905
6	0.9403111	0.8718449	0.0244552	0.0513549
7	0.9381835	0.8672685	0.0217476	0.0456290
8	0.9424388	0.8763086	0.0277205	0.0577409
9	0.9403111	0.8716520	0.0357365	0.0751735
10	0.9381835	0.8671788	0.0330970	0.0694612

Exercise 4 (Concept Checks)

[**1 point each**] Answer the following questions based on your results from the three exercises. Format your answer to this exercise as a table with one column indicating the part, and the other column for your answer. See the `rmarkdown` source for a template of this table.

(a) What value of k is chosen for KNN without predictor scaling?

```
##     k
## 12 23
```

(b) What is the cross-validated accuracy for KNN without predictor scaling?

```
##    k  Accuracy     Kappa AccuracySD   KappaSD
## 1 23 0.8976664 0.7706512 0.01603824 0.0362239
```

(c) What is the test accuracy for KNN without predictor scaling?

```
## [1] 0.86
```

(d) What value of k is chosen for KNN **with** predictor scaling?

```
##   k
## 2 3
```

(e) What is the cross-validated accuracy for KNN **with** predictor scaling?

```
##   k  Accuracy     Kappa AccuracySD    KappaSD
## 1 3 0.9552276 0.9035641  0.0138546 0.02921515
```

(f) What is the test accuracy for KNN **with** predictor scaling?

```
## [1] 0.88
```

(g) Do you think that KNN is performing better with or without predictor scaling?

(h) What value of `mtry` is chosen for the random forest?

```
rf_default$bestTune
```

(i) Using the random forest, what is the (estimated) probability that the 10th observation of the test data is a cancerous tumor?

```
predict(rf_default,newdata = tst[10,-1], type = 'prob')
```

(j) Using the random forest, what is the (test) sensitivity?

```
pred = predict(rf_default, newdata = tst)
tst_tab = table(predicted = pred, actual = tst$class)
tst_tab[4]/(tst_tab[3] + tst_tab[4])
```

```
## [1] 0.875
```

(k) Using the random forest, what is the (test) specificity?

```
tst_tab[1]/(tst_tab[1] + tst_tab[2])
```

(l) Based on these results, is the random forest or KNN model performing better?

```
(tst_tab[1] + tst_tab[4]) / (tst_tab[1] + tst_tab[2] + tst_tab[3] + tst_tab[4])
```

```
## [1] 0.93
```

```
get_best_result(knn_scale)$Accuracy
```

```
## [1] 0.9552276
```

```
get_best_result(knn_mod)$Accuracy
```

```
## [1] 0.8976664
```

part	answer
A	23
B	0.8976664
C	0.86
D	3
E	0.9552276
F	0.88
G	KNN is performing better with predictors scaling since the cross-validated accuracy and test accuracy is both greater when predictors are scaled.
H	4
I	0.04
J	0.875
K	0.967
L	knn with scale performs better since the accuracy of knn with scale is greater than KNN without scale and random forest.