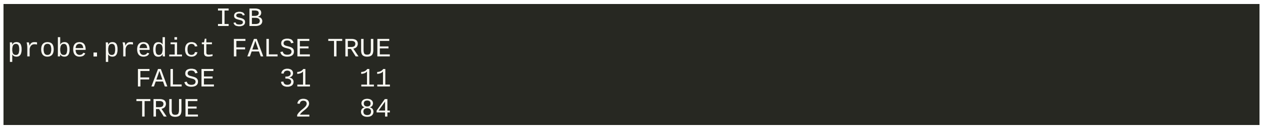
Charles Valentine
Homework 13
12/01/2016

Note:
All code is in script hw13.R --- I have made use of cat and print statements to display information easily!

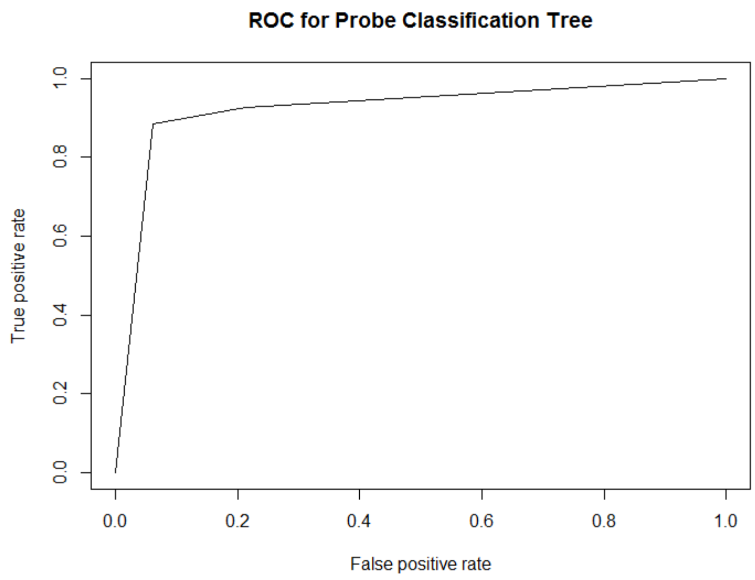
Problem 1

a) See code for implementation.
b) The confusion matrix and ROC curve are printed here for genes *39317_at* and *38018_g_at* as predicted by a classification tree.

OUTPUT

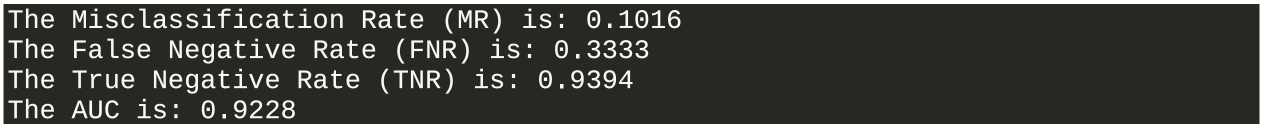


```
             IsB
probe.predict FALSE TRUE
        FALSE    31   11
        TRUE      2   84
```



c) Statistics for the classification tree fit to the above genes are presented here.

OUTPUT



```
The Misclassification Rate (MR) is: 0.1016
The False Negative Rate (FNR) is: 0.3333
The True Negative Rate (TNR) is: 0.9394
The AUC is: 0.9228
```

d) A 10-fold cross-validation to estimate the real false negative rate (FNR) was run and a value of 0.2833 was returned.

OUTPUT

```
The average FNR of a 10-fold cross-validation is: 0.2833
```

e) A logistic regression model using the above genes to predict B-cell patients was used and an 80% confidence interval was computed for the coefficient of gene *39317_g_at*.

OUTPUT

```
   10 %    90 %
-1.4274 -0.6048
```

f) N-fold cross-validation was used to estimate the MCR of the logistic regression classifier.

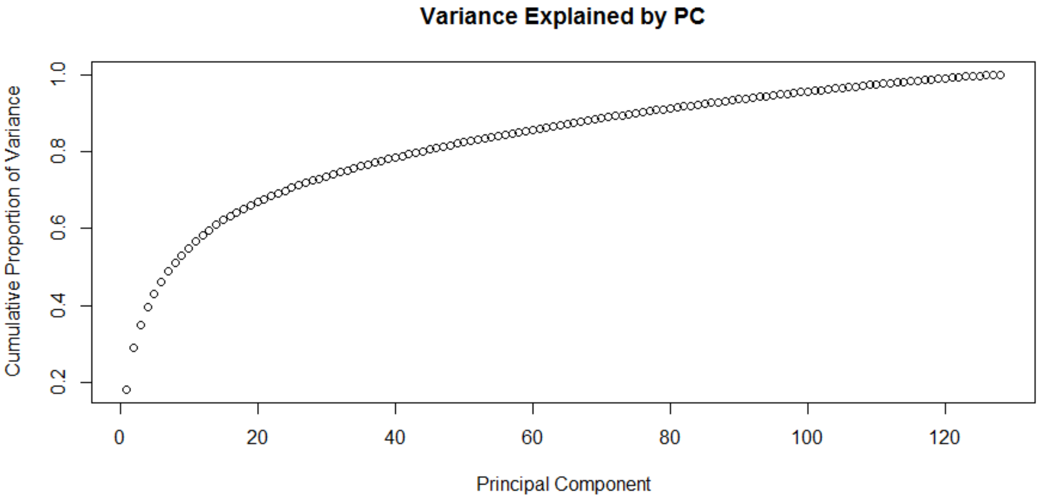
OUTPUT

```
The average MCR of a n-fold cross-validation is: 0.09375
```

g) A PCA was conducted on the scaled variables of the entire gene expression dataset ALL. It was determined that 97 principal components were needed to represent 95% of the variation in the dataset as can be show in the figure below.

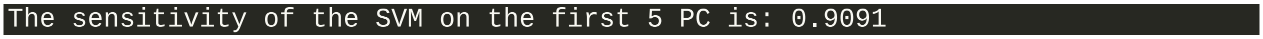
OUTPUT

```
It takes 97 principal components to explain 95% of the variance.
```



h) A SVM classifier was used on the first five PCs from Problem 1g. It was determined that the sensitivity or true positive rate is 0.9091.

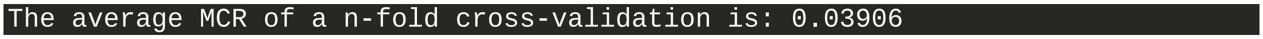
OUTPUT



```
The sensitivity of the SVM on the first 5 PC is: 0.9091
```

i) N-fold cross-validation was used on the same SVM classifier as in Problem 1h to estimate the misclassification rate (MCR). It was determined that the MCR is 0.03906.

OUTPUT



```
The average MCR of a n-fold cross-validation is: 0.03906
```

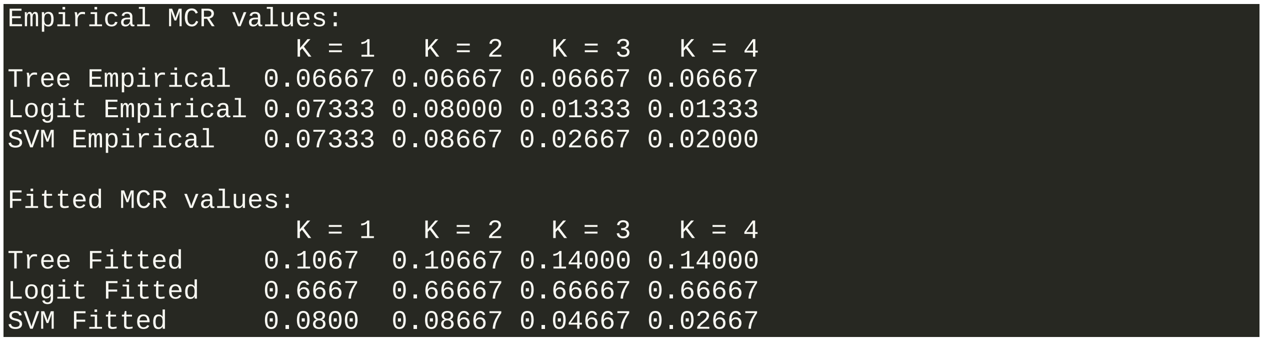
j) The support vector machine as implemented in Problem 1g uses linear planes to separate classification regions. The logistic map uses a simple logistic equation to assign probability of any one data point belonging to a class. The support vector machine is a much stronger statistical learning approach as it is less prone to overfitting as it is a more general model relying only on linear planes and not probability boundaries.

Problem 2

Using the *Iris* dataset built into R we ran three statistical learning models to determine if we could classify samples based only on three measurements. The three learning models we used were a classification tree, a logistic regression (multinomial), and a support vector machine. In order to simply the data fed into the models we only considered the data that is composed within the first four principal components independently.

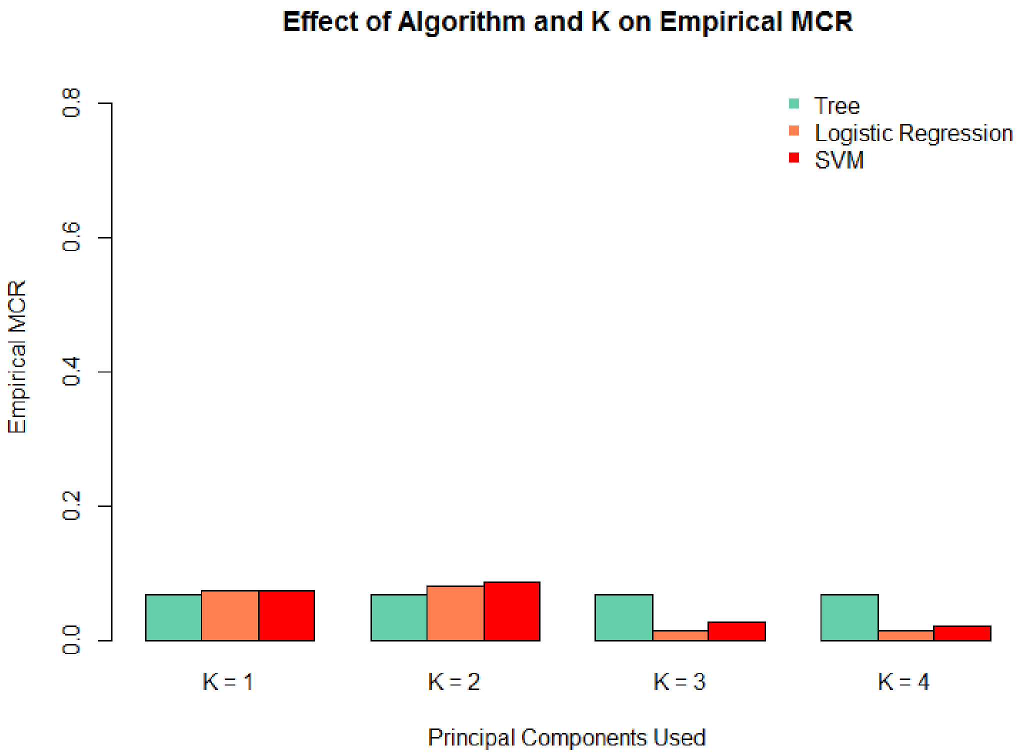
This has allowed us to compare algorithms and the best set of principal components to use. Figures on the next page indicate that all models over fit unless a cross-validation scheme is used. The logistic map is the worst culprit of overfitting as it's MCR increased an order of magnitude using N-fold cross-validation. For the cross-validated data it appears as if all algorithms perform the same regardless of PCs used except for the SVM which improves as the number of PCs increases (increase noticeable at K = 3). Therefore, the ideal K value in which all algorithms perform well while simultaneously using as few data as possible is K = 2. SVM is the best performer here.

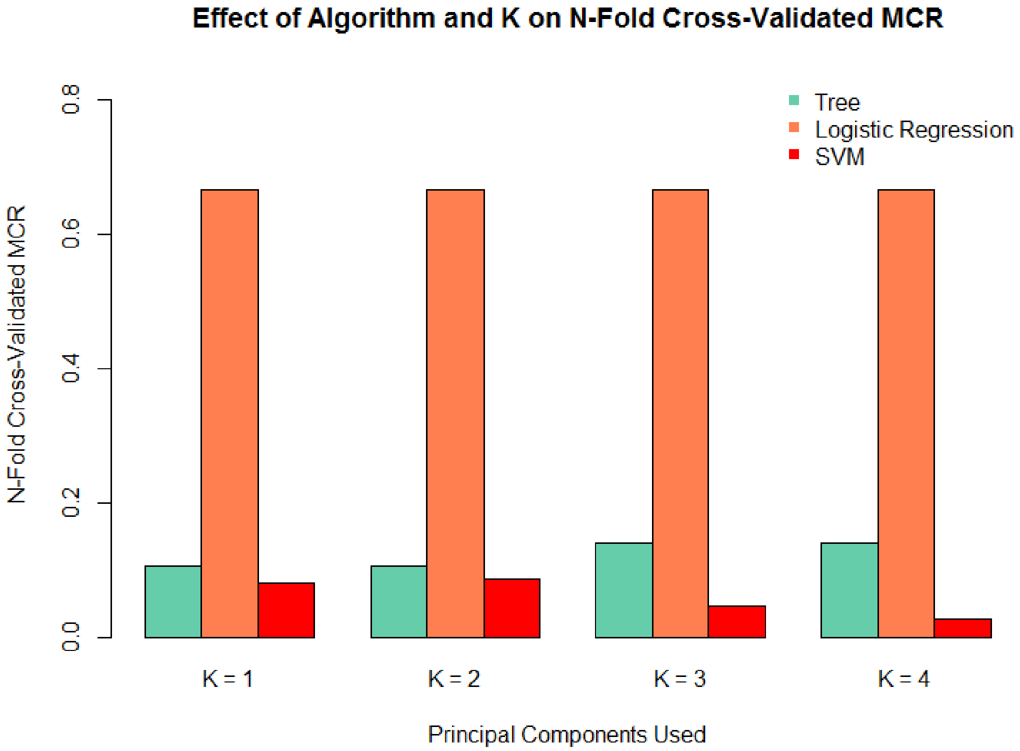
OUTPUT



```
Empirical MCR values:
                  K = 1   K = 2   K = 3   K = 4
Tree Empirical  0.06667 0.06667 0.06667 0.06667
Logit Empirical 0.07333 0.08000 0.01333 0.01333
SVM Empirical   0.07333 0.08667 0.02667 0.02000

Fitted MCR values:
                  K = 1   K = 2   K = 3   K = 4
Tree Fitted     0.1067  0.10667 0.14000 0.14000
Logit Fitted    0.6667  0.66667 0.66667 0.66667
SVM Fitted      0.0800  0.08667 0.04667 0.02667
```


Effect of Algorithm and K on Empirical MCR
Empirical MCR
0.0
0.2
0.4
0.6
0.8
Tree
Logistic Regression
SVM
K = 1
K = 2
K = 3
K = 4
Principal Components Used


Effect of Algorithm and K on N-Fold Cross-Validated MCR
N-Fold Cross-Validated MCR
0.0
0.2
0.4
0.6
0.8
Tree
Logistic Regression
SVM
K = 1
K = 2
K = 3
K = 4
Principal Components Used