DSCI445 - Homework 6

Your Name

Due 11/20/2019 by 4pm

Be sure to `set.seed(445)` at the beginning of your homework.

```
#reproducibility
set.seed(445)
```

1. We will explore the maximal margin classifier on a toy data set.

   a) We are given $n = 7$ observations in $p = 2$ dimensions. For each observation, there is an associated class label.

Obs	X_1	X_2	Y
1	3	4	Red
2	2	2	Red
3	4	4	Red
4	1	4	Red
5	2	1	Blue
6	4	3	Blue
7	4	1	Blue

   Sketch the observations.

   b) Sketch the optimal separating hyperplane and provide the equation for this hyperplane.

   c) Describe the classification rule for the maximal marginal classifier. It should be along the lines of "Classify to Red if $\beta_0 + \beta_1 X_1 + \beta_2 X_2 > 0$, and classify as Blue otherwise. Provide the values of $\beta_0, \beta_1, \beta_2$.

   d) On your sketch, indicate the margin for the maximal margin classifier.

   e) Indicate the support vectors for the maximal margin classifier.

   f) Argue that a slight movement of the seventh observation would not affect the maximal margin hyperplane.

   g) Draw an additional observation on the plot so that the two classes are no longer separable by a hyperplane.

2. We have seen that we can fit an SVM with a non-linear kernel in order to perform classification using a non-linear decision boundary. We will now see that we can also obtain a non-linear decision boundary by performing logistic regression using non-linear transformations of the features.

   a) Generate a data set with $n = 500$ and $p = 2$, such that the observations belong to two classes with a quadratic decision boundary. E.g.,

```
n <- 500
x1 <- runif(n) - 0.5
```

```r
x2 <- runif(n) - 0.5
y <- as.numeric(x1^2 - x2^2 > 0)
```

b) Plot the observations, colored according to their class labels.

c) Fit a logistic regression model to the data using X_1 and X_2 as predictors.

d) Apply this model to the training data in order to obtain a predicted class label for each training observation. Plot the observations, colored according to the *predicted* class labels. What shape is the decision boundary?

e) Now fit a logistic regression model to the data using non-linear functions of X_1 and X_2 as predictors (e.g., $X_1^2, X_1 \times X_2, \log(X_2)$, etc.)

f) Apply this model to the training data in order to obtain a predicted class label for each training observation. Plot the observations, colored according to the *predicted* class labels. What shape is the decision boundary? Repear a)- e) until you come up with an example in which the predicted class labels are obviously non-linear.

g) Fit a support vector classifier with X_1 and X_2 as predictors. Obtain a class predictor for each training observation. Plot the observations, colored according to the *predicted* class labels.

h) Fit an SVM using a non-linear kernel to the data with X_1 and X_2 as predictors. Obtain a class predictor for each training observation. Plot the observations, colored according to the *predicted* class labels.

i) Comment on your results.

3. In this problem, you will use support vector approaches to predict whether a given car gets high or low gas mileage based on the `Auto` data set in the `ISLR` package.

   a) Create a binary variable that takes value 1 for gas mileage above the median and 0 for cars below the median.

   b) Fit a support vector classifier to the data with various values of `cost`, in order to predict whether a car gets high or low gas mileage (be sure not to include the original gass mileage variable – no cheating!). Report the cross-validation errors associated with different values of this parameter, comment on your results.

   c) Now repeat (b) using SVMs with radial and polynomial basis kernels, with different values of `gamma`, `degree`, and `cost`. Report on your results.

   d) Make some plots to back up your assertions in b) and c).

4. This problem involves the `OJ` data set in the `ISLR` package.

   a) Create a training set containing a random sample of 900 observations and a test set containg the remaining observations.

   b) Fit a support vecotr classifier to the training set using `cost = 0.01` with `Purchase` as the response and the other variables as predictors. Use the `summary()` function to produce summary statistics and describe the results obtained.

   c) What are the training and test error rates?

   d) Use the `tune()` function to select an optimal `cost`. Consider values between 0.01 and 10.

   e) Compute the training and test error rates using this new value for `cost`.

   f) Repeat b) through e) using a support vector machine with a radial kernal and default value for `gamma`.

   g) Repeat b) through e) using a support vector machine with a polynomial kernal and `degree = 2`.

h) Which approach gives the best results on this data?