Lab 9: Support Vector Machines

```
library(tidyverse) ## data manipulation
library(knitr) ## tables

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

0.1 Data Preparation

We will make some simulated data to see how support vector classifiers and SVMs work.

Run the following code to create two datasets.

```
n1 <- 20
n2 <- 200
p <- 2

## training data sets
x_small <- matrix(rnorm(n1 * p), ncol = p)
x_large <- matrix(rnorm(n2 * p), ncol = p)
y_small <- c(rep(-1, n1/2), rep(1, n1/2))
y_large <- c(rep(1, n2/4*3), rep(2, n2/4))

## shift data farther apart
x_small[y_small == 1,] <- x_small[y_small == 1,] + 1
x_large[1:100,] <- x_large[1:100,] + 2
x_large[101:150,] <- x_large[101:150,] - 2

## put data into dataframes
df_small <- data.frame(x_small, y = as.factor(y_small))
df_large <- data.frame(x_large, y = as.factor(y_large))
```

1. Make two scatterplots to inspect the small and large training data sets. Describe what you see.

0.2 Support Vector Classifier

We will use the `e1071` library to fit the support vector classifier and the SVM. The `svm` function will fit both, with the `kernel` argument taking values in `"linear", "polynomial", "radial"`.

The `cost` argument allows us to specify the cost of violation to the margin. When the `cost` argument is small, margins will be wide.

```r
library(e1071) ## svm library
```

1. Use the `svm` function to fit a support vector classifier on the small data with $C = 10$ (use `scale = FALSE` to disallow rescaling of your data.
2. Inspect your model using `summary()`. How many support vectors were used to fit your classifier?
3. Predict (`predict()`) a grid of $\boldsymbol{X}$ values between the range of X_1 and X_2. Plot these predictions using `geom_tile()` to visualize the decision boundary and add a scatteplot of training data on top, colored by training label. Describe what you see.
4. Alter your plot from 2 to change the shape of the support vectors.
5. Use the `tune()` function to perform CV on the cost parameter. Which value of C would you choose?
6. Repeat 3. and 4. using your chosen C value. Describe what you see.

0.3 Support Vector Machines

1. Split the large data frame into 50% training and 50% test.
2. Fit a linear SVM, radial SVM with $\gamma = 1$, and polynomial SVM with $d = 3$ using `tune()` to choose the appropriate cost for each model.
3. Predict (`predict()`) a grid of $\boldsymbol{X}$ values between the range of X_1 and X_2. Plot these predictions using `geom_tile()` to visualize the decision boundary and add a scatteplot of training data on top, colored by training label. Describe what you see.
4. Predict your test data with your three models. Which model would you choose?