DSCI445 - Homework 2

Your Name

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

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

Regression

In this exercise you will create some simulated data and will fit simple linear regression models to it. Make sure to use `set.seed(445)` prior to starting part (a) to ensure reproducible results.

(a) Using the `rnorm()` function, create a vector x containing 100 observations drawn from a $N(0,1)$ distribution. This represents the feature, X.

(b) Using the `rnorm()` function, create a vector `eps` containing 100 observations drawn from a $N(0,0.25)$ distribution, i.e. a Normal distribution with mean zero and variance 0.25.

(c) Using `x` and `eps`, generate a vector `y` according to the model

$$Y = -1 + 0.5X + \epsilon.$$

What is the length of the vector `y`? What are the values of β_0 and β_1 in this linear model?

(d) Create a scatterplot displaying the relationship between `x` and `y`. Comment on what you observe.

(e) Fit a least squares linear model to predict `y` from `x`. Comment on the model obtained. How do $\hat{\beta}_0$ and $\hat{\beta}_1$ compare to β_0 and β_1?

(f) Display the least squares line on the scatterplot obtained in (d) in blue. Draw the population regression line on the plot in red. (See `geom_abline()` for how to add a line based on intercept and slope.)

(g) Now fit a polynomial regression model that predicts `y` using `x` and $\texttt{x}^2$. Is there evidence that the quadratic term improves the model fit? Explain your answer.

(h) Repeat (a)-(f) after modifying the data generation process in such a way that there is *less* noise in the data. The model should remain the same. You can accomplish this by changing the variance of the normal distribution used to generate the error term ϵ in (b). Describe your results.

(i) Repeat (a)-(f) after modifying the data generation process in such a way that there is *more* noise in the data. The model should remain the same. You can accomplish this by changing the variance of the normal distribution used to generate the error term ϵ in (b). Describe your results.

(j) What are the confidence intervals for β_0 and β_1 based on the original data set, the noisier data set, and the less noisy data sets? Comment on your results.

```r
# simulated regression task
```

Classification

1. When the number of features p is large, there tends to be a deterioration in the perforance of KNN and other *local* approaches that perform prediction using only observations that are *near* the test observation

for which a prediction must be made. This phenomenon is known as the *curse of dimensionality*, and it ties into the fact that non-parametric approaches often perform poorly when p is large. We will now investigae this curse.

(a) Suppose that we have a set of observations, each with measurements on $p = 1$ feature, X. We assume that X is uniformly (evenly) distributed on $[0, 1]$. Associated with each observation is a response value. Suppose that we wish to predict a test observation's response using only observations that are within 10% of the range of X closest to that test observation. For instance, in order to predict the response for a test observation with $X = 0.6$ we will use observations in the range $[0.55, 0.65]$. On average, what fraction of the available observations will we use to make this prediction?

(b) Now suppose that we have a set of observations, each with measurements on $p = 2$ features, X_1 and X_2. We assume (X_1, X_2) are uniformly distributed on $[0, 1] \times [0, 1]$. we wish to predict a test observation's response using only observations that are within 10% of the range of X_1 *and* within 10% of the range of X_2 closest to that test observation. For instance, in order to predict the response for a test observation with $X_1 = 0.6$ and $X_2 = 0.35$ we will use observations in the range $[0.55, 0.65]$ for X_1 and in the range $[0.3, 0.4]$ for X_2. On average, what fraction of the available observations will we use to make this prediction?

(c) Now suppose that we have a set of observations, each with measurements on $p = 100$ features. Again, the observations are uniformly distributed on each feature, and again each feature ranges from 0 to 1. we wish to predict a test observation's response using only observations that are within 10% of each feature's range that is closest to that test observation. What fraction of the available observations will we use to make this prediction?

(d) Using your answers to (a)–(c), argue that a drawback of KNN when p is large is that there are very few training observations *near* any given test observation.

(e) Now suppose that we wish to make a prediction for a test observation by creating a p-dimensional hypercube centered around the test observation that contains, on average, 10% of the training observations. For $p = 1, 2$, and 100, what is the length of each side of the hypercube? Comment on your answer.

[Hint: A hypercube is a generalization of a cube to an arbitrary number of dimensions. When $p = 1$, a hypercube is simple a line segment. When $p = 2$ it is a square, and when $p = 100$ it is a 100-dimensional cube.]

2. Suppose we collect data for a group of students in a statistics class with variables $X_1 =$ hours studied, $X_2 =$ undergrad GPA, and $Y =$ receive an A. We fit a logistic regression and produce estimated coefficients, $\hat{\beta}_0 = -6, \hat{\beta}_1 = 0.05, \widehat{beta}_2 = 1$.

   (a) Estimate the probability that a student who studies for 40h and has an undergrad GPA of 3.5 gets an A in the class.

   (b) How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class?

3. This question should be answered using the `Weekly` data set, which is part of the `ISLR` package. This data contains weekly percentage returns for the S&P 500 stock index between 1990 and 2010.

   (a) Produce some numerical and graphical summaries of the `Weekly` data. Do there appear to be any patterns?

   (b) Use the full data set to perform a logistic regression with `Direction` as the response and the five lag variables plus `Volume` as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?

   (c) Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.

(d) Now fit the logistic regression model using a traning data period from 1990 to 2009 with `Lag2` as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is the data from 2010).

(e) Repeat (d) using LDA.

(f) Repeat (d) using KNN with K = 1.

(g) Which of these methods appears to provide the best results on this data?

(h) Experiment with different combinations of predictors, including possible transformations and interactions, for each of the methods. Report the variables, method, and associated confusion matrix that appears to provide the best results on the held out data. Note that you can experiment with values for K in the KNN classifier.

```r
## load the data
library(ISLR)

## take a look
head(Weekly)
```

```
##   Year   Lag1   Lag2   Lag3   Lag4   Lag5    Volume  Today Direction
## 1 1990  0.816  1.572 -3.936 -0.229 -3.484 0.1549760 -0.270      Down
## 2 1990 -0.270  0.816  1.572 -3.936 -0.229 0.1485740 -2.576      Down
## 3 1990 -2.576 -0.270  0.816  1.572 -3.936 0.1598375  3.514        Up
## 4 1990  3.514 -2.576 -0.270  0.816  1.572 0.1616300  0.712        Up
## 5 1990  0.712  3.514 -2.576 -0.270  0.816 0.1537280  1.178        Up
## 6 1990  1.178  0.712  3.514 -2.576 -0.270 0.1544440 -1.372      Down
```

4. Using the `Boston` data set, fit classification models in order to predict whether a given suburb has a crime rate above or below the median. Explore logistic regression, LDA, and KNN models using various subsets of the predictors. Describe your findings.

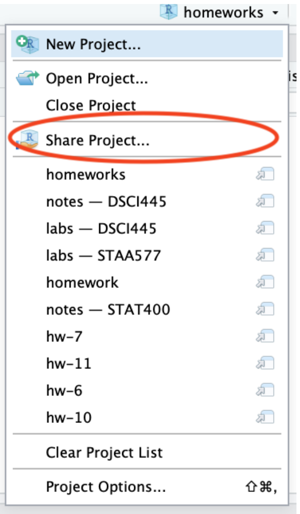
```r
## load the data
library(MASS)

## take a look
head(Boston)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
##   medv
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

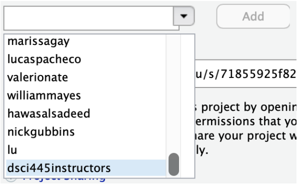
Turn in in a pdf of your homework to canvas using the provided Rmd file as a template. Your Rmd file on the server will also be used in grading, so be sure they are identical.

Be sure to share your server project with the instructor and grader. You only need to do this once per semester.

1. Open your `homeworks` project on liberator.stat.colostate.edu
2. Click the drop down on the project (top right side) > Share Project…



3. Click the drop down and add "dsci445instructors" to your project.



This is how you **receive points** for reproducibility on your homework!