**Stat 627/427 (Statistical Machine Learning)**

# Homework # 4

Multicollinearity (Sec. 3.3.3). Classification methods: logistic regression and K-nearest neighbor (sec. 2.2.3, 3.5, 4.1-4.3)

## Multicollinearlity

1. **(Page 125, chap. 3, #14)**. This problem focuses on multicollinearity.

   (a) Perform the following commands in R:

   ```
   > set.seed (1)
   > x1 = runif (100)
   > x2 = 0.5*x1 + rnorm(100)/10
   > y = 2 + 2*x1 + 0.3*x2 + rnorm(100)
   ```

   The last line corresponds to creating a linear model in which $y$ is a function of $x_1$ and $x_2$. Write out the form of the linear model. What are the regression coefficients?

   (b) What is the correlation between $x_1$ and $x_2$? Create a scatterplot displaying the relationship between the variables.

   (c) Using this data, fit a least squares regression to predict $y$ using $x_1$ and $x_2$. Describe the results obtained. What are $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$? What are the true $\beta_0$, $\beta_1$, and $\beta_2$? Can you reject the null hypothesis $H_0 : \beta_1 = 0$? How about the null hypothesis $H_0 : \beta_2 = 0$?

   (d) Now fit a least squares regression to predict $y$ using only $x_1$. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

   (e) Now fit a least squares regression to predict y using only $x_2$. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_2 = 0$?

   (f) Do the results obtained in (c)–(e) contradict each other? Explain your answer.

   (g) Now suppose we obtain one additional observation, which was unfortunately mismeasured. Use the following R code.

   ```
   > x1=c(x1, 0.1)
   > x2=c(x2, 0.8)
   > y=c(y,6)
   ```

   Re-fit the linear models from (c) to (e) using this new data. What effect does this new observation have on the each of the models? In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers. How do the slopes from all the considered models react on the newly added data point?

   (h) What are standard errors of estimated regression slopes in (a), (d), and (e)? Which models produce more stable and therefore, more reliable estimates?

   (i) Compute both VIF in question (a) and relate them to your answer to question (h).

2. **(for Stat-627 only)** Consider the simple linear regression model $Y_i = \beta_0 + \beta_1 X_1 + \varepsilon_i$, $i = 1, ..., n$, where the errors are independently and identically distributed random variables with mean zero and common variance $\sigma^2$, and all $X_i$ are different constants. Consider the following estimate of the regression slope,

$$\tilde{\beta}_1 = \frac{1}{n(n-1)} \sum_{i=0}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{Y_i - Y_j}{X_i - X_j}.$$

Is $\tilde{\beta}_1$ unbiased? Which one of $\tilde{\beta}_1$ and the least squares estimator $b_1$ would you prefer? Why?

## Logistic Regression

3. **Page 170, chap. 4, #9.** This problem has to do with *odds*.

(a) On average, what fraction of people with an odds of 0.37 of defaulting on their credit card payment will in fact default?

(b) Suppose that an individual has a 16% chance of defaulting on her credit card payment. What are the odds that she will default?

4. **Page 170, chap. 4, #6.** 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, $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = 1$.

(a) Estimate the probability that a student who studies for 40 hours 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% (predicted) chance of getting an A in the class?

## KNN

5. **Pages 53-54, chap. 2, #7.** The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

| Obs. | $X_1$ | $X_2$ | $X_3$ | $Y$ |
|---|---|---|---|---|
| 1 | 0 | 3 | 0 | Red |
| 2 | 2 | 0 | 0 | Red |
| 3 | 0 | 1 | 3 | Red |
| 4 | 0 | 1 | 2 | Green |
| 5 | −1 | 0 | 1 | Green |
| 6 | 1 | 1 | 1 | Red |

Suppose we wish to use this data set to make a prediction for $Y$ when $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors.

(a) Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$.

(b) What is our prediction with $K = 1$? Why?

(c) What is our prediction with $K = 3$? Why?

(d) If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for $K$ to be large or small? Why?