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

# Homework # 3

Regression, Partial F-tests (3.3). Classification: Logistic Regression, KNN (2.2.3, 4.1-4.3, 3.2, 3.5)

*Due February 13 by 12:00 noon, uploaded on Blackboard. Quiz 3 is on February 14*

## Multiple Linear Regression

1. **Page 123, chap. 3, ≈#10.** Consider the following variables from the *Carseats* data set.

| | |
|---|---|
| Sales | Unit sales (in thousands) at each location |
| Price | Price company charges for car seats at each site |
| Urban | A factor with levels No and Yes to indicate whether the store is in an urban or rural location |
| US | A factor with levels No and Yes to indicate whether the store is in the US or not |

(a) Fit a multiple regression model to predict Sales using Price, Urban, and US.

(b) Write out the model in equation form, being careful to handle the qualitative variables properly. *That is, write a separate model for each category.*

(c) For which of the predictors the null hypothesis $H_0 : \beta_j = 0$ is <u>not</u> rejected? *Verify your conclusion with the appropriate partial F-test (you can find examples in HW2 and the regression handout). Compare p-values of this test and the corresponding t-test. Also, compare the partial F-statistic with the squared t-statistic.*

(d) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

(e) Test linearity of Sales as a function of Price, according to our regression model. Is there a significant lack of fit?

## Logistic Regression

2. **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?

3. **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

4. **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?

5. **(For Stat-627 students only)** When the number of features $p$ is large, there tends to be a deterioration in the performance 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 investigate 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 observations 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 the 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 that $(X_1, X_2)$ are uniformly distributed on $[0, 1] \times [0, 1]$. We wish to predict a test observations 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 the prediction?

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

(d) Using your answers to parts (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.

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