STAT 673, Autumn 2009
Sat, Dec. 5th., 2009

Take Home Final Exam

There are three questions in this take home final. You must work on the final exam **alone**. Except for me, you **cannot** discuss this exam with anyone else. If you have questions email me at herbei@stat.osu.edu. Treat this like any other exam (not like a homework!). You may ask me questions of clarification, but I will not do the exam for you.

Your solution to this exam is due on Wed, Dec 9^{th} 2009 at 5:00pm on Carmen (pdf format). Internet connection issues CANNOT be claimed as a reason for not submitting the assignment on time.

Academic misconduct: Cheating, plagiarism and other forms of academic dishonesty will not be tolerated. Any violation will be prosecuted to the fullest extent as set out in University Rule 3335-31-02.

Include **relevant** printouts of your output windows with your submission. Attach the code with your submission as well. Your code must have comments, explaining what you are doing. Graphs should be labeled and titled properly, with a brief caption describing what is being plotted. The answer to each problem should be presented as a report, very briefly describing the problem itself, the methodology that is being used, and a detailed part containing the solution and conclusions.

All the data sets can be found at:
http://www.stat.osu.edu/~herbei/courses/673/Data/

Question 1 (5 points)

Write a 2-page report on a scientific article you have read during this Winter 09 quarter.

Question 2 (20 points)

The `ironslag` (see the website above) data has 53 measurements of iron content by two methods :`chemical` and `magnetic`. A scatterplot of the data suggests that the chemical and magnetic variables are positively correlated, but the relationship may not be linear. It appears that a quadratic polynomial or possibly an exponential or logarithmic model might fit the data better than a line. There are several steps to model selection but we will focus on prediction error in this case. The proposed models for predicting the magnetic measurement (Y) from the chemical measurement X are

1. Linear :$Y = \beta_0 + \beta_1 X + \epsilon$.
2. Quadratic: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$
3. Exponential: $\log(Y) = \beta_0 + \beta_1 X + \epsilon$
4. Log-Log: $\log(Y) = \beta_0 + \beta_1 \log(X) + \epsilon$.

(a) (8 points) Fit the four models above and find estimates for β_0, β_1 and β_2 in each case. Produce a scatter plot of the data and overlay the fitted line(curve) for each model.

(b) (10 points) For each of the models above, estimate the **prediction error** as follows:

- For $k = 1, \ldots, n$ let the observation (x_k, y_k) be the test point and use the remaining observations to fit the model.
- Fit the model(s) using only the $n-1$ observations $(x_i, y_i), i \neq k$.
- compute the predicted response(s) $\hat{y}_k$ for each model.
- Compute the prediction error $e_k = y_k - \hat{y}_k$.
- Estimate the mean squared prediction error

$$\hat{\sigma}_\epsilon^2 = \frac{1}{n} \sum_{i=1}^{n} e_k^2$$

(c) (2 points) Which of the four models above achieves the smallest prediction error ?

Question 3 (25 points)

`Iris Virginica`, `Iris Versicolor` and `Iris Setosa` are species of Iris. Although there are approx. 300 species of Iris known so far, we will assume that all irises are either "setosa", "virginica" or "versicolor".

The data set `iris` contains a sample of 150 specimens (50 of each). In R, use `data(iris)` to load the entire data set. The variables are `sepal.length, sepal.width, petal.length, petal.width` and `species`. When given a new Iris specimen, we would like to predict the `species`, having measurements of the other four variables above.

Perform the following tasks:

Figure 1: Iris setosa (left), virginica (middle) and Iris versicolor (right).

- Exploratory data analysis.

  (a) (5 points) Produce scatter plots for each pair of the four numerical variables in the data set `iris`. Color the points in each scatter plot according to the variable `species`. That is, your plot should look like Figure 2. Based on the scatter plots, comment on whether these variables have discrimination power when trying to distinguish the three categories of Iris. (Hint: you may find the R function `points` useful for this part).

  (b) (5 points) Produce density estimates for each of the four numerical variables in the data set `iris`. A multimodal density estimate for either of these variables, might be indicative of different categories of irises. Again, comment on the discrimination power of these numerical variables.

- Classifying irises.

  (c) (2 points) Define the distance between two specimens $S_i, i = 1, 2$ as

  $$d(S_1, S_2) = \sqrt{(SA_1 - SA_2)^2 + (PA_1 - PA_2)^2},$$

  where SA_1, SA_2, PA_1, PA_2 are the sepal area and petal area respectively for the two specimens (the area can be computed as `length` $\times$ `width`. Write a function that will compute this distance, when given two Iris specimens.

  (d) (3 points) Produce a scatter plot of `Sepal area` vs. `Petal Area`. Color the points according to the variable `species`. What do you observe regarding the discriminatory power of the two new variables (sepal area and petal area) ?

  (e) (5 points) Remove all the `Setosa` irises from the data set. Given a new specimen

  $$S^* = (\text{Sepal area } = 20.0, \quad \text{Petal Area} = 6.0, \quad \text{Species=???})$$

  predict whether it is, "virginica" or "versicolor" by searching in the remaining sample for the **five irises that "closest" to S^* in terms of the distance defined above**. The variable `Species` for S^* is determined by a majority vote between the five irises which are closest to S^*. For example, if the closest five irises are of type ("virginica", "virginica", "versicolor", "virginica", "versicolor"), then we set the variable `species` for S^* to "virginica".

  (f) (5 points) In this part we will continue to work with the reduced data set, obtained by removing al the "Iris setosa" from the original data set. The procedure described in part (e) is not perfect. We can estimate its accuracy by computing an error rate, as follows:

  For every specimen $S_i = (SepalArea_i, PetalArea_i, Species_i)$ in the given sample,

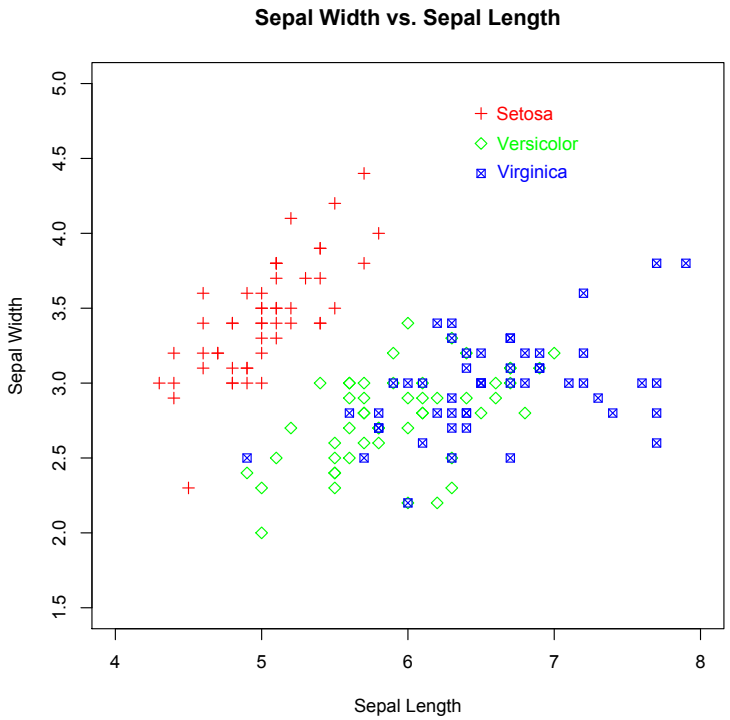

Figure 2: Sepal length vs. Sepal Width for the `iris` data set.

- create a new data set `DataNew` by removing case i from the original data.
- Predict the variable `Species` for case i as in part (e): search in `DataNew` for the five irises closest to S_i and use a majority vote between their `Species` as the predicted value for S_i.
- Compare the predicted value found above with the actual value $Species_i$. Record whether you made a mistake or not.
- The error rate of the procedure is

$$\frac{\#\text{ of mistakes}}{\text{sample size}}.$$

Find the error rate of the procedure described in part (e) for the given sample.