CSC334/424: Assignment #2

Due: Thursday, February 4th, 2016, by 10:00pm
Total: 65 points **(no late assignments for this assignment)**

Notes on putting together the assignment:

1. For problems #1, #2, and #3: Include the SPSS (or other software) output tables that support your reporting and interpretation.
2. When you write the solution for each problem, you need to break it up by the parts of the question instead of one big paragraph.

Problem #1 (Regression analysis - 20 points) The Housing dataset (housing.data) contains housing values in the suburbs of Boston. The detailed explanation concerning the input and output variables can be fetched from the UCI machine learning repository http://archive.ics.uci.edu/ml/datasets/Housing:

1. CRIM: per capita crime rate by town
2. ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS: proportion of non-retail business acres per town
4. CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5. NOX: nitric oxides concentration (parts per 10 million)
6. RM: average number of rooms per dwelling
7. AGE: proportion of owner-occupied units built prior to 1940
8. DIS: weighted distances to five Boston employment centers
9. RAD: index of accessibility to radial highways
10. TAX: full-value property-tax rate per $10,000
11. PTRATIO: pupil-teacher ratio by town
12. B: 1000(Bk - 0.63)^2 where Bk is the proportion of African Americans by town
13. LSTAT: % lower status of the population
14. MEDV: Median value of owner-occupied homes in $1000's (output variable)

a. Fit a linear regression model and report goodness of fit, the utility of the model, the estimated coefficients, their standard errors, and statistical significance. Use the default method for running regression analysis in SPSS and interpret your results.

b. Perform a feature selection on this data by using the forward selection method of the regression analysis. Analyze the output in terms of the order in which the variables are included in the regression model.

Problem #2 (Canonical Correlation Analysis – 20 points): Water, soil, and mosquito fish samples were collected at *n* = 165 sites/stations in the marshes of southern Florida. The following water variables were measured:

MEHGSWB	Methyl Mercury in surface water, ng/L
TURB	In situ surface water turbidity
DOCSWD	Dissolved Organic Carbon in surface water, mg/L
SRPRSWFB	Soluble Reactive Phosphorus in surface water,mg/L or ug/L
THGFSFC	Total Mercury in mosquitofish (*Gambusia affinis*), average of 7 individuals, ug/kg

In addition, the following soil variables were measured:

THGSDFC	Total Mercury in soil, ng/g
TCSDFB	Total Carbon in soil, %
TPRSDFB	Total Phosphorus in soil, ug/g

Perform a canonical correlation analysis, describing the relationships between the soil and water variables using the data[1] found in data_marsh_cleaned_hw2 (both xls and spss files).

1. Answer the following questions regarding the canonical correlations.
    a. Test the null hypothesis that the canonical correlations are all equal to zero. Give your test statistic, d.f., and p-value.
    b. Test the null hypothesis that the second and third canonical correlations equal zero. Give your test statistic, d.f., and p-value.
    c. Test the null hypothesis that the third canonical correlation equals zero. Give your test statistic, d.f., and p-value.
    d. Present the three canonical correlations, together with their standard errors. (Report the standard errors only if you are using SAS; SPSS will not output the standard errors)
    e. What can you conclude from the above analyses?
2. Answer the following questions regarding the canonical variates.
    a. Give the formulae for the significant canonical variates for the soil and water variables.
    b. Give the correlations between the significant canonical variates for soils and the soil variables, and the correlations between the significant canonical variates for water and the water variables.
    c. What can you conclude from the above analyses?

Problem 3 (Principal Component Analysis - 20 points): The data given in the file 'problem3.txt'[2] is the percentage of people employed in different industries in European countries during 1979. Techniques such as Principal Component Analysis (PCA) can be used to examine which countries have similar employment patterns. There are 26 countries in the file and 10 variables as follows:

Variable Names:

1. Country: Name of country
2. Agr: Percentage employed in agriculture
3. Min: Percentage employed in mining
4. Man: Percentage employed in manufacturing
5. PS: Percentage employed in power supply industries
6. Con: Percentage employed in construction
7. SI: Percentage employed in service industries
8. Fin: Percentage employed in finance
9. SPS: Percentage employed in social and personal services

[1] http://www.epa.gov/region4/sesd/reports/epa904r07001.html

[2] http://lib.stat.cmu.edu/DASL/Datafiles/EuropeanJobs.html

10. TC: Percentage employed in transport and communications.

Perform a principal component analysis using the covariance matrix:

a. How many principal components are required to explain 90% of the total variation for this data?
b. For the number of components in part a, give the formula for each component and a brief interpretation.
c. What countries have the highest and lowest values for each principal component (only include the number of components specified in part a). For each of those countries, give the principal component scores (again only for the number of components specified in part a).
d. Include and interpret the scatter plot of the data using the first two principal components.

Problem 4 (overview – 5 points): Briefly describe the similarities and differences between:

a. Linear regression and canonical correlation
b. Canonical correlation and principal component analysis