**CSC334/424: Assignment #3**

Due: Wednesday, July 12, 2017, by 6pm

Total: 35 points **NOTE-NO ALLOWANCE FOR LATE SUBMISSIONS ON THIS ASSIGNMENT!!**

**Problem #1 (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 (covariates) 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 (dependent 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 (dependent) and water (covariate, independent) variables using the data[[1]](#footnote-1) found in data\_marsh\_cleaned\_homework#2 (both xls and spss files under the course documents in the content folder of D2L).

1. Answer the following questions regarding the canonical correlations. Be sure to copy/past the relevant SPSS output for your answer. (a – f are worth 2 points each)  
   1. How many ‘levels’ canonical correlations were found?

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Canonical Correlations** | | | | | | | |
|  | Correlation | Eigenvalue | Wilks Statistic | F | Num D.F | Denom D.F. | Sig. |
| 1 | .386 | .175 | .696 | 4.052 | 15.000 | 433.809 | .000 |
| 2 | .345 | .135 | .818 | 4.176 | 8.000 | 316.000 | .000 |
| 3 | .268 | .077 | .928 | 4.087 | 3.000 | 159.000 | .008 |
| H0 for Wilks test is that the correlations in the current and following rows are zero | | | | | | | |

So, from the above table we can see that there are 3 levels of significant canonical correlations which are present for this dataset.

* 1. Are there any meaningful canonical correlations? Test the null hypothesis that all the canonical correlations are statistically equal to zero. Give your test statistic and p-value.   
       
     As we can see from the above table that the 3 canonical correlations that atre present are all meaningful.

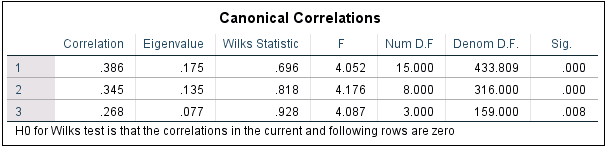
In this the Null hypothesis is that the all the canonical correlations are equal to zero but by looking at the F-value and the associated p-value (which is <0.05) we can say that the three canonical correlations are significant and hence we reject the null hypothesis.

* 1. Test the null hypothesis that, when considered apart from the first canonical correlation, the second and third canonical correlations equal zero. Give your test statistic and p-value.

Null hypothesis is that the second and third canonical correlations are equal to zero. But the F-value for these two canonical correlations is not zero and the p-value associated with these canonical correlations is < 0.05 which means we reject the null hypothesis.

* 1. Test the null hypothesis that the third canonical correlation equals zero. Give your test statistic and p-value.

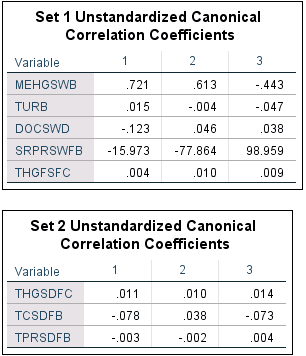
Null hypothesis is that the 3rd dimension is equal to zero. But we know from the table posted in a bit that the F-value and p-value associated with the F-value tell us that we can reject the null hypothesis.

* 1. Present the three canonical correlations (copy/paste the SPSS table)  
     
  2. What can you conclude about the canonical correlations from the above analyses?

From the above analyses, we can say that the correlations are significant and are meaningful and we can perform further analyses using the canonical variates.

1. Answer the following questions regarding the canonical variates. (a – d are worth 2 points each)  
   1. How many canonical variates are there?   
      There are 3 Canonical Variates.
   2. Give the formulae (predictive equations) for each of the significant canonical variates for the soil (dependent) and water variables (covariates, independent). Be sure to include the source data/tables from the CANCORR output in SPSS that supports your answer.

Here all the coefficients have been taken from the UNSTANDARDISED CANONICAL CORRELATION COEFFICIENTS



For Dependent Variables (Soil Indicators) I.E. SET 2:

SoilVariate\_1 = 0.11\*THGSDFC – 0.78\*TCSDFB – 0.003\*TPRSDFB

SoilVariate\_2 = 0.010\*THGSDFC + 0.38\*TCSDFB – 0.002\*TPRSDFB

SoilVariate\_3 = 0.014\*THGSDFC – 0.073\*TCSDFB + 0.004\*TPRSDFB

For Independent variables (Water Indicators) I.E. SET 1:

WaterVariate\_1 = 0.721\*MEHGSWB + 0.015\*TURB – 0.123\*DOCSWD – 15.973\*SRPRSWFB + 0.004\*THGFSFC

WaterVariate\_2 = 0.613\*MEHGSWB – 0.004\*TURB + 0.046\*DOCSWD – 77.864\*SRPRSWFB + 0.010\*THGFSFC

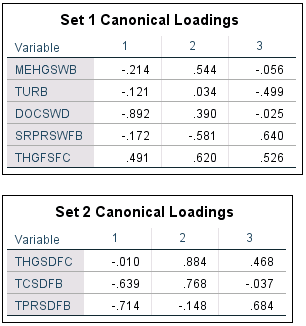
WaterVariate\_3 = - 0.443\*MEHGSWB - 0.047\*TURB + 0.038\*DOCSWD + 98.959\*SRPRSWFB + 0.009\*THGFSFC

* 1. 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.

To see the correlations between the significant canonical variates for soils and the soil variables we would check the canonical loadings for Set 2 since it is related to the soil variates and variables.

Similarly, for water variates and water variable correlation we would infer to the Set 1 Canonical Loadings tables.

The values present in the Canonical Loading table show how much the variable is correlated to the Canonical Variate. The rows represent the correlation value of a variable and the columns are the canonical variate functions.



* 1. What can you conclude from the above analyses?

To understand which variable is the strongest contributor for a given variate we would refer to the Canonical Loadings for each set which is given above. For the first canonical variate, the water variables DOCSWD and THGFSFC contribute the most. For the second Canonical water variate, the SRPRSWFB and THGFSFC are the strongest contributor and similarly for the third variable SRPRSWFB and THGFSFC.

For the first canonical variate for soil the soil variable that is the strongest contributor is TPRSDFB. For second soil variate the soil variable that is the strongest contributor is THGSDFC and similarly for third soil variate the strongest contributor is TPRSDFB.

The following equations which are given below are the equations for the variates:

SoilVariate\_1 = 0.11\*THGSDFC – 0.78\*TCSDFB – 0.003\*TPRSDFB

WaterVariate\_1 = 0.721\*MEHGSWB + 0.015\*TURB – 0.123\*DOCSWD – 15.973\*SRPRSWFB + 0.004\*THGFSFC

So, for SoilVariate\_1 we can say that a unit increase in THGSDFC would result in 0.11 increase in the value for SoilVariate\_1.

A unit increase in TCSDFB would result in a decrease of 0.78 in the value of SoilVariate\_1 when keeping all the other coefficients constant.

A unit increase in the TPRSDFB would result in a decrease of 0.003 for the value of SoilVariate\_1. The same logic can be applied to all the other Variates equations.

**Problem 2 (8 points):** An analysis of German Credit data was based on observations of 30 variables for 1000 past applicants for credit. Each applicant was rated as “good credit” (700 cases) or “bad credit” (300 cases). New applicants for credit can also be evaluated on these 30 "predictor" variables. This data was used to develop a credit scoring rule that can be used to determine if a new applicant is a good credit risk or a bad credit risk, based on values for one or more of the predictor variables.

Misclassification can have adverse impact on profitability, especially when what are actually bad risk applicants are classified as good risk. The consequences of misclassification have been assessed as follows: the costs of a false positive (incorrectly saying an applicant is a good credit risk) outweigh the cost of a false negative (incorrectly saying an applicant is a bad credit risk) by a factor of five. This can be summarized in Figure 1.

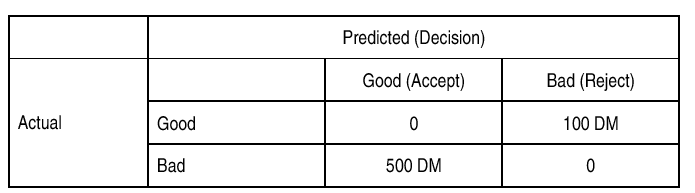
****

Figure 1: Opportunity cost table (in Deutch Marks)

The opportunity cost table was derived from the average net profit per loan as shown below in Figure 2:

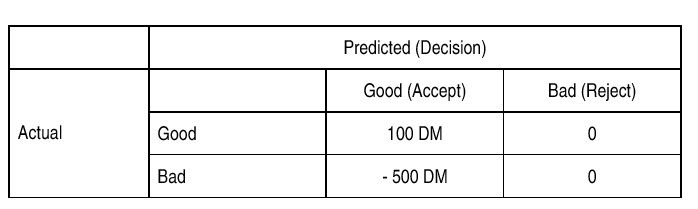


Figure 2: Table for Problem 2

The 1000 rows of data was divided randomly into training (60% “cases selected”) and validation (40% “cases not selected”) partitions and fitted with a linear discriminant classification model. The resulting confusion matrix is below.

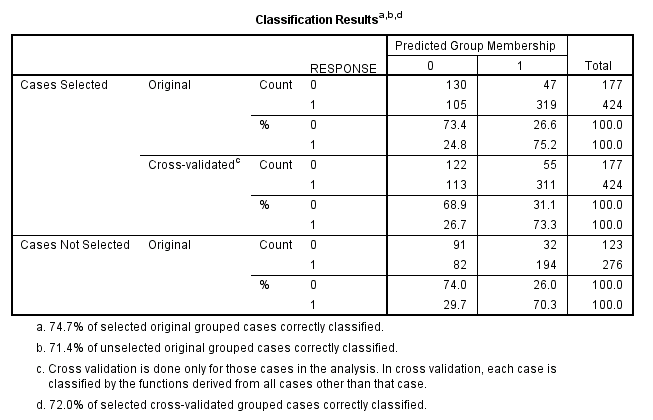
1. Use the data in Figure 2 and the confusion matrix (below) to compute the cost/gain matrix for the validation (“cases not selected”) data. Interpret your results – what do the net profitability figures tell you about the effectiveness of the decision model?

So, based on the question we must compute the cost/gain matrix for the validation (“cases not selected”) data. The table is as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| COST/GAIN MATRIX |  | GOOD(ACCEPT) | BAD(REJECT) |
| ACTUAL | GOOD | 194\*100 = 19,400 | 0 |
|  | BAD | 32\*-500 = -16,000 | 0 |

Profit = 19400-16000 = 3400

Based on the above matrix we can say that the model is slightly profitable. Since we are getting a profit of 3400 DM its up to the Organization to decide whether to continue with this model or not because the penalty incurred due to wrong classification is far greater than the profit made from correct prediction.



**Problem 3 (7 points):** Briefly describe the hierarchical and partitioning clustering approaches, including the advantages and disadvantages of each one of them. Using a field of interest to you, give an example of how clustering can be applied in reducing the number of features, or identifying he characteristics of homogeneous groups, or other aspects of your selected field of interest.

Partitioning Clustering: This clustering technique decomposes a data set into a set of disjoint clusters. In this we specify the value of k which is the final number of clusters that must be present in the output. Each cluster in the output satisfies the following requirements which are as follows: i) Each cluster contains at least one data point ii) each data point belongs to only one group. K-Means and K-Medoids are a couple examples of Partitioning clustering techniques. In K-Means clustering a random point is chosen to be the center of the cluster and then for each data point the distance from the cluster centers is computed and based on the distance metric the data points will be assigned to the clusters. Once assigned the cluster’s center/mean is again computed and the whole process is repeated. Once there is no change in the cluster means the algorithm is stopped. The stopping criteria can be the number of iteration the algorithm must run or if there are no more changes to the cluster means. The cluster means in K-Means is not a data point but may lie on a data point but it is not mandatory. In K-Medoids the process is also similar k-means but the difference is that the mean/center of the cluster must be an actual data point. The disadvantage of this method is that the value of k must be chosen before the start but there is no concrete method for determining the value of k. This method is also prone to noisy data.

Partitioning clustering can be used to create clusters of peoples for a given survey and people belonging to the same cluster will be all have some underlying pattern which common for everyone in the group.

Hierarchical Clustering: In this the clusters have a predetermined ordering from top to bottom an example of this can be the organization of folders on the hard drives. Hierarchical clustering is of two types Divisive (Top Down) and Agglomerative (Bottom Up) and both come up with dendrograms as the final output. In Divisive approach, we start with all observations in a single cluster and based on the distance measurement we do the split till we have the desired number of clusters or till no further splitting can’t be made. In agglomerative clustering, we start with each observation as an individual cluster and we combine the clusters based on the distance metric. The combination of clusters is done till we arrive at the desired number of clusters or when no more grouping can be carried out or when we have our desired cluster. Hierarchical clustering is unstable clustering and hence Partitioning clustering is preferred whenever possible. In Hierarchical clustering deciding when to cut the tree is also difficult. The advantage of using Hierarchical clustering is that only a distance matrix is needed.

1. <http://www.epa.gov/region4/sesd/reports/epa904r07001.html> [↑](#footnote-ref-1)