Software Development for Data Analysis

Discriminant Analysis

Preliminaries

- Under the name of *discriminant analysis* there are reunited various explicative, descriptive and predictive methods designed to study a class or category based population.
- *Discriminant analysis* belongs to the class of supervised learning type of problems, which implies the machine learning task of learning a function that maps an input to an output based on example input-output pairs.
- Each individual observation is characterized by **a set of independent predictor variables and one qualitative variable** whereby the class it belongs to is identified.

Discriminant Analysis

Preliminaries

- The population is divided in 2 subsets:
  - *a)* ***the base sample***, for which the qualitative variable value is known, hence the observations are categorized;
  - *b)* ***the uninvestigated sample***, case in which the observations are not categorized, and the qualitative variable value is not known.

Discriminant Analysis

Preliminaries

- Discriminant analysis intends to:
    a) identify the rules based on which the individual observations can be classified, placed in certain classes or categories,
    b) and, on the other hand, to reduce the number of necessary variables for categorization, or for making the discrimination.
- The first aspect highlights the predictive, decisional character of discriminant analysis, while the second one rather reveals the descriptive aspect of the discriminant analysis.

Discriminant Analysis

Preliminaries

- Discriminant analysis is frequently applied in fields and problems, such as:
  - pattern recognition,
  - financial sector, credit institutions, in order to predict the behavior of credit solicitants,
  - medicine, based on laboratory results, there is to be identified a function for estimating the type of symptoms associated to a disease or its probable evolution,
  - meteorology, the prediction of avalanche, based on the weather related variables, snowfalls etc.

Discriminant Analysis

Notations

- *Observation matrix:*

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots & & & \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix},$$

where n is the number of observations, and m is the number of predictor variables (independent variables).

Discriminant Analysis

Notations

- *<u>Discriminant variable</u>*:

$$Y = \begin{bmatrix} y_1 \\ \cdots \\ y_n \end{bmatrix}.$$

It is a qualitative variable. A value y_i, i = 1,n represents the class (group or category) the observation i belongs to.

There could be $q \neq$ n number of groups, classes or categories.

Discriminant Analysis

Notations

- *Observation vectors*:

  w_i, i=1,n , where w_i is the row i of matrix X.
- *Variable vectors:*

  x_j, j =1,m , where x_j is the column j of matrix X.
- *Group centers matrix*:

$$G = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1m} \\ g_{21} & g_{22} & \cdots & g_{2m} \\ \cdots & & & \\ g_{q1} & g_{q2} & \cdots & g_{qm} \end{bmatrix} \quad ,$$

where q is the number of groups, classes or categories.

Discriminant Analysis

Notations

- A value g_{kj} represent the mean of predictor variable j for the k group.
- Group center vectors G_k, k=1,q, $G_k = \begin{bmatrix} g_{k1} \\ \dots \\ g_{km} \end{bmatrix}$.
- The overall mean: $\overline{X} = \begin{bmatrix} \overline{x}_1 \\ \dots \\ \overline{x}_m \end{bmatrix}$.
- The diagonal matrix of group frequencies:

$$\mathrm{D}_G = \begin{bmatrix} n_1 & 0 & \dots & 0 \\ 0 & n_2 & \dots & 0 \\ \dots & & & \\ 0 & 0 & \dots & n_q \end{bmatrix}.$$

Discriminant Analysis

Variability indicators and dispersion (scatter)

- Discrimination among groups is achieved with variability and dispersion indicators.

A. Scatter matrices (*sum of square and cross product)*:
   - reflect the scatter level associated to the whole collectivity (*SST*),
   - within the groups (*SSW*), and
   - the scatter of groups among each other (*SSB*).

- *SST* is the scatter matrix of the whole collectivity, and shows the scatter level around the overall mean.

Discriminant Analysis

Variability indicators and dispersion (scatter)

- The general term of *SST* matrix is:

$$\mathrm{SST}_{jl} = \sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)\left(x_{il} - \bar{x}_l\right) = \sum_{k=1}^{q}\sum_{i\in k}\left(x_{ij} - \bar{x}_j\right)\left(x_{il} - \bar{x}_l\right) =$$

$$= \sum_{k=1}^{q}\sum_{i\in k}\left(x_{ij} - g_{kj} + g_{kj} - \bar{x}_j\right)\left(x_{il} - g_{kl} + g_{kl} - \bar{x}_l\right) =$$

$$= \sum_{k=1}^{q}\sum_{i\in k}\left(x_{ij} - g_{kj}\right)\left(x_{il} - g_{kl}\right) + \sum_{k=1}^{q}\sum_{i\in k}\left(g_{kj} - \bar{x}_j\right)\left(g_{kl} - \bar{x}_l\right) +$$

$$+ \sum_{k=1}^{q}\sum_{i\in k}\left(x_{ij} - g_{kj}\right)\left(g_{kl} - \bar{x}_l\right) + \sum_{k=1}^{q}\sum_{i\in k}\left(g_{kj} - \overline{x_j}\right)\left(x_{il} - g_{kl}\right)$$

Discriminant Analysis

Variability indicators and dispersion (scatter)

- The general term of *SST* matrix is:

$$\mathrm{SST}_{jl} = \sum_{k=1}^{q}\sum_{i\in k}\left(x_{ij} - g_{kj}\right)\left(x_{il} - g_{kl}\right) + \sum_{k=1}^{q} n_k\left(g_{kj} - \bar{x}_j\right)\left(g_{kl} - \bar{x}_l\right) +$$

$$+ \sum_{k=1}^{q}\left(g_{kl} - \bar{x}_l\right)\sum_{i\in k}\left(x_{ij} - g_{kj}\right) + \sum_{k=1}^{q}\left(x_{il} - g_{kl}\right)\sum_{i\in k}\left(g_{kj} - \bar{x}_j\right)$$

a) The first sum, $\sum_{k=1}^{q}\sum_{i\in k}\left(x_{ij} - g_{kj}\right)\left(x_{il} - g_{kl}\right)$, represents the general term of the scatter matrix within the groups, *SSW*.

Discriminant Analysis

Variability indicators and dispersion (scatter)

b) The second sum, $\sum_{k=1}^{q} n_k \left(g_{kj} - \bar{x}_j\right)\left(g_{kl} - \bar{x}_l\right)$, is general term of the scatter matrix among the groups, *SSB*.

c) The third and the forth sums have the value 0 (zero), because the simple sums of the deviation from the groups means are 0:

$$\sum_{i \in k} \left(x_{ij} - g_{kj}\right) = 0 \text{ and } \sum_{i \in k} \left(x_{il} - g_{kl}\right) = 0$$

- Therefore: $SST_{jl} = SSW_{jl} + SSB_{jl}$, $j=1,m$, $l=1,m$.
- And in terms of matrices: $SST = SSW + SSB$.

Discriminant Analysis

Variability indicators and dispersion (scatter)

- If there are taken into account the degrees of freedom, then the scatter matrices are computed as follows:

$$MST = \frac{SST}{n-1} \;,\; \mathrm{MSW} = \frac{SSW}{n-q} \;,\; \mathrm{MSB} = \frac{SSB}{q-1} \;.$$

Discriminant Analysis

Variability indicators and dispersion (scatter)

B. Covariance matrices:

– The general terms of the covariance matrices differ from those of the scatter matrices by the fact that they are computed as mean values.

– Overall, total covariance:

$$T_{jl} = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \overline{x}_j\right)\left(x_{il} - \overline{x}_l\right) \text{ , } j=1,m \text{ , } l=1,m$$

– Intra-group covariance (within the groups)

$$W_{jl} = \sum_{k=1}^{q}\frac{n_k}{n}\cdot\frac{1}{n_k}\sum_{i\in k}\left(x_{ij} - g_{kj}\right)\left(x_{il} - g_{kl}\right) \text{ , } j=1,m \text{ , } l=1,m$$

Discriminant Analysis

Variability indicators and dispersion (scatter)

– Inter-group covariance (among the groups):

$$B_{jl} = \sum_{k=1}^{q} \frac{n_k}{n} \left(g_{kj} - \overline{x}_j \right)\left(g_{kl} - \overline{x}_l \right), \; j=1,m \;,\; l=1,m$$

- The relation between terms is the same as in the case of scatter matrices:

  $T_{jl} = W_{jl} + B_{jl}$, $j=1,m$, $l=1,m$.
- And in terms of matrices: $T = W + B$.

Discriminant Analysis

Variability indicators and dispersion (scatter)

C. Total variance:

- It is given by the values contained on the principal diagonals of the covariance and scatter matrices:

  $VT = \text{Trace}(T)$ is the general (overall) total variance,

  $VW = \text{Trace}(W)$ is the total intra-group variance,

  $VB = \text{Trace}(B)$ is the total inter-group variance.

- Or, with scatter matrices:

  $VT = \text{Trace}(SST)$,

  $VW = \text{Trace}(SSW)$,

  $VB = \text{Trace}(SSB)$.

Discriminant Analysis

Variability indicators and dispersion (scatter)

D. *Generalized variance:*

– It is computed as determinant of covariance and scatter matrices:

$VGT = |T|$,

$VGW = |W|$,

$VGB = |B|$.

– Or, with scatter matrices:

$VGT = |SST|$,

$VGW = |SSW|$,

$VGB = |SSB|$.

Discriminant Analysis

Model significance. Statistical tests.

- Model testing is executed in 2 phases:
  - One Fisher test, based on Wilks statistics, which shows if the set of predictor variables can make the discrimination on the groups of instances as a whole;
  - Individual statistical tests for each predictor variables whereby is decided whether such a variable can be a good predictor.

a) *<u>The global F test</u>*:

H0: $G_1 = G_2 = ... = G_q$

H1: $\exists$ two groups i, k such that $G_i \neq G_k$

Discriminant Analysis

Model significance. Statistical tests.

- It is computed the following lambda indicator:

$$\Lambda = \frac{|SSB|}{|SSB + SSW|}$$

- The greater Λ is, the more likely that the H0 hypothesis is to be rejected.
- The test statistics is a Fisher value computed as follows:

$$F = \frac{1 - \Lambda^{1/b}}{\Lambda^{1/b}} \frac{ab - c}{m(q-1)}$$

where *a*, *b* and *c* are computed as:

Discriminant Analysis

Model significance. Statistical tests.

where a, b and c are computed as:

$$a = n - q - \frac{m - q + 2}{2}$$

$$b = \begin{cases} \sqrt{\dfrac{m^2(q-1)-4}{m^2+(q-1)^2-5}} & \text{, if } m^2+(q-1)^2-5>0 \\ 1 & \text{, if } m^2+(q-1)^2-5\le 0 \end{cases}$$

$$c = \frac{m(q-1)-2}{2}$$

Discriminant Analysis

Model significance. Statistical tests.

- If $F^{Computed} > F^{Critic}_{m(q-1),ab-c;\alpha}$,

then the null hypothesis (H0) is rejected with a degree of credence 1-α.

b) *<u>Individual statistical tests, for each predictor variable</u>*:

- A predictor variable is considered a good predictor if is able to separate the groups as clear as possible.
- Therefore, the ratio between the inter-group variance and the intra-group variance is as great as possible.

Discriminant Analysis

Model significance. Statistical tests.

- Having the ratio between variances, the test F is to be applied.
- Therefore, for a given predictor variable j, the null and the alternative hypothesis are:

  H0: $g_{1j} = g_{2j} = ... = g_{qj}$

  H1: $\exists\ k, i$ two groups such that $g_{kj} \neq g_{ij}$

- The statistics of the test is: $F_j = \dfrac{SSB_j}{SSW_j}$.
- The critical value for q-1 and n-q degrees of freedom and a significance threshold α is:

  $F^{Critic}_{q-1;n-q;\alpha}$.

Discriminant Analysis

Model significance. Statistical tests.

- If $F_j^{Computed} > F_{q-1;n-q;\alpha}^{Critic}$,

then the null hypothesis (H0) is rejected with level of trust 1-α.