Linear Discriminant Analysis Numerical Example

By Kardi Teknomo, PhD (../../copyright.html) .

< Previous (LDA.html) | Next (Cluster%20and%20discriminant%20analysis.html) | Index (index.html) >

Numerical Example of Linear Discriminant Analysis (LDA)

Here is an example of LDA. We are going to solve linear discriminant using MS excel. You can download the worksheet companion of this numerical example here. (../../download/download.php?file=LDA)

Factory "ABC" produces very expensive and high quality chip rings that their qualities are measured in term of curvature and diameter. Result of quality control by experts is given in the table below.

| Curvature	Diameter	Quality Control Result
2.95	6.63	Passed
2.53	7.79	Passed
3.57	5.65	Passed
3.16	5.47	Passed
2.58	4.46	Not Passed
2.16	6.22	Not Passed
3.27	3.52	Not Passed

As a consultant to the factory, you get a task to set up the criteria for automatic quality control. Then, the manager of the factory also wants to test your criteria upon new type of chip rings that even the human experts are argued to each other. The new chip rings have curvature 2.81 and diameter 5.46.

Can you solve this problem by employing Discriminant Analysis?

Solutions

When we plot the features, we can see that the data is linearly separable. We can draw a line to separate the two groups. The problem is to find the line and to rotate the features in such a way to maximize the distance between groups and to minimize distance within group.

original data
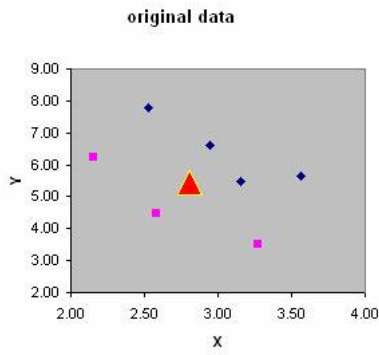
$\mathbf{x}$ = features (or independent variables) of all data. Each row (denoted by k) represents one object; each column stands for one feature.

$\mathbf{y}$ = group of the object (or dependent variable) of all data. Each row represents one object and it has only one column.

In our example,
$$\mathbf{x} = \begin{bmatrix} 2.95 & 6.63 \\ 2.53 & 7.79 \\ 3.57 & 5.65 \\ 3.16 & 5.47 \\ 2.58 & 4.46 \\ 2.16 & 6.22 \\ 3.27 & 3.52 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 2 \\ 2 \\ 2 \end{bmatrix}$$
and

$\mathbf{x}_k$ = data of row k. For example, $\mathbf{x}_3 = \begin{bmatrix} 3.57 & 5.65 \end{bmatrix}$, $\mathbf{x}_7 = \begin{bmatrix} 3.27 & 3.52 \end{bmatrix}$

g = number of groups in $\mathbf{y}$. In our example, $g = 2$

$\mathbf{x}_i$ = features data for group i. Each row represents one object; each column stands for one feature. We separate $\mathbf{x}$ into several groups based on the number of category in $\mathbf{y}$.

$$\mathbf{x}_1 = \begin{bmatrix} 2.95 & 6.63 \\ 2.53 & 7.79 \\ 3.57 & 5.65 \\ 3.16 & 5.47 \end{bmatrix} \quad \mathbf{x}_2 = \begin{bmatrix} 2.58 & 4.46 \\ 2.16 & 6.22 \\ 3.27 & 3.52 \end{bmatrix}$$

μ_i = mean of features in group i, which is average of $\mathbf{x}_i$

$\mu_1 = \begin{bmatrix} 3.05 & 6.38 \end{bmatrix}$, $\mu_2 = \begin{bmatrix} 2.67 & 4.73 \end{bmatrix}$

μ = global mean vector, that is mean of the whole data set.

In this example, $\mu = \begin{bmatrix} 2.88 & 5.676 \end{bmatrix}$

$\mathbf{x}_i^o$ = mean corrected data, that is the features data for group i, $\mathbf{x}_i$, minus the global mean vector μ

$$\mathbf{x}_i^o = \begin{bmatrix} 0.060 & 0.951 \\ -0.357 & 2.109 \\ 0.679 & -0.025 \\ 0.269 & -0.209 \end{bmatrix} \quad \mathbf{x}_2^o = \begin{bmatrix} -0.305 & -1.218 \\ -0.732 & 0.547 \\ 0.386 & -2.155 \end{bmatrix}$$

$\mathbf{c}_i = \dfrac{\left(\mathbf{x}_i^o\right)^T \mathbf{x}_i^o}{n_i}$ = covariance matrix of group i

$$\mathbf{c}_1 = \begin{bmatrix} 0.166 & -0.192 \\ -0.192 & 1.349 \end{bmatrix} \quad \mathbf{c}_2 = \begin{bmatrix} 0.259 & -0.286 \\ -0.286 & 2.142 \end{bmatrix}$$

$C(r,s) = \dfrac{1}{n}\sum_{i=1}^{g} n_i \cdot c_i(r,s)$ = pooled within group covariance matrix. It is calculated for each entry (r,s) in the matrix. In our example, $\frac{4}{7}\cdot 0.166 + \frac{3}{7}\cdot 0.259 = 0.206$, $\frac{4}{7}(-0.192) + \frac{3}{7}(-0.286) = -0.233$ and $\frac{4}{7}\cdot 1.349 + \frac{3}{7}\cdot 2.142 = 1.689$, therefore

$$C = \begin{bmatrix} 0.206 & -0.233 \\ -0.233 & 1.689 \end{bmatrix}$$

The inverse of the pooled covariance matrix is $C^{-1} = \begin{bmatrix} 5.745 & 0.791 \\ 0.791 & 0.701 \end{bmatrix}$

P = prior probability vector (each row represent prior probability of group i). If we do not know the prior probability, we just assume it is equal to total sample of each group divided by the total samples, that is $p_i = \frac{n_i}{N}$

$$p = \begin{bmatrix} 0.571 \\ 0.429 \end{bmatrix} = \begin{bmatrix} \frac{4}{7} \\ \frac{3}{7} \end{bmatrix}$$

Discriminant function

$$f_i = \mu_i C^{-1} x_k^T - \tfrac{1}{2} \mu_i C^{-1} \mu_i^T + \ln(p_i)$$

We should assign object k to group i that has maximum f_i

The results of our computation are given in MS Excel as shown in the figure below.

	A	B	C	D	E	F	G	H	I	J	K
1	Linear Discriminant Analysis: 2 Category 2 Dimensional Data										
2											
3	Training Data, D				Mean Corrected Data			Discriminant function			Results
4	class	X1	X2		X1o	X2o		f1	f2		Classificatio
5	1	2.95	6.63		0.060	0.951		55.220	53.071		1
6	1	2.53	7.79		-0.357	2.109		53.774	51.394		1
7	1	3.57	5.65		0.679	-0.025		62.476	59.589		1
8	1	3.16	5.47		0.269	-0.209		51.953	50.764		1
9	2	2.58	4.46		-0.305	-1.218		32.028	34.313		2
10	2	2.16	6.22		-0.732	0.547		34.554	35.757		2
11	2	3.27	3.52		0.386	-2.155		41.174	42.414		2
12	prediction	2.81	5.46		-0.078	-0.219		44.049	44.085		2
13											
14	average	2.888	5.676								
15	std dev	0.454	1.300								

The discriminant function is our classification rules to assign the object into separate group. If we input the new chip rings that have curvature 2.81 and diameter 5.46, reveal that it does not pass the quality control.

Transforming all data into discriminant function (f_1, f_2) we can draw the training data and the prediction data into new coordinate. The discriminant line is all data of discriminant function $X = \{f_1, f_2\}$ and $Y = \{f_1, f_2\}$. In MS Excel, you can hold CTRL key wile dragging the second region to select both regions.
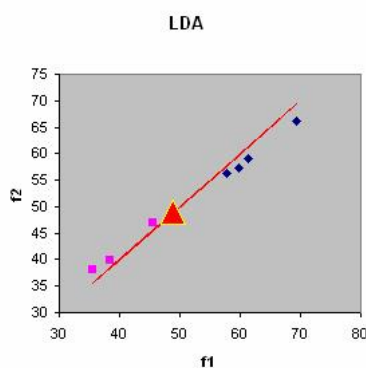
< Previous (LDA.html) | Next (Cluster%20and%20discriminant%20analysis.html) | Index (index.html) >

Preferable reference for this tutorial is

Teknomo, Kardi (2015) Discriminant Analysis Tutorial. http://people.revoledu.com/kardi/ tutorial/LDA/

Immune Repertoire Profiling

Sequencing within reach, for any application.

Illumina

New Crystal 4K TV
Samsung.com