MS-E2112 Multivariate Statistical Analysis (5cr)
Lecture 9: Discriminant Analysis and Classification

Lecturer: Pauliina Ilmonen
Slides: Ilmonen/Kantala

Contents

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Discriminant Analysis

Discriminant Analysis

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

The aim in discriminant analysis is to find a way to separate two or more classes of objects or events. That is then used in classification of new observations.

Discriminant Analysis

Consider $g, \; g > 1$, categories (populations or groups). The object in discriminant analysis is to allocate an individual to one of these g groups based on his measurements. For example, the population might consist of different diseases and the measurement is the symptoms of a patient. Thus one is trying to diagnose a patient's disease based on his symptoms.

Discriminant Analysis, Examples

- Two bivariate normally distributed populations.
- More complicated bivariate settings.

Discriminant Analysis, Normal Variables

Discriminant Analysis, Normal Variables

Let $n \times p$ matrix

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_g \end{bmatrix},$$

where each X_i, $i \in 1, ..., g$, is an $n_i \times p$ data matrix corresponding to group/population i coming from normal distribution $N(\mu_i, \Sigma_i)$. We here assume that the covariance matrices Σ_i are always of full rank.

Discriminant Analysis, Normal Variables

The probability density function of $N(\mu, \Sigma)$ distributed variables (with full rank covariance matrix) can be given as

$$(2\pi)^{-p/2} det(\Sigma)^{-1/2} exp(-1/2((x-\mu)^T \Sigma^{-1}(x-\mu)))$$

and the parameters μ and Σ can be estimated consistently by the sample mean vector and the sample covariance matrix, respectively.

Discriminant Analysis, Normal Variables

Under the assumption of normal distributions, an observation x can be allocated to one of the g groups on the basis of estimated probability density functions. Let $S_i = cov(X_i)$, and let $\bar{x}_i = mean(X_i)$. The observation x is allocated to group j, if

$$ln(det(S_j)) + (x - \bar{x}_j)^T S_j^{-1}(x - \bar{x}_j) < ln(det(S_i)) + (x - \bar{x}_i)^T S_i^{-1}(x - \bar{x}_i), \text{ for all } i \neq j.$$

Discriminant Analysis, Normal Variables

If the g groups are assumed to come from normal distributions with equal covariance matrices, then a consistent estimate of the common covariance matrix Σ is given by

$$S = \frac{1}{n-g}\sum_{i=1}^{g}(n_i - 1)S_i.$$

An observation x is allocated to group j, if

$$(x-\bar{x}_j)^T S^{-1}(x-\bar{x}_j) < (x-\bar{x}_i)^T S^{-1}(x-\bar{x}_i), \text{ for all } i \neq j.$$

Fisher's Linear Discriminant Function

Fisher's Linear Discriminant Function

Let $n \times p$ matrix

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_g \end{bmatrix},$$

where each X_i, $i \in 1, ..., g$, is an $n_i \times p$ data matrix corresponding to group/population i.

Fisher's Linear Discriminant Function

Let

$$W = \sum_{i=1}^{g} (n_i - 1) S_i,$$

where $S_i = cov(X_i)$, and let

$$B = \sum_{i=1}^{g} n_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T.$$

The matrix W measures within group dispersions and the matrix B measures dispersion between groups.

Fisher's Linear Discriminant Function

Fisher's linear discriminant function is the linear function $a^T x$, where a is the maximizer of

$$\frac{a^T Ba}{a^T Wa}.$$

Thus Fisher's linear discriminant function is a linear function that maximizes the ratio of between groups dispersion and within group dispersions.

The solution is obtained by setting a to be equal to the eigenvector of $W^{-1}B$ that corresponds to the largest eigenvalue.

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Fisher's Linear Discriminant Function

Once linear discriminant function has been calculated, an observation x can be allocated to one of the g groups on the basis of its discriminant score $a^T x$. The observation x is allocated to the population whose mean score is closest to the $a^T x$. That is, x is allocated to group j, if

$$|a^T x - a^T \bar{x}_j| < |a^T x - a^T \bar{x}_i|, \text{ for all } i \neq j.$$

Fisher's Linear Discriminant Function

Fisher's linear discriminant function is most important in the special case of $g = 2$ groups. Then the matrix B has rank 1, and it can be written as

$$B = \frac{n_1 n_2}{n} dd^T,$$

where $d = \bar{x}_1 - \bar{x}_2$. Thus, $W^{-1}B$ has only one non-zero eigenvalue and that equals to

$$tr(W^{-1}B) = \frac{n_1 n_2}{n} d^T W^{-1} d.$$

The corresponding eigenvector is

$$a = W^{-1}d.$$

Fisher's LDA, Example 1

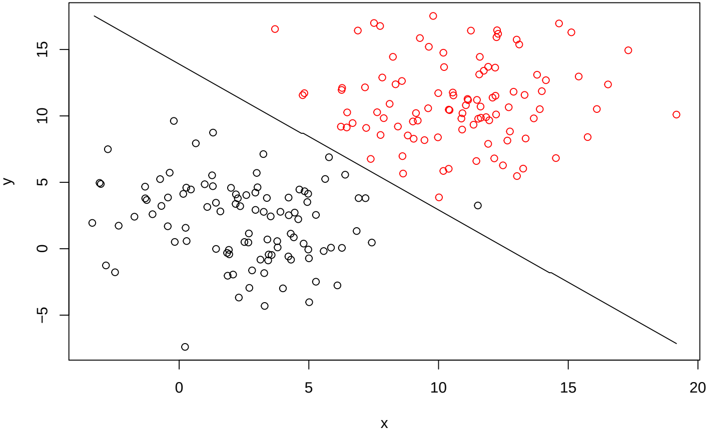

Figure: Fisher's linear discriminant analysis under normality (two groups).

Fisher's LDA, Example 2

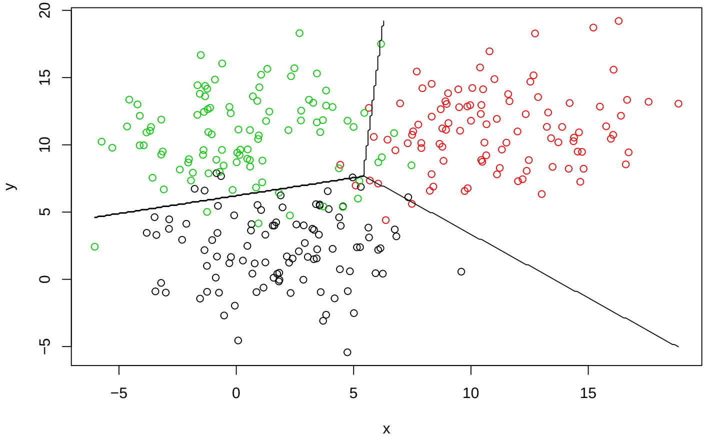

Figure: Pairwise Fisher's linear discriminant analysis under normality (three groups).

Statistical Depth

Statistical Depth

Let $S_n = \{x_1, ..., x_n\}$ denote a set of p variate observations from distribution F_x. Statistical depth $D(y, S_n)$ measures centrality of any p variate y with respect to S_n. The value of $D(y, S_n)$ is always between 0 and 1 and the larger the value of $D(y, S_n)$ is, the more central y is with respect to S_n.

Mahalanobis Depth

Let $S_n = \{x_1, ..., x_n\}$ denote a set of p variate observations from distribution F_x. The Mahalanobis depth $D_M(y, S_n)$ is defined as follows.

$$D_M(y, S_n) = \frac{1}{1 + d^2},$$

with

$$d = \sqrt{(y - \bar{x})^T C^{-1} (y - \bar{x})},$$

where $\bar{x}$ is the sample mean vector and C the sample covariance matrix calculated from the sample S_n.
Similar depth functions may be constructed by replacing the sample mean vector with some other location vector and the sample covariance matrix by some other scatter matrix.

Mahalanobis Depth, population version

Let x denote a p variate random variable with cumulative distribution function F_x. The population Mahalanobis depth $D_M(y, F_x)$ is defined as follows.

$$D_M(y, F_x) = \frac{1}{1 + d^2},$$

with

$$d = \sqrt{(y - \mu)^T \Sigma^{-1} (y - \mu)},$$

where $\mu = \mu(F_x)$ is the mean vector and $\Sigma = \Sigma(F_x)$ is the covariance matrix of the random variable x.

Half Space Depth

Let $S_n = \{x_1, ..., x_n\}$ denote a set of p variate observations from distribution F_x. The half space depth $D_H(y, S_n)$ is defined as follows.

$$D_H(y, S_n) = \min_{u \in U} \frac{1}{n} |\{x_i \in S_n \mid u^T(x_i - y) \geq 0\}|,$$

where U denotes the unit sphere in $\mathbb{R}^p$.

Half Space Depth, Example

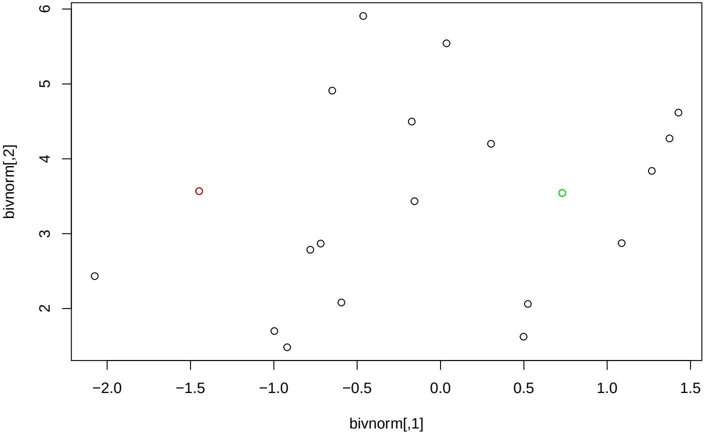

Figure: Bivariate normal distribution. The half space depth value of the red point is $2/20 = 0.1$. The half space depth value of the green point is $5/20 = 0.25$.

Half Space Depth, Population Version

Let x denote a p variate random variable with cumulative distribution function F_x. The population half space depth $D_H(y, F_x)$ is defined as follows.

$$D_H(y, F_x) = \inf_{u \in U} P(u^T(x - y) \geq 0),$$

where U denotes the unit sphere in $\mathbb{R}^p$.

Depth Functions

Mahalanobis depth and half space depth are just two examples of statistical depth functions. There are several other depth functions that have been presented in the literature.
Let x denote a p variate random variable with cumulative distribution function F_x. In general, depth functions should fulfill the following properties (Zuo and Serfling):

- Affine invariance: For any p vector b and any $p \times p$ matrix A, $D(y, F_x) = D(Ay + b, F_{Ax+b})$.
- Maximality at center: If there exist a unique point of symmetry θ such that $\theta + x$ is distributed as $\theta - x$, then $D(\theta, F_x) = \sup_y D(y, F_x)$.
- Monotonicity with respect to the deepest point: If there exist a deepest point α, then for any p vector v $D(\alpha + tv, F_x)$ is monotonically decreasing function of $t > 0$.
- Vanishing at infinity:$D(y, F_x) \to 0$, as $||y|| \to \infty$.

Classification Based on Statistical Depth

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Classification Based on Statistical Depth

Consider two samples $S_n = \{x_1, ..., x_n\}$ and $T_m = \{z_1, ..., z_m\}$ from distributions F_x and F_z, respectively. A new observation y can now be allocated as coming from F_x or F_z by using a depth function. If $D(y, S_n) \geq D(y, T_m)$, the observation y is allocated as coming from F_x, and otherwise it is allocated as coming from F_z.

The procedure generalizes naturally to several distributions. The observation is allocated as coming from the distribution F_w that corresponds to the largest depth value for y.

Other Approaches

Other Approaches

- Classification based "closest neighbors" or on local depths.
- Context related classification.
- ...

Misclassification Rates

Misclassification Rates

In discriminant analysis, it is desirable to find such classification rules that reduce misclassification as much as possible. In practice one can also take into account the costs of misclassification. For example, it can be worse not to detect an illness than to classify a healthy individual as ill.

Misclassification Rates

Calculating exact misclassification rates can be difficult or even impossible when exact underlying distributions are not known.

Misclassification Rates

Misclassification rates are often estimated by calculating sample misclassification rates. After defining a classification rule, the data is classified according to that rule, and sample misclassification rate is obtained. Note that estimated misclassification rates obtained this way grossly underestimate the true misclassification rates - even when sample sizes n_i are large. The problem comes from the fact that the same sample is used to construct the rule and also to test the quality of the classification

Misclassification Rates, Training Sample

Misclassification rates can also be estimated by dividing the original sample into two parts. A training sample (for example 80% of the observations) is used to construct the rule. The rest of the sample is used in approximating the misclassification rate. However, this approach requires large sample sizes and the evaluated classification rule is not the same rule as the one that would be obtained using the entire original sample.

Example

Three different viruses were spreading in the city. Viruses B and C are typically not lethal, whereas virus A requires immediate medical attention, as it is lethal if untreated.
The next slide contains symptoms and laboratory test results of the virus type for 20 patients.

Example

P. num	Age	Gender	Fever(l)	Fever(h)	Rash	Sore throat	Headache	Nausea	Diarrhea	Sleepy	Muscle cramps	Vomiting	Virus
1	19	F	x		x	x		x	x			x	C
2	54	M		x	x					x	x	x	B
3	86	M		x	x	x					x	x	B
4	47	M		x	x		x			x	x	x	B
5	11	F		x		x		x	x			x	C
6	32	-		x			x			x	x	x	A
7	66	F		x	x	x				x	x		B
8	12	M		x		x		x	x			x	C
9	33	F	x		x		x			x	x	x	C
10	18	F	x					x	x			x	C
11	48	M		x	x		x			x	x	x	B
12	78	M	x		x	x		x	x			x	C
13	90	F		x	x					x	x		B
14	36	M		x			x	x		x	x	x	A
15	9	F		x	x	x				x	x	x	C
16	30	F		x	x					x	x	x	B
17	25	F	x		x	x	x	x	x			x	C
18	6	M		x	x	x		x				x	C
19	21	F	x		x		x			x	x	x	B
20	17	-	x		x	x		x	x			x	C

Example

Typical symptoms of the viruses.

Virus	A	B	C
Symptoms	High-grade fever, headache, sleepy, muscle cramps, vomiting	High-grade fever, headache, rash, sleepy, muscle cramps, vomiting	Fever, sore throat, nausea, diarrhea, vomiting, rash

Example

Three new patients have the following symptoms. For each patient — determine the virus that is causing the patient's symptoms.

	Age	Gender	Symptoms
1	23	M	High-grade fever, headache, rash, sleepy, muscle cramps, vomiting
2	49	F	Low-grade fever, sore throat, nausea, diarrhea, vomiting, rash
3	17	F	Headache, nausea, diarrhea, vomiting, rash
4	55	M	High-grade fever, sore throat, nausea, diarrhea, vomiting, rash

Example

"Correct" answers:

Patient	Virus	Reason
1	B	Very typical symptoms of virus B.
2	C	Typical symptoms of virus C.
3	C	Headache is not a typical symptom of virus C. However, one can reason that diarrhea and vomiting may cause dehydration which then causes headache. Patient 3 does not have fever, but it was noted earlier that virus C does not always cause fever.
4	B + C	Did you notice, that none of the adults with virus C had high-grade fever? Here, it was something else that caused the high-grade fever — virus B.

Next Week

Next week we will talk about clustering.

References

References I

K. V. Mardia, J. T. Kent, J. M. Bibby, Multivariate Analysis, Academic Press, London, 2003 (reprint of 1979).

References II

- R. V. Hogg, J. W. McKean, A. T. Craig, Introduction to Mathematical Statistics, Pearson Education, Upper Sadle River, 2005.
- R. A. Horn, C. R. Johnson, Matrix Analysis, Cambridge University Press, New York, 1985.
- R. A. Horn, C. R. Johnson, Topics in Matrix Analysis, Cambridge University Press, New York, 1991.

References III

R. Y. Liu, J. M. Parelius, K. Singh, Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference (with discussion), The Annals of Statistics, 27, 783–858, 1999.

Y. Zuo, R. Serfling, General notions of statistical depth function, The Annals of Statistics, 28, 461–482, 2000.