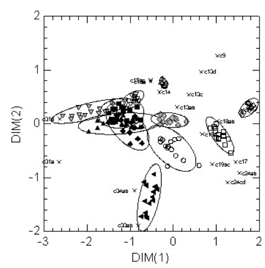
MULTIVARIATE STATISTICAL ANALYSIS FOR FOOD SCIENCE AND AGRICULTURE: AN INTRODUCTION
8. SUPERVISED PATTERN RECOGNITION

Prof. Eugenio Parente
Scuola di Scienze Agrarie- Università della Basilicata



Outline

- Classification problems
- Discriminant analysis (DA)
    - linear discriminant analysis (LDA)
        - discrimination functions and statistics
        - LDA example: the Iris dataset
        - Stepwise LDA
        - Classification matrix and diagnostic tools
        - Cross-validation
        - Classification of unknown objects
    - quadratic discriminant analysis
- Classification (CT) trees (outline)
- Supervised artificial neural networks (sANN) (outline)

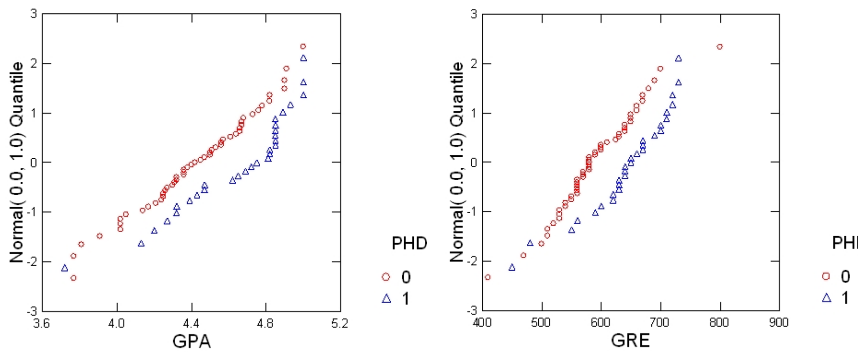
The classification problem(s)

- **Classification problems**
  - One grouping variable (y), 2 or more (usually several) (normally distributed) continuous variables (**X**) (DA, CT, sANN)
  - One grouping variable (y), 2 or more (usually several) categorical (nominal, ordinal) variables (**X**) (CT, sANN)
  - One grouping variable (y), 2 or more (usually several) categorical (nominal, ordinal) variables (**X**) (sANN)
- **Applications**
  - **Classification in taxonomy**: find a method/function to obtain the best classification of biological specimens for which the a priori membership is known into known groups (species); use the method/function to classify unknown specimens
  - **Diagnostics in medicine** (or plant pathology): : find a method/ function to obtain the best classification of specimen/subjects on the basis of symptoms for which the a priori membership is known into known groups (disease or lack thereof); use the method for diagnosis
  - **Food authenticity**: discriminating PDO cheese from imitation cheese from multivariate data sets (gross composition, proteolysis, etc) (PLS-DA works better, though)

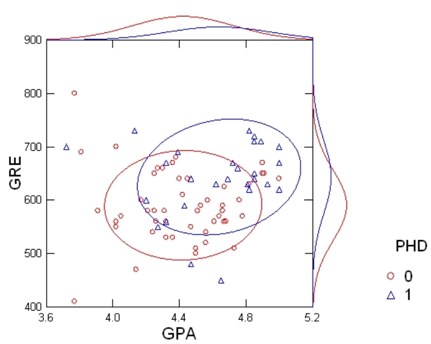
A simple problem

- A University is trying to reduce failures in obtaining a PhD degree
- For a number of students who have obtained or not the degree (i.e. prior group membership is known) the Grade Point Average (GPA) in previous degrees and the score for the Graduate Record Examination (GRE) are collected
- The data are in the file admit.syz
- Which is the best classification rule?

Exploring the admit dataset: probability plots

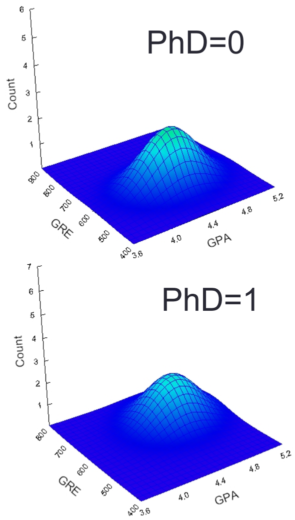


Exploring the admit dataset: scatterplot



- Group means are different but confidence ellipses (p=0.68) on samples overlap
- The covariance matrix is likely to be different (different tilt of ellipses)

Density plots, the covariance matrix, and the discriminant function



Allocation of individuals to group G_i on the basis of multivariate normal distributions can be done using Mahalanobis distance between the individual x vector and the vector for the G_i mean

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x}-\boldsymbol{\mu})'\mathbf{S}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

Fisher's linear discriminant function

- A linear function (discriminant function, z) of the variables is calculated in such a way that the ratio of between group variance to within group variance of the function is maximized
- Cases are allocated to groups on the basis of the discriminant score and group means for discriminant scores: if $\mu_1<\mu_2$ then x_1 belongs to G_1 if $z_1<0.5*(\mu_1+\mu_2)$

$$z = a_1x_1 + a_2x_2 + \ldots + a_px_p$$

$$\max(V) = \frac{\mathbf{a}'\mathbf{B}\mathbf{a}}{\mathbf{a}'\mathbf{S}\mathbf{a}} \rightarrow (\mathbf{B}-\lambda\mathbf{S})\mathbf{a} = 0$$

The output for the admit file

Group Frequencies

0	1
51	29

The group frequencies are unequal; if true prior probabilities are known they can be used to adjust the covariance

Group Means

	0	1
GPA	4.42255	4.63862
GRE	590.49020	643.44828

Between-Groups F-matrix
df: 2 77

	0	1
0	0.00000	
1	9.46905	0.00000

Wilks's Lambda

Lambda	:	0.80260
df	:	(2, 1, 78)
Approx. F-Ratio	:	9.46905
df	:	(2, 77)
p-Value	:	0.00021

Wilk's lambda is a MANOVA statistics to test significance of difference of group means

The output for the admit file

Classification Functions

	0	1
CONSTANT	-133.66677	-150.55267

Canonical scores of group means

	0	1
GPA	44.81788	46.92002
GRE	0.11554	0.12656

Coefficients for the classification function

Variable		F-to-Remove	Tolerance	Variable		F-to-Enter	Tolerance
2	GPA	6.61985	0.99443				
5	GRE	8.90074	0.99443				

F values for testing inequality of group means

Classification Matrix (Cases in row categories classified into columns)

	0	1	%correct
0	45	6	88
1	12	17	59
Total	57	23	78

Classification matrix: true classification on colums, classification from LDA on rows

Jackknifed Classification Matrix

	0	1	%correct
0	44	7	86
1	12	17	59
Total	56	24	76

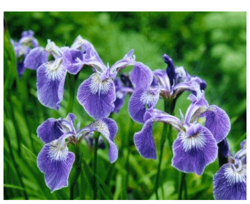
Same, but calculated on the results of the procedure repeated n times, leaving each time one case out

More statistics for the classification matrix

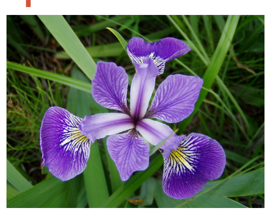
		True group mebership			
		No PhD	PhD	% correct	
Estimated group membership	No PhD	TP=44	FP=7	86	PPV
	PhD	FN=12	TN=17	59	NPV
	Total	56	24	76	
		Sens=79	Spec=71		

- PPV=positive predictive value=TP/(TP+FP)
- NPV=negative predictive value=TN/(TN+FN)
- Sensitivity=TP/(TP+FN) (high=low type II error)
- Specificity=TN/(FP+TN) (high=low type I error)

The Iris example



Iris setosa

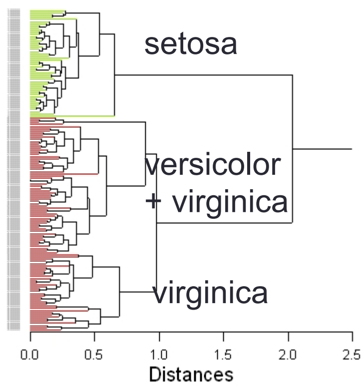


Iris versicolor

Iris virginica

- Classical example for linear discriminant analysis
- 4 measurements (petal lenght and width, sepal length and width), with biological variation
- Devise a classification function to classify new specimens in existing species

Is cluster analysis appropriate?



CLUSTERAVEEU(rows) by SPECIES_IRIS(columns)

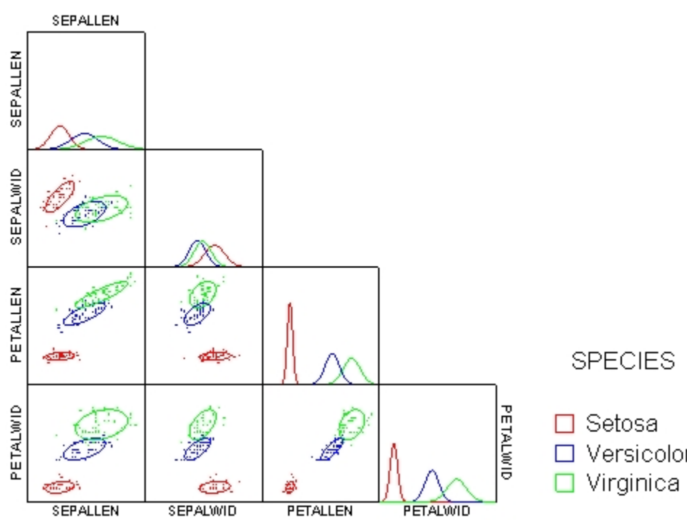
	Setosa	Versicolor	Virginica	Total
1	50	0	0	50
2	0	50	14	64
3	0	0	36	36
Total	50	50	50	150

Percents of Total Count

CLUSTERAVEEU(rows) by SPECIES_IRIS(columns)

	Setosa	Versicolor	Virginica	Total	N
1	33.33333	0.00000	0.00000	33.33333	50.00000
2	0.00000	33.33333	9.33333	42.66667	64.00000
3	0.00000	0.00000	24.00000	24.00000	36.00000
Total	33.33333	33.33333	33.33333	100.00000	
N	50.00000	50.00000	50.00000		150.00000

SPLOM on Iris data



The output from Iris data

Group Frequencies

Setosa	Versicolor	Virginica
50	50	50

Group Means

	Setosa	Versicolor	Virginica
SEPALLEN	5.00600	5.93600	6.58800
SEPALWID	3.42800	2.77000	2.97400
PETALLEN	1.46200	4.26000	5.55200
PETALWID	0.24600	1.32600	2.02600

Between-Groups F-matrix
df: 4 144

	Setosa	Versicolor	Virginica
Setosa	0.00000		
Versicolor	550.18889	0.00000	
Virginica	1,098.27375	105.31265	0.00000

Wilks's Lambda

Lambda : 0.02344
df : (4, 2, 147)
Approx. F-Ratio : 199.14534
df : (8, 288)
p-Value : 0.00000

Look at the F matrix to see which groups are best separated

04/02/2013 Multistat 3 cfu, Dec 2012 - Jan 2013

The output from Iris data

Classification Functions

	Setosa	Versicolor	Virginica
CONSTANT	-86.30847	-72.85261	-104.36832

	Setosa	Versicolor	Virginica
SEPALLEN	23.54417	15.69821	12.44585
SEPALWID	23.58787	7.07251	3.68528
PETALLEN	-16.43064	5.21145	12.76654
PETALWID	-17.39841	6.43423	21.07911

	Variable	F-to-Remove	Tolerance		Variable	F-to-Enter	Tolerance
2	SEPALLEN	4.72115	0.34799				
3	SEPALWID	21.93593	0.60886				
4	PETALLEN	35.59017	0.36513				
5	PETALWID	24.90433	0.64931				

Look at the F values to see which variables are worst or best at performing the separation

The output from Iris data

Classification Matrix (Cases in row categories classified into columns)

	Setosa	Versicolor	Virginica	%correct
Setosa	50	0	0	100
Versicolor	0	48	2	96
Virginica	0	1	49	98
Total	50	49	51	98

Jackknifed Classification Matrix

	Setosa	Versicolor	Virginica	%correct
Setosa	50	0	0	100
Versicolor	0	48	2	96
Virginica	0	1	49	98
Total	50	49	51	98

The output from the Iris data

Eigenvalues

32.19193	0.28539

Canonical Correlations

0.98482	0.47120

Cumulative Proportion of Total Dispersion

0.99121	1.00000

Examine values for canonical variables to see which is explaining best the difference among groups

Test Statistic

Statistic	Value	Approx. F-Ratio	df		p-Value
Wilks's Lambda	0.02344	199.14534	8	288	0.00000
Pillai's Trace	1.19190	53.46649	8	290	0.00000
Lawley-Hotelling Trace	32.47732	580.53210	8	286	0.00000

The output from the Iris data

Canonical Discriminant Functions

	1	2
Constant	2.10511	-6.66147

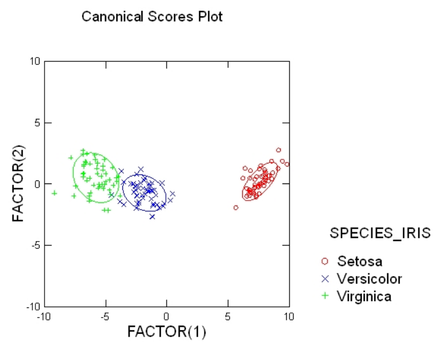
	1	2
SEPALLEN	0.82938	0.02410
SEPALWID	1.53447	2.16452
PETALLEN	-2.20121	-0.93192
PETALWID	-2.81046	2.83919

Canonical Discriminant Functions: Standardized by Within Variances

	1	2
SEPALLEN	0.42695	0.01241
SEPALWID	0.52124	0.73526
PETALLEN	-0.94726	-0.40104
PETALWID	-0.57516	0.58104

Canonical Scores of Group Means

	1	2
Setosa	7.60760	0.21513
Versicolor	-1.82505	-0.72790
Virginica	-5.78255	0.51277



Malahanobis distances and probabilities

	SPECIE...	PREDICTD	MISCLASS	DISTAN...	DISTAN...	DISTAN...	PROB(1)	PROB(2)	PROB(3)
124	3.000	3.000	0.000	144.798	8.038	3.579	0.000	0.097	0.903
125	3.000	3.000	0.000	179.432	19.841	1.173	0.000	0.000	1.000
126	3.000	3.000	0.000	171.098	17.364	5.525	0.000	0.003	0.997
127	3.000	3.000	0.000	137.521	6.823	3.902	0.000	0.188	0.812
128	3.000	3.000	0.000	137.031	7.043	3.315	0.000	0.134	0.866
129	3.000	3.000	0.000	199.868	23.382	0.887	0.000	0.000	1.000
130	3.000	3.000	0.000	155.549	13.399	9.085	0.000	0.104	0.896
131	3.000	3.000	0.000	196.345	23.442	5.754	0.000	0.000	1.000
132	3.000	3.000	0.000	176.557	26.779	11.656	0.000	0.001	0.999
133	3.000	3.000	0.000	208.571	27.319	1.894	0.000	0.000	1.000
134	3.000	2.000	1.000	133.067	5.253	7.236	0.000	0.729	0.271
135	3.000	3.000	0.000	175.566	21.132	15.833	0.000	0.066	0.934
136	3.000	3.000	0.000	215.460	34.805	8.707	0.000	0.000	1.000
137	3.000	3.000	0.000	206.848	34.312	6.443	0.000	0.000	1.000
138	3.000	3.000	0.000	161.274	13.325	3.160	0.000	0.006	0.994
139	3.000	3.000	0.000	134.191	6.961	4.094	0.000	0.193	0.807
140	3.000	3.000	0.000	166.607	16.533	2.344	0.000	0.001	0.999
141	3.000	3.000	0.000	207.918	31.749	4.451	0.000	0.000	1.000

Identification of new cases or crossvalidation by subsampling (1)

SPECIE...	SEPALLEN	SEPALWID	PETALLEN	PETALWID	CLUST...	CLUST...	WEIGHT1	SPECIES2
1.000	5.100	3.500	1.400	0.200	1.000	1.000	0.171	1.000
1.000	4.900	3.000	1.400	0.200	1.000	1.000	0.158	1.000
1.000	4.700	3.200	1.300	0.200	1.000	1.000	0.601	1.000
1.000	4.600	3.100	1.500	0.200	1.000	1.000	0.909	.
1.000	5.000	3.600	1.400	0.200	1.000	1.000	0.710	1.000
1.000	5.400	3.900	1.700	0.400	1.000	1.000	0.301	1.000
1.000	4.600	3.400	1.400	0.300	1.000	1.000	0.920	.
1.000	5.000	3.400	1.500	0.200	1.000	1.000	1.000	.
1.000	4.400	2.900	1.400	0.200	1.000	1.000	0.135	1.000
1.000	4.900	3.100	1.500	0.100	1.000	1.000	0.234	1.000
1.000	5.400	3.700	1.500	0.200	1.000	1.000	0.402	1.000
1.000	4.800	3.400	1.600	0.200	1.000	1.000	0.662	1.000
1.000	4.800	3.000	1.400	0.100	1.000	1.000	0.880	.
1.000	4.300	3.000	1.100	0.100	1.000	1.000	0.822	.

- Use a random number generator to select 65-80% of cases
- Use the weight function to calculate the discriminant function only for selected cases or use the new variable for calculating LDA and save distances
- Calculate the classification matrix on the remaining cases

Iris data: non validated vs. cross-validated

Non validated

Classification Matrix (Cases in row categories classified into columns)

	Setosa	Versicolor	Virginica	%correct
Setosa	50	0	0	100
Versicolor	0	48	2	96
Virginica	0	1	49	98
Total	50	49	51	98

Jackknifed Classification Matrix

	Setosa	Versicolor	Virginica	%correct
Setosa	50	0	0	100
Versicolor	0	48	2	96
Virginica	0	1	49	98
Total	50	49	51	98

Crossvalidated

Classification Matrix (Cases in row categories classified into columns)

	Setosa	Versicolor	Virginica	%correct
Setosa	35	0	0	100
Versicolor	0	35	2	95
Virginica	0	0	36	100
Total	35	35	38	98

Classification of Cases with zero weight or frequency

	Setosa	Versicolor	Virginica	%correct
Setosa	15	0	0	100
Versicolor	0	12	1	92
Virginica	0	1	13	93
%correct	15	13	14	95

Prof. Eugenio Parente
Dip. Biologia, Università della Basilicata

Identification of new cases

	SPECIE...	WEIGHT2	PREDICTD	MISCLASS	DISTAN...	DISTAN...	DISTAN...	PROB(1)	PROB(2)	PROB(3)
118	3.000	1.000	3.000	0.000	263.476	46.107	11.495	0.000	0.000	1.000
119	3.000	0.000	3.000	0.000	371.771	81.356	24.130	0.000	0.000	1.000
120	3.000	1.000	3.000	0.000	200.986	14.290	9.368	0.000	0.081	0.919
121	3.000	1.000	3.000	0.000	247.185	35.071	2.240	0.000	0.000	1.000
122	3.000	1.000	3.000	0.000	209.530	21.240	5.254	0.000	0.000	1.000
123	3.000	0.000	3.000	0.000	307.312	54.546	14.800	0.000	0.000	1.000
124	3.000	1.000	3.000	0.000	184.639	11.681	3.161	0.000	0.014	0.986
125	3.000	0.000	3.000	0.000	222.488	25.113	1.327	0.000	0.000	1.000
126	3.000	1.000	3.000	0.000	215.385	22.527	5.594	0.000	0.000	1.000
127	3.000	1.000	3.000	0.000	174.060	9.633	3.763	0.000	0.052	0.948
128	3.000	1.000	3.000	0.000	169.871	8.951	4.130	0.000	0.084	0.916
129	3.000	1.000	3.000	0.000	250.315	30.651	1.423	0.000	0.000	1.000
130	3.000	1.000	3.000	0.000	198.178	17.808	9.131	0.000	0.013	0.987
131	3.000	1.000	3.000	0.000	254.663	33.998	6.618	0.000	0.000	1.000
132	3.000	1.000	3.000	0.000	221.587	33.372	11.322	0.000	0.000	1.000
133	3.000	1.000	3.000	0.000	261.448	35.671	2.459	0.000	0.000	1.000
134	3.000	0.000	2.000	1.000	166.363	6.719	8.549	0.000	0.720	0.280
135	3.000	0.000	3.000	0.000	212.711	22.941	18.628	0.000	0.106	0.894

- To identify new cases use missing values on the grouping variable or use 0 weights, run the model and save distance and data
- Look at distance and probabilities in the saved file

The ourworld file

- Contains economic and social indicator for world countries, classified a priori in 3 groups: Europe, Islamic, Newworld
- The file can be used to illustrate problems for MLR, multicollinearity, use of PCA, need for transformation
- In terms of discriminant analysis it may be used to look at which variables are more important in correctly discriminating the three group of countries (by using stepwise LDA) and if a quadratic model is appropriate
- For this example, the Systat output will be presented

Quadratic discriminant analysis

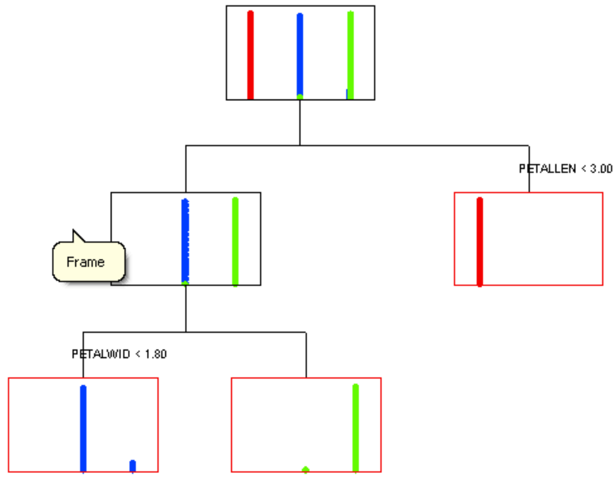
$$\mu_1'\left(\Sigma_2^{-1}-\Sigma_1^{-1}\right)\mathbf{x}-2\mathbf{x}'\left(\Sigma_2^{-1}\mu_2-\Sigma_1^{-1}\mu_1\right)+\left(\mu_2'\Sigma_2^{-1}\mu_2-\mu_1'\Sigma_1^{-1}\mu_1\right)\geq \ln\frac{|\Sigma_1|}{|\Sigma_2|}+2\cdot\ln\left(\frac{\pi_1}{\pi_2}\right)$$

- A quadratic discriminant analysis is necessary
  - If the covariance matrices of the groups are different
  - If the mean vectors are equal but the covariance matrices are different
- The model is more complex than LDA
  - Both linear and quadratic terms are present in the discriminant function
  - There are more coefficients in the discriminant function (linear, interaction and quadratic terms) and therefore more cases are necessary for the estimation (if possible n must be >> than the number of coefficients)
  - Overfitting is more likely than with LDA

Classification and regression trees

- Alternative to Discriminant Analysis for supervised pattern recognition
- Operate iteratively on the data file to find the optimal partition values of continuous (regression trees) or discontinuous (classification trees) depdendent variables to maximize the discrimination of cases into groups whose membership is known a priori
- The output is composed of
  - A classification tree with the data partition
  - Statistics on goodness of classification
- CT have analogies with ANOVA, DA and cluster analysis

Classification tree for the Iris data



04/02/2013 Multistat 3 cfu, Dec 2012 - Jan 2013

More on supervised pattern recognition

- PLS-DA
  - similar to DA (it is a PLS model using a categorical response variable)
  - more flexible
  - useful with datasets with more variables than observations (PLC components, rather than the original variables are used)
  - the model is more parsimonious than the corresponding DA model
- Supervised Artifical Neural Networks
  - See lecture 6
  - Very flexible, can handle noisy and incomplete data
  - Can use both categorical and continuous predictors
  - Needs careful planning (size and composition of training and test set, planning of the network architecture, pretreatment of variables, etc.)