# Financial classification models

Advanced Financial Accounting II

School of Business and Economics at

Åbo Akademi University

# Contents

# The classification problem

- In a traditional classification problem the main purpose is to assign one of *k labels* (or classes) to each of *n objects*, in a way that is consistent with some observed data, i.e. to **determine the class of an observation based on** a set of variables known as **predictors** or **input variables**
- Typical classification problems in finance are for example
  - Financial failure/bankruptcy prediction
  - Credit risk rating

# Classification methods

- There are several statistical and mathematical methods for solving the classification problem, e.g.
  - Discriminant analysis
  - Logistic regression
  - The recursive partitioning algorithm (RPA)
  - Mathematical programming
    - Linear programming models
    - Quadratic programming models
  - Neural network classifiers
  - New methods are continuously being developed
- The lecture notes describe these methods, with focus laid on discriminant analysis.

# Discriminant analysis

- Discriminant analysis is the most common technique for classifying a set of observations into predefined classes
- The model is based on a set of observations for which the classes are known
- This set of observations is sometimes referred to as the **training set** or **estimation sample**

# Discriminant Analysis – Historical background

- Discriminant analysis is concerned with the problem to assign or allocate an object (e.g. a firm) to its correct population
- The statistical method originated with Fisher (1936)
- The theoretical framework was derived by Anderson (1951)
- The term discriminant analysis emerged from the research by Kendall & Stuart (1968) and Lachenbruch & Mickey (1968)
- Discriminant analysis was originally applied in accounting by Altman (1968) using U.S. data and by Aatto Prihti (1975) on Finnish data

# Discriminant analysis...

- Based on the training set, the technique constructs a set of linear functions of the predictors, known as discriminant functions, such that

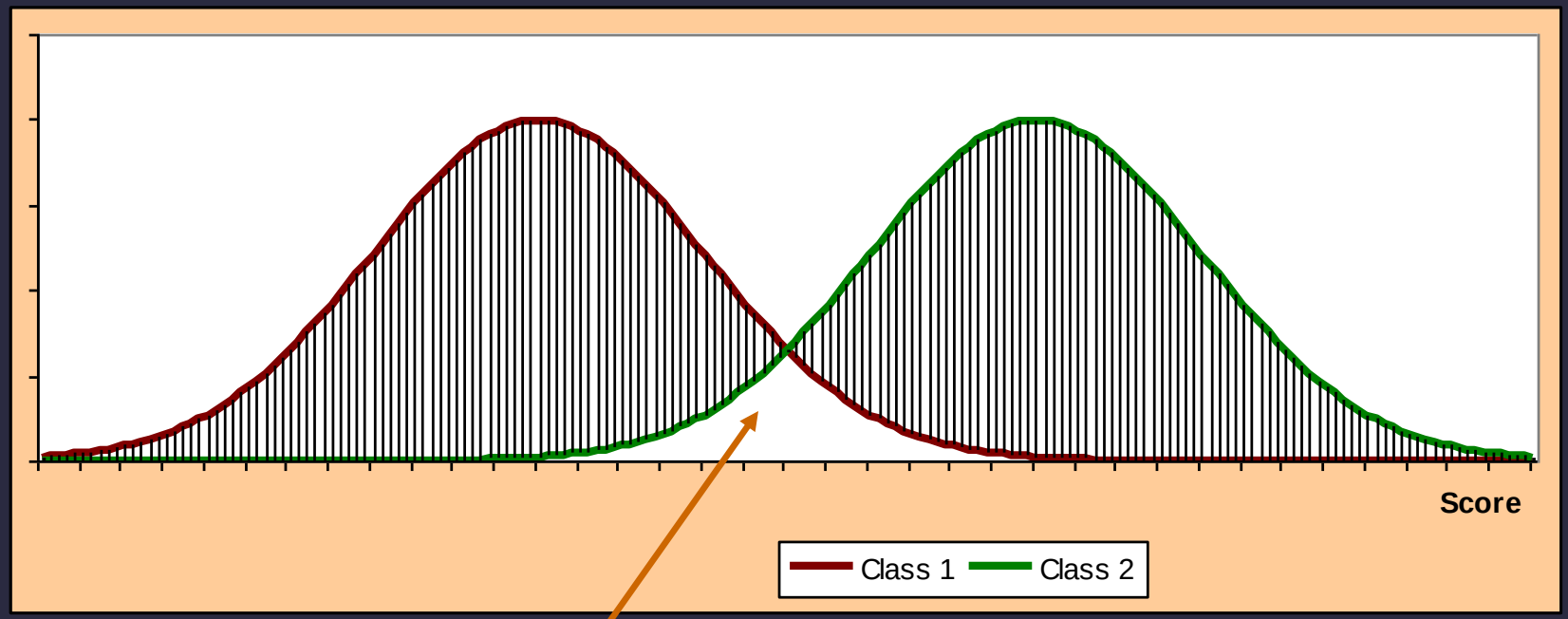
$$L = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + c,$$

  where the $\beta$'s are **discriminant coefficients**, the *x*'s are the input variables or predictors and *c* is a constant.
- Two types of discriminant functions are discussed later
  - Canonical discriminant functions (*k*-1)
  - Fisher's discriminant functions (*k*)

# Discriminant functions

- The discriminant functions are optimized to provide a classification rule that minimizes the probability of misclassification
  - See figure on the next page
- In order to achieve optimal performance, some statistical assumptions about the data must be met
  - Each group must be a sample from a multivariate normal population
  - The population covariance matrices must all be equal
- In practice the discriminant has been shown to perform fairly well even though the assumptions on data are violated

# Distributions of the discriminant scores for two classes




**A discriminant function is optimized to minimize the common area for the distributions**

# Canonical discriminant functions

- A canonical discriminant function is a linear combination of the discriminating variables which are formed to satisfy certain conditions
  - The coefficients for the **first function** are derived so that the **group means** on the function are **as different as possible**
  - The coefficients for the **second function** are derived to maximize the difference between group means under the **added condition** that values on the second function are **not correlated with** the values on the **first function**
  - A third function is defined in a similar way having coefficients which maximize the group differences while being uncorrelated with the previous function and so on
- The maximum number on unique functions is Min(Groups – 1, No of discriminating variables)

# Fisher's Discriminant functions

- The discriminant functions are used to predict the class of a new observation with unknown class
- For a *k* class problem, *k* discriminant functions are constructed
- Given a new observation, all the *k* discriminant functions are evaluated and the **observation** is **assigned to class** *i* if the *i*:th discriminant function has **the highest value**

# Interpretation of the Fisher's discriminant function coefficients

- The discriminant functions are used to compute the discriminant score for a case in which the original discriminating variables are in standard form
- The discriminant score is computed by multiplying each discriminating variable by its corresponding coefficient and adding together these products
- There will be a separate score for each case on each function
- The coefficients have been derived in such a way that the discriminant scores produced are in standard form
- Any single score represents the number of standard deviations the case is away from the mean for all cases on the given discriminant function

# Interpretation of the Fisher's discriminant function coefficients…

- The standardized discriminant function coefficients are of great analytical importance
- When the sign is ignored, each coefficient represents the relative contribution of its associated variable for that function
- The sign denotes whether the variable is making a positive or negative contribution
- The interpretation is analogous to the interpretation of beta weights in multiple regression
- As in factor analysis, the coefficients can be used to "name" the functions by identifying the dominant characteristics they measure

# Variable selection: Analyzing group differences

- Although the variables are interrelated and the multivariate statistical techniques such as discriminant analysis incorporate these dependencies, it is often helpful to begin analyzing the differences between groups by examining univariate statistics
- The first step is to compare the group means of the predictor variables
  - A significant inequality in group means indicates the predictor variable's ability to separate between the groups
  - The significance test for the equality of the group means is an *F*-test with 1 and *n*-*g* degrees of freedom
  - If the observed significance level is less than 0.05, the hypothesis of equal group means is rejected

# Analyzing group differences: Wilks' Lambda

- Another statistic used to analyze the univariate equality of group means is Wilks' Lambda, sometimes called the U-statistic
- Lambda is the ratio of the within-groups sum of squares to the total sum of squares
- Lambda has values between 0 and 1
- A lambda of 1 occurs when all observed group means are equal
- Values close to 0 occur when within-groups variability is small compared to total variability
- Large values of lambda indicate that group means do not appear to be different while small values indicate that group means do appear to be different

# Multivariate Wilks' Lambda statistic

- In the case of several variables $\{X_1, X_2,...,X_p\}$, the total variability is expressed by the total cross product matrix ***T***
- The sum of cross-product matrix ***T*** is decomposed into the within-group sum of cross- product matrix ***W*** and the between-group sum of cross-product matrix ***B*** such that

$$\boldsymbol{T} = \boldsymbol{W} + \boldsymbol{B} \Leftrightarrow \boldsymbol{W} = \boldsymbol{T} - \boldsymbol{B}$$

# Multivariate Wilks' Lambda statistic...

- For the set of the X variables, the multivariate global Wilks' Lambda is defined as

$$\Lambda_p = |\boldsymbol{W}| / |\boldsymbol{W} + \boldsymbol{B}| = |\boldsymbol{W}| / |\boldsymbol{T}| \sim \Lambda(p,m,n)$$

where

$|\boldsymbol{W}|$ = the determinant of the within-group SSCP matrix

$|\boldsymbol{B}|$ = the determinant of the between-groups SSCP matrix

$|\boldsymbol{T}|$ = the determinant of the total sum of cross product matrix

$\Lambda(p,m,n)$ = Wilks' Lambda distribution

For large *m,* Bartlett's (1954) approximation allows Wilks' lambda to be approximated by a Chi-square distribution

$$\left(\frac{p-n+1}{2}-m\right)\log\Lambda(p,m,n) \sim \chi^2_{np}$$

# Variable selection: Correlations between predictor variables

- Since interdependencies among the variables affect most multivariate analyses, it is worth examining the correlation matrix of the predictor variables
- Including highly correlated variables in the analysis should be avoided as correlations between variables affect the magnitude and the signs of the coefficients
- If correlated variables are included in the analysis, care should be exercised when interpreting the individual coefficients

# Case: Bankruptcy prediction in the Spanish banking sector

- Reference: Olmeda, Ignacio and Fernández, Eugenio: "Hybrid classifiers for financial multicriteria decision making: The case of bankruptcy prediction", *Computational Economics* **10**, 1997, 317-335.
- Sample: 66 Spanish banks
  - 37 survivors
  - 29 failed
- Sample was divided in two sub-samples
  - Estimation sample, 34 banks, for estimating the model parameters
  - Holdout sample, 32 banks, for validating the results

# Case: Bankruptcy prediction in the Spanish banking sector

Input variables

- Current assets/Total assets
- (Current assets-Cash)/Total assets
- Current assets/Loans
- Reserves/Loans
- Net income/Total assets
- Net income/Total equity capital
- Net income/Loans
- Cost of sales/Sales
- Cash flow/Loans

# Empirical results

- Analyzing the total set of 66 observations
  - Group statistics – comparing the group means
  - Testing for the equality of group means
  - Correlation matrix
- Classification with different methods
  - Estimating classification models using the estimation sample of 34 observations
  - Checking the validity of the models by classifying the holdout sample of 32 observations

# Group statistics

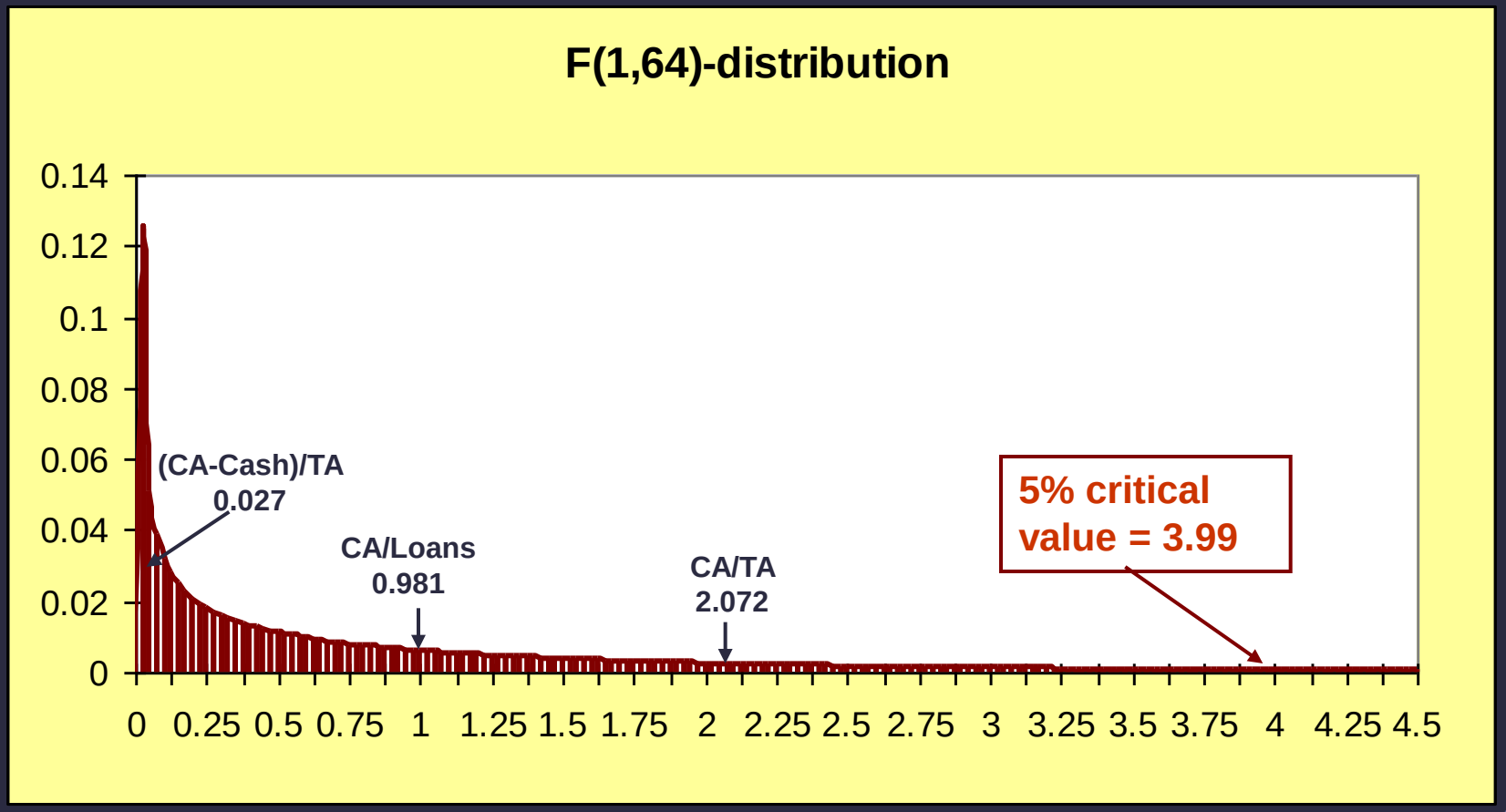
| | Class 0 N=37 | | Class 1 N=29 | | Total N=66 | |
|---|---|---|---|---|---|---|
| | Mean | St.dev | Mean | St.dev | Mean | St.dev |
| CA/TA | ,410 | ,114 | ,370 | ,108 | ,393 | ,112 |
| (CA-Cash)/TA | ,268 | ,089 | ,264 | ,092 | ,266 | ,089 |
| CA/Loans | ,423 | ,144 | ,390 | ,117 | ,409 | ,133 |
| Reserves/Loans | ,038 | ,054 | ,016 | ,012 | ,028 | ,043 |
| NI/TA | ,008 | ,005 | -,003 | ,019 | ,003 | ,014 |
| NI/TEC | ,167 | ,082 | -,032 | ,419 | ,079 | ,299 |
| NI/Loans | ,008 | ,005 | -,003 | ,020 | ,003 | ,015 |
| CofS/Sales | ,828 | ,062 | ,957 | ,188 | ,885 | ,147 |
| CF/Loans | ,018 | ,029 | ,004 | ,012 | ,012 | ,024 |

# Tests of equality of group means

| Insignificant difference | Wilks' Lambda | F | df1 | df2 | Sig. |
|---|---|---|---|---|---|
| CA/TA | ,969 | 2,072 | 1 | 64 | ,155 |
| (CA-Cash)/TA | 1,000 | ,027 | 1 | 64 | ,871 |
| CA/Loans | ,985 | ,981 | 1 | 64 | ,326 |
| Reserves/Loans | ,932 | 4,667 | 1 | 64 | ,034 |
| NI/TA | ,864 | 10,041 | 1 | 64 | ,002 |
| NI/TEC | ,889 | 8,011 | 1 | 64 | ,006 |
| NI/Loans | ,863 | 10,149 | 1 | 64 | ,002 |
| CofS/Sales | ,805 | 15,463 | 1 | 64 | ,000 |
| CF/Loans | ,918 | 5,713 | 1 | 64 | ,020 |

**No significant difference in group means**

# *F*(1,64)-distribution



# Tests of equality of group means

- The tests of equality of the group means indicate that for the three first predictor variables there does not seem to be any significant difference in group means
  - *F*-values < 3.99, the 5 % critical value for *F*(1,64)
  - Significance > 0.05
- The result is confirmed by the Wilks' lambda values close to 1
- As the results indicate low univariate discriminant power for these variables, some or all of them may be excluded from analysis in order to get a parsimonious model

# Pooled Within-Groups Correlation Matrix

| | CA/TA | (C-C)/TA | CA/Loa | Res/Loa | NI/TA | NI/TEC | NI/Loa | CS/Sal | CF/Loa |
|---|---|---|---|---|---|---|---|---|---|
| CA/TA | 1,000 | | | | | | | | |
| (C-C)/TA | ,760 | 1,000 | | | | | | | |
| CA/Loa | ,917 | ,641 | 1,000 | | | | | | |
| Res/Loa | ,013 | -,230 | ,099 | 1,000 | | | | | |
| NI/TA | ,038 | -,007 | ,058 | ,174 | 1,000 | | | | |
| NI/TEC | -,023 | -,016 | -,035 | ,033 | ,956 | 1,000 | | | |
| NI/Loa | ,048 | -,015 | ,072 | ,194 | ,999 | ,947 | 1,000 | | |
| CS/Sal | -,087 | -,147 | -,104 | -,288 | -,565 | -,419 | -,570 | 1,000 | |
| CF/Loa | -,007 | -,013 | ,014 | ,116 | ,223 | ,181 | ,225 | -,372 | 1,000 |

# Correlations between predictor variables

- The variables *Current assets/Total assets* and *Current assets/Loans* are highly correlated (Corr = 0,917)
- The variables explain the same variation in the data
- Including both the variables in the discriminant function does not improve the explanation power but may lead to multicollinearity problem in estimation
- Only one of the variables should be selected into the set of explanatory variables
- For the same reason, only one of the variables *Net income/Total assets*, *Net income/Total equity capital* and *Net income/Loans* should be selected

# Summary of Canonical Discriminant Functions

Eigenvalues

| Function | Eigenvalue | % of Variance | Cumulative % | Canonical Correlation |
|---|---|---|---|---|
| 1 | ,417[a] | 100,0 | 100,0 | ,542 |

a. First 1 canonical discriminant functions were used in the analysis.

Wilk's Lambda

| Test of Function(s) | Wilk's Lambda | Chi-square | df | Sig. |
|---|---|---|---|---|
| 1 | ,706 | 20,899 | 8 | ,007 |

# Canonical Discriminant Function Coefficients

| | Function 1 | |
|---|---|---|
| | Standardized | Unstandardized |
| CA/TA | -1,318 | -11,825 |
| (CA-Cash)/TA | ,625 | 6,940 |
| CA/Loans | ,612 | 4,601 |
| Reserves/Loans | -,228 | -5,510 |
| NI/TA | 1,134 | 85,998 |
| NI/TEC | -1,264 | -4,456 |
| CofS/Sales | ,780 | 5,884 |
| CF/Loans | -,180 | -7,864 |
| Constant | | -3,957 |

Relative contribution of each variable to discriminant function

# Functions at group centroids

| Class | Function 1 |
|---|---|
| 0 | -,563 |
| 1 | ,718 |

Unstandardized canonical discriminant functions evaluated at group means

# Example of classifying an observation by the canonical discrimiant function

| | Obs. 1 | **Coeff.** | **Score** |
|---|---|---|---|
| Constant | | -3,957 | -3,957 |
| CA/TA | 0.4611 | -11,825 | -5,453 |
| CA_Cash/TA | 0.3837 | 6,940 | 2,663 |
| CA/Loans | 0.4894 | 4,601 | 2,252 |
| Res/Loans | 0.0077 | -5,510 | -0,042 |
| NI/TA | 0.0057 | 85,998 | 0,490 |
| NI/TEC | 0.0996 | -4,456 | -0,444 |
| CofS/Sales | 0.8799 | 5,884 | 5,177 |
| CF/Loans | 0.0092 | -7,864 | -0,072 |
| **Total Score** | | | **0,614** |

Distance to group centroid for Group 1 (Failed), 0,718, smaller than distance to group centroid for Group 0 (Survived), -0,563 ⇨ Classification to the closest group **Failed**

# Fisher's discriminant function coefficients

| | Survived | Failed |
|---|---|---|
| Constant | -66,485 | -71,653 |
| CA/TA | 15,352 | ,207 |
| CA_Cash/TA | 82,320 | 912,208 |
| CA/Loans | -29,866 | -23,973 |
| Res/Loans | 81,189 | 74,071 |
| NI/TA | 2006,853 | 2116.987 |
| NI/TEC | -65,300 | -71,007 |
| CofS/Sales | 126,771 | 134,307 |
| CF/Loans | 185,726 | 175,654 |

# Example on classifying an observation by Fisher's discriminant functions

| | Obs. 1 | Survived | Score | Failed | Score |
|---|---|---|---|---|---|
| Constant | | -66,485 | -66,485 | -71,653 | -71,653 |
| CA/TA | 0.4611 | 15,352 | 7,079 | ,207 | 0,095 |
| CA_Cash/TA | 0.3837 | 82,320 | 31,586 | 912,208 | 34,997 |
| CA/Loans | 0.4894 | -29,866 | -14,616 | -23,973 | -11,732 |
| Res/Loans | 0.0077 | 81,189 | 0,625 | 74,071 | 0,570 |
| NI/TA | 0.0057 | 2006,853 | 11,439 | 2116.987 | 12,067 |
| NI/TEC | 0.0996 | -65,300 | -6,054 | -71,007 | -7,072 |
| CofS/Sales | 0.8799 | 126,771 | 111,546 | 134,307 | 118,177 |
| CF/Loans | 0.0092 | 185,726 | 1,709 | 175,654 | 1,616 |
| Total Score | | | 76,378 | | 77,064 |

Larger score $\Rightarrow$
Classification: Failed

# Confusion matrix – Classification results

| | | Predicted class | |
|---|---|---|---|
| | | Survived | Failed |
| True class | Survived | 28 | 9 |
| | | 75,7 % | 24,3 % |
| | Failed | 4 | 25 |
| | | 13,8 % | 86,2 % |

# Summary of classifications with different classification methods(Estimation sample)

| Method | Correct class | Errors | | Total number | Percents | | |
|---|---|---|---|---|---|---|---|
| | | SW | NE | | Correct | SW | NE |
| RPA | 30 | 1 | 3 | 34 | 88.24 % | 2.94 % | 8.82 % |
| MDA | 30 | 0 | 4 | 34 | 88.24 % | 0.00 % | 11.76 % |
| MDA-Q | 31 | 0 | 3 | 34 | 91.18 % | 0.00 % | 8.82 % |
| MDA-W | 31 | 0 | 3 | 34 | 91.18 % | 0.00 % | 8.82 % |
| LogR | 33 | 0 | 1 | 34 | 97.06 % | 0.00 % | 2.94 % |
| LP | 28 | 1 | 5 | 34 | 82.35 % | 2.94 % | 14.71 % |
| LP-Q | 34 | 0 | 0 | 34 | 100.00 % | 0.00 % | 0.00 % |
| LPG | 33 | 0 | 1 | 34 | 97.06 % | 0.00 % | 2.94 % |
| LPGQ | 34 | 0 | 0 | 34 | 100.00 % | 0.00 % | 0.00 % |
| Kohonen | 24 | 3 | 7 | 34 | 70.59 % | 8.82 % | 20.59 % |

# Summary of classifications (Holdout sample)

| Method | Correct class | Errors SW | Errors NE | Total number | Percents Correct | Percents SW | Percents NE |
|---|---|---|---|---|---|---|---|
| **RPA** | 27 | 2 | 3 | 32 | 84.38 % | 6.25 % | 9.38 % |
| **MDA** | 25 | 4 | 3 | 32 | 78.13 % | 12.50 % | 9.38 % |
| **MDA-Q** | 20 | 7 | 5 | 32 | 62.50 % | 21.88 % | 15.63 % |
| **MDA-W** | 25 | 5 | 2 | 32 | 78.13 % | 15.63 % | 6.25 % |
| **LogR** | 28 | 3 | 1 | 32 | 87.50 % | 9.38 % | 3.13 % |
| **LP** | 24 | 5 | 3 | 32 | 75.00 % | 15.63 % | 9.38 % |
| **LP-Q** | 21 | 7 | 4 | 32 | 65.63 % | 21.88 % | 12.50 % |
| **LPG** | 25 | 4 | 3 | 32 | 78.13 % | 12.50 % | 9.38 % |
| **LPGQ** | 21 | 7 | 4 | 32 | 65.63 % | 21.88 % | 12.50 % |
| **Kohonen** | 16 | 4 | 12 | 32 | 50.00 % | 12.50 % | 37.50 % |

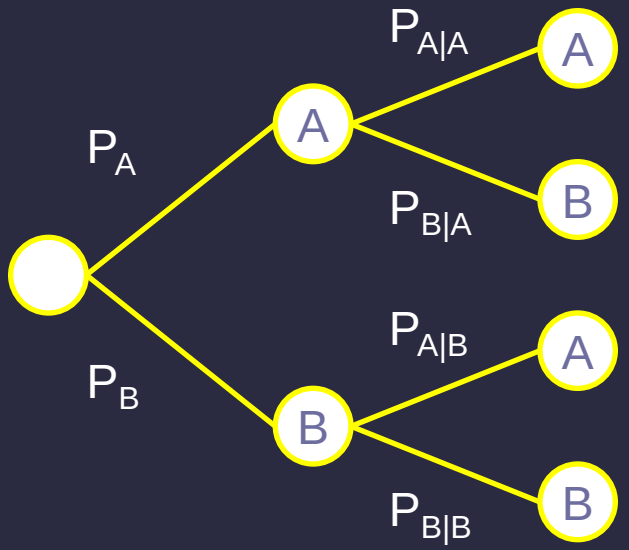
# Classification results – Error types

- The classification results for the different methods differ in
  - Total classification accuracy
    - Descriptive (Estimation sample)
    - Predictive (Holdout sample)
  - Error types
    - Classifying a survivor as failed
    - Classifying a failed as survivor
- Many methods may be calibrated to take into account the relative severity of the two types of errors

# The multiplication rule

The multiplication rule for probabilities is

(1) $P_{AB} = P_{A|B} \cdot P_B$ and $P_{BA} = P_{B|A} \cdot P_A$




← Correct classification

← Misclassification: Object A is allocated to group B

← Misclassification: Object B is allocated to group A

← Correct classification

# Probability of erroneous assignment

- Assume that we have a sample $X \in \Re^{T \times p}$ of random measurements and $k$ regions $R_i$, $i = 1,\ldots,k$. The probability distribution for region $i$ is $f_i(x)$.
- By the multiplication rule,

(2) $p_{ij} = p_{i|j}\, p_j, \qquad (i, j = 1,\ldots,k)$

is the probability of assigning an object belonging to population $j$ erroneously to group $i$.

# Probability of erroneous assignment…

- All we have to do in order to evaluate the probability of misclassifying an object belonging to population *j* is to sum (2) over all *k* non-overlapping regions

(3) $$e_j = \sum_{\substack{i=1 \\ i \neq j}}^{k} p_{ij} = p_j \sum_{\substack{i=1 \\ i \neq j}}^{k} p_{i|j} \quad (j = 1,\ldots,k)$$

- $p_{i|j}$ = the conditional probability of an object from *j* being assigned to group *i*. That is equivalent to the probability mass of $f_j$ over region $R_i$:

(4) $$p_{i|j} = \int_{R_i} f_j(x)dx$$

# Probability of correct classification

- Using (4), we may write (3) as:

(5) $$e_j = p_j \sum_{\substack{i=1 \\ i \neq j}}^{k} \int_{R_i} f_i(x)dx$$

- The probability of correct classification of an object is

(6) $$F_j = 1 - e_j = 1 - p_j \sum_{\substack{i=1 \\ i \neq j}}^{k} \int_{R_i} f_j(x)dx = p_j \int_{R_j} f_j(x)dx$$

# The maximization problem for optimal allocation

- We obtain the last equality because

$$\sum_{i=1}^{k} p_{i|j} = \sum_{\substack{i=1 \\ i \neq j}}^{k} p_{i|j} + p_{j|j} = 1$$

  and the probability distribution for region $R_j$ is obtained by substituting $p_{i|j}$ in (4)

- The allocation problem is to maximize $F_e = \sum_{i=1}^{k} F_j$ in (6) by choosing an optimal partition $(R_1, \ldots, R_k)$ of the sample space:

(7) Maximize $$F = \sum_{j=1}^{k} p_j \int_{R_j} f_j(x)dx$$

# Two populations and known distributions

- When the distributions are unknown, like in practice, they must be assumed/estimated
- The same formulae are still used
- When $\underline{k} = 2$, the maximization problem (7) becomes

(8) Maximize $$F = p_1 \int_{R_1} f_1(x)dx + p_2 \int_{R_2} f_2(x)dx$$

- Hogg and Craig (1978) used a similar proof as for the Newman-Pearson lemma for statistical tests of simple hypotheses to extract the optimal partitioning (maximum of (8))

# The optimal partitioning - Proof

(9) $$R_1 = \left\{ x \left| \frac{f_1(x)}{f_2(x)} \geq \frac{p_2}{p_1} \right. \right\}, R_2 = \left\{ x \left| \frac{f_1(x)}{f_2(x)} < \frac{p_2}{p_1} \right. \right\}$$

- We present the key steps of the proof below (cf. Karson, 1982)
- Let $_1$, $_2$ be arbitrary of the sample space $X$ such that $_1 \cup \ _2 = X$ and $_1 \cap \ _2 = \varnothing$.
- Let $R_1 = \{ x \mid \phi_1(x) \geq \phi_2(x) \}$ and

(10) $R_2 = \{ x \mid \phi_1(x) < \phi_2(x) \}$, where

$\phi_i(x)$, $i = 1,2$ are continuous functions in $X \in \Re^p$

- Then $R_1 \cup R_2 = X$ and $R_1 \cap R_2 = \varnothing$

# The optimal partitioning – Proof…

- Let

$$\tilde{I} = \int_{\tilde{R}_1} \phi_1(x)dx + \int_{\tilde{R}_2} \phi_2(x)dx$$

$$I = \int_{R_1} \phi_1(x)dx + \int_{R_2} \phi_2(x)dx$$

- Consider the difference

(11)

$$\Delta = I - \tilde{I}$$
$$= \int_{R_1} \phi_1(x)dx + \int_{R_2} \phi_2(x)dx - \int_{\tilde{R}_1} \phi_1(x)dx - \int_{\tilde{R}_2} \phi_2(x)dx$$

# The optimal partitioning – Proof…

- We know that $R_1 = (R_1 \cap \quad _2) \cup (R_1 \cap \quad _1)$

  $R_2 = (R_2 \cap \quad _1) \cup (R_2 \cap \quad _2)$

  $_1 = (\quad _1 \cap R_1) \cup (\quad _1 \cap R_2)$

  $_2 = (\quad _2 \cap R_1) \cup (\quad _2 \cap R_2)$

- We can therefore write (10) as

(12)
$$\Delta = \underbrace{\int_{R_1 \cap \tilde{R}_2} \phi_1(x)dx}_{①} + \underbrace{\int_{R_1 \cap \tilde{R}_1} \phi_1(x)dx}_{②} + \underbrace{\int_{R_2 \cap \tilde{R}_1} \phi_2(x)dx}_{③} + \underbrace{\int_{R_2 \cap \tilde{R}_2} \phi_2(x)dx}_{④} - \underbrace{\int_{\tilde{R}_1 \cap R_1} \phi_1(x)dx}_{⑤} - \underbrace{\int_{\tilde{R}_1 \cap R_2} \phi_1(x)dx}_{⑥} - \underbrace{\int_{\tilde{R}_2 \cap R_1} \phi_2(x)dx}_{⑦} - \underbrace{\int_{\tilde{R}_2 \cap R_2} \phi_2(x)dx}_{⑧}$$

# The optimal partitioning – Proof…

- We note that ② = ⑤ and ④ = ⑧, hence they are eliminated and (11) reduces to

(13)
$$\Delta = \underbrace{\int_{R_1 \cap \tilde{R}_2} \phi_1(x)dx}_{①} - \underbrace{\int_{\tilde{R}_1 \cap R_2} \phi_1(x)dx}_{⑥} + \underbrace{\int_{R_2 \cap \tilde{R}_1} \phi_2(x)dx}_{③} - \underbrace{\int_{\tilde{R}_2 \cap R_1} \phi_2(x)dx}_{⑦}$$

- By assembling the terms involving identical regions, i.e., ① & ⑦ and ③ & ⑥ respectively, we obtain

(14)
$$\Delta = \int_{R_1 \cap \tilde{R}_2} (\phi_1(x) - \phi_2(x))dx + \int_{R_2 \cap \tilde{R}_1} (\phi_2(x) - \phi_1(x))dx$$

# Some comments on hypothesis testing

Assume that we – as a bank institution – want to distinguish between non-distressed ($H_0$) vs. distressed ($H_1$) firms using a suitable financial ratio FR (for example based on the discriminant score), in order to reduce the financial risk in loan decisions

To do this, we need to compare the FR of a firm with a critical value $FR_c$

⇨ If $FR > FR_c$, then the firm is assumed to be distressed, otherwise not.

# Some comments on hypothesis testing…

There is a tension between type I and type II errors

The first type is smaller, the higher is the significance (i.e. the smaller is $\alpha$): The probability of rejecting $H_0$ falsely is smaller, the smaller is $\alpha$

Type I error is the probability of rejecting $H_0$ even if it is true

With $\alpha$ = 10% this probability is twice that of $\alpha$ = 5% and ten times that of $\alpha$ = 1%

We throw away a gold nugget among the rubbish in 10% of all cases by rejecting $H_0$ for firms that actually are non-distressed.

# Some comments on hypothesis testing…

If we get an extremely high FR for a firm, however, everybody will realize that the probability of that firm being non-distressed is practically negligible:

The probability of such an outcome being generated by chance is very low.

In such a case it is safe to conclude that the firm is financially distressed and, for example, to reject financing a project that the firm is contemplating.

On the other hand, the more we shift the critical significance level ($FR_c$) to the right, the less frequently we will reject $H_0$

If $FR_c$ is extremely high, we will accept $H_0$ almost always: Everybody will receive a loan from our bank.

# Some comments on hypothesis testing…

- But the more we shift the critical level $FR_c$ to the right, the more often we will accept $H_0$ even if it is false: there will be firms in our clientele that should not be there
- These firms are distressed, even though we have failed to detect this because of a high $FR_{c.}$ This latter error is denoted Type II
- Because of the high $FR_c$ the test has a low power: the probability of failing to reject a false null hypothesis is unduly high
- The probability of type I vs. type II errors depend on the significance level $\alpha$, the properties of the test statistic (here: FR) and the statistical properties of the database
- Statistical experts warn against a slavish usage of the standard type I significance test in a statistical context.

# Other techniques in financial classification

- Logistic regression
- The recursive partitioning algorithm (RPA)
- Mathematical programming
- Linear programming models
- Quadratic programming models
- Neural network classifiers

# Logistic Regression

- Logistic regression is part of a category of statistical models called generalized linear models
- Whereas discriminant analysis can only be used with continuous independent variables. Logistic regression allows one to predict a discrete outcome, such as group membership, from a set of variables that may be continuous, discrete, dichotomous, or a mix of any of these
- Generally, the dependent or response variable is dichotomous, such as presence/absence or success/failure.

# Logistic Regression...

- Even though the dependent variable in logistic regression is usually dichotomous, that is, the dependent variable can take the value 1 with a probability of success *q*, or the value 0 with probability of failure 1-*q*, applications of logistic regression have also been extended to cases where the dependent variable is of more than two cases

# Logistic Regression...

- The independent or predictor variables in logistic regression can take any form, i.e. logistic regression makes no assumption about the distribution of the independent variables
  - ⇨ They do not have to be normally distributed, linearly related or of equal variance within each group
- The relationship between the predictor and response variables is not a linear function, instead, the **logistic regression function** is used, which is the logit transformation of probability *q*

# Logistic Regression...

- **The Model:**

$$\theta = \frac{e^{(\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n)}}{1+e^{(\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n)}}$$

where $\alpha$ is the constant of the equation and, $\beta$:s are the coefficient of the predictor variables

- An alternative form of the logistic regression equation is:

$$\text{logit}[\theta(x)] = \log\left[\frac{\theta(x)}{1-\theta(x)}\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

# Logistic Regression...

- The goal of logistic regression is to correctly predict the category of outcome for individual cases using **the most parsimonious model**
- To accomplish this goal, a model is created that includes all predictor variables that are useful in predicting the response variable.
- Different methods for model creation
  - Stepwise regression
  - Backward stepwise regression

# Logistic Regression...

- Stepwise regression
  - Variables are entered into the model in the order specified by the researcher or logistic regression can test the fit of the model after each coefficient is added or deleted
  - Used in the exploratory phase of research where no a-priori assumptions regarding the relationships between the variables are made, thus the goal is to discover relationships

# Logistic Regression...

- Backward stepwise regression
  - The analysis begins with a full or saturated model and variables are eliminated from the model in an iterative process
  - The fit of the model is tested after the elimination of each variable to ensure that the model still adequately fits the data
  - When no more variables can be eliminated from the model, the analysis has been completed
  - The preferred method of exploratory analyses

# Logistic Regression...

- Two main uses of logistic regression
  - The prediction of group membership
    - Calculates **the probability** or success over the probability of failure
    - The results of the analysis are in the form of an odds ratio
    - For example, logistic regression is often used in epidemiological studies where the result of the analysis is the probability of developing cancer after controlling for other associated risks
  - Logistic regression also provides knowledge of the relationships and strengths among the variables

# Recursive Partitioning Algorithm (RPA)

- A decision tree model for classification
- For each independent variable the observations in each class are sorted in increasing order, and the cumulative density functions for each class are defined
- The maximum absolute difference between the cumulative functions defines the cutting variable and cutting point for a node in the decision tree

# Recursive Partitioning Algorithm, an example

- Assume that we have a sample of 9 cases of which 5 belong to class 1 and 4 to class 2. The cases are measured by two predictor variables $x_1$ and $x_2$. The input data is presented in the following table:

# Recursive Partitioning Algorithm, an example...

| Case | Class | $x_1$ | $x_2$ |
|---|---|---|---|
| 1 | 1 | 2 | 7 |
| 2 | 1 | 1 | 8 |
| 3 | 1 | 7 | 9 |
| 4 | 1 | 2 | 5 |
| 5 | 1 | 4 | 8 |
| 6 | 2 | 6 | 3 |
| 7 | 2 | 3 | 1 |
| 8 | 2 | 8 | 6 |
| 9 | 2 | 8 | 3 |

# Recursive Partitioning Algorithm, an example...

- The cases are first ordered in ascending order of the first predictor variable $x_1$
- Then, the empirical cumulative distributions $F_1(x_1)$ and $F_2(x_1)$ are computed, and the absolute difference $|F_1(x_1) - F_2(x_1)|$ is computed
- The results of the computations are presented in the following table:

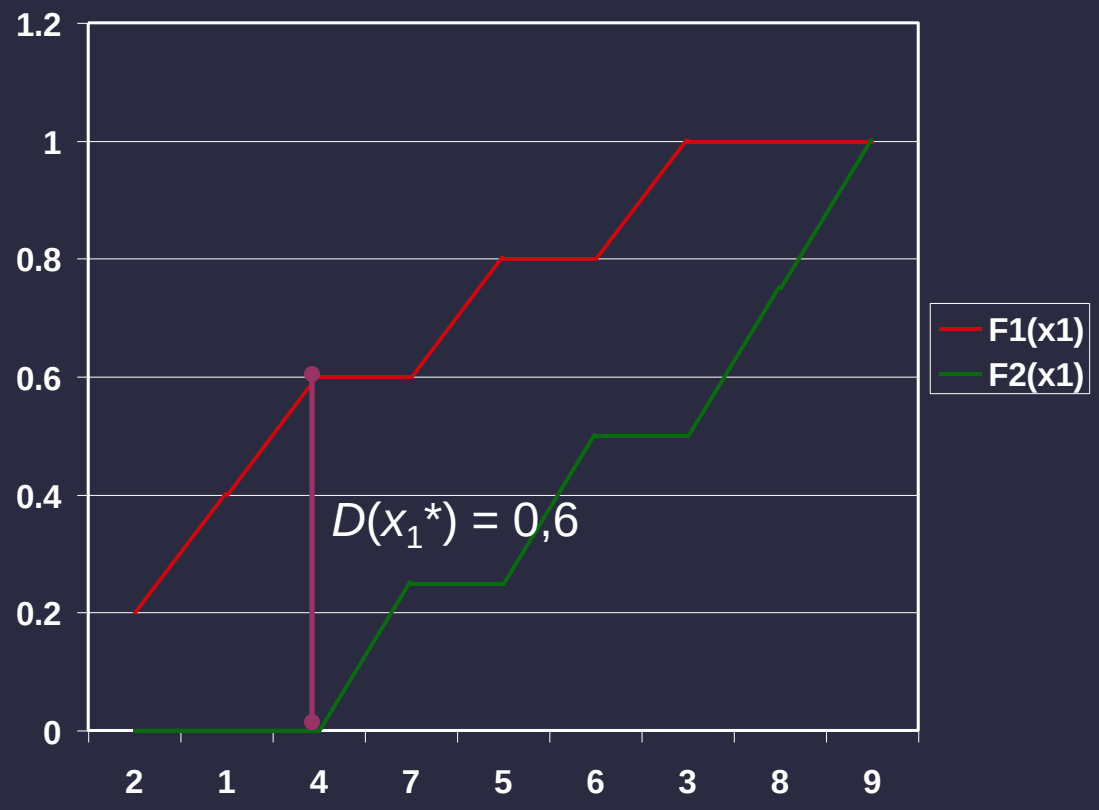
# Recursive Partitioning Algorithm, an example...

| Case | $x_1$ | Class | $F_1(x_1)$ | $F_2(x_1)$ | $\lvert F_1(x_1) - F_2(x_1) \rvert$ |
|---|---|---|---|---|---|
| 2 | 1 | 1 | 0,20 | 0,00 | 0,20 |
| 1 | 2 | 1 | 0,40 | 0,00 | 0,40 |
| 4 | 2 | 1 | 0,60 | 0,00 | **0,60** |
| 7 | 3 | 2 | 0,60 | 0,25 | 0,35 |
| 5 | 4 | 1 | 0,80 | 0,25 | 0,55 |
| 6 | 6 | 2 | 0,80 | 0,50 | 0,30 |
| 3 | 7 | 1 | 1,00 | 0,50 | 0,50 |
| 8 | 8 | 2 | 1,00 | 0,75 | 0,25 |
| 9 | 8 | 2 | 1,00 | 1,00 | 0,00 |

# Recursive Partitioning Algorithm, an example...

- The maximum value of the absolute difference between the cumulative distribution functions for the first predictor variable is 0.60, corresponding to value $x_1 = 2$.
- The best discrimination based on variable $x_1$ is achieved by assigning the three cases with the value of $x_1$ less than or equal to 2 to the class to which the majority of the cases in this subgroup, i.e. to class 1, and the six cases with $x_1$ greater than 2 to class
- Thus, two of the nine cases are misclassified by variable $x_1$

# Recursive Partitioning Algorithm, an example...



# Recursive Partitioning Algorithm, an example...

- The same procedure is then performed with the other predictor variable $x_2$, in order to find the best univariate discriminator
- The computational results and the corresponding graphs are presented below:

# Recursive Partitioning Algorithm, an example...

| Case | $x_2$ | Class | $F_1(x_2)$ | $F_2(x_2)$ | $\lvert F_1(x_2) - F_2(x_2) \rvert$ |
|---|---|---|---|---|---|
| 7 | 1 | 2 | 0,00 | 0,25 | 0,25 |
| 6 | 3 | 2 | 0,00 | 0,50 | 0,60 |
| 9 | 3 | 2 | 0,00 | 0,75 | 0,75 |
| 4 | 5 | 1 | 0,20 | 0,75 | 0,55 |
| 8 | 6 | 2 | 0,20 | 1,00 | **0,80** |
| 1 | 7 | 1 | 0,40 | 1,00 | 0,60 |
| 2 | 8 | 1 | 0,60 | 1,00 | 0,40 |
| 5 | 8 | 1 | 1,00 | 1,00 | 0,20 |
| 3 | 9 | 1 | 1,00 | 1,00 | 0,00 |

# Recursive Partitioning Algorithm, an example...

# Recursive Partitioning Algorithm, an example...

- The maximum value of the absolute difference between the cumulative distributions is now 0.8, corresponding to value $x_2 = 6$
- Thus the best discrimination based on variable $x_2$ is achieved by assigning the five cases with $x_2$ less than or equal to 6 into class 2 and the other four cases into class 1.
- By this partitioning, only one of the nine cases is misclassified, i.e. variable $x_2$ is superior to variable $x_1$, in terms of univariate discrimination power.

# Recursive Partitioning Algorithm, an example...

- Mathematically, the best univariate discriminator is found by comparing the maximum distances $D(x_1)$ and $D(x_2)$ and selecting the variable with the maximum $D(x_j)$
- As the maximum $D(x_j)$ is

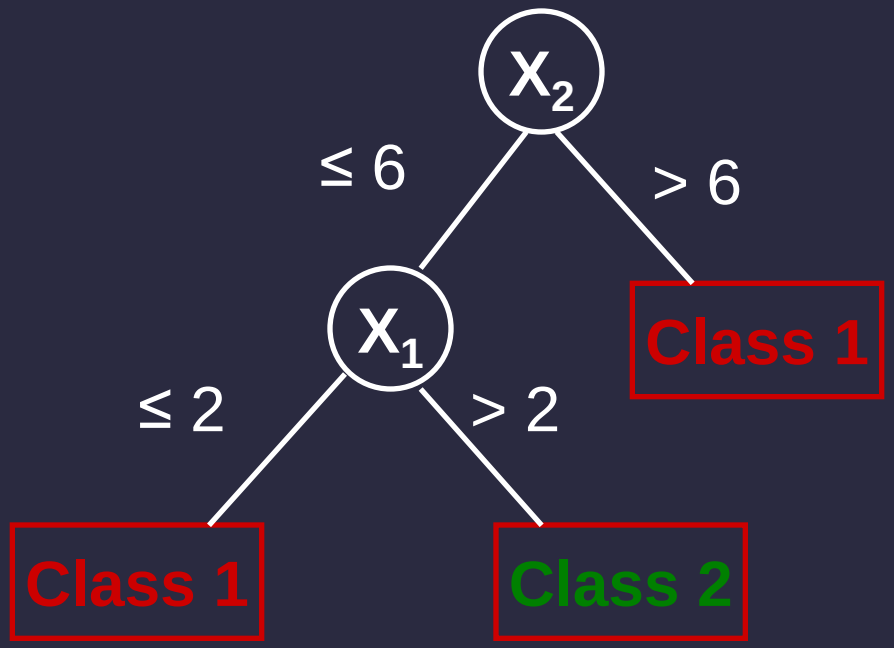
  $$\text{Max}(D(x_1),D(x_2)) = \text{Max}(0.6;0.8) = 0.8 = D(x_2)$$

  $x_2$ is the variable with the greatest univariate discrimination power and the first splitting is done in the way suggested by the second predictor variable

# Recursive Partitioning Algorithm, an example...

- As one of the two subgroups contains cases from both classes, an additional partitioning of the subgroup consisting of observations 4, 6, 7, 8 and 9 is possible
- The maximum distance in this second partitioning is 1.0 corresponding to value $x_1 = 2$
- The optimal partitioning now is to assign the case with $x_1$ equal to 2 into class 1 and the other four cases into class 2
- All the nine cases are now correctly assigned in pure classes

# Recursive Partitioning Algorithm, an example... The decision tree



# The Linear Programming classification model by Freed and Glover (1981)

- Given observations $x_i$ and groups $G_j$, find the linear transformation $a$, and the appropriate boundaries $b_j^L$ and $b_j^U$, to 'properly' categorize each $x_i$
- Bounds $b_j^L$ and $b_j^U$ represent respectively the lower and upper boundaries for points assigned to group $j$.
- Thus the task is to determine a linear predicting or weighting scheme $a$ and breakpoints $b_j^L$ and $b_j^U$, such that

$$b_j^L \leq x_k a \leq b_j^U \Leftrightarrow x_k \in G_j$$

and

$$b_1^L < b_1^U < b_2^L < b_2^U < \ldots < b_g^U$$

# The Linear Programming classification model by Freed and Glover (1981) …

- The points $x_i$ may of course be distributed in a way that makes complete group differentiation impossible
- Therefore, it becomes important to endow the weighting scheme with the power to establish the foregoing group differentiation with minimum exception
- This implies that we should determine a predictor $\boldsymbol{a}$ such that:

$$\boldsymbol{x}_i\boldsymbol{a} \geq b_j^L, \boldsymbol{x}_i\boldsymbol{a} \leq b_j^U \text{ for all } x_i \in G_j.$$

# The Linear Programming classification model by Freed and Glover (1981) …

- To ensure that the target is achieved as nearly as possible, we impose the following goal constraints:

$$b_j^U + \varepsilon < b_{j+1}^L + \alpha_j, \text{ for } j = 1,\ldots,g-1$$

where $g$ = number of groups and $0 < \varepsilon$.

- The objective function is defined as

$$\text{Minimize} \sum_{j=1}^{g-1} \alpha_j$$

# Neural Network classification

- Neural networks are computation models that mimic the human learning process (cf. Östermark [2009])
- A network is trained by
  - Giving one observation at a time as input
  - Computing the output value for the observation with the current net
  - Comparing the computed output value with the known correct result
  - Adjusting the net weights based on the difference between the computed and observed output values

# An example of a neural network classifier

**Classification**

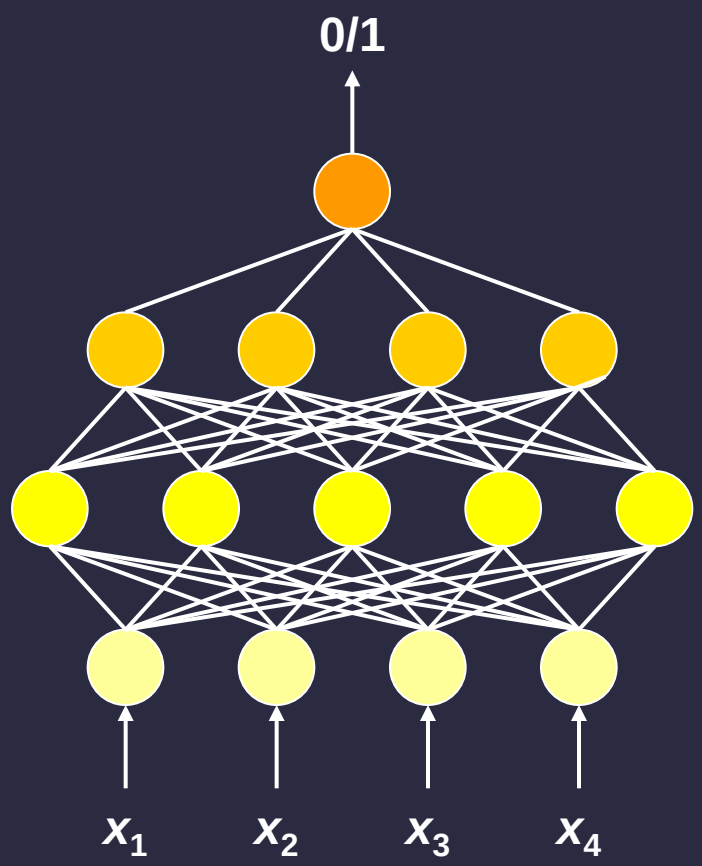
**Output layer**

**Second hidden layer**

**First hidden layer**

**Input layer**

**Predictor variables**



# 3. Case: Bankruptcy prediction in the Spanish banking sector

- Reference: Olmeda, Ignacio and Fernández, Eugenio: "Hybrid classifiers for financial multicriteria decision making: The case of bankruptcy prediction", *Computational Economics* **10**, 1997, 317-335.
- Sample: 66 Spanish banks
  - 37 survivors
  - 29 failed
- Sample was divided in two sub-samples
  - Estimation sample, 34 banks, for estimating the model parameters
  - Holdout sample, 32 banks, for validating the results

# Case: Bankruptcy prediction in the Spanish banking sector

Input variables

- Current assets/Total assets
- (Current assets-Cash)/Total assets
- Current assets/Loans
- Reserves/Loans
- Net income/Total assets
- Net income/Total equity capital
- Net income/Loans
- Cost of sales/Sales
- Cash flow/Loans

# Empirical results

- Analyzing the total set of 66 observations
  - Group statistics – comparing the group means
  - Testing for the equality of group means
  - Correlation matrix
- Classification with different methods
  - Estimating classification models using the estimation sample of 34 observations
  - Checking the validity of the models by classifying the holdout sample of 32 observations

# Confusion matrix – Classification results for the holdout sample using Logistic Regression

| | | Predicted class | |
|---|---|---|---|
| | | Survived | Failed |
| True class | Survived | 17 | 1 |
| | | 94.44 % | 5.56 % |
| | Failed | 3 | 11 |
| | | 21.43 % | 78.57 % |

# Summary of classifications (Estimation sample)

| Method | Correct class | Errors SW | Errors NE | Total number | Percents Correct | Percents SW | Percents NE |
|---|---|---|---|---|---|---|---|
| **RPA** | 30 | 1 | 3 | 34 | 88.24 % | 2.94 % | 8.82 % |
| **MDA** | 30 | 0 | 4 | 34 | 88.24 % | 0.00 % | 11.76 % |
| **MDA-Q** | 31 | 0 | 3 | 34 | 91.18 % | 0.00 % | 8.82 % |
| **MDA-W** | 31 | 0 | 3 | 34 | 91.18 % | 0.00 % | 8.82 % |
| **LogR** | 33 | 0 | 1 | 34 | 97.06 % | 0.00 % | 2.94 % |
| **LP** | 28 | 1 | 5 | 34 | 82.35 % | 2.94 % | 14.71 % |
| **LP-Q** | 34 | 0 | 0 | 34 | 100.00 % | 0.00 % | 0.00 % |
| **LPG** | 33 | 0 | 1 | 34 | 97.06 % | 0.00 % | 2.94 % |
| **LPGQ** | 34 | 0 | 0 | 34 | 100.00 % | 0.00 % | 0.00 % |
| **Kohonen** | 24 | 3 | 7 | 34 | 70.59 % | 8.82 % | 20.59 % |

# Summary of classifications (Holdout sample)

| Method | Correct class | Errors SW | Errors NE | Total number | Percents Correct | Percents SW | Percents NE |
|---|---|---|---|---|---|---|---|
| **RPA** | 27 | 2 | 3 | 32 | 84.38 % | 6.25 % | 9.38 % |
| **MDA** | 25 | 4 | 3 | 32 | 78.13 % | 12.50 % | 9.38 % |
| **MDA-Q** | 20 | 7 | 5 | 32 | 62.50 % | 21.88 % | 15.63 % |
| **MDA-W** | 25 | 5 | 2 | 32 | 78.13 % | 15.63 % | 6.25 % |
| **LogR** | 28 | 3 | 1 | 32 | 87.50 % | 9.38 % | 3.13 % |
| **LP** | 24 | 5 | 3 | 32 | 75.00 % | 15.63 % | 9.38 % |
| **LP-Q** | 21 | 7 | 4 | 32 | 65.63 % | 21.88 % | 12.50 % |
| **LPG** | 25 | 4 | 3 | 32 | 78.13 % | 12.50 % | 9.38 % |
| **LPGQ** | 21 | 7 | 4 | 32 | 65.63 % | 21.88 % | 12.50 % |
| **Kohonen** | 16 | 4 | 12 | 32 | 50.00 % | 12.50 % | 37.50 % |

# Classification results – Error types

- The classification results for the different methods differ in
  - Total classification accuracy
    - Descriptive (Estimation sample)
    - Predictive (Holdout sample)
  - Error types
    - Classifying a survivor as failed
    - Classifying a failed as survivor
- Many methods may be calibrated to take into account the relative severity of the two types of errors

# Fisher's discriminant function coefficients

| | Survived | Failed |
|---|---|---|
| Constant | -758.242 | -758.800 |
| CA/TA | 48.588 | 34.572 |
| CA_Cash/TA | 9.800 | 23.506 |
| CA/Loans | -18.031 | -16.947 |
| Res/Loans | 351.432 | 342.204 |
| NI/TA | -246 563.200 | -236 546.700 |
| NI/TEC | 774.368 | 740.035 |
| NI/Loans | 23 681.300 | 21 4974.000 |
| CofS/Sales | 1 499.659 | 1 505.547 |
| CF/Loans | 14 625.844 | 14 245.368 |

# Example on classifying an observation by discriminant functions

| | Obs. 1 | Survived | Score | Failed | Score |
|---|---|---|---|---|---|
| Constant | | -758.24 | -758.24 | -758.800 | -758.80 |
| CA/TA | 0.4611 | 48.59 | 22.40 | 34.572 | 15.94 |
| CA_Cash/TA | 0.3837 | 9.80 | 3.76 | 23.506 | 9.02 |
| CA/Loans | 0.4894 | -18.03 | -8.82 | -16.947 | -8.29 |
| Res/Loans | 0.0077 | 351.43 | 2.71 | 342.204 | 2.63 |
| NI/TA | 0.0057 | -246563.2 | -1405.41 | -236546.7 | -1348.32 |
| NI/TEC | 0.0996 | 774.37 | 77.13 | 740.035 | 73.71 |
| NI/Loans | 0.0061 | 23681.3 | 1364.46 | 214974.0 | 1311.34 |
| CofS/Sales | 0.8799 | 1499.66 | 1319.55 | 1505.547 | 1324.73 |
| CF/Loans | 0.0092 | 14625.84 | 134.56 | 14245.368 | 131.06 |
| Total Score | | | 752.08 | | 753.02 |

Larger score ⇒
Classification: Failed

# 4. Some comments on hypothesis testing

Assume that we – as a bank institution - want to distinguish between non-distressed (H0) vs. distressed (H1) firms using a suitable financial ratio FR (for example based on the discriminant score), in order to reduce the financial risk in loan decisions. To do this, we need to compare the FR of a firm with a critical value $FR_c$. If FR> $FR_c$, then the firm is assumed to be distressed, otherwise not.

There is a tension between type I and type II errors. The first type is smaller, the higher is the significance (i.e. the smaller is $\alpha$): The probability of rejecting H0 falsely is smaller, the smaller is $\alpha$. Type I error is the probability of rejecting H0 even if it is true. With $\alpha$=10% this probability is twice that of $\alpha$=5% and ten times that of $\alpha$=1%. We throw away a gold nugget among the rubbish in 10% of all cases by rejecting H0 for firms that actually are non-distressed.

If we get an extremely high FR for a firm, however, everybody will realize that the probability of that firm being non-distressed is practically negligible: The probability of such an outcome being generated by chance is very low. In such a case it is safe to conclude that the firm is financially distressed and, for example, to reject financing a project that the firm is contemplating.

# Some comments on hypothesis testing…

- On the other hand, the more we shift the critical significance level ($FR_c$) to the right, the less frequently we will reject H0. If $FR_c$ is extremely high, we will accept H0 almost always: Everybody will receive a loan from our bank.
- But the more we shift the critical level $FR_c$ to the right, the more often we will accept H0 even if it is false: there will be firms in our clientele that should not be there. These firms are distressed, even though we have failed to detect this because of a high $FR_c$.
- This latter error is denoted Type II. Because of the high $FR_c$ the test has a low power: the probability of failing to reject a false null hypothesis is unduly high.
- The probability of type I vs. type II errors depend on the significance level $\alpha$, the properties of the test statistic (here: FR) and the statistical properties of the database. Statistical experts warn against a slavish usage of the standard type I significance test in a statistical context.

# References


- Bartlett, M.S. (1954). "A note on multiplying factors for various χ2 approximations". *J. Royal Statist. Soc. Series* B 16, pp. 296–298.
- Balcaen S, Ooghe H (2006). “35 years of studies on business failure: on overview of the classic statistical methodologies and their related problems. The British Accounting Review 38, 69-93.
- Freed, N. and F. Glover: "Evaluating alternative Linear Programming models to solve the two-group discriminant problem", *Decision Sciences*, **17**, 1986, pp. 151-162.
- Frydman, H., E. T. Altman, and D. L. Kao: "Introducing recursive partitioning for financial classification: the case of financial distress", *The Journal of Finance*, **40**:1, March, 1985, 269-291
- Olmeda, Ignacio and Fernández, Eugenio: "Hybrid classifiers for financial multicriteria decision making: The case of bankruptcy prediction", *Computational Economics* **10**, 1997, 317-335.

# References


- Aziz M.A, Dar H. A: Predicting corporate bankruptcy: where we stand?. Corporate Governance, vol 6, No 1, 2006, 18-33.

# References


- Östermark, Ralf and Jaana Aaltonen: "Comparing mathematical, statistical and artificial intelligence based techniques in bankruptcy prediction", *Accounting & Business Review* **5**:1, 1998, 95-120.
- Östermark, Ralf and Rune Höglund; "Addressing the multigroup discriminant problem using multivariate statistics and mathematical programming ", *European Journal of Operational Research* **108**:1, 1998, 224-237.
- Östermark, R.: Geno-mathematical identification of the multi-layer perceptron. *Neural Computing and Applications* **18**:4, 2009, pp. 331-344. (**http://www.springerlink.com/openurl.asp?genre=article&id=doi:10.1007/s00521-008-0184-4**).

# References


- Bartlett, M.S. (1954). "A note on multiplying factors for various χ2 approximations". *J. Royal Statist. Soc. Series* B 16, pp. 296–298.
- Balcaen S, Ooghe H (2006). “35 years of studies on business failure: on overview of the classic statistical methodologies and their related problems. The British Accounting Review 38, 69-93.
- Freed, N. and F. Glover: "Evaluating alternative Linear Programming models to solve the two-group discriminant problem", *Decision Sciences*, **17**, 1986, pp. 151-162.
- Frydman, H., E. T. Altman, and D. L. Kao: "Introducing recursive partitioning for financial classification: the case of financial distress", *The Journal of Finance*, **40**:1, March, 1985, 269-291
- Olmeda, Ignacio and Fernández, Eugenio: "Hybrid classifiers for financial multicriteria decision making: The case of bankruptcy prediction", *Computational Economics* **10**, 1997, 317-335.

# References


- Aziz M.A, Dar H. A: Predicting corporate bankruptcy: where we stand?. Corporate Governance, vol 6, No 1, 2006, 18-33.
- Östermark, Ralf and Jaana Aaltonen (1998): "Comparing mathematical, statistical and artificial intelligence based techniques in bankruptcy prediction", *Accounting & Business Review* **5**:1, 95-120.
- Östermark, Ralf and Rune Höglund (1998): "Addressing the multigroup discriminant problem using multivariate statistics and mathematical programming ", *European Journal of Operational Research* **108**:1, 224-237.
- Östermark, R (2009): Geno-mathematical identification of the multi-layer perceptron. *Neural Computing and Applications* **18**:4, pp. 331-344. (**http://www.springerlink.com/openurl.asp?genre=article&id=doi:10.1007/s00521-008-0184-4**).