Klasyfikatory liniowe
Linear classifiers

JERZY STEFANOWSKI
Institute of Computing Sciences,
Poznań University of Technology

UMiSN – slajdy wykładu
Wersja 2010

Plan

1. Liniowe klasyfikatory
2. Klasyczne liniowa analiza dyskryminacyjna
3. Podejścia probabilistyczne
4. Inne zagadnienia
5. Oprogramowanie

Formalizacja problemu klasyfikacji

- W przestrzeni danych (ang. measurement space) Ω znajdują się wektory danych **x** stanowiące próbkę uczącą *D*, należące do dwóch lub więcej *K* klas

$$D = \left\{ \left(\mathbf{x}_i, c_i \right) \middle| \, x_i \in R^p, c_i \in \{C_1, \ldots C_k\} \right\}_{i=1}^{N}$$

- Klasyfikacja jest dokonywana na podstawie funkcji będącej liniową kombinacją *p* cech i parametrów

$$y = f(\mathbf{x}, \mathbf{w})$$

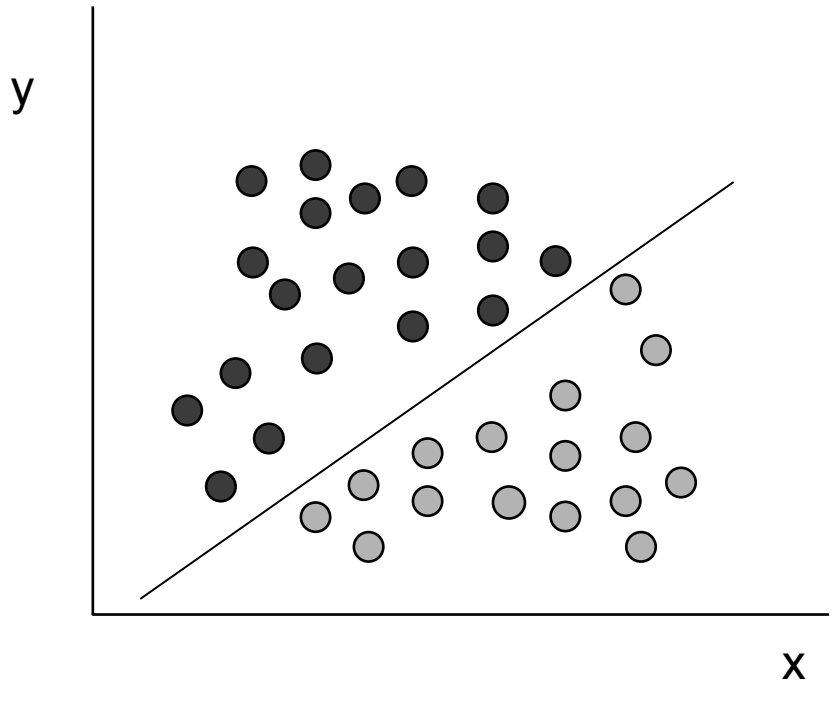
- Dążymy do sytuacji

$$y_i = f(\mathbf{x}_i, \mathbf{w}) = c_i$$

- i/lub minimalizacji błędów klasyfikacji

$$y_i \neq c_i$$

Liniowa funkcja separująca (graniczna)

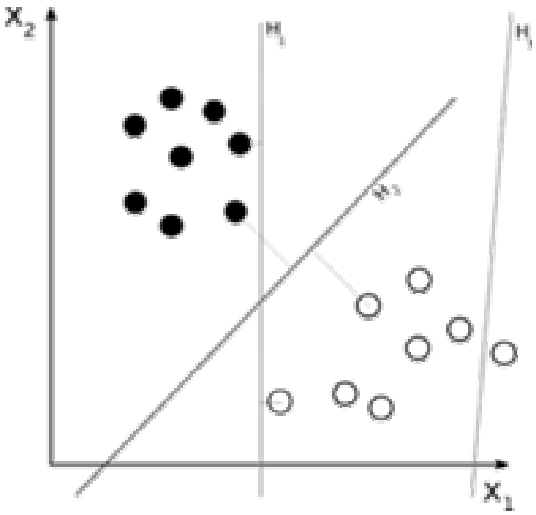


- Szukamy klasyfikatora pozwalającego na podział całej przestrzeni na obszary odpowiadającej klasom (dwie lub więcej) oraz pozwalającego jak najlepiej klasyfikować nowe obiekty x do klas
- Podejście opiera się na znalezieniu tzw. granicy decyzyjnej między klasami $\rightarrow f(\mathbf{x})=\mathbf{w}^T\cdot\mathbf{x}$

$$y=\begin{cases} f(\mathbf{x}_i)>T & \mathbf{x}_i\in C_1 \\ f(\mathbf{x}_i)<T & \mathbf{x}_i\in C_2 \end{cases}$$

Różne podejścia do budowy klasyfikatorów liniowych

- Podejścia generatywne (probabilistyczne)
  - Analiza dyskryminacyjna (związ. z rozkładem normalnym)
  - Wersja klasyfikacji Bayesowskiej (dwumianowy rozkład)
- Podejścia wykorzystujące własności zbioru uczącego
  - Perceptron liniowy Rosenblata (iteracyjne poprawki wag)
  - Metoda wektorów nośnych (max. marginesu klasyfikatora)
  - Regresja logistyczna (EM estymacja)



Co jest celem analizy dyskryminacyjnej

- Podejście statystyczne do problemów klasyfikowania obiektów (term. ang. *Discriminant Analysis*)
  - Oryginalnie wprowadzona przez R.A.Fishera (1936) dla funkcji liniowych (2 klasy),
  - Metody probabilistyczne – B.Welch .
- Dostępna w wielu programach, np. SAS, SPSS, R lub Statistica,…
- Liczne zastosowania
- …

Liniowa analiza dyskryminacyjna

- Problem wprowadzony przez R.A.Fishera w 1936 dla wielowymiarowej przestrzeni atrybutów (zmiennych liczbowych) – dyskryminacja 2 klas
- Fisher oryginalnie zaproponował poszukiwanie kierunku projekcji, na którym można dobrze rozdzielić zrzutowane obie klasy
  - Średnie w klasach są dostatecznie oddalone od siebie
  - Obszary rozrzutu (rozproszenia, zmienności) obu klas nie nakładają się zbyt mocno.

Intuicja projekcji w Fisher's Linear Discriminant [EST}

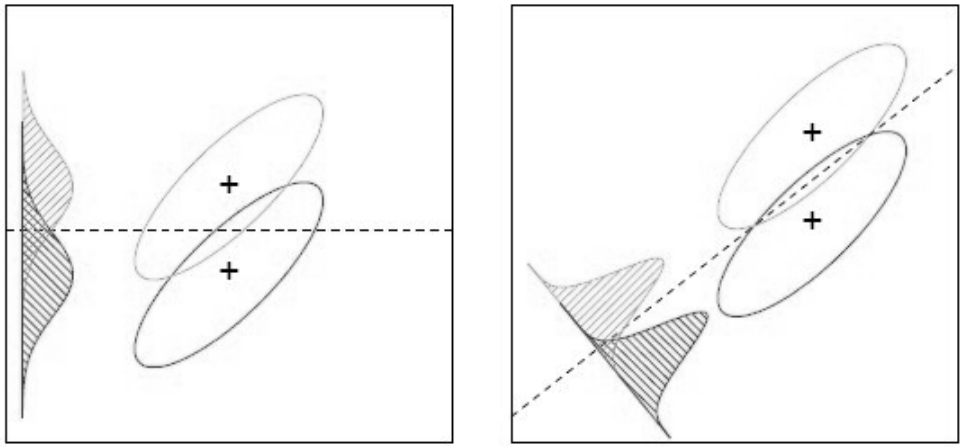


Figure 4.9: *Although the line joining the centroids defines the direction of greatest centroid spread, the projected data overlap because of the covariance (left panel). The discriminant direction minimizes this overlap for Gaussian data (right panel).*

„From training set we want to find out a direction where the separation between the class means is high and overlap between the classes is small"

Trochę uwag matem. o projekcji

- Dysponujemy przykładami uczącymi opisanymi *p* -cechami $\mathbf{x}=[x_1,x_2,\ldots,x_p]^T$ należącymi do dwóch klas C_1 i C_2 (odpowiednio n_1 i n_2)
- Wektory p-wymiarowe **x** są zrzutowane na prostą (kierunek związany z parametrami **w**). Algebraicznie odpowiada to zastąpieniu ich skalarem $z = \mathbf{w}^T \cdot \mathbf{x}$. Celem jest taki dobór w aby na podstawie nowej zmiennej *z* przykłady z obu klas były jak najlepiej rozdzielone.

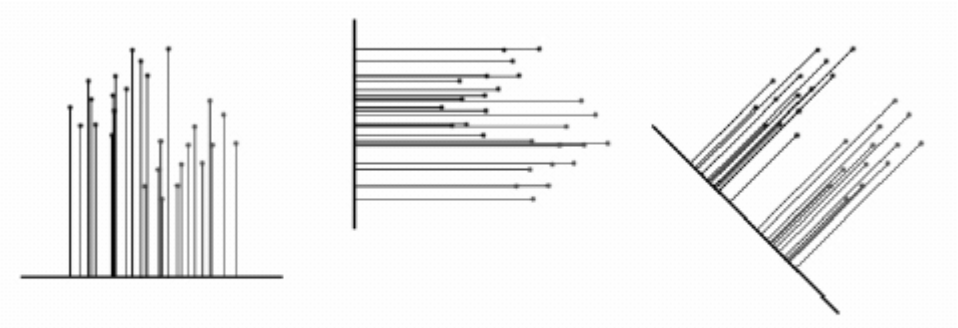


Figure 1: Projection of data from two clases onto various lines.

Założenia co do danych

- Fisher – dość ograniczone założenia: wektor *p* wartości oczekiwanych E(**x**) oraz rozproszenie charakteryzowane przez macierz kowariancji Σ=Cov(**x**)=E[(**x**-E(**x**))·[(**x**-E(**x**))^T]
- Estymatory

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})\cdot(x_i - \bar{x})^T$$

- Wariancja po rzutowaniu **x** na prosta o wektorze kierunkowym **w**

$$Var(\mathbf{w}^T\mathbf{x}) = \mathbf{w}^T\Sigma\mathbf{w}$$

Sformułowanie problemu Fisher LDA

Cel

- Maksymalizuj odległość zrzutowanych średnich klas
- Minimalizuj wariancje wewnątrz klasową

- Odległość między rzutami średnich

$$(\mathbf{w}^T \bar{\mathbf{x}}_1 - \mathbf{w}^T \bar{\mathbf{x}}_2)^2$$

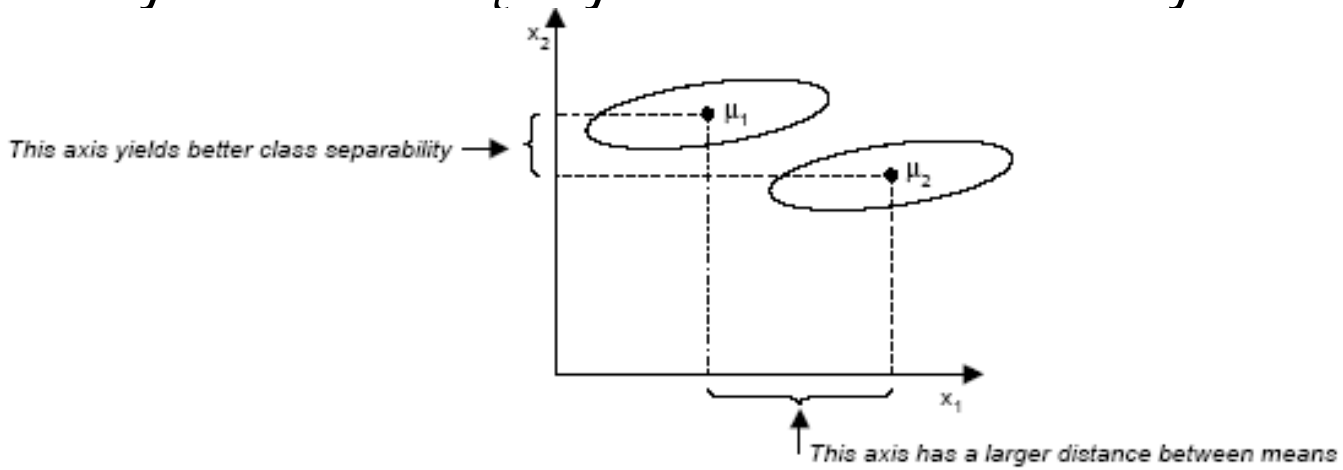
- Fisher założył, że obie klasy mają taką samą macierz kowariancji $S=S_1+S_2$. Dlatego wskaźnik zmienności wewnątrzgrupowej (wspólnej dla obu klas) zdefiniowany jest jako:

$$S_W = \frac{1}{n-2} \sum_{k=1}^{2} (n_k - 1) S_k$$

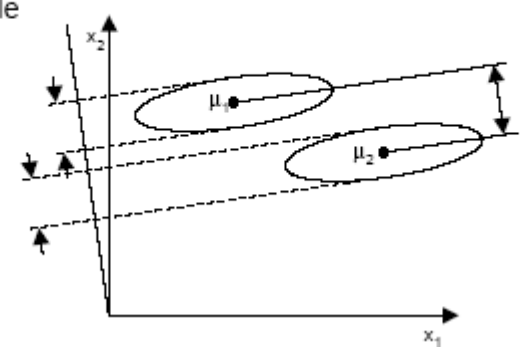
- Pamiętaj, że po rzutowaniu mamy $\mathbf{w}^T S_W \mathbf{w}$

Co optymalizować?

- Czy różnica między rzutami średnich wystarcza?



Therefore, we will be looking for a projection where examples from the same class are projected very close to each other and, at the same time, the projected means are as farther apart as possible



Sformułowanie problemu Fisherian LDA

- W celu maksymalizacji odległości rzutów średnich klas i minimalizacji wariancji wewnątrzklasowej należy poszukiwać wektora w który maksymalizuje następujące wyrażenie:

$$J(\mathbf{w}) = \frac{\left(\mathbf{w}^T \overline{\mathbf{x}}_1 - \mathbf{w}^T \overline{\mathbf{x}}_2\right)^2}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}$$

- Po znalezieniu kierunku maksymalizujące J(**w**) można stosować zasadę klasyfikacji na rzutowanej prostej. Przydziel **x** do klasy *j* dla której

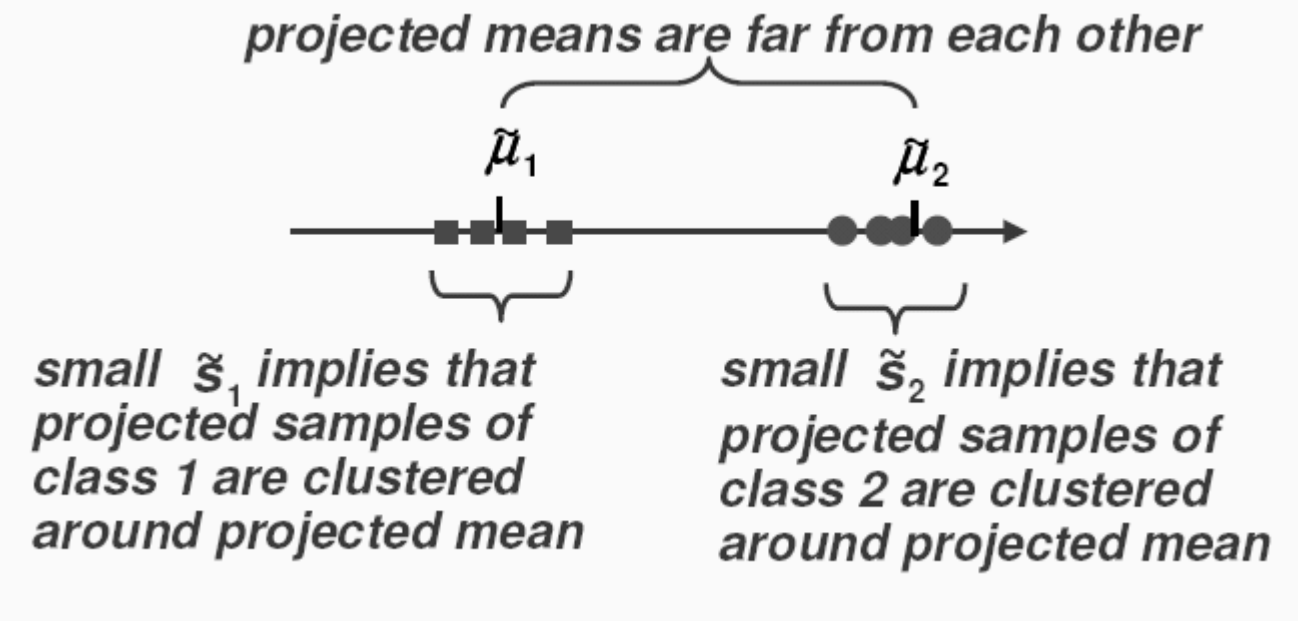
$$\left|\tilde{\mathbf{w}}^T \mathbf{x} - \tilde{\mathbf{w}}^T \overline{\mathbf{x}}_j\right| < \left|\tilde{\mathbf{w}}^T \mathbf{x} - \tilde{\mathbf{w}}^T \overline{\mathbf{x}}_k\right|$$

- Można wykazać, że ten wektor jest proporcjonalny

$$\tilde{\mathbf{w}} \propto S_W^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$$

Uwagi o konstrukcji wskaźnika

- Zwiększanie J(w) ma gwarantować dobrą separacje klas i ich rzutów



Hiperpłaszczyzna separująca

- Wyraz wolny to środek odcinka między rzutami średnich

$$m = \frac{1}{2}(\mathbf{w}^T \bar{\mathbf{x}}_1 - \mathbf{w}^T \bar{\mathbf{x}}_2) = \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T S_W^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$$

- Stąd liniowa funkcja dyskryminacyjna Fishera

$$y = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T S_W^{-1}[\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)]$$

- Więcej informacji, np.
  - J.Koronacki, J.Cwik: Statystyczne systemy uczące się
  - M.Krzyśko et al.: Systemy uczące się

Przypadek wielu klas (K > 2)

- Rozwiązanie Fishera uogólniono dla większej liczby K klas (C.Rao 1948)
- Średnia w próbie uczącej $\bar{\mathbf{x}} = \frac{1}{n}\sum_{j=1}^{K} \bar{\mathbf{x}}_j$
- Macierz zmienności wewnątrzklasowej

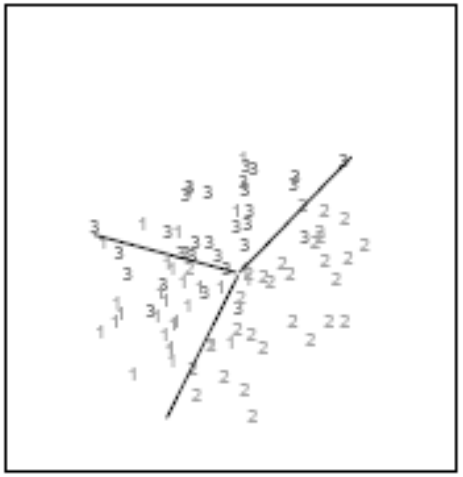
$$S_W = \frac{1}{n-K}\sum_{j=1}^{K}\sum_{i \in C_j}(x_i - \bar{x}_j)(x_i - \bar{x}_j)^T$$

- Macierz zmienności międzyklasowej

$$S_B = \frac{1}{K-1}\sum_{j=1}^{K} n_j(\bar{x}_j - \bar{x})(\bar{x}_j - \bar{x})^T$$

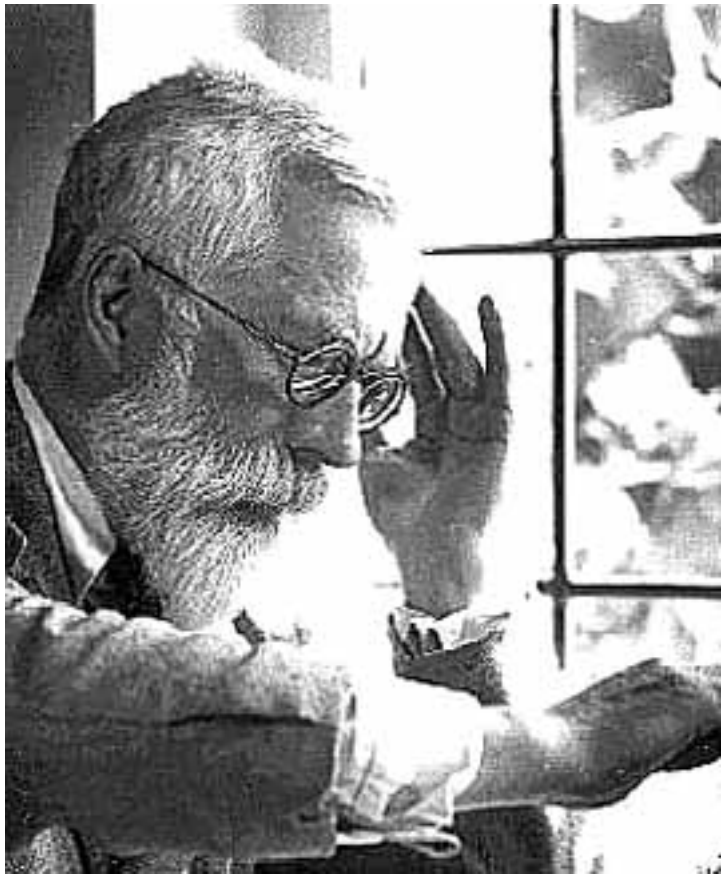
- Poszukuj wektora **w** maksymalizującego

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}$$



O autorze

- Ronald A. Fisher, 1890-1962



“The elaborate mechanism built on the theory of infinitely large samples is not accurate enough for simple laboratory data. Only by systematically tackling small sample problems on their merits does it seem possible to apply accurate tests to practical data.”

1936

Podejścia opisowe i probabilistyczne

- Stochastyczne / probabilistyczne
  - Zbiór obserwacji jest próbą losową pobraną z *k* podpopulacji $\pi 1, \pi 2, \ldots, \pi k$; celem jest taki podział aby podpopulacje odpowiadały właściwym *k* klasom $C1, C2, \ldots, Ck$
- Opisowe
  - Nie rozważa się losowości próby, zakłada się że posiadany zbiór zawiera przykłady z *k* klas $C1, C2, \ldots, Ck$; zadanie polega na poprawnym podziale zbioru na klasy

Sformułowanie probabilistyczne z Tw. Bayesa

- Obiekty $\mathbf{x} \in R^p$ i wielowymiarowy rozkład prawdopodobieństwa – funkcja gęstości $f(\mathbf{x}|C_i)$
- Każda klasa Ci opisana prawdopodobieństwa apriori p_i
- Bayesowska reguła klasyfikowania

$$P(C \mid x) = P(x \mid C)P(C) / P(x)$$

- Przydziel nowy obiekt **x** do tej klasy Ci dla której prawdopodobieństwo a posteriori jest największe:

$$P(C_j \mid \underline{x}) = p_j \cdot f(\underline{x} \mid C_j) \Big/ \sum_{i=1}^{K} p_i \cdot f(\underline{x} \mid C_i)$$

Rozwiązanie probabilistycznej reguły klasyfikacji

- Załóżmy, że rozkłady wektora x w poszczególnych klasach są p-wymiarowymi rozkładami normalnymi:

$$f(\underline{x} \mid C_i) = (2\pi)^{-0,5p} \left|\Sigma_i\right|^{-0,5} \exp\left[-0,5(\underline{x} - \underline{\mu}_i)^T \Sigma_i^{-1} (\underline{x} - \underline{\mu}_i)\right]$$

- Wykorzystując logarytmiczne przekształcenie twierdzenia Bayesa, obiekt x jest przydzielany do tej klasy Cj dla której funkcja dyskryminująca osiąga maksimum:

$$\delta_j(x) = -0,5(\underline{x} - \underline{\mu}_j)^T \Sigma_j^{-1} (\underline{x} - \underline{\mu}_j) - 0,5 \log\left|\Sigma_j\right| + \log p_j$$

- Jest to kwadratowa funkcja dyskryminująca (QDA)

Liniowa funkcja

- Założenie równości macierzy kowariancji Σ

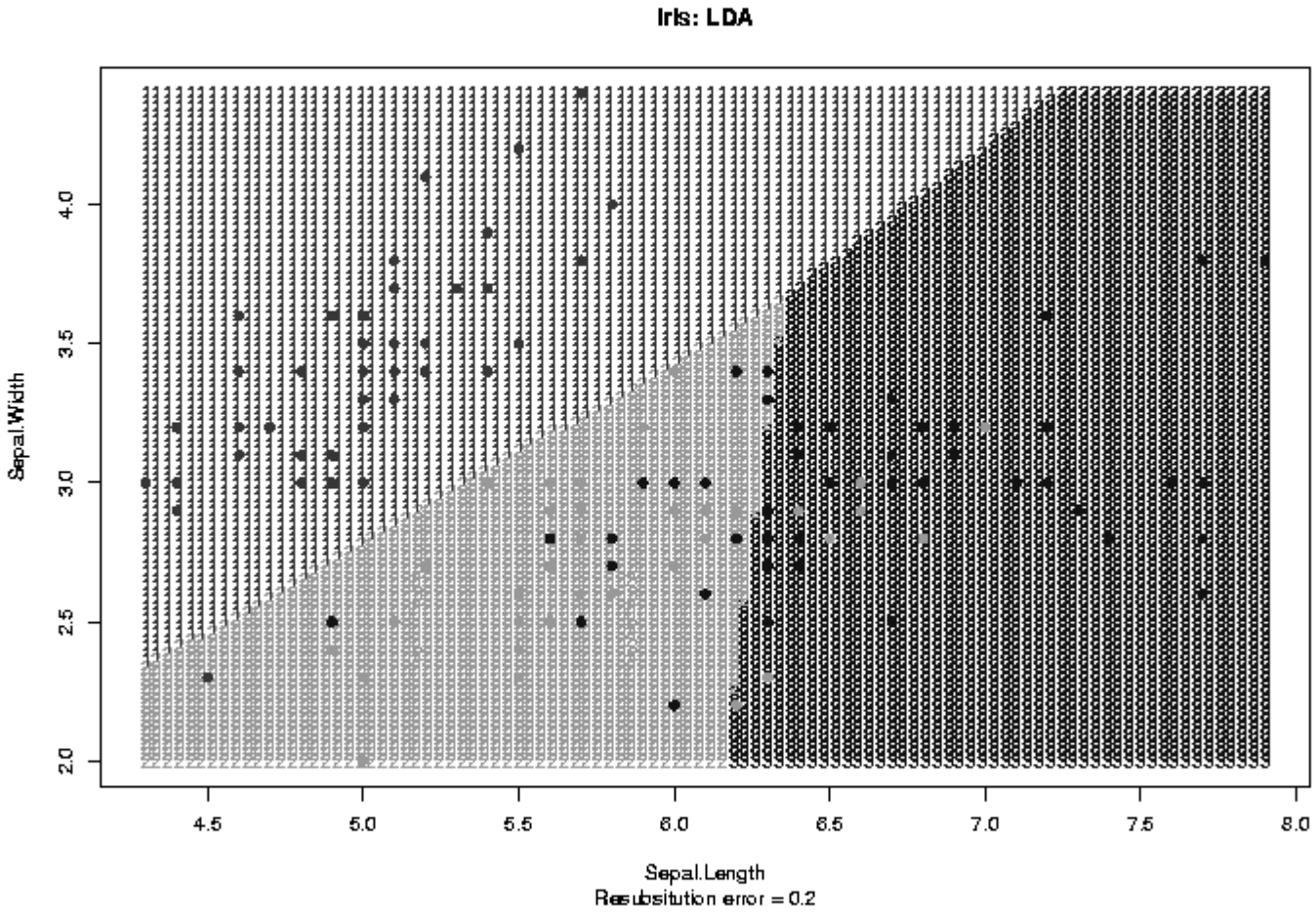
$$\delta_j(x) = \underline{x}^T \Sigma^{-1} \underline{\mu}_j - 0{,}5 \mu_j^T \Sigma^{-1} \underline{\mu}_j + \log\, p_j$$

- Dla dwóch klas – przekształcenie log-ratio

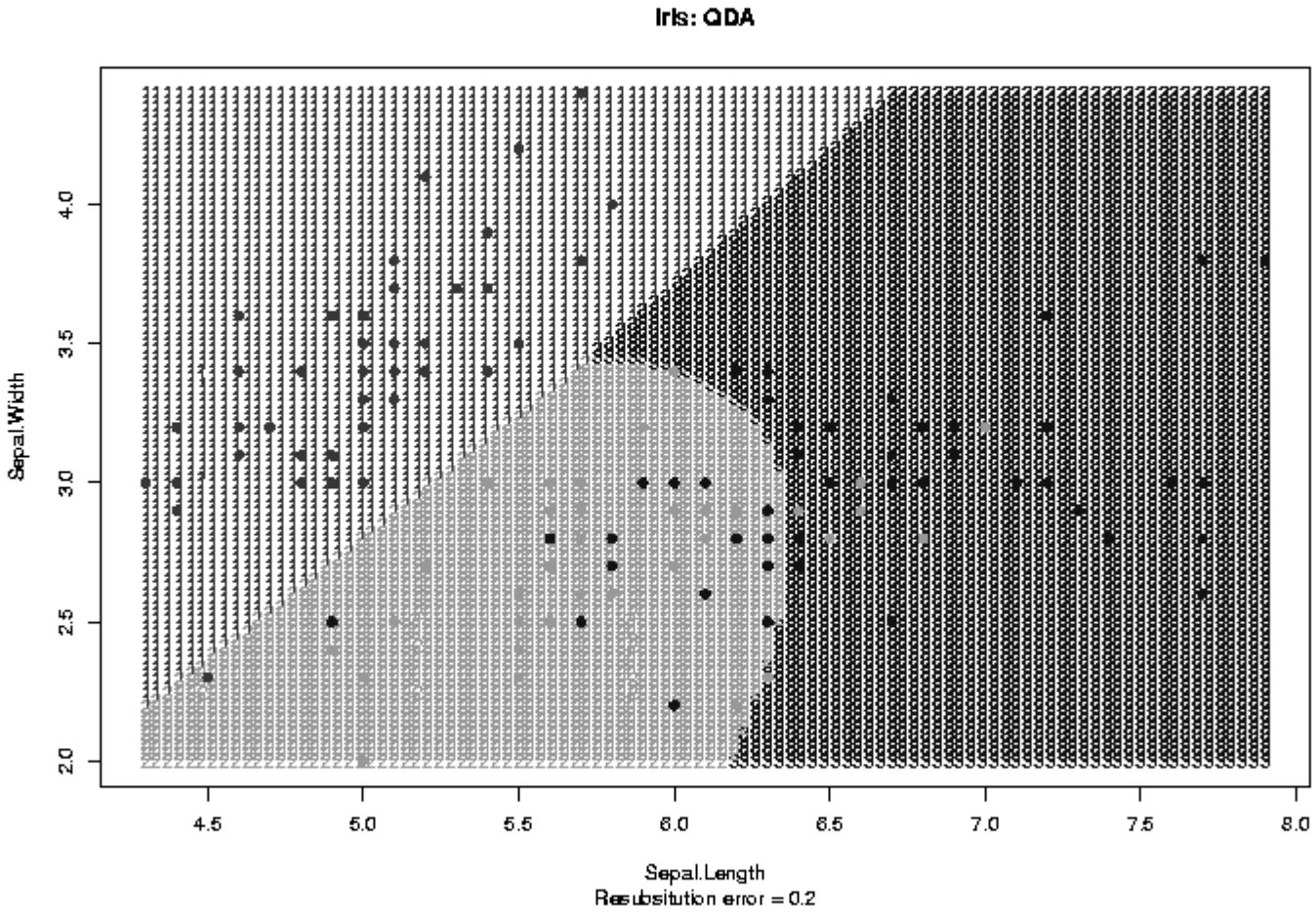
$$\begin{aligned} \log \frac{\Pr(y=k|\mathbf{x})}{\Pr(y=l|\mathbf{x})} &= \log \frac{f_k(\mathbf{x})}{f_l(\mathbf{x})} + \log \frac{p_k}{p_l} \\ &= \log \frac{p_k}{p_l} - \frac{1}{2}(\mu_k + \mu_l)^T \Sigma^{-1} (\mu_k - \mu_l) \\ &\quad + \mathbf{x}^T \Sigma^{-1} (\mu_k - \mu_l) \end{aligned}$$

- Więcej w Krzyśko … lub Hastie et al. Elements of Statistical Learning

Example: Linear discriminant analysis



Example: Quadratic discriminant analysis



Porównanie rozwiązań LDA i QDA

- Wybrany zbiór danych (za Hastie et al. Elements of Statistical Learning)

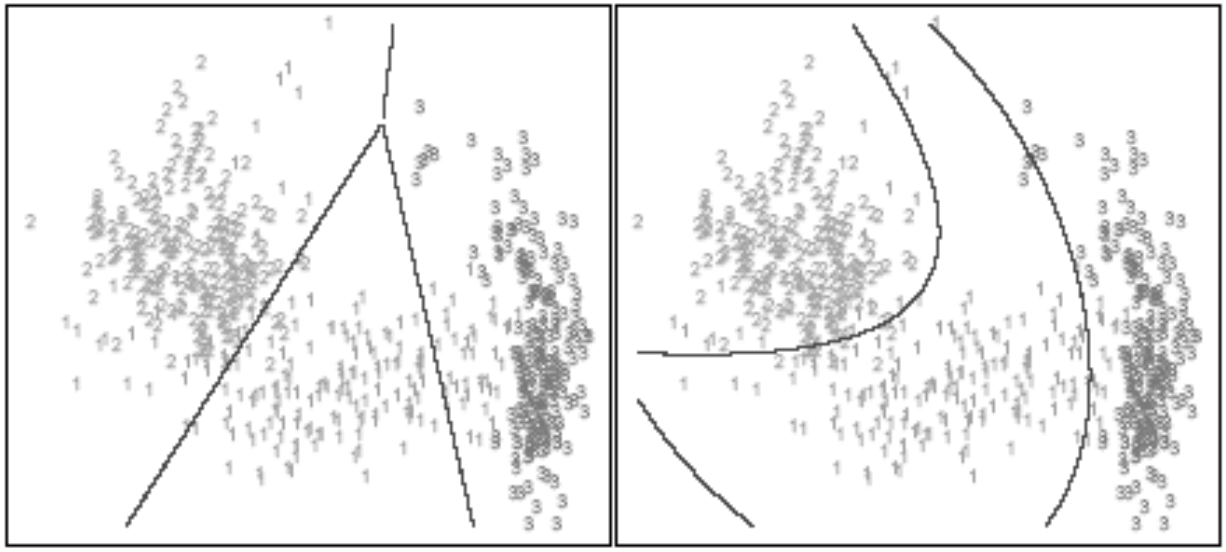


FIGURE 4.1. *The left plot shows some data from three classes, with linear decision boundaries found by linear discriminant analysis. The right plot shows quadratic decision boundaries. These were obtained by finding linear boundaries in the five-dimensional space* $X_1, X_2, X_1X_2, X_1^2, X_2^2$. *Linear inequalities in this space are quadratic inequalities in the original space.*

Wymogi stosowania modeli AD

- Zmienne wyrażone na skalach liczbowych
  - Specjalne podejścia dla zmiennych jakościowych (binaryzacja, model lokacyjny,…)
- Zmienne mają wielowymiarowy rozkład normalnych
- Macierze kowariancji dla poszczególnych klas są równe $\rightarrow$ jeśli nie, to bardziej złożone funkcje kwadratowe dyskryminujące.
- Problem doboru właściwych zmiennych.

Selekcja zmiennych

- W funkcji dyskryminującej uwzględniaj zmienne o dobrych właściwościach dyskryminujących
- Przykład kryterium jakości dyskryminacji:

$$\lambda = \frac{|S_w|}{|S_W + S_B|}$$

gdzie macierz zmienności wewnątrzklasowej

$$S_W = \frac{1}{n-k} \sum_{j=1}^{k} \sum_{i \in C_j} (x_i - \bar{x}_j)(x_i - \bar{x}_j)^T$$

a macierz zmienności międzyklasowej

$$S_B = \frac{1}{k-1} \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x})(\bar{x}_j - \bar{x})^T$$

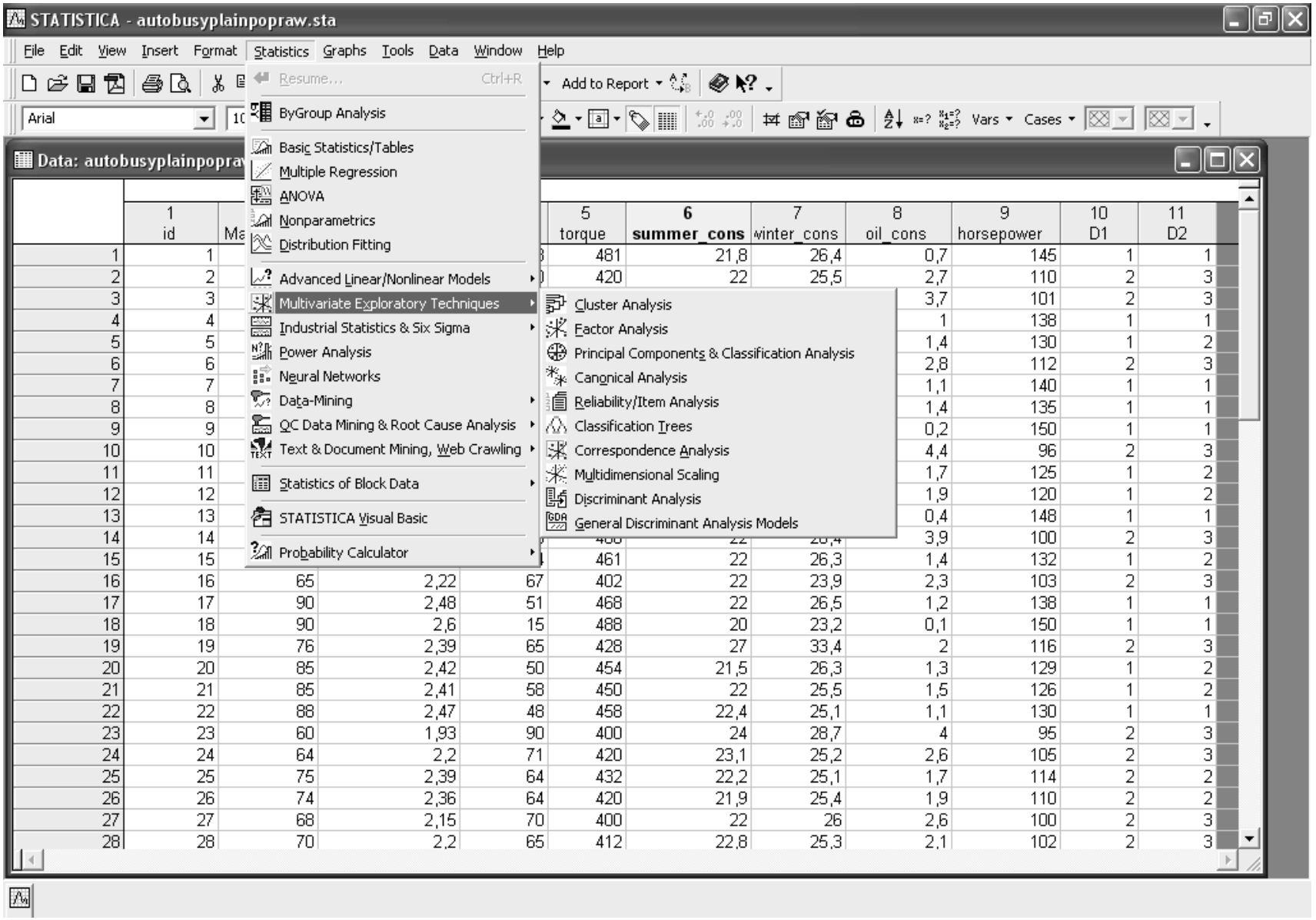
Inne zagadnienia

- Pojęcie zmiennych kanonicznych – kierunki które dobrze separują k klasy (także ich wizualizacja)
- Dyskryminacja oparta na regresji liniowej i logistycznej
- Uogólnienie modeli liniowych – elastyczna dyskryminacja (FDA)
- Ad a metoda wektorów nośnych (SVM)
- Powiązanie z metodą PCA
- Odniesienia do Analizy Korespondencji

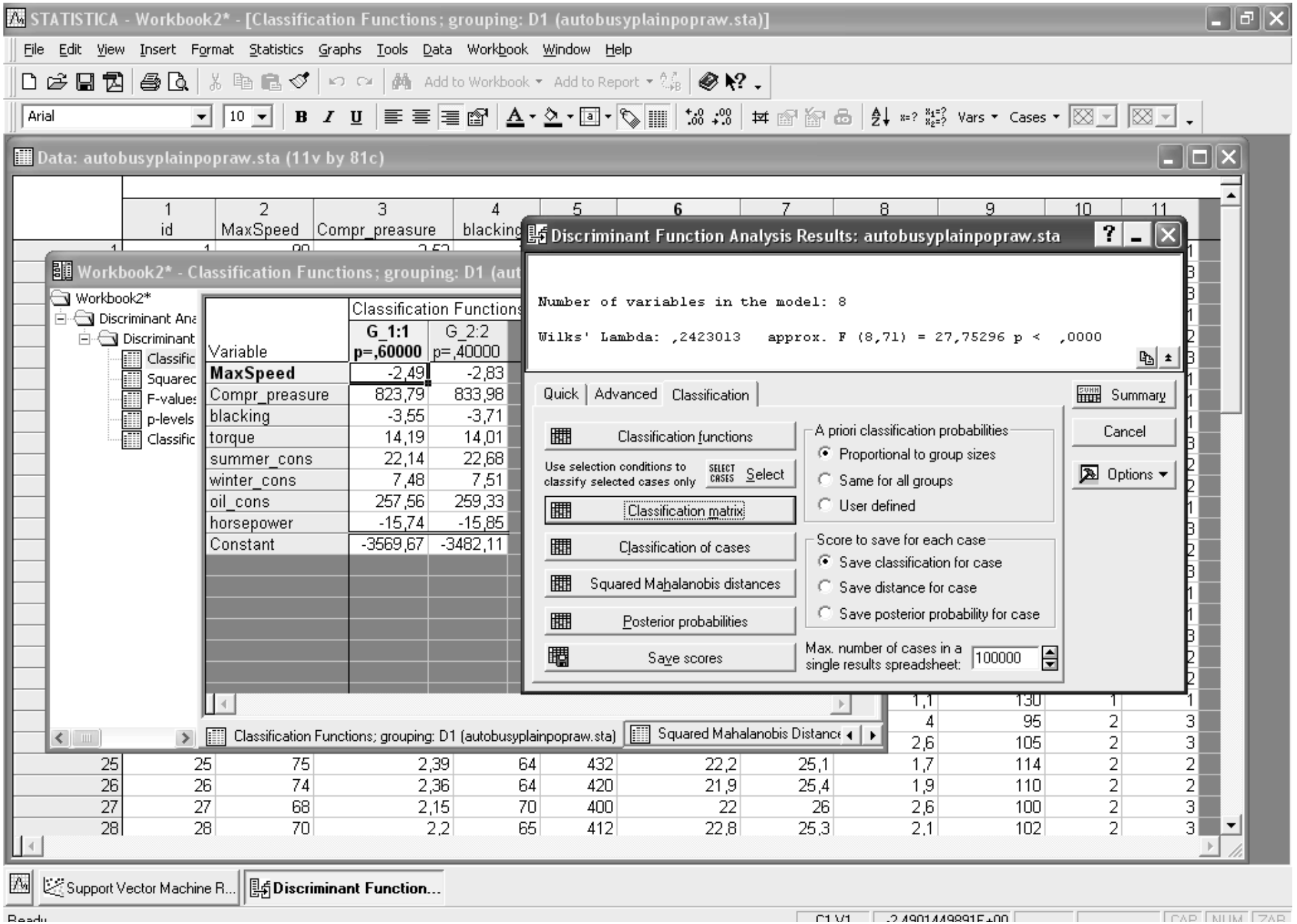
Typowe obszary zastosowań

- Analiza danych finansowych (zwłaszcza banki, polityka kredytowa, predykcja bankructw)
- Badania marketingowe
  - Także identyfikacja czynników różnicujących klasy klientów
- Badania danych medycznych, biologicznych lub innych powiązanych nauk
- Rozpoznawania twarzy na obrazach

Implementacje np. Statistica



	1 id	Ma				5 torque	6 summer_cons	7 winter_cons	8 oil_cons	9 horsepower	10 D1	11 D2
1	1					481	21,8	26,4	0,7	145	1	1
2	2					420	22	25,5	2,7	110	2	3
3	3								3,7	101	2	3
4	4								1	138	1	1
5	5								1,4	130	1	2
6	6								2,8	112	2	3
7	7								1,1	140	1	1
8	8								1,4	135	1	1
9	9								0,2	150	1	1
10	10								4,4	96	2	3
11	11								1,7	125	1	2
12	12								1,9	120	1	2
13	13								0,4	148	1	1
14	14								3,9	100	2	3
15	15					461	22	26,3	1,4	132	1	2
16	16	65	2,22	67		402	22	23,9	2,3	103	2	3
17	17	90	2,48	51		468	22	26,5	1,2	138	1	1
18	18	90	2,6	15		488	20	23,2	0,1	150	1	1
19	19	76	2,39	65		428	27	33,4	2	116	2	3
20	20	85	2,42	50		454	21,5	26,3	1,3	129	1	2
21	21	85	2,41	58		450	22	25,5	1,5	126	1	2
22	22	88	2,47	48		458	22,4	25,1	1,1	130	1	1
23	23	60	1,93	90		400	24	28,7	4	95	2	3
24	24	64	2,2	71		420	23,1	25,2	2,6	105	2	3
25	25	75	2,39	64		432	22,2	25,1	1,7	114	2	2
26	26	74	2,36	64		420	21,9	25,4	1,9	110	2	2
27	27	68	2,15	70		400	22	26	2,6	100	2	3
28	28	70	2,2	65		412	22,8	25,3	2,1	102	2	3


STATISTICA - Workbook2* - [Classification Functions; grouping: D1 (autobusyplainpopraw.sta)]
File Edit View Insert Format Statistics Graphs Tools Data Workbook Window Help
Add to Workbook
Add to Report
Arial
10
Vars
Cases
Data: autobusyplainpopraw.sta (11v by 81c)
1 id
2 MaxSpeed
3 Compr_preasure
4 blacking
5
6
7
8
9
10
11
Workbook2* - Classification Functions; grouping: D1 (aut
Workbook2*
Discriminant Ana
Discriminant
Classific
Squarec
F-values
p-levels
Classific
Classification Functions
Variable
G_1:1 p=,60000
G_2:2 p=,40000
MaxSpeed -2,49 -2,83
Compr_preasure 823,79 833,98
blacking -3,55 -3,71
torque 14,19 14,01
summer_cons 22,14 22,68
winter_cons 7,48 7,51
oil_cons 257,56 259,33
horsepower -15,74 -15,85
Constant -3569,67 -3482,11
Classification Functions; grouping: D1 (autobusyplainpopraw.sta)
Squared Mahalanobis Distance
Discriminant Function Analysis Results: autobusyplainpopraw.sta
Number of variables in the model: 8
Wilks' Lambda: ,2423013 approx. F (8,71) = 27,75296 p < ,0000
Quick
Advanced
Classification
Summary
Cancel
Options
Classification functions
Use selection conditions to classify selected cases only
Select
Classification matrix
Classification of cases
Squared Mahalanobis distances
Posterior probabilities
Save scores
A priori classification probabilities
Proportional to group sizes
Same for all groups
User defined
Score to save for each case
Save classification for case
Save distance for case
Save posterior probability for case
Max. number of cases in a single results spreadsheet:
100000
25 25 75 2,39 64 432 22,2 25,1 1,7 114 2 2
26 26 74 2,36 64 420 21,9 25,4 1,9 110 2 2
27 27 68 2,15 70 400 22 26 2,6 100 2 3
28 28 70 2,2 65 412 22,8 25,3 2,1 102 2 3
Support Vector Machine R...
Discriminant Function...
Ready

Więcej

Przeczytaj literaturę

- T.Hastie, R.Tibshirani, J.Friedman: The Elements of Statistical Learning. Springer (zwłaszcza rozdz. 4) $\rightarrow$ poszukaj wersji elektronicznej pdf
- J.Koronacki, J.Ćwik: Statystyczne systemy uczące się (rozdz. 1 oraz o FDA w rozdz. 6)
- M.Krzyśko, W.Wołyński, T.Górecki,M.Skorzybut: Systemy uczące się. + wcześniejsze prace M.Krzyśko o analizie dyskryminacyjnej
- Angielska Wikipedia „Linear discriminant analysis”
- McLachlan, G. J. (2004). Discriminant Analysis and Statistical Pattern Recognition. Wiley.
- Duda, R. O.; Hart, P. E.; Stork, D. H. (2000). Pattern Classification (2nd ed.). Wiley