Analiza dyskryminacyjna
Wybrane podejścia i zastosowania

Katarzyna Stąpor
Politechnika Śląska, Instytut Informatyki
Akademicka 16, 44-100 Gliwice
email Katarzyna.Stapor@polsl.pl

PLAN

Zadanie klasyfikacji/dyskryminacji - podstawowe pojęcia

Liniowa analiza dyskryminacyjna w ujęciu bayesowskim – klasyfikator Bayesa i jego optymalność

Liniowa dyskryminacja Fishera

Zastosowanie ekonomiczne – przewidywanie zdolności kredytowej

Zastosowanie biologiczne – przewidywanie typu oddziaływań białek

Zadanie klasyfikacji

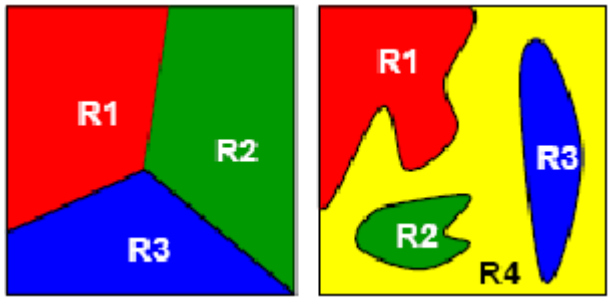
Rozpoznawanie = klasyfikacja = analiza dyskryminacyjna

Klasyfikacja/analiza dyskryminacyjna - problem predykcji etykiety klasy na podstawie obserwacji (wektora cech)

Klasyfikator Ψ – funkcja odwzorowująca przestrzeń E reprezentacji obiektów w zbiór decyzji (etykiet klasowych)

$$\Psi: \quad E \rightarrow I = \{1, 2, ..., c\}$$

$\Psi(x)$ prognoza etykiety klasowej wektora x



przestrzeń reprezentacji
z **obszarami decyzyjnymi**

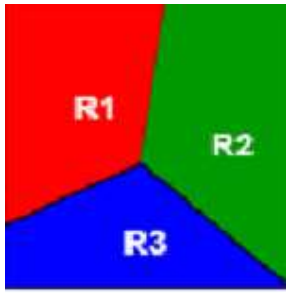
Wyróżniamy dwa etapy:

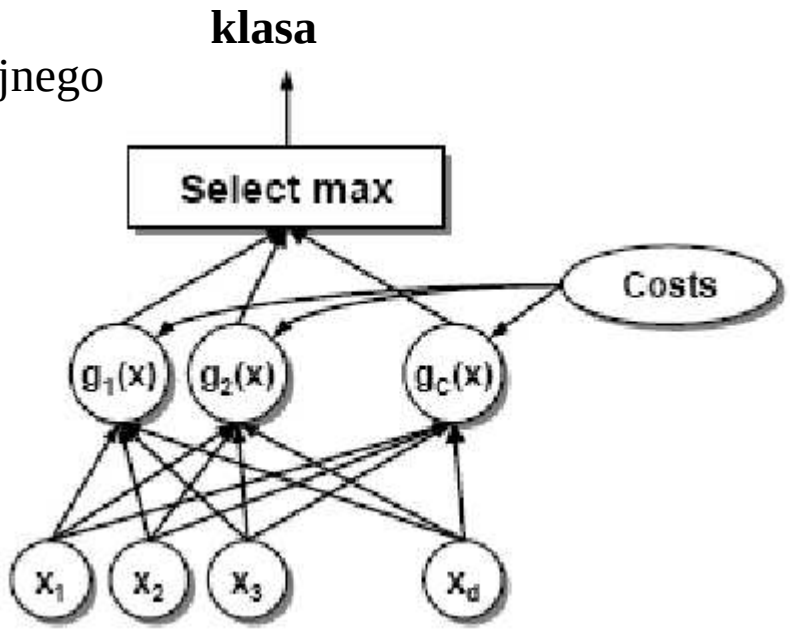
- **etap uczenia / budowy klasyfikatora** – w którym znajdujemy reguły klasyfikacyjne w oparciu o tak zwany **zbiór uczący**
- **etap klasyfikacji / wykorzystania modelu** – w którym dokonujemy klasyfikacji zasadniczego zbioru obiektów, których przynależność jest nam nieznana, w oparciu o znalezione charakterystyki klas

Funkcje dyskryminacyjne klasyfikatora

Funkcja dyskryminacyjna g_k k-tej klasy
„funkcja charakterystyczna” k-tego obszaru decyzyjnego

$$\forall_{\substack{x \in O_k \\ j=1,\ldots,c,\ j \neq k}} \quad g_k(x) > g_j(x)$$





Klasyfikator Ψ

$$\Psi(x) = k \quad \text{jeżeli} \quad \forall_{\substack{j=1,\ldots,c \\ j \neq k}} g_k(x) > g_j(x)$$

Klasyfikator – funkcja/reguła decyzyjna

Podejście bayesowskie

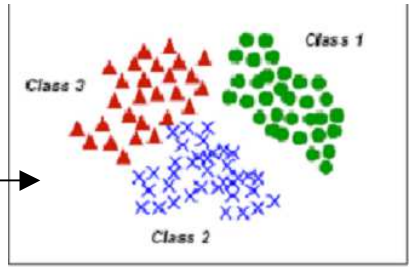
Model probabilistyczny zadania klasyfikacji

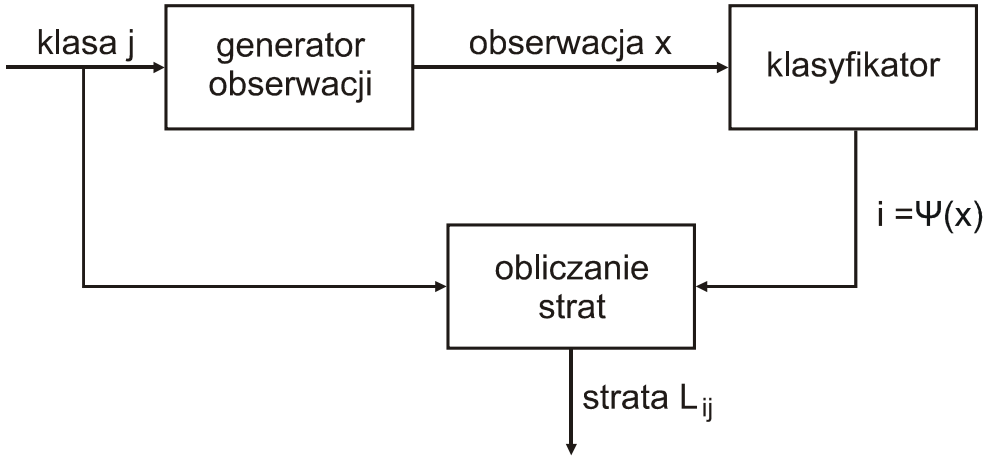
Zbiór obiektów dzieli się na klasy. Wartości cech obiektów - realizacje wielowymiarowej **ciągłej zmiennej losowej *X*,** ich klasy – realizacje **dyskretnej zmiennej losowej *Y***.
Wartości numerów klas – losowane zgodnie z **prawdopodobieństwami a priori klas**:

$$P(j) \quad j \in I = \{1, \ldots, c\}$$

Rozkład cech w klasie *j* opisane **gęstością warunkową** klasy *j:*

$$f(x \mid j) = f_j(x) \quad x \in R^d, \quad j \in I$$





Klasyfikator przyporządkowuje obserwacji *x* numer klasy *i*.
Pociąga to za sobą ewentualną **stratę** polegającą na zaliczeniu obiektu do klasy *i*, gdy jego prawdziwą klasą jest klasa *j*

$$0 \le L_{ij} < \infty$$

Miarą jakości klasyfikatora Ψ jest **rzeczywisty poziom błędu**:

$$e(\Psi) = P(\Psi(X) \neq Y)$$

Klasyfikator Bayesa (bayesowski)

Przypadek 2 klas $(X,Y) \sim (\mu, r) \qquad R^d \times I = \{0,1\}$

r - funkcja regresji, prawdopodobieństwo a posteriori:

$$\forall_{x \in D_X \subset R^d} \quad r(x) = E(Y \mid X = x) = 1 \cdot P(Y = 1 \mid X = x) + 0 \cdot P(Y = 0 \mid X = x) =$$

$$= P(Y = 1 \mid X = x)$$

Z twierdzenia Bayesa mamy:

$$r(x) = P(Y = 1 \mid X = x) = \frac{f(x \mid Y = 1)P(Y = 1)}{f(x \mid Y = 1)P(Y = 1) + f(x \mid Y = 0)P(Y = 0)} =$$

$$= \frac{f_1(x)P(1)}{f_1(x)P(1) + f_0(x)P(0)}$$

Klasyfikator Bayesa:

$$\Psi_B(x) = \begin{cases} 1 & r(x) > \frac{1}{2} \\ 0 & \textit{poza tym} \end{cases} \quad \Leftrightarrow \quad \Psi_B(x) = \begin{cases} 1 & P(1) \cdot f_1(x) > P(0) \cdot f_0(x) \\ 0 & \textit{poza tym} \end{cases}$$

Przypisuje obiekt *x* do klasy najbardziej prawdopodobnej

Klasyfikator Bayesa (bayesowski)

$$\Psi_B(x) = \begin{cases} 1 & P(1)\cdot f_1(x) > P(0)\cdot f_0(x) \\ 0 & \textit{poza tym} \end{cases}$$

Przypadek wielu klas $I = \{1, 2, ..., c\}$

$$\Psi_B(x) = \arg\max_k P(Y = k \mid X = x) = \arg\max_k P(k)\, f_k(x)$$

$g_i(x) = P(i)\cdot f_i(x)$ funkcja dyskryminacyjna i-tej klasy

Optymalność klasyfikatora Bayesa

Twierdzenie T1

Klasyfikator bayesowski jest **optymalny**, tj. jeżeli Ψ jest jakimkolwiek innym klasyfikatorem, to:

$$e(\Psi_B) \leq e(\Psi)$$

lub równoważnie

$$P(\Psi_B(x) = Y) \geq P(\Psi(x) = Y)$$

$e(\Psi) = P(\Psi(X) \neq Y)$ rzeczywisty poziomem błędu klasyfikatora

Klasyfikator Bayesa dla klas gaussowskich

Rozkłady cech normalne:

$$f(x \mid i) = f_i(x) = \frac{1}{(2\pi)^{d/2} \mid \Sigma_i \mid^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i) \right]$$

Σ macierz kowariancji μ wektor wartości średnich

$g_i(x) = \ln f_i(x) + \ln P(i)$ **funkcja dyskryminacyjna i-tej klasy**

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i) - \frac{1}{2}\ln \mid \Sigma_i \mid + \ln P(i)$$

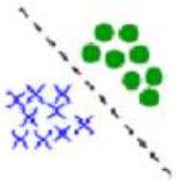
Macierze kowariancji w poszczególnych klasach są identyczne:

$$\Sigma_i = \Sigma \quad i = 1, \ldots, c$$

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \Sigma^{-1} (x-\mu_i) + \ln P(i)$$

Uproszczenie poprzez eliminację czynnika $x^T \Sigma^{-1} x$ stałego dla każdej z klas:

$$g_i(x) = x^T \Sigma^{-1} \mu_i - \frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i + \ln P(i)$$



Funkcje dyskryminacyjne - liniowe, powierzchnie rozdzielające obszary decyzyjne - hiperpłaszczyzny

Empiryczny klasyfikator gaussowski

μ_i, Σ_i w rzeczywistych warunkach najczęściej nieznane, zastępujemy je **estymatorami.** **Bayes „plug-in” = empiryczny klasyfikator Bayesa**

Estymatory MNW:

Średnia próbkowa

$$\hat{\mu}_i \equiv \overline{x}_i = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} x_{i,j}$$

$x_{i,j}$ j-ty wektor/obserwacja z i-tej klasy

N_i liczba obserwacji w i-tej klasie

Próbkowa macierz kowariancji

$$\hat{\Sigma}_i \equiv S_i = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (x_{i,j} - \overline{x}_i)(x_{i,j} - \overline{x}_i)^T$$

Estymacja parametrów = **uczenie klasyfikatora** gaussowskiego na podstawie zbioru uczącego (próby losowej)

Empiryczny klasyfikator gaussowski

Estymacja parametrów = **uczenie klasyfikatora** gaussowskiego na podstawie zbioru uczącego (próby losowej)

estymatory

$$\hat{g}_i(x) = x^T \hat{\Sigma}^{-1} \hat{\mu}_i - \frac{1}{2} \hat{\mu}_i^T \hat{\Sigma}^{-1} \hat{\mu}_i + \ln \hat{P}(i)$$

Funkcja dyskryminacyjna i-tej klasy empirycznego klasyfikatora gaussowskiego

Podejście fisherowskie

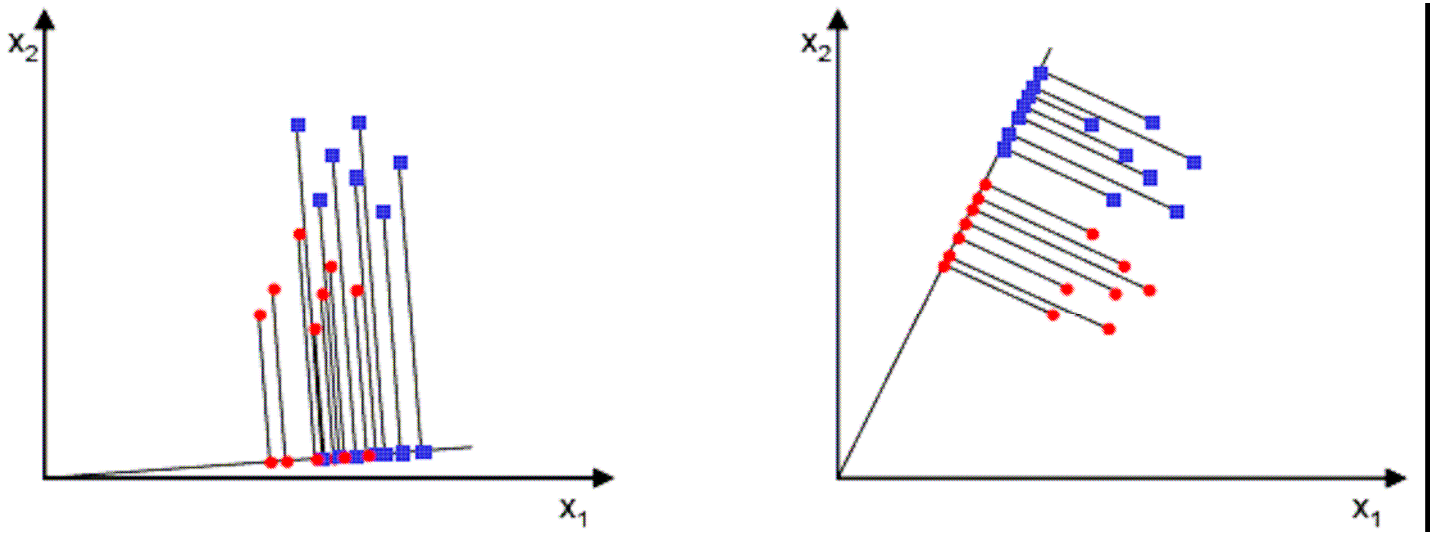
Fisherowska analiza dyskryminacyjna
Fisher's Linear Discriminant Analysis (LDA)

Cel LDA – redukcja wymiarowości przy zachowaniu mocy dyskryminacyjnej

Zbiór uczący (2-klasowy) w *D*-wymiarowej przestrzeni – celem jest znalezienie kierunku *a*, tak by rzuty elementów z obu na ten kierunek:

$$y = a^T x$$

były maksymalnie rozdzielone (d=1)



Ilustracja idei LDA dla 2-wymiarów (D=2), 2 klas

LDA dla wielu klas

Mamy: $X = (X_1, ..., X_d)^T$ d-wymiarowa zmienna losowa

$G_1, ..., G_c$ realizacje pochodzą z ***c*** klas

$\mu_1, ..., \mu_c$ $\Sigma_1, ..., \Sigma_c$ Środki i macierze kowariancji w klasach

ZAŁOŻENIE: macierze kowariancji równe i pełnego rzędu

$$\Sigma_1 = \Sigma_2 = ... = \Sigma_c = \Sigma$$

Średnia ogólna $$\bar{\mu} = \frac{1}{c}\sum_{i=1}^{c} \mu_i$$

Międzygrupowa macierz kowariancji:

$$B = \sum_{i=1}^{c} (\mu_i - \bar{\mu})(\mu_i - \bar{\mu})^T$$

Kombinacja liniowa $Y = a^T X$
(**rzut na kierunek a**)

LDAdla wielu klas

$$Y = a^T X$$

$E(Y) = a^T E(X \mid g_i) = a^T \mu_i = \mu_{iY}$ Wartość oczekiwana *i*-tej klasy

$Var(Y) = a^T Cov(X) a = a^T \Sigma a = \sigma_Y^2$ **Wariancja wewnątrzgrupowa**

$\bar{\mu}_Y = \frac{1}{c}\sum_{i=1}^{c} \mu_{iY} = \frac{1}{c}\sum_{i=1}^{c} a^T \mu_i = a^T (\frac{1}{c}\sum_{i=1}^{c} \mu_i) = a^T \bar{\mu}$ Średnia ogólna

Iloraz: (**wariancja międzygrupowa (suma kwadratów odległości środków klas od średniej ogólnej)) / (wariancja Y):**

$$\frac{\sum_{i=1}^{c} (\mu_{iY} - \bar{\mu}_Y)^2}{\sigma_Y^2} = \frac{\sum_{i=1}^{c} \left(a^T \mu_i - a^T \bar{\mu}\right)^2}{a^T \Sigma a} =$$

Mierzy zmienność międzygrupową w relacji do wewnątrzgrupowej

$$= \frac{a^T \left(\sum_{i=1}^{c} (\mu_i - \bar{\mu})(\mu_i - \bar{\mu})^T \right) a}{a^T \Sigma a} = \frac{a^T B a}{a^T \Sigma a}$$

Znajdź takie ***a***, które maksymalizuje iloraz

Próbkowa wersja LDA

Σ, μ_i, B nieznane ---- konieczna estymacja ilorazu !

Próbkowa międzygrupowa macierz kowariancji:

$$S_b = \sum_{i=1}^{c} n_i (\overline{x}_i - \overline{x})(\overline{x}_i - \overline{x})^T$$

$$\overline{x} = \frac{\sum_{i=1}^{c} n_i \overline{x}_i}{\sum_{i=1}^{c} n_i} = \frac{\sum_{i=1}^{c} \sum_{j=1}^{n_i} x_{ij}}{\sum_{i=1}^{c} n_i}$$

Średnia próbkowa

$$\overline{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

Średni wektor w i-tej klasie

n_i Liczba elementów w i-tej populacji

x_{ij} j-ty element w *i*-tej populacji

Próbkowa wersja LDA

Σ, μ_i, B nieznane ---- konieczna estymacja ilorazu !

Próbkowa wewnątrzgrupowa macierz kowariancji

$$S_w = \sum_{i=1}^{c} (n_i - 1) S_i = \sum_{i=1}^{c} \sum_{j=1}^{n_i} \left(x_{ij} - \overline{x}_i \right) \left(x_{ij} - \overline{x}_i \right)^T$$

$$\frac{a^T B_\mu a}{a^T \Sigma a}$$ próbkowy odpowiednik: $$\frac{a^T S_b a}{a^T S_w a}$$

Liniowa Analiza Dyskryminacyjna (LDA) Fishera

Twierdzenie 1

Niech $\lambda_1 > \ldots > \lambda_s > 0^{(*)}$ niezerowych **wartości własnych** $S_w^{-1} S_b$

$v_1 > \ldots > v_s$ odpowiadające **wektory własne** wyskalowane tak, by: $v^T S_w v = 1$

Wektor $\boldsymbol{a}$ maksymalizujący iloraz:

$$\frac{a^T S_b a}{a^T S_w a} = \frac{a^T \left[\sum_{i=1}^{c} n_i (\overline{x}_i - \overline{x})(\overline{x}_i - \overline{x})^T \right] a}{a^T \left[\sum_{i=1}^{c} \sum_{j=1}^{n_i} \left(x_{ij} - \overline{x}_i \right) \left(x_{ij} - \overline{x}_i \right)^T \right] a} \qquad (1)$$

jest dany jako $a_1 = v_1$, kontynuując, $a_k = v_k$ przy warunku $Cov(v_k^T X, v_i^T X) = 0 \quad i < k$ maksymalizuje iloraz (1).

$v_k X$ to ***k*-ta zmienna dyskryminacyjna**

(*) można pokazać z prawd. 1: $s \leq \min(c-1,\ d)$

Fisherowska analiza dyskryminacyjna

Umożliwia redukcję wymiarowości !!!

Można pokazać, że z prawdopodobieństwem 1:

$$s \leq \min (c-1,\ d)$$

d - wymiarowość wejściowej przestrzeni
c - liczba klas

Prawdziwa wartość s – liczba efektywnych kierunków jest trudna do zdefiniowania

Testy statystyczne (**Wilk's lambda**) dla testowania
istotności zmiennych dyskryminacyjnych !!!

Klasyfikator Fishera

Nowa obserwacja ***x*** przypisywana jest do klasy G_k jeśli

$$D_k(x) = \min_{j=1,\ldots,c} \quad D_j(x)$$

gdzie

$$D_j^2(x) = [V^T(x-\mu_j)]^T[V^T(x-\mu_j)] - 2\log\pi_j = \sum_{i=1}^{s}(y_i - \mu_{jY_i})^2 - 2\log\pi_j$$

$$V = (v_1,\ldots,v_s) \quad d \times s$$

μ_{jY_i} i-ta składowa środka j-tej grupy w nowej przestrzeni dyskryminacyjnej

odległość euklidesowa od środka j-tej grupy w **nowej przestrzeni dyskryminacyjnej**

$\mathrm{cov}(V^T X) = I$ kierunki wyznaczające nową przestrzeń dyskryminacyjną są **nieskorelowane**

Podejście Fukunagi

Rozwiązanie problemu optymalizacyjnego:

$$J_{Fuk}(A) = tr((AS_w A^T)^{-1}(AS_b A^T))$$

przy warunku $A^T A = I$

A – macierz przekształcenia
równoważne jest znalezieniu takich *a*, które spełniają

$$S_b a = \gamma S_w a \quad \gamma \neq 0$$

Uogólnione zagadnienie własne

K. Fukunaga.(1990). *Introduction to statistical pattern recognition*. New York, Academic Press.

ZASTOSOWANIA

Ocena zdolności kredytowej przedsiębiorstw

- **metody klasyczne:** opisowe, punktowe, mieszane
- **metody nieklasyczne:** proste i **złożone (metoda dyskryminacyjna)**

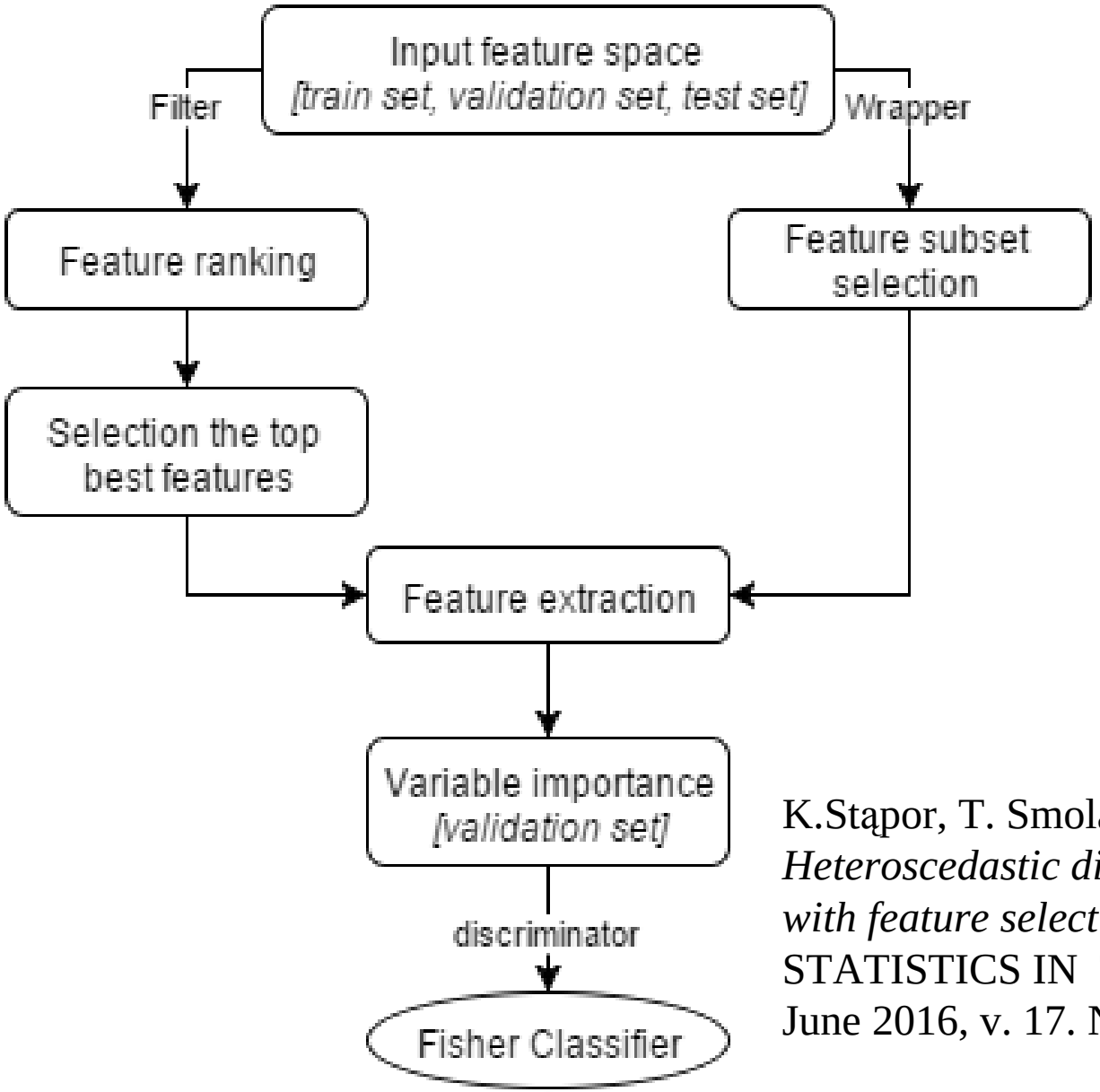
Parametry metod złożonych mogą być określane drogą analizy systematycznej
bądź też zaproponowane przez ekspertów zajmujących się analizą ryzyka kredytowego.

HETEROSCEDASTIC DISCRIMINANT ANALYSIS COMBINED WITH FEATURE SELECTION FOR CREDIT SCORING

K. Stąpor, T. Smolarczyk, P. Fabian

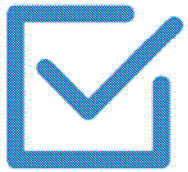
STATISTICS IN TRANSITION new series, June 2016
Vol. 17, No. 2, pp. 1–16

Our CS model architecture



K.Stąpor, T. Smolarczyk, P. Fabian: *Heteroscedastic discriminant analysis combined with feature selection for credit scoring*. STATISTICS IN TRANSITION New Series, June 2016, v. 17. Nr 2, pp. 1-16.

Credit Scoring model overview



Feature selection

- SFFS
- CFS
- Fisher Score
- mRMR
- GRASP
- Memetic
- Tabu

Feature extraction

- FDA
- Heteroscedastic extension of FDA (FDA_Cher)

Classification

- Fisher classifier – nearest mean in new discriminant space

Problem of discrimination between good and bad clients

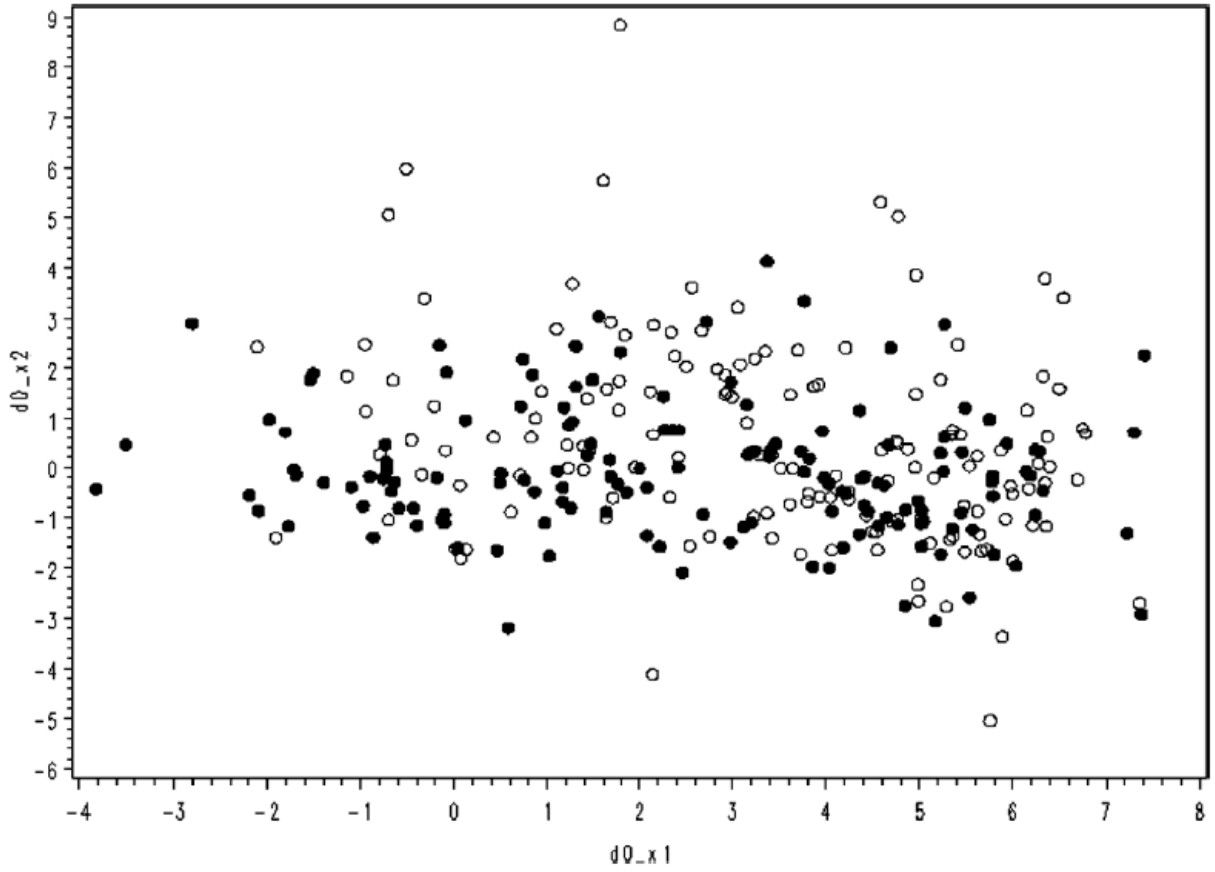


Fig. 1. Illustration of the poor separability of the credit data. Partial least squares was used to transform data for 100 good cases (black) and 100 bad cases (white) selected randomly from the data into two factors given as the *x* and *y* axis of the graph.

German credit dataset

Attribute	Description	Values
1.	Status of existing checking account (qualitative)	A11 : ... < 0 DM A12 : 0 <= ... < 200 DM A13 : ... >= 200 DM /salary assignments for at least 1 year A14 : no checking account
2.	Duration in month (numerical)	
3.	Credit history (qualitative)	A30 : no credits granted/all credits paid back duly A31 : all credits at this bank paid back duly A32 : existing credits paid back duly until now A33 : delay in paying off in the past A34 : critical account/other credits existing (not at this bank)
4.	Purpose (qualitative)	A40 : car (new) A41 : car (used) A42 : furniture/equipment A43 : radio/television A44 : domestic appliances A45 : repairs A46 : education A47 : (vacation - does not exist?) A48 : retraining A49 : business A410 : others

German credit dataset

5.	Credit amount (numerical)	
6.	Savings account/bonds (qualitative)	A61 : ... < 100 DM A62 : 100 <= ... < 500 DM A63 : 500 <= ... < 1000 DM A64 : .. >= 1000 DM A65 : unknown/ no savings account
7.	Present employment since (qualitative)	A71 : unemployed A72 : ... < 1 year A73 : 1 <= ... < 4 years A74 : 4 <= ... < 7 years A75 : .. >= 7 years
8.	Instalment rate in percentage of disposable income (numerical)	
9.	Personal status and sex (qualitative)	A91 : male : divorced/separated A92 : female : divorced/separated/married A93 : male : single A94 : male : married/widowed A95 : female : single
10.	Other debtors / guarantors (qualitative)	A101 : none A102 : co-applicant A103 : guarantor
11.	Present residence since (numerical)	

UCI Machine Learning Repository

Nominal features were replaced by a number of binary features representing every possibility.

12.	Property (qualitative)	A121 : real estate A122 : if not A121 : building society savings agreement/life insurance A123 : if not A121/A122 : car or other, not in attribute 6 A124 : unknown / no property
13.	Age in years (numerical)	
14.	Other instalment plans (qualitative)	A141 : bank A142 : stores A143 : none
15.	Housing (qualitative)	A151 : rent A152 : own A153 : for free
16.	Number of existing credits at this bank (numerical)	
17.	Job (qualitative)	A171 : unemployed/ unskilled - non-resident A172 : unskilled - resident A173 : skilled employee / official A174 : management/ self-employed/highly qualified employee/ officer
18.	Number of people being liable to provide maintenance (numerical)	
19.	Telephone (qualitative)	A191 : none A192 : yes, registered under the customer's name
20.	Foreign worker (qualitative)	A201 : yes A202 : no

German credit dataset

UCI Machine Learning Repository

Heteroscedastic Discriminant Analysis

The concept of Directed Distance Matrices (DDM)

If there is discriminatory information present because of the heteroscedasticity of the data, then this should become apparent in the DDM !!!

This extra distance because of the heteroscedasticity, is, in general, in different directions than the eigenvector v, which separates the means and so DDM should have more than one nonzero eigenvalues !!!

The specific DDM is based on the Chernoff distance between two normally distributed densities:

$$\text{dist} - \text{Cher}(d_1, d_2) = -\log \int d_1^{\alpha}(x) d_2^{1-\alpha}(x) dx \qquad \alpha \in (0,1)$$

$$S_C = S^{-\frac{1}{2}}(m_1 - m_2)(m_1 - m_2)^T S^{-\frac{1}{2}} + \frac{1}{p_1 p_2}(\log S - p_1 \log S_1 - p_2 \log S_2) \qquad \text{DDM}$$

$\text{Trace}(S_C) = \text{dist} - \text{Cher}$

S_B is replaced by S_C in optimization process

Loog, Duin, *Non-iterative heteroscedastic linear dimension reduction for two-class data: from Fisher to Chernoff.* Proc. 4th Int. Workshop S+SSPR, 508-517

Chernoff criterion

Heteroscedastic Discriminant Analysis

$$J_C(A) = tr\left((AS_W A^T)^{-1} A(m_1 - m_2)(m_1 - m_2)^T A^T - AS_W^{\frac{1}{2}} \frac{p_1 \log\left(S_W^{-\frac{1}{2}} S_1 S_W^{-\frac{1}{2}}\right) + p_2 \log\left(S_W^{-\frac{1}{2}} S_2 S_W^{-\frac{1}{2}}\right)}{p_1 p_2} S_W^{\frac{1}{2}} A^T \right)$$

This is maximized by determining an eigenvalue decomposition of:

$$S_W^{-1}\left(S_B - S_W^{\frac{1}{2}} \frac{p_1 \log\left(S_W^{-\frac{1}{2}} S_1 S_W^{-\frac{1}{2}}\right) + p_2 \log\left(S_W^{-\frac{1}{2}} S_2 S_W^{-\frac{1}{2}}\right)}{p_1 p_2} S_W^{\frac{1}{2}} \right)$$

and taking the rows of the transform L to equal d eigenvectors corresponding to the d largest eigenvalues

Feature selection

Fisher Score

- Fisher Score algorithm is designed to find subset of features that will maximize the distance between instances from different classes and, at the same time, minimize the distances within the same class

- n_i is the number of instances of class
- μ_i and σ_i is the mean and variance of class *i*, corresponding to the *r*-th feature
- μ and σ are the mean and variance of the whole dataset respectively

$$F_r = \frac{\sum_{i=1}^{c} n_i (\mu_i - \mu)^2}{\sum_{i=1}^{c} n_i \sigma_i^2}$$

Attribute importance analysis
Wrapper feature selection frequency

- Attribute 1 - status of existing checking account (as in filter)
- Attribute 2 - duration in month
- Attribute 3 - credit history
- Attribute 5 - savings account / bond
- Attribute 22 – job (unemployed/ unskilled)

Algorithm \ Attribute	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Average FDA	91	57	50	27	51	36	23	19	24	30	17	9	13	13	0	1	0	6	3	6	11	1	9	11
Average FDA_Cher	81	54	69	51	66	56	47	50	57	60	59	51	43	51	57	57	56	44	61	56	64	37	41	44
Average All	86	56	59	39	59	46	35	34	41	45	38	30	28	32	29	29	28	25	32	31	38	19	25	28

Experimental Results

Algorithm \ dataset	FDA			FDA_Cher				FDA_Cher (1 direction)		
	Accuracy rate (%)		Number of selected features	Accuracy rate (%)		Number of selected features	Number of directions[1]	Accuracy rate (%)		Number of selected features
	Avg.	Std.	Median	Avg.	Std.	Median		Avg.	Std.	Median
All features	30.20%	0.42%	24	70.00%	3.16%	24	3	69.40%	3.63%	24
CFS	62.30%	4.85%	3	74.00%	4.00%	22	5	73.40%	4.27%	20
Fisher Score	66.10%	2.33%	5	76.30%	3.56%	18	3	74.90%	2.42%	17
MRMR	62.50%	5.60%	2	74.80%	4.34%	23	2	72.60%	5.87%	16
SFFS	62.30%	4.32%	3	69.00%	4.85%	4	3	67.60%	7.11%	2
GRASP	59.50%	2.95%	4	69.20%	5.63%	9	2	68.90%	5.38%	10
Memetic	61.10%	3.25%	7	69.40%	5.82%	11	3	67.40%	5.62%	11
Tabu	63.40%	2.50%	10	70.60%	5.58%	14	3	69.40%	4.50%	15

[1] The best number of directions selected from tests on ⟨1,5⟩

CISIM 2016 16th International Conference on Computer Information Systems and Industrial Management Applications

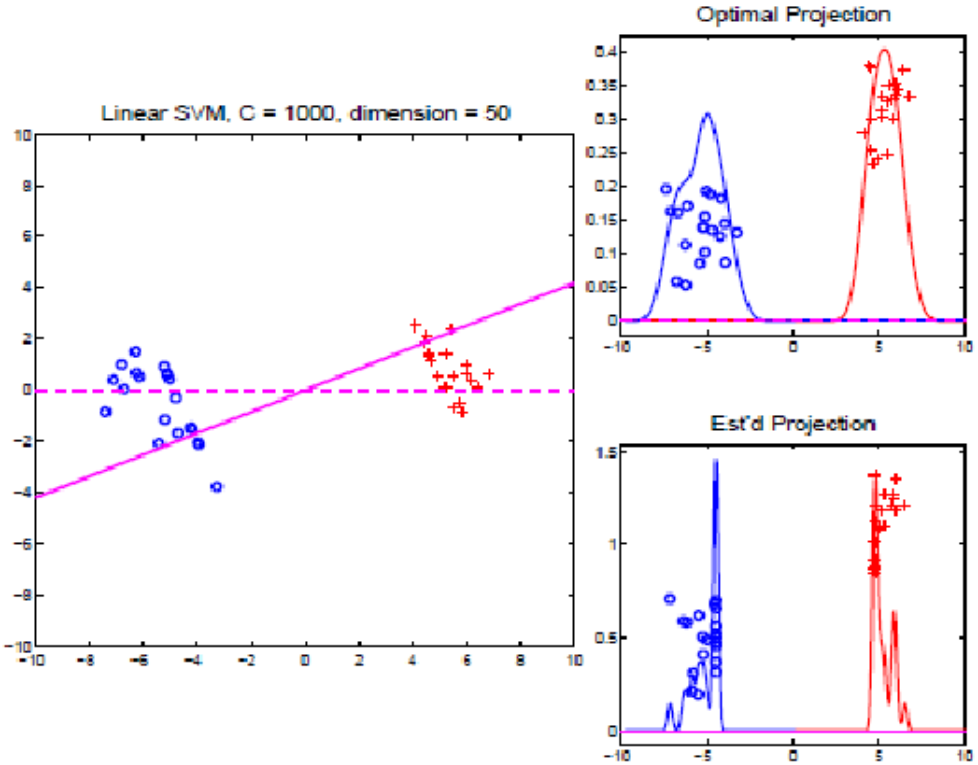
Classification of protein interactions based on sparse discriminant analysis and energetic features

Katarzyna Stąpor and Piotr Fabian
Silesian Technical University, Faculty of Computer Science, Gliwice, Poland

LDA in HDLSS - problems

When **the number of variables exeeds the sample size** (i.e. in **HDLSS** settings), the within-class covariance matrix is **singular** and the classical LDA breaks down.

Different methods have been proposed but they suffer from **data piling problem**
Sparsification is needed !!!!



Toy example illustrating **data piling for SVM**
SVM direction – solid line
Optimal direction – dashed line
In many applications of SVM in HDLSS large portion of data – support vectors !
Projecting them on the normal SVM hyper plane – **many of the projections are identical –** which we call **data pilling**
It indicates **influence of noise** and causes **overfitting** – **high out-of-sample clasification error !**

-

Regularized sparse LDA (rSLDA) – link of generalized eigenvalue problem to regression

Theorem

S_w - positive definite matrix, $S_w = R_w^T R_w$ its Cholesky decomposition.

H_b $k \times p$ matrix,

$V_1,\ldots,V_q$ - eigenvectors of $S_W^{-1} S_B$ corresponding to the q largest eigenvalues $\lambda_1 \geq \ldots \geq \lambda_q$ $(q \leq \min(p, k-1))$

$A = [\alpha_1,\ldots,\alpha_q]$, $B = [\beta_1,\ldots,\beta_q]$

solution to the following problem:

$$S_b \beta = \mu S_w \beta$$

$$\beta_j = V_j$$

$$\min_{A,B} \sum_{i=1}^{k} \left\| R_w^{-T} H_{b,i} - A B^T H_{b,i} \right\|^2 + \lambda \sum_{j=1}^{q} \beta_j^T (S_w) \beta_j$$

subject to $A^T A = I_{p \times q}$, $\lambda > 0$

where:

$H_{b,i} = \sqrt{n_i}(\bar{x}_i - \bar{x})^T$ i-th row of the matrix:

$$H_b = \left(\sqrt{n_1}(\bar{x}_1 - \bar{x}),\ldots,\sqrt{n_k}(\bar{x}_k - \bar{x}) \right)^T,$$

e^{n_i} vector of ones with length n_i,

It first relates the discriminant vector to a regression coefficient vector by **transforming the generalized eigenvalue problem to a regression type problem**

Then $\hat{\beta}_j$, $j = 1,\ldots,q$, span the same linear space as V_j, $j = 1,\ldots,q$.

The Lasso* -- sparsity and penalized regression

Lasso for linear models

$$\hat{\beta}(\lambda) = \text{argmin}_{\beta}(n^{-1}\|\mathbf{Y} - \mathbf{X}\beta\|^2 + \underbrace{\lambda}_{\geq 0} \underbrace{\|\beta\|_1}_{\sum_{j=1}^{p}|\beta_j|})$$

$\rightsquigarrow$ convex optimization problem

- Lasso does variable selection some of the $\hat{\beta}_j(\lambda) = 0$ (because of "ℓ_1-geometry")

L_1 **penalty function**

balances the **fit** of the model with its **complexity**

The larger λ the **more sparse** the final solution β

Sparsity is important for **predictive accuracy** and **interpretation** of the final model

* R. Tibshirani *Regression shrinkage and selection via the lasso*. J. Royal Statistical Society, B, 58, 267-288, 1996.

rSLDA algorithm ----- regularization and optimization for extraction of discriminant vectors β_j

the first q sparse discriminant directions $\beta_1,\ldots,\beta_q$ are defined as the solutions to the following optimization problem:

$$\min_{A,B} \sum_{i=1}^{k} \left\| R_w^{-T} H_{b,i} - AB^T H_{b,i} \right\|^2 + \lambda \sum_{j=1}^{q} \beta_j^T \left(S_w + \gamma \frac{tr(S_w)}{p} I \right) \beta_j + \sum_{j=1}^{q} \lambda_{1,j} \left\| \beta_j \right\|_1$$

subject to $A^T A = I_{p\times q}$, where $B = [\beta_1,\ldots,\beta_q]$, $\left\|\beta_j\right\|_1$ is the 1-norm of the vector β_j, the same λ is used for all q directions, different $\lambda_{1,j}$'s are allowed to penalize different discriminant directions.

γ - regularization parameter

The above probem can be numerically solved by alternating optimization over A and B

* Qiao Z., Zhou L., Huang J. (2009) *Sparse linear discriminant analysis with applications to high dimensional low sample size data*. IAENG Int. Journal of Applied Mathematics, 39, 1.

rSLDA algorithm ----- regularization and optimization for extraction of discriminant vectors β_j

$$\min_{A,B} \sum_{i=1}^{k} \left\| R_w^{-T} H_{b,i} - AB^T H_{b,i} \right\|^2 + \lambda \sum_{j=1}^{q} \beta_j^T \left(S_w + \gamma \frac{tr(S_w)}{p} I \right) \beta_j + \sum_{j=1}^{q} \lambda_{1,j} \left\| \beta_j \right\|_1$$

Producing sparse discriminant vectors in HDLSS:

1) as in the Lasso by adding L_1 penalty to the objective function in the regression problem (**sparsity**)

2) To stabilize the solution (singularity of S_W) adding positive multiple of identity matrix **(special regularization**)

Regularized sparse LDA (rSLDA) algorithm*

1. Form the matrices from the input data:

$$H_w = X - \begin{pmatrix} e^{n_1}(\bar{x}_1)^T \\ \dots \\ e^{n_K}(\bar{x}_K)^T \end{pmatrix}$$

$$H_b = \left(\sqrt{n_1}(\bar{x}_1 - \bar{x}), \dots, \sqrt{n_k}(\bar{x}_k - \bar{x})\right)^T$$

2. Compute upper triangular matrix R_w from the Cholesky decomposition of:

$\left(S_w + \gamma \frac{tr(S_w)}{p} I\right)$ **such that** $\left(S_w + \gamma \frac{tr(S_w)}{p} I\right) = R_w^T R_w$

3. Solve the *q* independent optimization problems

$$\min_{\beta_j} \beta_j^T (\tilde{W}^T \tilde{W}) \beta_j - 2\tilde{y}^T \tilde{W} \beta_j + \lambda_1 \left\|\beta_j\right\|_1 \quad j = 1, \dots, q$$

where

$$\tilde{W}_{(n+p)\times p} = \begin{pmatrix} H_b \\ \sqrt{\lambda} \cdot R_w \end{pmatrix} \quad \tilde{y}_{(n+p)\times 1} = \begin{pmatrix} H_b R_w^{-1} \alpha_j \\ 0 \end{pmatrix}$$

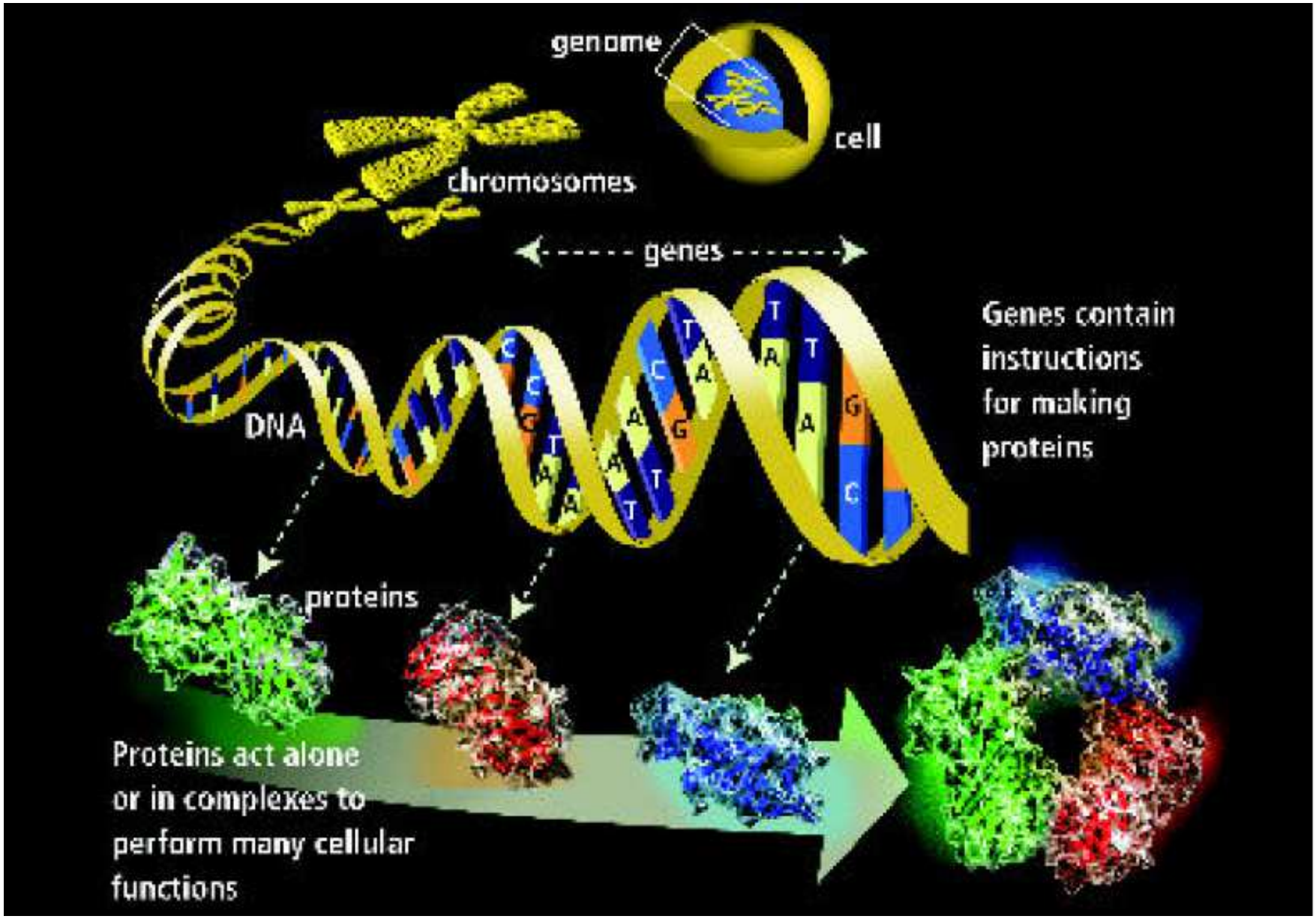
4. Compute SVD:

$$R_w^{-T}\left(H_B^T H_B\right) B = UDV^T \text{ and let } A = UV^T$$

5. Repeat steps 3 and 4 until converges.

* Qiao Z., Zhou L., Huang J. (2009) *Sparse linear discriminant analysis with applications to high dimensional low sample size data*. IAENG Int. Journal of Applied Mathematics, 39, 1.

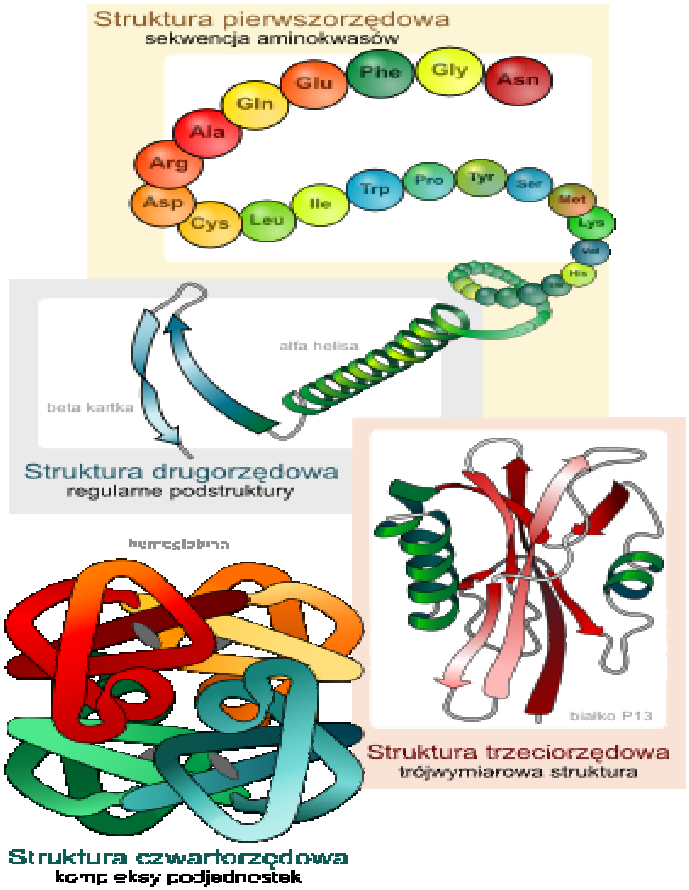
From chromosomes to proteins



PROTEINS

Proteins are **large biomolecules** found in all organisms and **form the very basis of life**

Proteins are **long polypeptide chains** consisting of **amino acid residues connected by peptide bonds**



Proteins **perform many functions within living organisms**:

- catalyzing metabolic reactions
- DNA replication
- responding to stimuli
- transporting molecules from one location to another

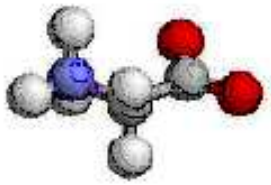
Basics of **protein structure**:

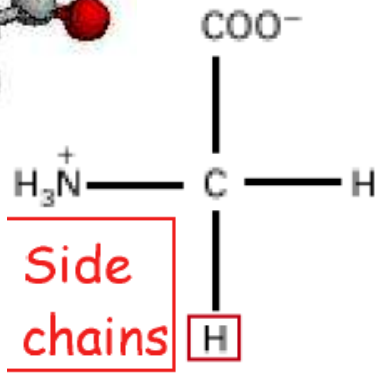
- **Primary structure**
- **Secondary structure**
- **Tertiary structure**
- **Quaternary structure**

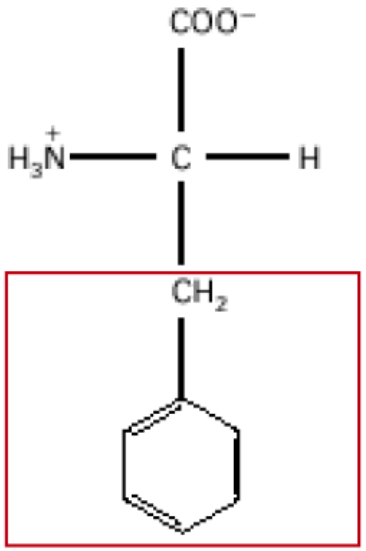
Primary structure – the protein sequence

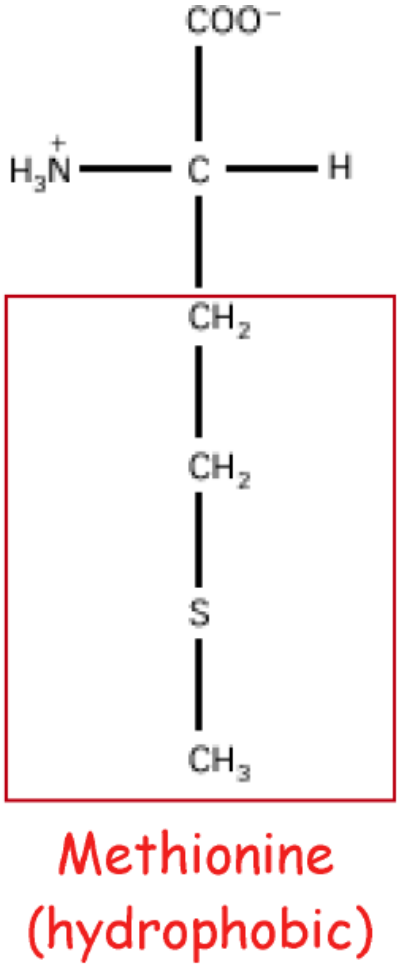
Amino acids – the building blocks of proteins

20 amino acids with different characteristics:
small, large, polar, charged, hydrophobic, ...





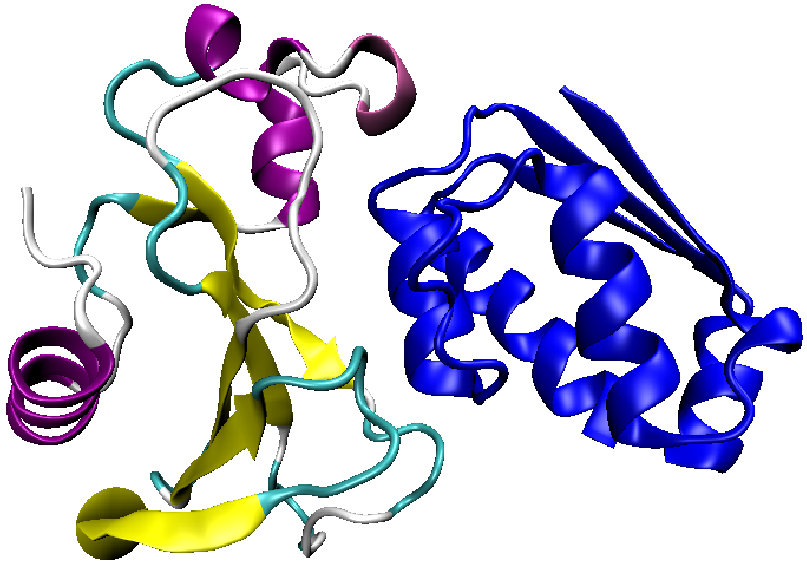




The genetic code: each amino acid is coded by 3 nucleotides (codon)

Protein-protein interaction

Protein–protein interactions (**PPI**s) refer to **physical contacts** established between two or more proteins as a result of biochemical events and/or electrostatic forces



Examples of PPI

- Signal transduction
- Transport across membranes
- Cell metabolism
- Muscle contraction

Protein-protein interaction

Homo-oligomers / hetero-oligomers
constituted by only one type/distinct of protein subunit

Non-obligate / Obligate protein complex
Can form stable crystal structure of its own (without any other associated protein) *in vivo*/ can't be found to create a crystal structure alone, but can be found as a part of a protein complex which creates a stable crystal structure.

Transient vs permanent/stable protein complex
form and break down transiently *in vivo*, whereas permanent complexes have a relatively long half-life.

Typically, the obligate interactions (protein-protein interactions in an obligate complex) **are permanent**,
whereas non-obligate interactions have been found to be either permanent or transient.[

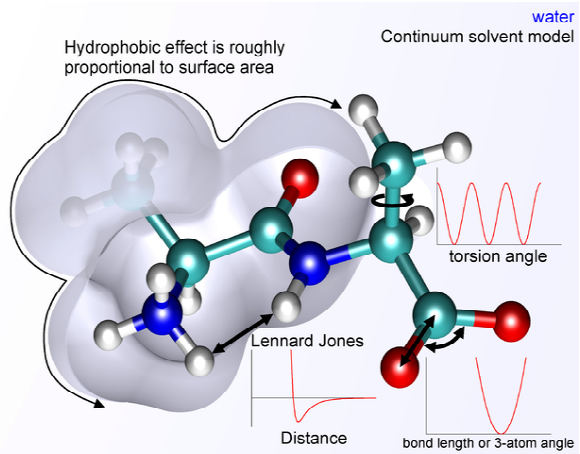
Protein force fields

Force field - functional form and parameter sets used to calculate the **potential energy** of a system of atoms or coarse-grained particles in molecular mechanics and molecular dynamics simulations.
The parameters of the energy functions can be derived from experimental work and quantum mechanical calculations.
"**All-atom**" force fields provide parameters for every type of atom in a system.
"**Coarse-grained**" potentials, frequently used in simulations of proteins provide even more crude representations for increased computational efficiency.

The **basic functional form of potential energy** includes:

- **bonded terms** for interactions of atoms that are linked by covalent bonds
- **nonbonded** ("noncovalent") terms that describe the long-range electrostatic and van der Waals forces:

E_total = E_bonded + E_nonbonded



Protein force fields

E_total = E_bonded + E_nonbonded

E_bonded = E_bond + E_angle + E_dihedral

E_nonbonded = E_elektrostatic + E_van-der-Vaals

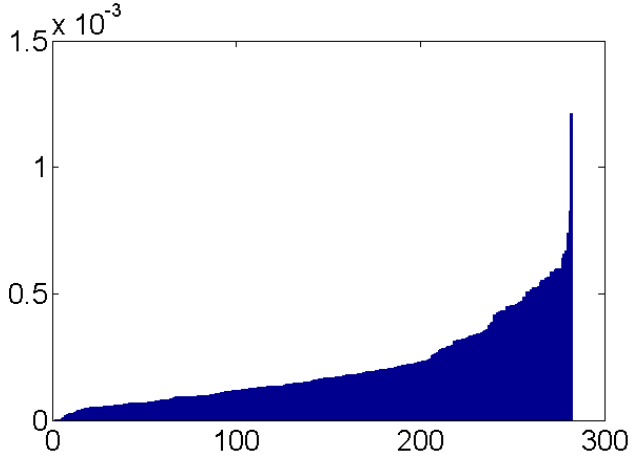
The nonbonded forces are only applied to atom pairs separated
by at least three bonds

A general framework used to predict PPI types

- PDB → creating PPI dataset
- (energetic) feature extraction using FastContact
- feature selection via **sparse rLDA**
- classification (nearest mean)
- evaluation and analysis

Variable selection via sparse rLDA

s p a r s i f i c a t i o n



Components of a vector β_1 obtained by **rSLDA algorithm** sorted in ascending order (|.|)

m **---** the number of significant variables involved in specifying the discriminant direction

Varying values of *m* --- only the *m* maximum values of the coordinates of the vector β_1 are left, the rest is zeroed.

Only these m values are used to project 282-dimensional vector of samples form protein dataset onto a one-dimensional space.

Experimental results

The **error rate** grows rapidly and then decreases with the rise of *m*, up to **m=28** (error = ~25% ±5) → **classifier performance ~75% ±5**

Then, for bigger values of *m*, almost a constant error rate was observed.

28 input variables "selected" by the rslda algorithm are the most significant for classification:

Among these 28 features – **13 are from the receptor residues contributing to the desolvation free energy**, but these are not from the beginning of the above list !.
In each of the 7 groups of energetic features – only **features with extreme** (min or max) **contribution** to the energy are always selected.
The features from the **beginning of the list** are those **from the receptor residues contributing to the electrostatics energy.**

Electrostatic energy is the most important in the prediction of obligate/non-obligate protein-protein interactions !!!