Linear Methods for Classification

4.1 - 4.3

Adrian Płoszczyca

Wprowadzenie

Dane:

$$\mathbf{x}_{11}, \mathbf{x}_{12}, ..., \mathbf{x}_{1n_1} \quad \text{z grupy } 1$$

$$\mathbf{x}_{21}, \mathbf{x}_{22}, ..., \mathbf{x}_{2n_2} \quad \text{z grupy } 2$$

$$...$$

$$\mathbf{x}_{K1}, \mathbf{x}_{K2}, ..., \mathbf{x}_{Kn_K} \quad \text{z grupy } K$$

gdzie $\mathbf{x}_{ki} = (x_{ki}^{(1)}, x_{ki}^{(2)}, ..., x_{ki}^{(p)})$ jest i-tą obserwacją z k-tej grupy o wartościach w p-wymiarowym zbiorze $\mathcal{X}$, $\mathbf{x}_{ki} \in \mathcal{X}$.

$\mathcal{G}$ – zbiór etykiet klas od 1 do K

Plan

- wstęp
- regresja liniowa
- LDA

Liniowe granice decyzyjne

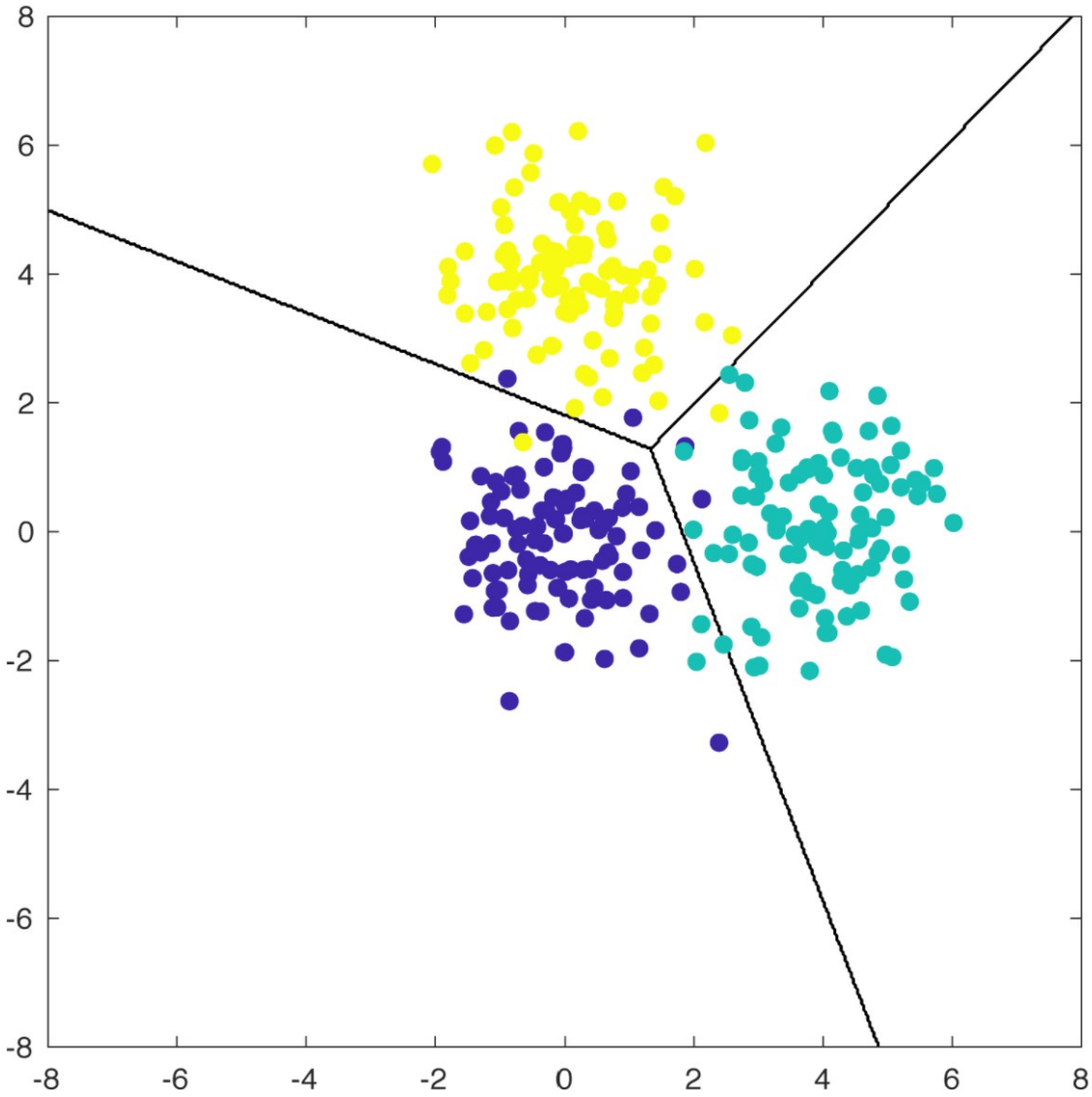
- $k = 1, ..., K$
- $\hat{f}_k(x) = \hat{\beta}_{k0} + \hat{\beta}_k^T x$
- $\hat{f}_k(x) = \hat{f}_l(x)$

Liniowe granice decyzyjne

- $\delta_k(x)$ - funkcje dyskryminacyjne
- Chcemy podać taką regułę decyzyjneą, która przypisze obserwacjom $\mathbf{x} \in \mathcal{X}$ jakąś klasę ze zbioru klas $\mathcal{G}$.

$$\delta(\mathbf{x}): \quad \mathcal{X} \to \mathcal{G}.$$

Liniowe granice decyzyjne



Metody modelujące prawdopodobieństwo

$$P(G = 1|X = x) = \frac{exp(\beta_0 + \beta^T x)}{1 + exp(\beta_0 + \beta^T x)}$$

$$P(G = 2|X = x) = \frac{1}{1 + exp(\beta_0 + \beta^T x)}$$

- logit: $\log(\frac{p}{1-p})$

$$\log \frac{P(G = 1|X = x)}{P(G = 2|X = x)} = \beta_0 + \beta^T x.$$

Linear Regression of an Indicator Matrix

- $K = 2$
- $\mathcal{G} = \{0, 1\}$

$$E(y|X = \mathbf{x}) = P(y = 1|X = \mathbf{x})$$

Dyskryminacja oparta na regresji liniowej

- Estymujemy funkcję $E(y|X = \mathbf{x})$ za pomocą liniowej funkcji zmiennych $x^{(j)}, \quad j = 1, 2, ..., p$
- Etykieta ma postać

$$\mathbf{y} = (y^{(1)}, y^{(2)}, ..., y^{(K)})$$

- w przypadku klasy k:

$$\mathbf{y} = (0, ..., 0, 1, 0, ..., 0)$$

Zapis macierzowy

- Próbę $(\mathbf{x}_1, \mathbf{y}_1), ..., (\mathbf{x}_n, \mathbf{y}_n)$ możemy zapisać jako

$$\mathbf{X}_{(n,p+1)} = \begin{bmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(p)} \\ 1 & x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(p)} \\ & & \dots & & \\ 1 & x_n^{(1)} & x_n^{(2)} & \dots & x_n^{(p)} \end{bmatrix} \quad \mathbf{Y}_{(n,g)} = \begin{bmatrix} y_1^{(1)} & y_1^{(2)} & \dots & y_1^{(p)} \\ y_2^{(1)} & y_2^{(2)} & \dots & y_2^{(p)} \\ & \dots & & \\ y_n^{(1)} & y_n^{(2)} & \dots & y_n^{(p)} \end{bmatrix}$$

Model liniowy

- wektor zmiennych objaśniających

$$\mathbf{x} = (x^{(1)}, x^{(2)}, ..., x^{(p)})$$

- wektor zmiennych objaśnianych

$$\mathbf{y} = (y^{(1)}, y^{(2)}, ..., y^{(K)})$$

- mamy $K(p+1)$ parametrów
- macierz parametrów: $\mathbf{B}_{(p+1,K)}$

Model liniowy

- Metoda najmniejszych kwadratów

$$\min_{\mathbf{B}} \sum_{i=1}^{N} ||\mathbf{y}_i - [1, \mathbf{x}_i]\mathbf{B}||^2$$

- Rozwiązanie:

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

lub równoważnie $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}$

gdzie $\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$

Cechy modelu liniowego

- Dla nowej obserwacji mamy

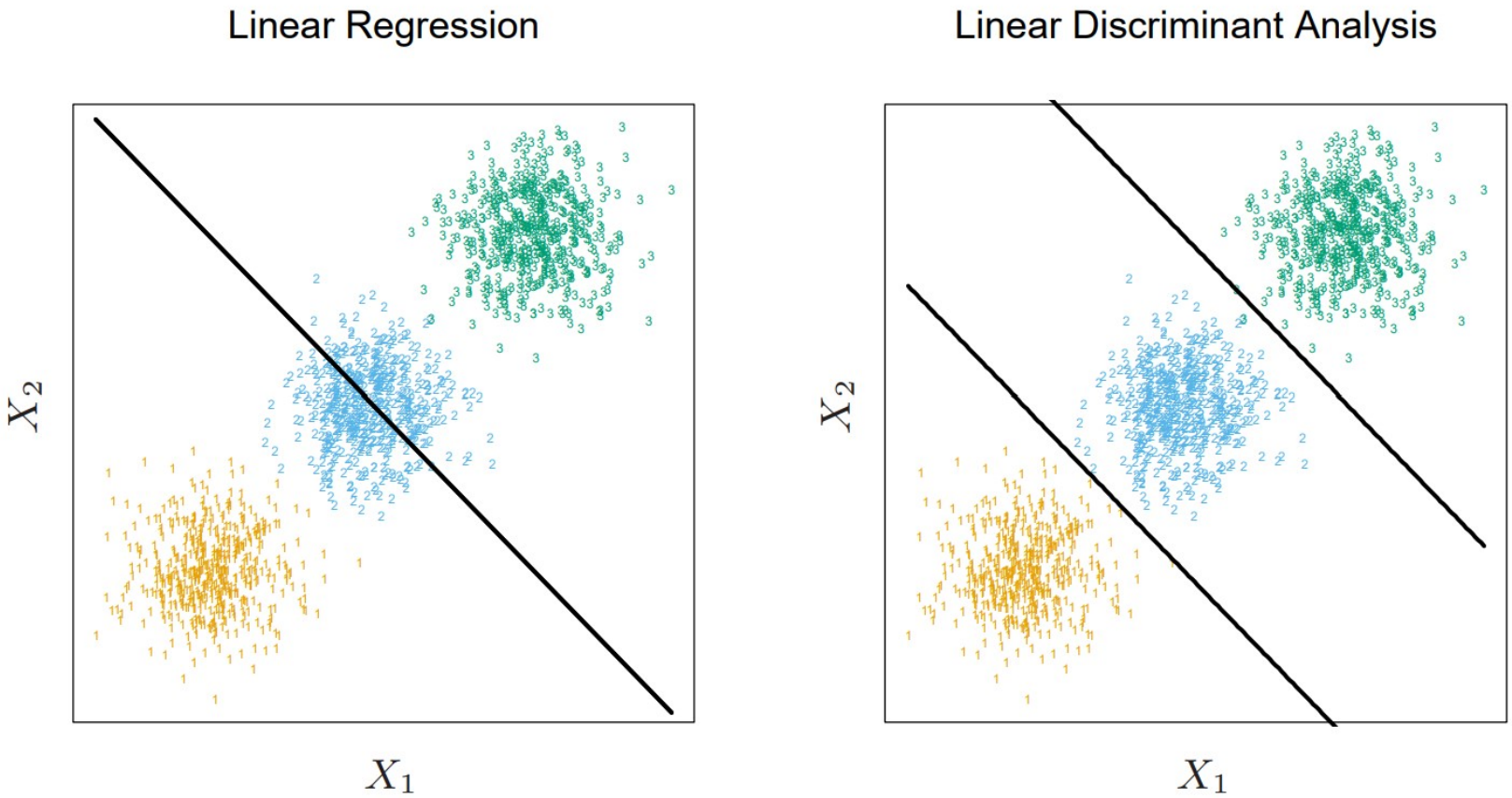
$$\hat{\mathbf{y}}(\mathbf{x}) = [1, \mathbf{x}]\hat{\mathbf{B}}$$

- $P(k|X = \mathbf{x}) = E(y^{(k)}|X = \mathbf{x})$
- dla każdej wartości $\mathbf{x}$ mamy $\sum_{k=1}^{K} \hat{y}^{(k)}(\mathbf{x}) = 1$.

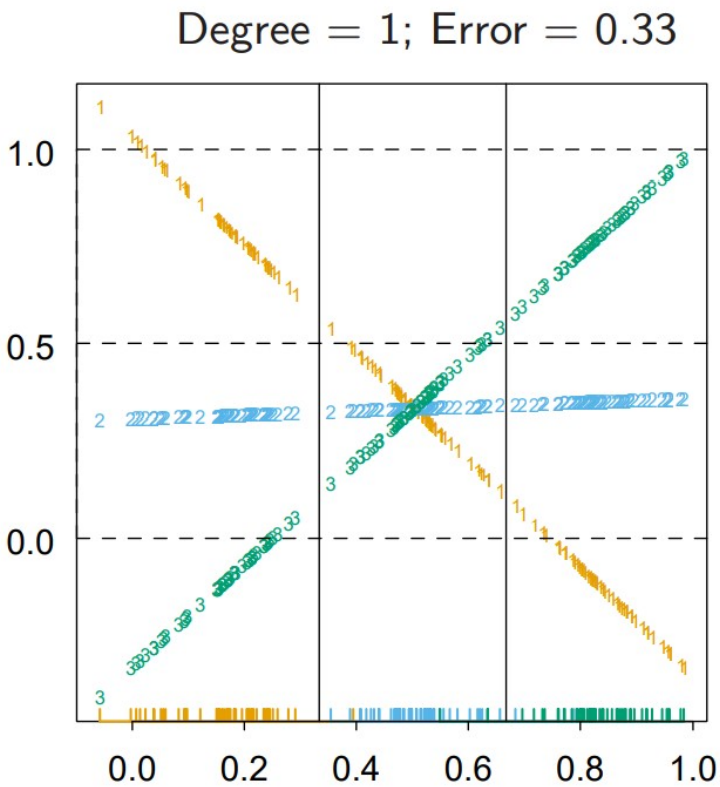
Reguła dyskryminacyjna

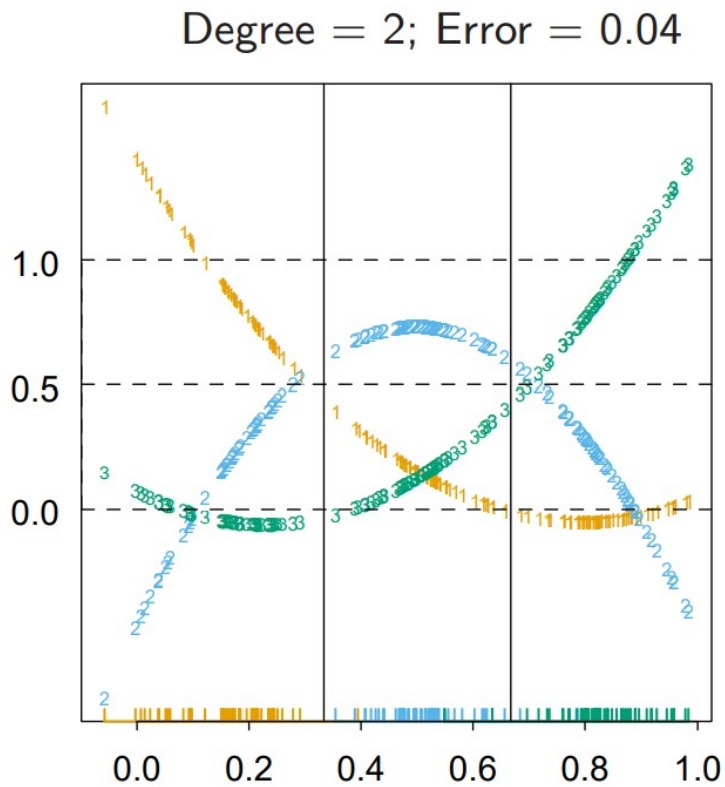
- obliczane jest dopasowanie, tzn liczymy wartość funkcji $\hat{f}(x)^T = (1, x^T)\hat{\mathbf{B}}$
- identyfikowany jest największy element i klasyfikowany jako $\hat{G}(x) = argmax_{k \in \mathcal{G}} \hat{f}_k(x)$

Maskowanie klas



Maskowanie klas





Porównanie metod

Technique	Error Rates	
	Training	Test
Linear regression	0.48	0.67
Linear discriminant analysis	0.32	0.56
Quadratic discriminant analysis	0.01	0.53
Logistic regression	0.22	0.51

Linear Discriminant Analysis

- Chcemy znaleźć taki kierunek a w przestrzeni $\mathcal{X}$, który najlepiej rozdziela klasy

$$E(\mathbf{x}) \equiv E\mathbf{x} = [E(x^{(1)}), E(x^{(2)}), ..., E(x^{(p)})]^T$$

$$\Sigma \equiv \mathrm{Cov}(\mathbf{x}) = E[(\mathbf{x} - E\mathbf{x})(\mathbf{x} - E\mathbf{x})^T] = (\sigma_{ij})_{i,j=1}^{p}$$

$$\sigma_{ij} = E[(x^{(i)} - Ex^{(i)})(x^{(j)} - Ex^{(j)})]$$

Linear Discriminant Analysis

- wariancja zmiennej losowej $\mathbf{a}^T\mathbf{x}$

$$\text{Var}(\mathbf{a}^T\mathbf{x}) = E(\mathbf{a}^T\mathbf{x} - E(\mathbf{a}^T\mathbf{x}))^2 = E[\mathbf{a}^T(\mathbf{x} - E\mathbf{x})(\mathbf{x} - E\mathbf{x})^T\mathbf{a}] =$$

$$= \mathbf{a}^T\text{Cov}(\mathbf{x})\mathbf{a} = \mathbf{a}\Sigma\mathbf{a}$$

- średnie

$$\bar{\mathbf{x}}_k = \frac{1}{n_k}\sum_{i=1}^{n_k}\mathbf{x}_{ki}, \qquad k = 1, 2.$$

Linear Discriminant Analysis

- założenie: klasy charakteryzują się taką samą macierzą kowariancji

$$\mathbf{W} = \frac{1}{n-2}\sum_{k=1}^{2}(n_k - 1)\mathbf{S}_k = \sum_{k=1}^{2}\left[\sum_{i=1}^{n_k}(x_{ki} - \bar{\mathbf{x}}_k)(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)^T\right],$$

$$n = n_1 + n_2$$

gdzie $\mathbf{S}_k$ dla $k = 1, 2$ są próbkowymi macierzami kowariancji w klasach 1 i 2

- miara zmienności wewnątrzgrupowej: $\mathbf{a}^T\mathbf{W}\mathbf{a}$

Linear Discriminant Analysis

- Znajdujemy kierunek $\mathbf{a}$ w $\mathcal{X}$, który najlepiej rozdziela obydwie podpróby uczące

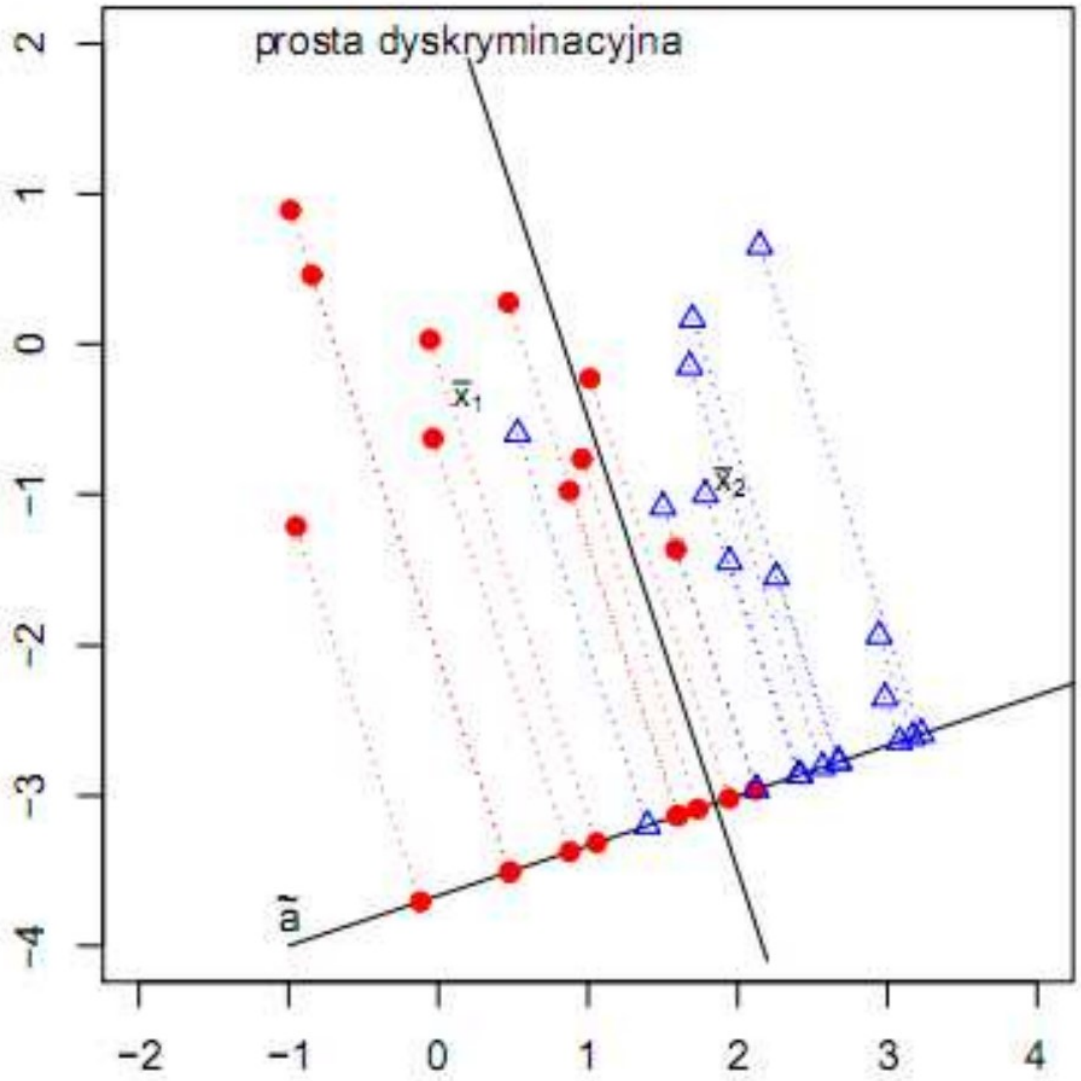
$$\text{argmax}_a \frac{(\mathbf{a}^T\bar{\mathbf{x}}_2 - \mathbf{a}^T\bar{\mathbf{x}}_1)^2}{\mathbf{a}^T\mathbf{W}\mathbf{a}}$$

- Rzutujemy ortogonalnie obie średnie klas oraz nową obserwację $\mathbf{x}$ o nieznanej klasie na ten kierunek i klasyfikujemy $\mathbf{x}$ do klasy j, jeżeli zachodzi

$$|\mathbf{a}^T\mathbf{x} - \mathbf{a}^T\bar{\mathbf{x}}_j| < |\mathbf{a}^T\mathbf{x} - \mathbf{a}^T\bar{\mathbf{x}}_k|$$

dla $k \neq j,\ j, k \in \{1, 2\}$.

Linear Discriminant Analysis



Linear Discriminant Analysis

Hiperpłaszczyzna rozdzielająca klasy:

$$(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1)^T \mathbf{W}^{-1} [\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_2 + \bar{\mathbf{x}}_1)] = 0.$$

Linear Discriminant Analysis

- Chcemy znać $P(G|X)$
- $f_k(x)$ - gęstość X w klasie $G = k$
- π_k - prawdopodobieństwo a priori dla klasy k
- $\sum_{k=1}^{K} \pi_k = 1$

Mamy:

$$P(G = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_k}$$

Linear Discriminant Analysis

- Gęstości klas:

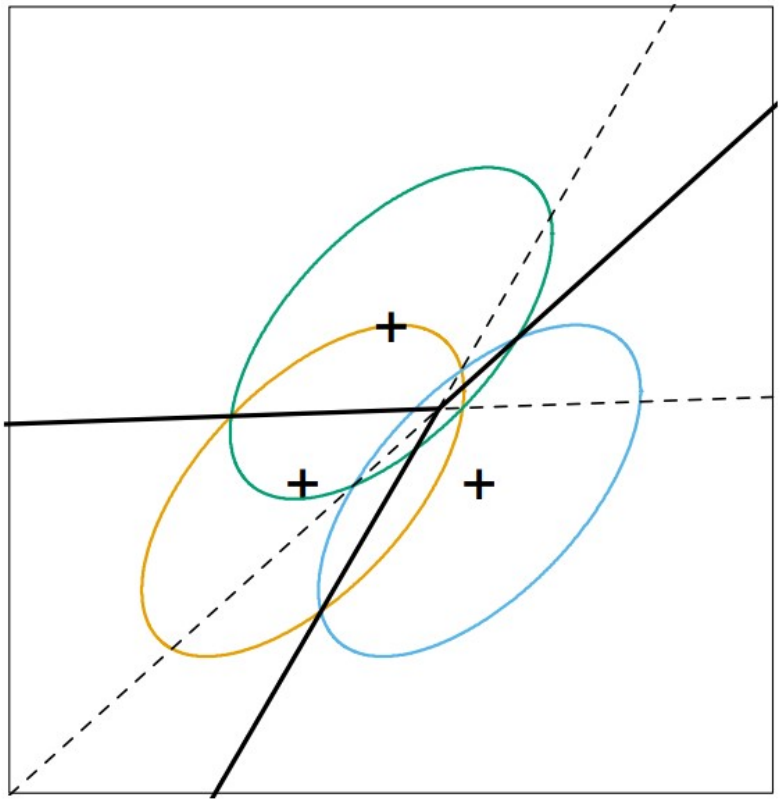
$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)}$$

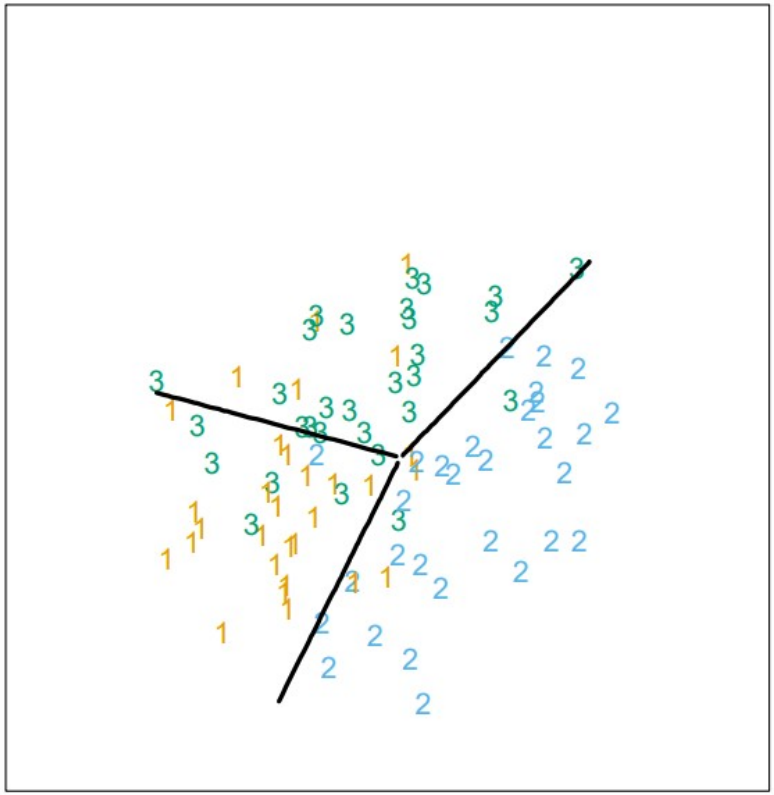
- $\Sigma_k = \Sigma \ \forall k$

$$\log \frac{P(G=k|X=x)}{P(G=l|X=x)} = \log \frac{f_k(x)}{f_l(x)} + \log \frac{\pi_k}{\pi_l}$$

$$= \log \frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k + \mu_l)^T \Sigma^{-1}(\mu_k - \mu_l) + x^T \Sigma^{-1}(\mu_k - \mu_l)$$

Linear Discriminant Analysis





Linear Discriminant Analysis

W wyniku LDA dostajemy funkcje dyskryminacyjne postaci

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$

i odpowiadającą im regułę decyzyjną $G(x) = \text{argmax}_k \delta_k(x)$.

Linear Discriminant Analysis

Estymacje:

- $\hat{\pi}_k = \frac{N_k}{N}$, gdzie N_k jest liczbą obserwacji w klasie k,
- $\hat{\mu}_k = \sum_{g_i=k} \frac{x_i}{N_k}$,
- $\hat{\Sigma} = \sum_{k=1}^{K} \sum_{g_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T/(N-K)$.

W przypadku dwóch klas LDA klasyfikuje do klasy 2 jeśli

$$X^T \hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) > \frac{1}{2}\hat{\mu}_2^T \hat{\Sigma}^{-1}\hat{\mu}_2 - \frac{1}{2}\hat{\mu}_1^T \hat{\Sigma}^{-1}\hat{\mu}_1 + \log(N_1/N) - \log(N_2/N)$$

Quadratic Discriminant Analysis

- LDA:

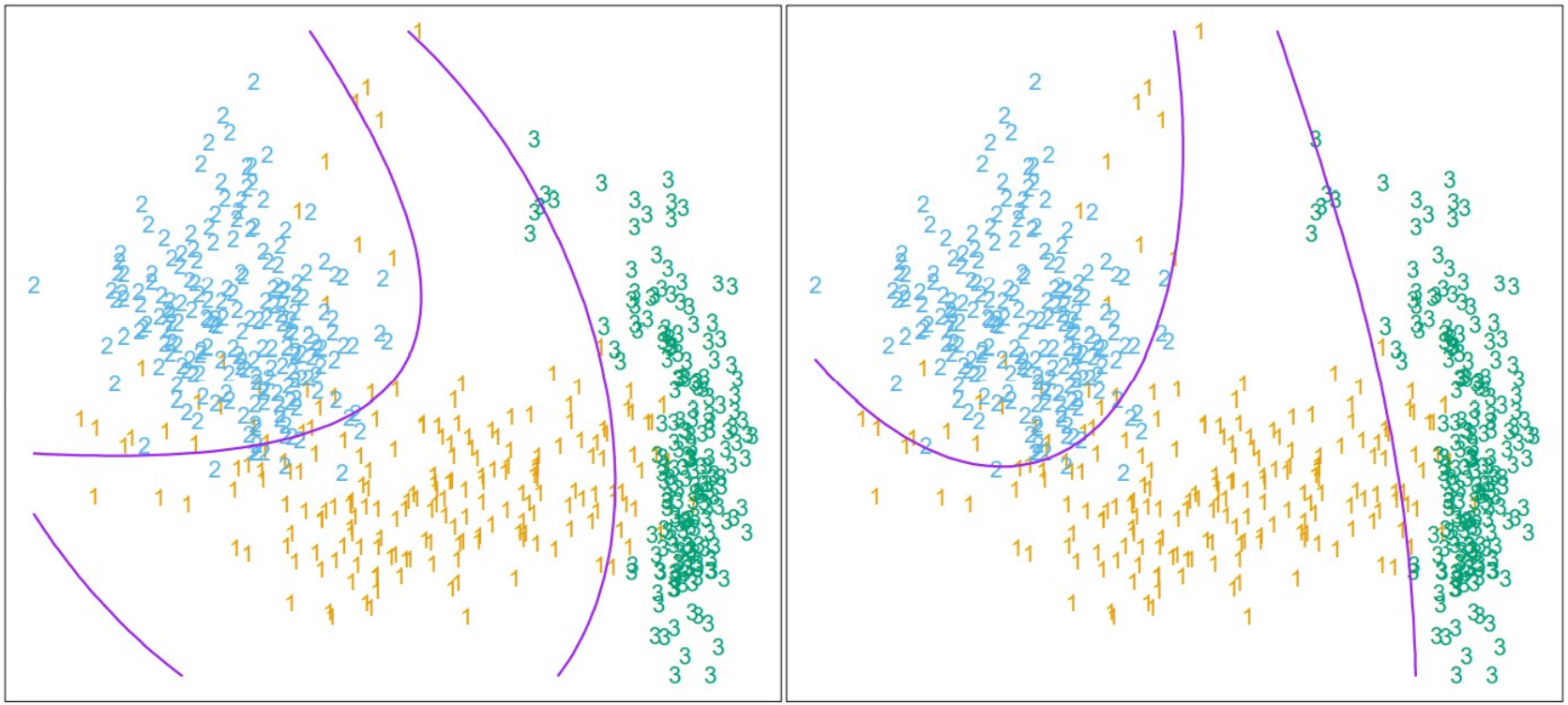
$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$

- QDA:

$$\delta_k(x) = -\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) + \log \pi_k$$

- $\{x : \delta_k(x) = \delta_l(x)\}$

LDA vs QDA



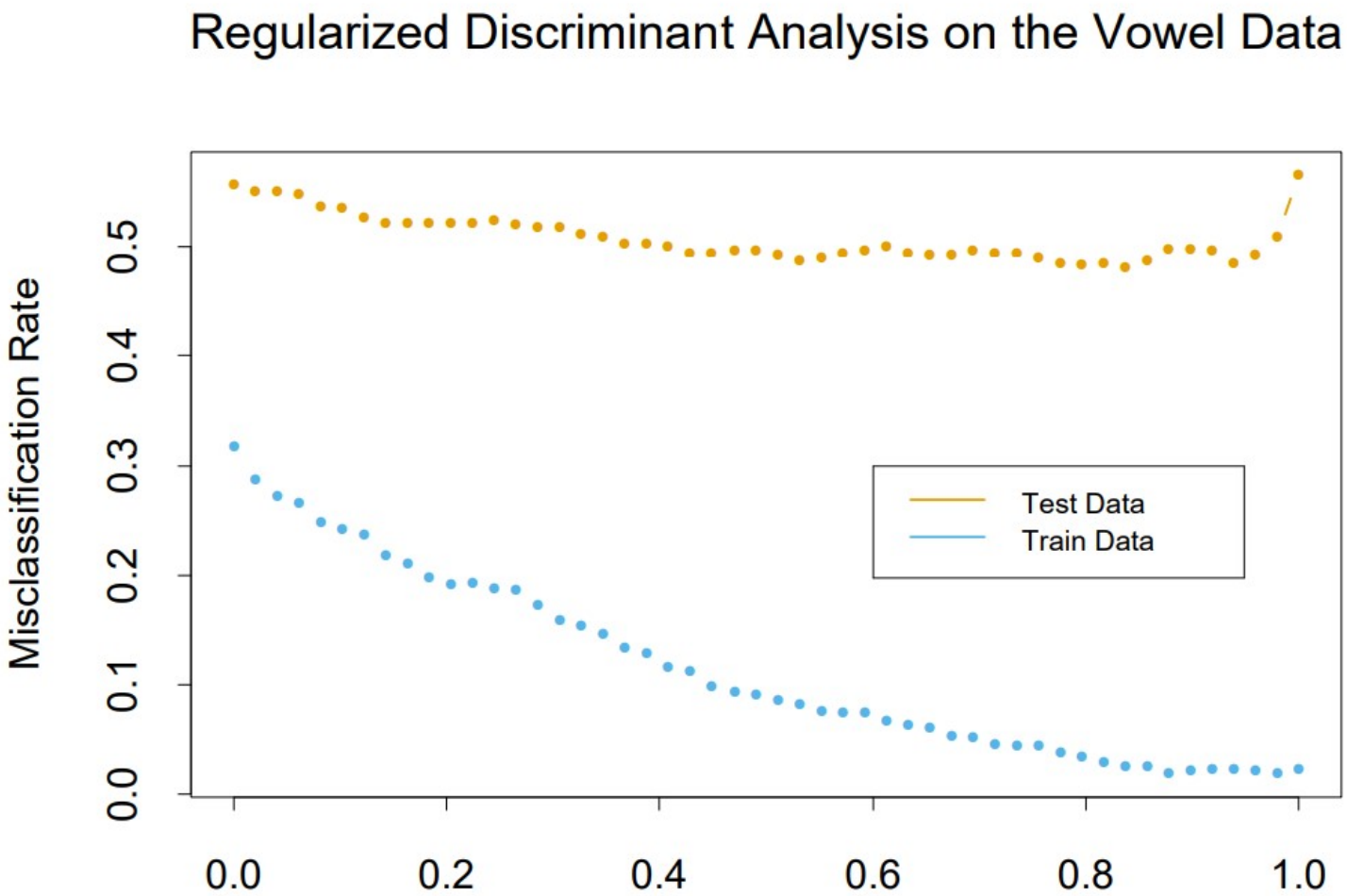
Regularized Discriminant Analysis

- Regularyzowane macierze kowariancji są postaci

$$\hat{\Sigma}_k(\alpha) = \alpha\hat{\Sigma}_k + (1-\alpha)\hat{\Sigma}$$

$$\alpha \in [0,1]$$

Regularized Discriminant Analysis



Computations for LDA

$$\hat{\Sigma}_k = U_k D_k U_k^T$$

- U_k - ortonormalna macierz $p \times p$
- D_k jest diagonalną macierzą z dodatnimi wartościami własnymi

Wtedy składnikami potrzebnymi do obliczenia $\delta_k(x)$ są

- $(x - \hat{\mu}_k)^T \hat{\Sigma}_k^{-1} (x - \hat{\mu}_k) = [U_k^T (x - \hat{\mu}_k)]^T D_k^{-1} [U_k^T (x - \hat{\mu}_k)]$
- $\log |\hat{\Sigma}_k| = \sum_l \log d_{kl}$

Computations for LDA

- przetransformować dane względem wspólnego estymatora kowariancji $\hat{\Sigma}$: $X^* \leftarrow D^{-\frac{1}{2}} U^T X$, gdzie $\hat{\Sigma} = UDU^T$ w taki sposób, że estymowana macierz kowariancji X^* będzie identycznościowa
- zaklasyfikować do klasy o najbliższym centroidzie w tej przetransformowanej przestrzeni, modulo efekt prawdopodobieństwa a priori π_k tej klasy

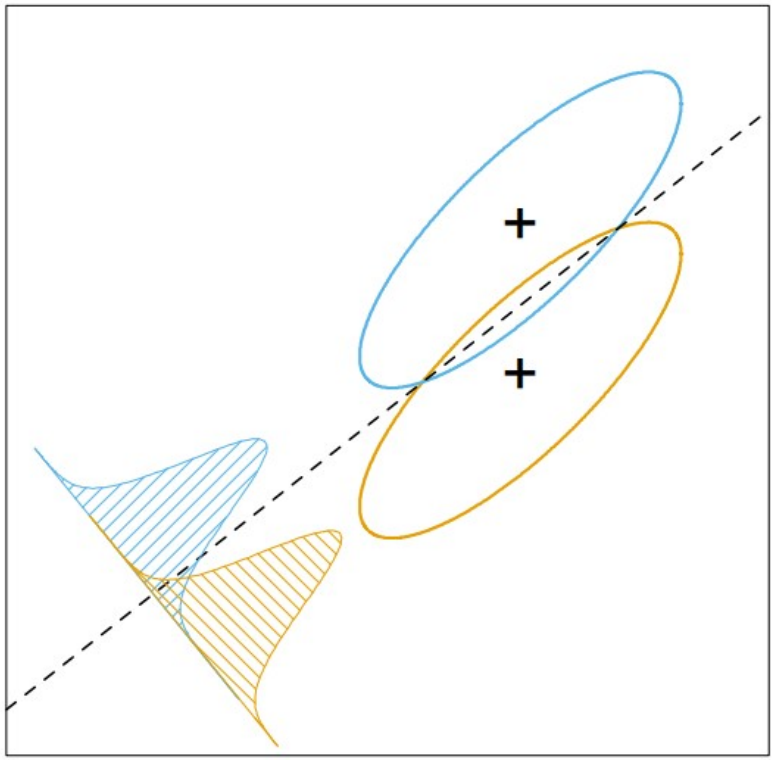
Reduced-Rank LDA

- rzutujemy X^* na tą podprzestrzeń H_{K-1}
- chcemy mieć taki wymiar $L < K - 1$ że podprzestrzeń $H_L \subseteq H_{K-1}$ będzie dla LDA w jakimś sensie optymalna

Reduced-Rank LDA

- obliczyć macierz centroidów klas M wymiaru $K \times p$ i macierz kowariancji W (kowariancja wewnątrzgrupowa)
- obliczyć $M^* = MW^{-\frac{1}{2}}$
- obliczyć macierz kowariancji M^* czyli B^* (kowariancja międzygrupowa)
- obliczyć $B^* = V^* D_B V^{*T}$
- $Z_l = v_l^T X$, gdzie $v_l = W^{-\frac{1}{2}} v_l^*$

Reduced-Rank LDA



Reduced-Rank LDA

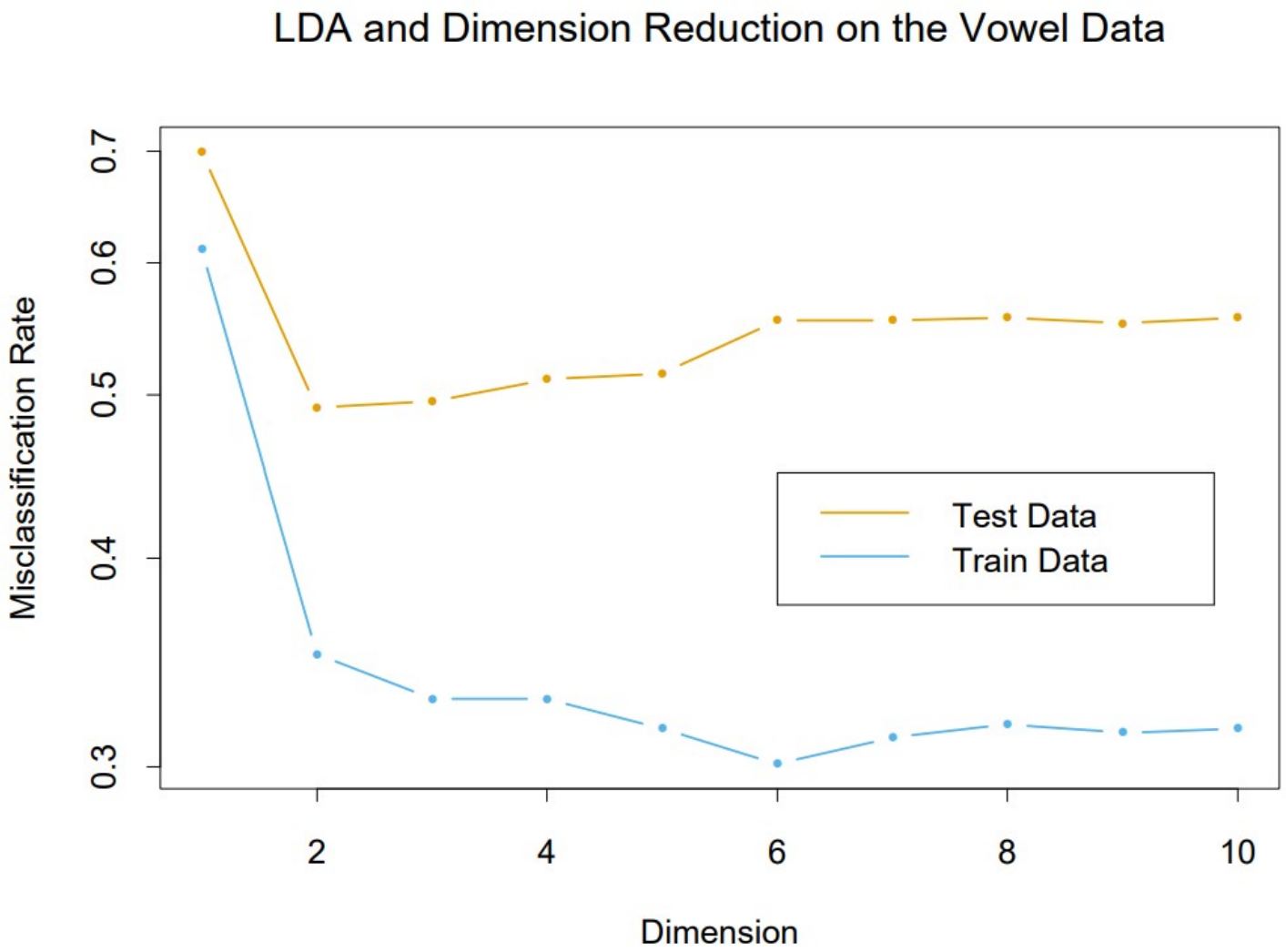
Problem sprowadza się do maksymalizacji

$$\max_a \frac{a^T B a}{a^T W a}$$

Reduced-Rank LDA

- klasyfikacja ze wspólnymi macierzami kowariancji prowadzi do liniowych granic decyzyjnych. Klasyfikację można osiągnąć przez przekształcenie danych tak, aby miały identycznościową macierz kowariancji i klasyfikację do najbliższego centroidu (modulo $\log \pi_k$)
- Ponieważ liczą się tylko odległości względne do centroidów, można ograniczyć dane do podprzestrzeni rozpiętej przez centroidy [w przestrzeni sferycznej].
- Ta podprzestrzeń może być dalej rozłożona na kolejne podprzestrzenie optymalne z punktu widzenia separacji środka ciężkości

Reduced-Rank LDA



Reduced-Rank LDA