SYSTEMY UCZĄCE SIĘ

WYKŁAD 5. LINIOWE METODY KLASYFIKACJI

Dr hab. inż. Grzegorz Dudek
Wydział Elektryczny
Politechnika Częstochowska

Częstochowa 2014

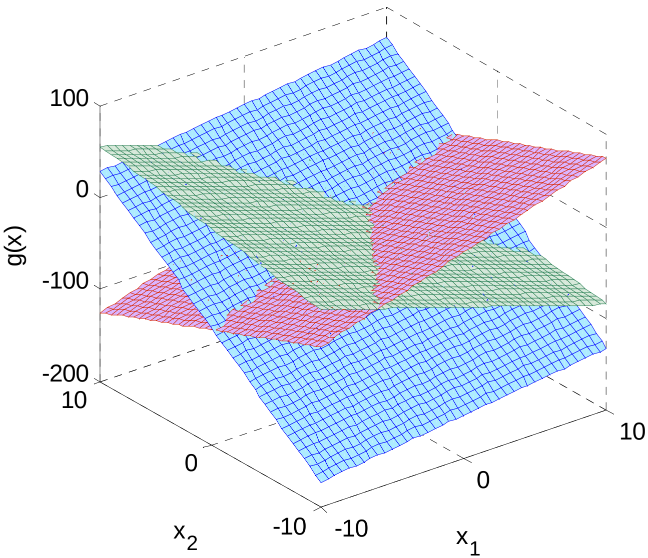
Cel klasyfikacji liniowej

Wyznaczyć w przestrzeni przykładów *X* powierzchnię dyskryminacyjną (liniową) rozdzielającą przykłady z różnych klas.

Powierzchnie dyskryminacyjne można określić pośrednio na podstawie porównań wartości *K* funkcji dyskryminacyjnych $g_1(\mathbf{x})$, $g_2(\mathbf{x})$, ..., $g_K(\mathbf{x})$ zdefiniowanych dla każdej klasy.

Funkcja dyskryminacyjna $g_i(\mathbf{x})$ ma tę własność, że dla wszystkich przykładów należących do klasy *i* zachodzi:

$$g_i(\mathbf{x}) > g_j(\mathbf{x}), \quad i, j = 1,2,\ldots,K, i \neq j$$



FUNKCJE DYSKRYMINACYJNE I MASZYNA LINIOWA

Funkcje dyskryminacyjne $g_i(\mathbf{x})$ i $g_j(\mathbf{x})$ przyległych obszarów decyzyjnych definiują powierzchnię dyskryminacyjną rozdzielająca te obszary:

$$g_i(\mathbf{x}) = g_j(\mathbf{x})$$

Przyjmijmy, że każda klasa reprezentowana jest przez punkt prototypowy $\mathbf{P}_i = [p_{i,1}\ p_{i,2} \ldots p_{i,n}]$, np. środek klasy:

$$\mathbf{m}_k = \frac{1}{N_k}\sum_{l=1}^{N_k} \mathbf{x}_l^k, \quad k = 1, 2, \ldots, K \text{ (numer klasy)}$$

gdzie N_k jest liczbą przykladów z klasy k; liczbę wszystkich przykładów oznaczmy N.

Maszyna liniowa oblicza odległość pomiędzy przykładem **x**, a każdym punktem prototypowym. Przykład zostaje zaliczony do tej klasy, którą reprezentuje najbliższy punkt prototypowy.

Niech miarą odległości będzie kwadrat odległości euklidesowej:

$$\| \mathbf{x} - \mathbf{P}_i \|^2 = \left(\sqrt{(\mathbf{x} - \mathbf{P}_i)^T (\mathbf{x} - \mathbf{P}_i)}\right)^2 = \mathbf{x}^T\mathbf{x} - 2\mathbf{P}_i^T\mathbf{x} + \mathbf{P}_i^T\mathbf{P}_i$$

FUNKCJE DYSKRYMINACYJNE I MASZYNA LINIOWA

Niezależny od klasy składnik $\mathbf{x}^T\mathbf{x}$ można pominąć i przyjąć, że pozostałość ze znakiem przeciwnym podzielona przez 2 będzie pełnić rolę funkcji dyskryminacyjnej:

$$g_i(\mathbf{x}) = \mathbf{P}_i^T\mathbf{x} - 0.5\mathbf{P}_i^T\mathbf{P}_i$$

Jest to funkcja liniowa ze współczynnikiem kierunkowym $\mathbf{a}_i = \mathbf{P}_i$ i wyrazem wolnym $a_{i,0} = -0.5\mathbf{P}_i^T\mathbf{P}_i$.

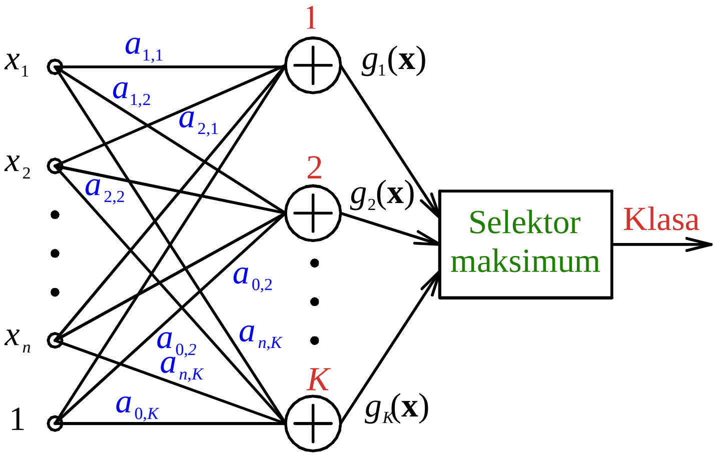
Powierzchnię dyskryminacyjną rozdzielającą dwa przyległe obszary opisuje równanie:

$$\mathbf{P}_i^T\mathbf{x} - 0.5\mathbf{P}_i^T\mathbf{P}_i = \mathbf{P}_j^T\mathbf{x} - 0.5\mathbf{P}_j^T\mathbf{P}_j$$

W przypadku przykładów dwuwymiarowych $\mathbf{x} = [x_1, x_2]^T$ linia decyzyjna pomiędzy klasą 1 i 2 ma postać:

$$p_{1,1}x_1 + p_{1,2}x_2 - 0.5(p_{1,1}^2 + p_{1,2}^2) = p_{2,1}x_1 + p_{2,2}x_2 - 0.5(p_{2,1}^2 + p_{2,2}^2)$$

$$x_2 = \frac{(p_{2,1} - p_{1,1})x_1 + 0.5(p_{1,1}^2 + p_{1,2}^2 - p_{2,1}^2 - p_{2,2}^2)}{p_{1,2} - p_{2,2}}$$



Liniowa separowalność

Jeśli istnieje K liniowych funkcji dyskryminacyjnych $g_i(\mathbf{x})$ o postaci $\mathbf{a}^T\mathbf{x} + a_0$, takich że:

$$g_i(\mathbf{x}) > g_j(\mathbf{x}) \text{ dla każdego } \mathbf{x} \in \text{ klasy } i,\ i, j = 1,2,\ldots,K, i \neq j$$

to obszary zawierające wyłącznie punkty z jednej klasy (tzw. obszary decyzyjne) są liniowo separowalne.

Dane są punkty prototypowe: $\mathbf{P}_1 = [2, 10]^T$, $\mathbf{P}_2 = [-5, 2]^T$, $\mathbf{P}_3 = [5, -5]^T$. Zaprojektuj maszynę liniową.

Wagi funkcji dyskryminacyjnych:

$$a_{1,1} = 2,\ a_{1,2} = 10,\ a_{1,0} = -0.5\mathbf{P}_1^T\mathbf{P}_1 = -52 \quad \rightarrow \quad g_1(\mathbf{x}) = 2x_1 + 10x_2 - 52$$

$$a_{2,1} = -5,\ a_{2,2} = 2,\ a_{2,0} = -0.5\mathbf{P}_2^T\mathbf{P}_2 = -14.5 \quad \rightarrow \quad g_2(\mathbf{x}) = -5x_1 + 2x_2 - 14.5$$

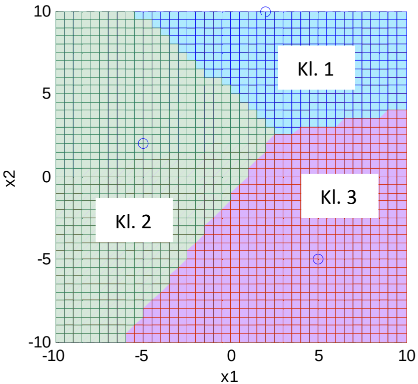
$$a_{3,1} = 5,\ a_{3,2} = -5,\ a_{3,0} = -0.5\mathbf{P}_3^T\mathbf{P}_3 = -25 \quad \rightarrow \quad g_3(\mathbf{x}) = 5x_1 - 5x_2 - 25$$

Linie decyzyjne:

$$S_{1-2} : g_1(x) - g_2(x) = 0 \rightarrow 7x_1 + 8x_2 - 37.5 = 0$$

$$S_{1-3} = g_1(x) - g_3(x) = 0 \rightarrow -3x_1 + 15x_2 - 27 = 0$$

$$S_{2-3} = g_2(x) - g_3(x) = 0 \rightarrow -10x_1 + 7x_2 + 10.5 = 0$$



Klasyfikacja nowego przykładu $\mathbf{x}^* = [0, 2]$:

$$g_1(\mathbf{x}^*) = -32,\ g_2(\mathbf{x}^*) = -10.5,\ g_3(\mathbf{x}^*) = -35 \rightarrow \mathbf{x} \in \text{klasy 2}$$

Parametry maszyny liniowej wyznacza się na podstawie punktów prototypowych w sposób analityczny. W regule perceptronowej wartości współczynników hiperpłaszczyzny dyskryminacyjnej uzyskuje się w procesie uczenia z nauczycielem na podstawie zbioru trenującego.

Reguła klasyfikacji w przypadku dwóch klas ma postać (tzw. dychotomizator):

$$\mathbf{a}^T\mathbf{x} > 0 \quad \forall \mathbf{x} \in \text{klasy } 1$$
$$\mathbf{a}^T\mathbf{x} < 0 \quad \forall \mathbf{x} \in \text{klasy } 2$$

gdzie: $\mathbf{x} = [x_1, x_{2, \ldots,} x_n, 1]^T$, $\mathbf{a} = [a_1, a_{2, \ldots,} a_n, a_0]^T$

Szukamy takich współczynników, które minimalizują kryterium:

$$J(\mathbf{a}) = \sum_{\mathbf{x} \in Z} \delta_{\mathbf{x}} \mathbf{a}^T \mathbf{x}$$

gdzie: Z – podzbiór przykładów niepoprawnie klasyfikowanych, $\delta_{\mathbf{x}} = -1$, jeśli $\mathbf{x} \in$ klasy 1 i $\delta_{\mathbf{x}} = +1$, jeśli $\mathbf{x} \in$ klasy 2.

Reguła perceptronowa

Do znalezienia minimum można zastosować algorytm największego spadku gradientu. W kolejnych iteracjach tego algorytmu modyfikujemy współczynniki, do momentu osiągnięcia minimum kryterium (poprawnej klasyfikacji wszystkich przykładów uczących).

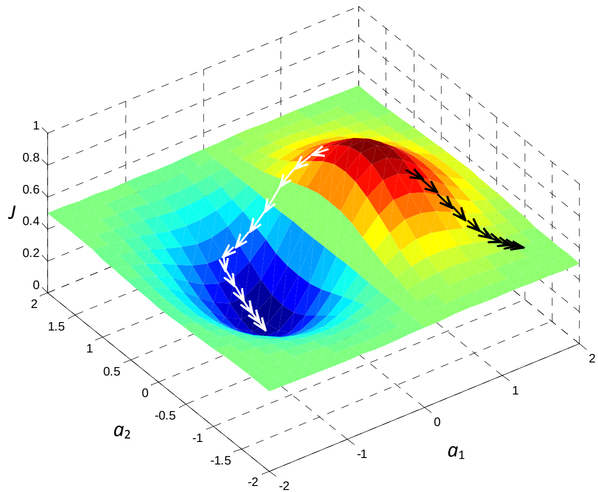
Gradient $\nabla J(\mathbf{a}) = \left[\frac{\partial J}{\partial a_1}, \frac{\partial J}{\partial a_2}, \ldots, \frac{\partial J}{\partial a_0}\right]$ ze znakiem ujemnym wskazuje kierunek "przesunięcia" współczynników:

$$\mathbf{a} \leftarrow \mathbf{a} - \eta \nabla J(\mathbf{a})$$

gdzie $\eta > 0$ współczynnikiem uczenia.

Ponieważ $\nabla J(\mathbf{a}) = \sum_{\mathbf{x} \in Z} \delta_{\mathbf{x}} \mathbf{x}$ perceptronową regułę uczenia możemy zapisać:

$$\mathbf{a} \leftarrow \mathbf{a} - \eta \sum_{\mathbf{x} \in Z} \delta_{\mathbf{x}} \mathbf{x}$$



1. Wybierz losowo **a**, ustal η.
2. Powtarzaj
   2.1. Z = $\emptyset$
   2.2. Powtarzaj dla *i* =1, 2, ..., *N*
      2.2.1. Jeśli $\delta_{\mathbf{x}}\mathbf{a}^T\mathbf{x}_i \geq 0$, to $Z = Z \cup \{ \mathbf{x}_i \}$
   2.3. Jeśli Z = $\emptyset$, to zakończ
   2.4. $\mathbf{a} \leftarrow \mathbf{a} - \eta \sum_{\mathbf{x} \in Z} \delta_{\mathbf{x}} \mathbf{x}$

- Algorytm przerywa działanie, gdy znajdzie jakąkolwiek płaszczyznę separującą klasy.
- Jeśli przykłady są liniowo separowalne, algorytm zawsze znajduje rozwiązanie w skończonej liczbie kroków (jest zbieżny).

Reguła perceptronowa – algorytm dla wielu klas

W przypadku K klas liniowo separowalnych oczekujemy:

$$\mathbf{a}_i^T\mathbf{x} > \mathbf{a}_j^T\mathbf{x} \quad \text{dla każdego } \mathbf{x} \in \text{ klasy } i,\ i, j = 1,2,\ldots,K,\ i \neq j$$

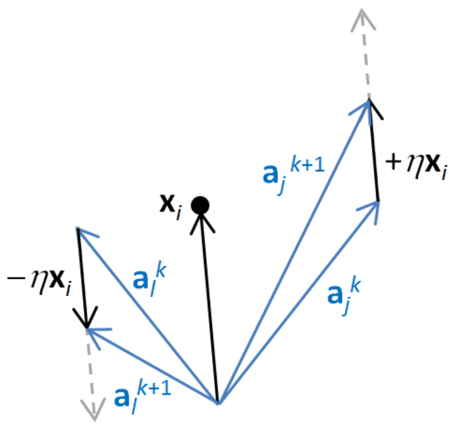
1. Wybierz losowo $\mathbf{a}_j$ dla j = 1, 2, ..., K, ustal η.
2. Powtarzaj

   2.1. Powtarzaj dla i =1, 2, ..., N

   2.1.1. Jeśli $\boldsymbol{x}_i$ ma klasę j i dla pewnych l zachodzi $\mathbf{a}_j^T\mathbf{x}_i \le \mathbf{a}_l^T\mathbf{x}_i$ (błędna klasyfikacja), to:

   $$\mathbf{a}_j \leftarrow \mathbf{a}_j + \eta\mathbf{x}_i$$
   $$\mathbf{a}_l \leftarrow \mathbf{a}_l - \eta\mathbf{x}_i$$

   2.2. Jeśli w pętli 2.1 nie nastąpiła modyfikacja żadnych współczynników **a**, to zakończ.



REGUŁA PERCEPTRONOWA – PRZYKŁAD

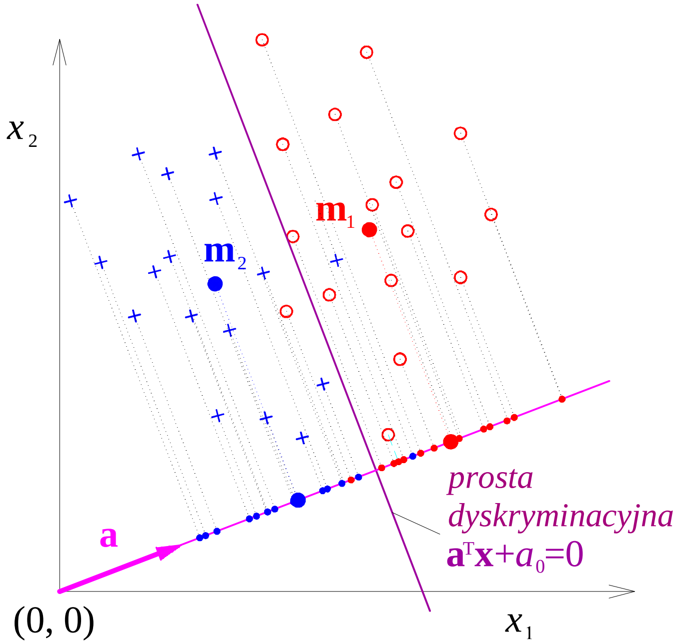
Przeprowadź trening perceptronu dla przykładów należących do trzech klas: $\mathbf{x}_1 = [10, 2, 1]^T$, $\mathbf{x}_2 = [2, -5, 1]^T$, $\mathbf{x}_3 = [-5, 5, 1]^T$. Przyjmij $\eta = 1$ i współczynniki startowe $\mathbf{a}_1 = [1, -2, 0]^T$, $\mathbf{a}_2 = [0, -1, -2]^T$, $\mathbf{a}_3 = [1, 3, 1]^T$.

Przedstaw decyzje perceptronu i współczynniki w kolejnych iteracjach. Przedstaw linie dyskryminacyjne. Zobrazuj rozwiązanie na wykresie.

Idea klasyfikacji Fishera

Znajdź kierunek **a** w przestrzeni przykładów *X*, który po zrzutowaniu na niego przykładów pozwala najlepiej rozdzielić przykłady obu klas. Miara separowalności uwzględnia odległości między klasami i rozrzut przykładów wewnątrz klas.

Kierunek **a** to wektor współczynników stojących przy zmiennych w równaniu prostej w postaci ogólnej. Prosta dyskryminacyjna jest prostopadła do tego kierunku.



Tok potępowania

Metoda wymaga podziału zbioru trenującego na dwa podzbiory: A – przykłady z klasy 1, B – przykłady z klasy 2. Zbiór przykładów należących do danej klasy reprezentuje środek klasy **m** i macierz kowariancji **S** informująca o ich rozrzutach w różnych kierunkach.

1. Wyznaczamy środki klas jako średnie wektorowe przykładów należących do poszczególnych klas:

$$\mathbf{m}_k = \frac{1}{N_k} \sum_{l=1}^{N_k} \mathbf{x}_l^k, \quad k = 1, 2 \text{ (numer klasy)}$$

   gdzie N_k jest liczbą przykladów z klasy k; liczbę wszystkich przykładów oznaczmy N.

2. Jako miary rozrzutu przykładów w klasach wyznaczamy macierze kowariancji:

$$\mathbf{S}_k = \frac{1}{N_k - 1} \sum_{l=1}^{N_k} (\mathbf{x}_{k,l} - \mathbf{m}_k)(\mathbf{x}_{k,l} - \mathbf{m}_k)^T, \quad k = 1, 2$$

3. Wyznaczamy macierz kowariancji wspólnej dla obu klas:

$$\mathbf{W} = \frac{(N_1 - 1)\mathbf{S}_1 + (N_2 - 1)\mathbf{S}_2}{N - 2}$$

Miarą rozproszenia przykładów wzdłuż kierunku **a** jest wielkość: $\mathbf{a}^T \mathbf{W} \mathbf{a}$.

4. Znajdź kierunek $\tilde{\mathbf{a}}$, dla którego wyrażenie

$$\frac{(\mathbf{a}^T \mathbf{m}_2 - \mathbf{a}^T \mathbf{m}_1)^2}{\mathbf{a}^T \mathbf{W} \mathbf{a}} \qquad (*)$$

osiąga wartość maksymalną. Zachodzi to wtedy, gdy kwadrat odległości pomiędzy średnimi klas $\mathbf{m}_1$ i $\mathbf{m}_2$ zrzutowanymi na ten kierunek (licznik) jest jak największy, a rozrzut przykładów wzdłuż tego kierunku (mianownik) jest jak najmniejszy.

5. Reguła klasyfikacji: zrzutuj średnie $\mathbf{m}_1$ i $\mathbf{m}_2$ oraz nowy przykład **x** na kierunek $\tilde{\mathbf{a}}$, zmierz odległości pomiędzy **x** i środkami $\mathbf{m}_k$ wzdłuż kierunku $\tilde{\mathbf{a}}$: $|\, \tilde{\mathbf{a}}^T \mathbf{x} - \tilde{\mathbf{a}}^T \mathbf{m}_k \,|$ i zaklasyfikuj **x** do tej klasy, której środek leży bliżej.

Prosta (płaszczyzna, hiperpłaszczyzna) dyskryminacyjna wynika z p. 5 – jest to płaszczyzna prostopadła do kierunku $\tilde{\mathbf{a}}$, przechodząca przez środek odcinka łączącego $\tilde{\mathbf{a}}^T\mathbf{m}_1$ i $\tilde{\mathbf{a}}^T\mathbf{m}_2$. Wyznacza ją równość:

$$\left|\tilde{\mathbf{a}}^T\mathbf{x}-\tilde{\mathbf{a}}^T\mathbf{m}_1\right|=\left|\tilde{\mathbf{a}}^T\mathbf{x}-\tilde{\mathbf{a}}^T\mathbf{m}_2\right|$$

Przyrównując pochodne kryterium (*) do zera można znaleźć kierunek optymalny: $\tilde{\mathbf{a}}=\mathbf{W}^{-1}(\mathbf{m}_2-\mathbf{m}_1)$, wobec czego równanie płaszczyzny dyskryminacyjnej można zapisać:

$$(\mathbf{m}_2-\mathbf{m}_1)\mathbf{W}^{-1}[\mathbf{x}-0.5(\mathbf{m}_1+\mathbf{m}_2)]=0$$

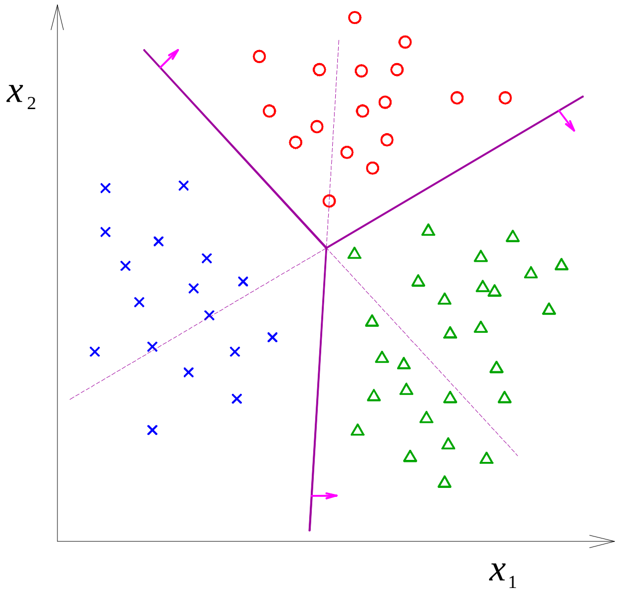
Czyli jeśli dla naszego przykładu **x** zachodzi: $(\mathbf{m}_2-\mathbf{m}_1)\mathbf{W}^{-1}[\mathbf{x}-0.5(\mathbf{m}_1+\mathbf{m}_2)]>0$ klasyfikujemy go do klasy 2. Gdy zachodzi nierówność odwrotna przykład klasyfikujemy do klasy 1.

Równanie prostej dyskryminacyjnej w przypadku dwuwymiarowym (prosta o współczynnikach $\tilde{\mathbf{a}}$, przechodząca przez środek odcinka łączącego $\mathbf{m}_1$ i $\mathbf{m}_2$):

$$x_2=-\frac{\tilde{a}_1x_1+\tilde{a}_0}{\tilde{a}_2},\ \text{gdzie}\ \tilde{a}_0=-\tilde{\mathbf{a}}^T\,\frac{\mathbf{m}_1+\mathbf{m}_2}{2}$$

Problem wielu klas

Dla wszystkich możliwych zestawów par klas tworzy się odrębną płaszczyznę dyskryminacyjną. W obliczeniach przyjmuje się wspólną dla wszystkich klas macierz kowariancji wewnątrzgrupowej:



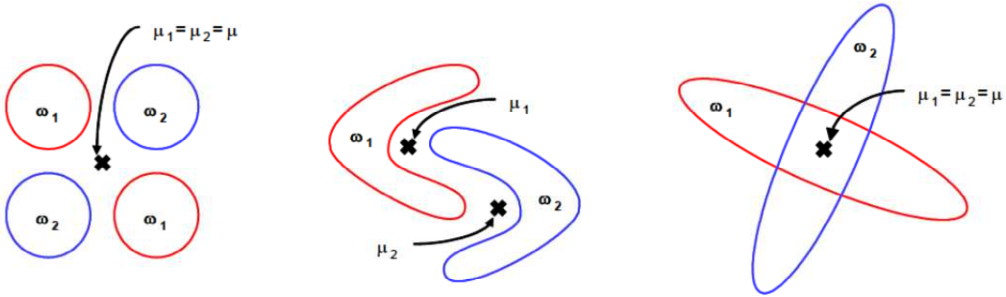
$$\mathbf{W} = \frac{1}{N-K}\sum_{k=1}^{K}(N_k - 1)\mathbf{S}_k$$

gdzie *K* to liczba klas.

Każda hiperpłaszczyzna klasyfikuje przykład do jednej klasy. Jako klasę przykładu wybiera się klasę większościową.

Ograniczenia

- Metoda Fishera zakłada, że macierze kowariancji wewnątrz klas są identyczne, a rozkłady klas są gaussowskie i jednomodalne:



- Metoda zawodzi, gdy więcej informacji o położeniu klas zawartych jest nie w ich środkach, ale w wariancji: