Statystyczna Eksploracja Danych

Wykład 6 - maszyny wektorów podpierających (SVM)

dr inż. Julian Sienkiewicz

4 kwietnia 2019

Cechy

Cechy

- maszyny wektorów podpierających - *support vector machines* (SVM),

Cechy

- maszyny wektorów podpierających - *support vector machines* (SVM),
- klasyfikacja pod nadzorem,

Cechy

- maszyny wektorów podpierających - *support vector machines* (SVM),
- klasyfikacja pod nadzorem,
- dość nowa metoda, zaproponowana przez Władimira Wapnika w latach 90-tych XX w.,

Cechy

- maszyny wektorów podpierających - *support vector machines* (SVM),
- klasyfikacja pod nadzorem,
- dość nowa metoda, zaproponowana przez Władimira Wapnika w latach 90-tych XX w.,
- nowe spojrzenie na zadanie wyboru najlepszej **hiperpłaszczyzny** dyskryminacyjnej,

Cechy

- maszyny wektorów podpierających - *support vector machines* (SVM),
- klasyfikacja pod nadzorem,
- dość nowa metoda, zaproponowana przez Władimira Wapnika w latach 90-tych XX w.,
- nowe spojrzenie na zadanie wyboru najlepszej **hiperpłaszczyzny** dyskryminacyjnej,
- następnie pomysł wzbogacenia przestrzeni obserwacji i szukania hiperpłaszczyzny dyskryminacyjnej w **nowej przestrzeni**,

Cechy

- maszyny wektorów podpierających - *support vector machines* (SVM),
- klasyfikacja pod nadzorem,
- dość nowa metoda, zaproponowana przez Władimira Wapnika w latach 90-tych XX w.,
- nowe spojrzenie na zadanie wyboru najlepszej **hiperpłaszczyzny** dyskryminacyjnej,
- następnie pomysł wzbogacenia przestrzeni obserwacji i szukania hiperpłaszczyzny dyskryminacyjnej w **nowej przestrzeni**,
- opiera się na rozwiązaniu prostej optymalizacji kwadratowo-liniowej,

Cechy

- maszyny wektorów podpierających - *support vector machines* (SVM),
- klasyfikacja pod nadzorem,
- dość nowa metoda, zaproponowana przez Władimira Wapnika w latach 90-tych XX w.,
- nowe spojrzenie na zadanie wyboru najlepszej **hiperpłaszczyzny** dyskryminacyjnej,
- następnie pomysł wzbogacenia przestrzeni obserwacji i szukania hiperpłaszczyzny dyskryminacyjnej w **nowej przestrzeni**,
- opiera się na rozwiązaniu prostej optymalizacji kwadratowo-liniowej,
- często wykorzystywana do:

Cechy

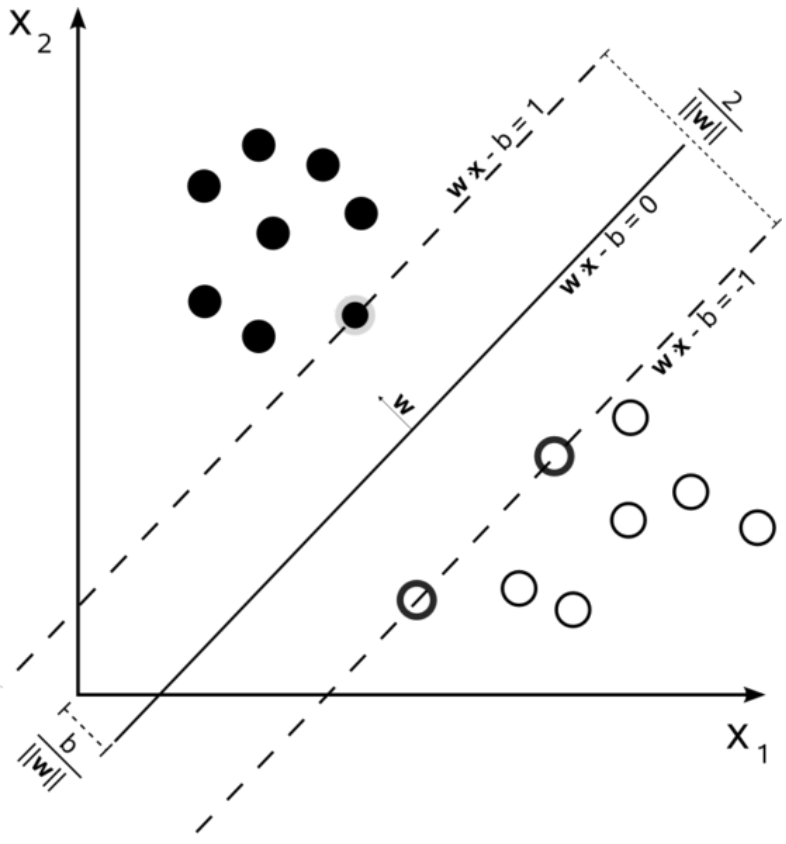
- maszyny wektorów podpierających - *support vector machines* (SVM),
- klasyfikacja pod nadzorem,
- dość nowa metoda, zaproponowana przez Władimira Wapnika w latach 90-tych XX w.,
- nowe spojrzenie na zadanie wyboru najlepszej **hiperpłaszczyzny** dyskryminacyjnej,
- następnie pomysł wzbogacenia przestrzeni obserwacji i szukania hiperpłaszczyzny dyskryminacyjnej w **nowej przestrzeni**,
- opiera się na rozwiązaniu prostej optymalizacji kwadratowo-liniowej,
- często wykorzystywana do: kategoryzacji tekstu,

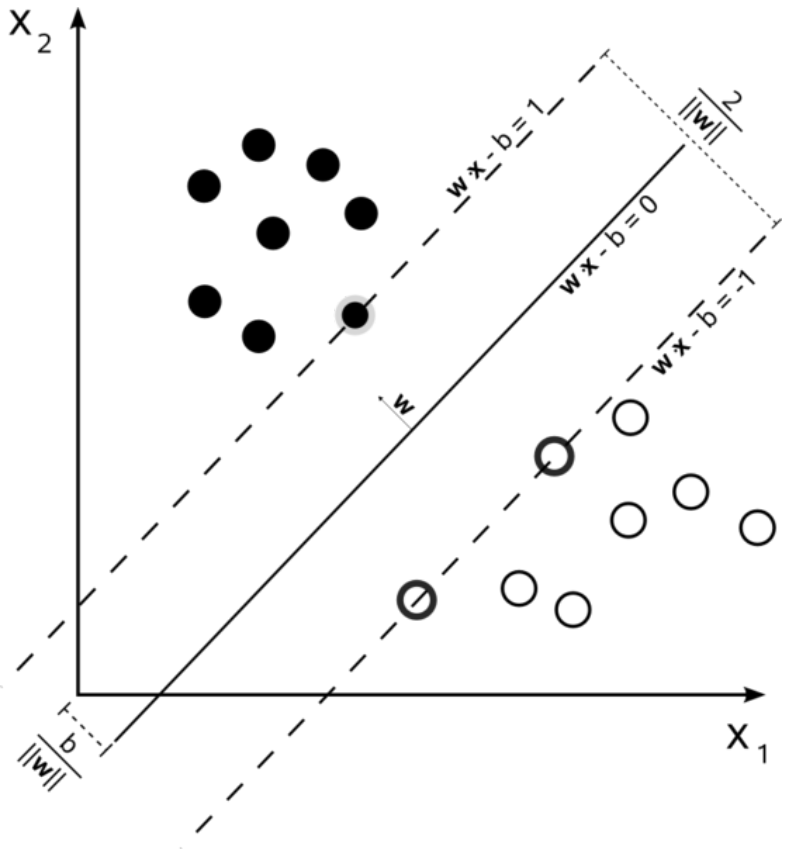
Cechy

- maszyny wektorów podpierających - *support vector machines* (SVM),
- klasyfikacja pod nadzorem,
- dość nowa metoda, zaproponowana przez Władimira Wapnika w latach 90-tych XX w.,
- nowe spojrzenie na zadanie wyboru najlepszej **hiperpłaszczyzny** dyskryminacyjnej,
- następnie pomysł wzbogacenia przestrzeni obserwacji i szukania hiperpłaszczyzny dyskryminacyjnej w **nowej przestrzeni**,
- opiera się na rozwiązaniu prostej optymalizacji kwadratowo-liniowej,
- często wykorzystywana do: kategoryzacji tekstu, klasyfikacji obrazów,

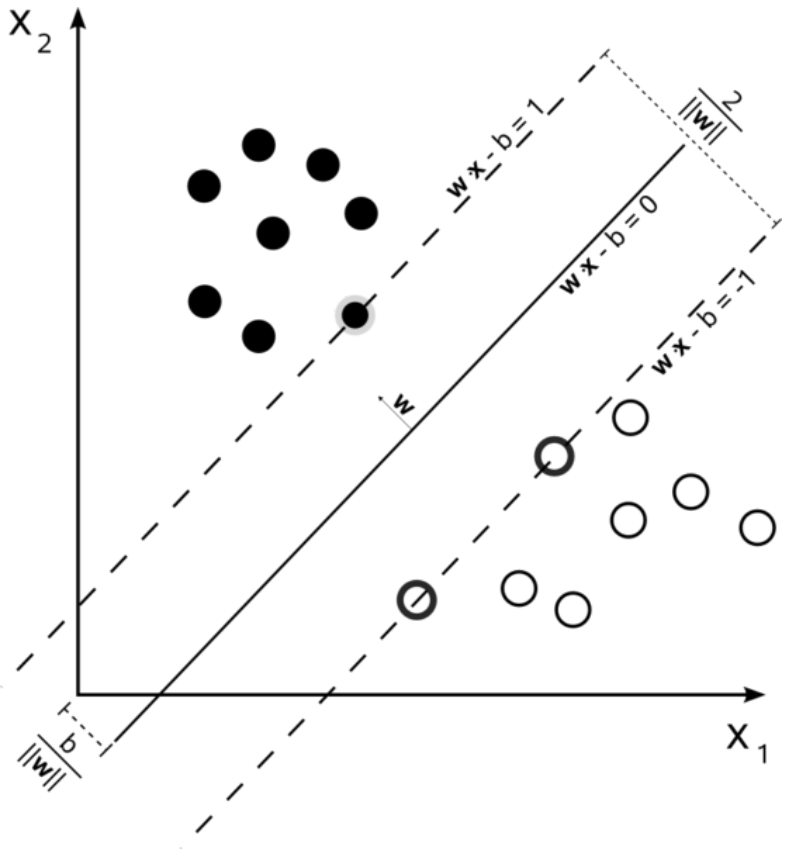
Cechy

- maszyny wektorów podpierających - *support vector machines* (SVM),
- klasyfikacja pod nadzorem,
- dość nowa metoda, zaproponowana przez Władimira Wapnika w latach 90-tych XX w.,
- nowe spojrzenie na zadanie wyboru najlepszej **hiperpłaszczyzny** dyskryminacyjnej,
- następnie pomysł wzbogacenia przestrzeni obserwacji i szukania hiperpłaszczyzny dyskryminacyjnej w **nowej przestrzeni**,
- opiera się na rozwiązaniu prostej optymalizacji kwadratowo-liniowej,
- często wykorzystywana do: kategoryzacji tekstu, klasyfikacji obrazów, rozpoznawania pisma odręcznego.

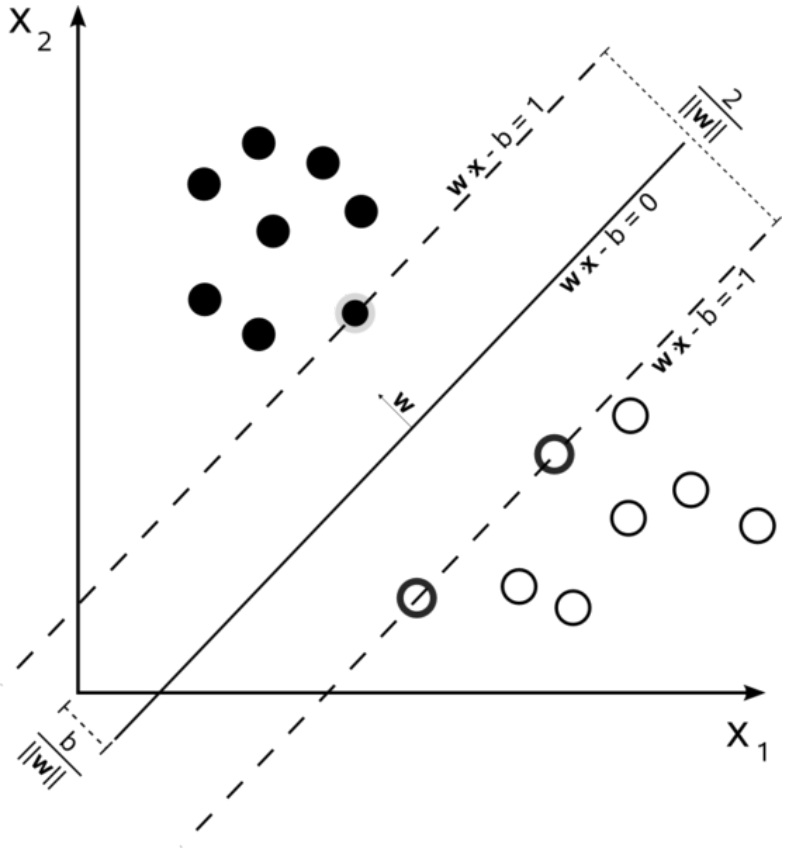

X2
X1
wx - b = 1
wx - b = 0
wx - b = -1
2/||w||
w
b/||w||



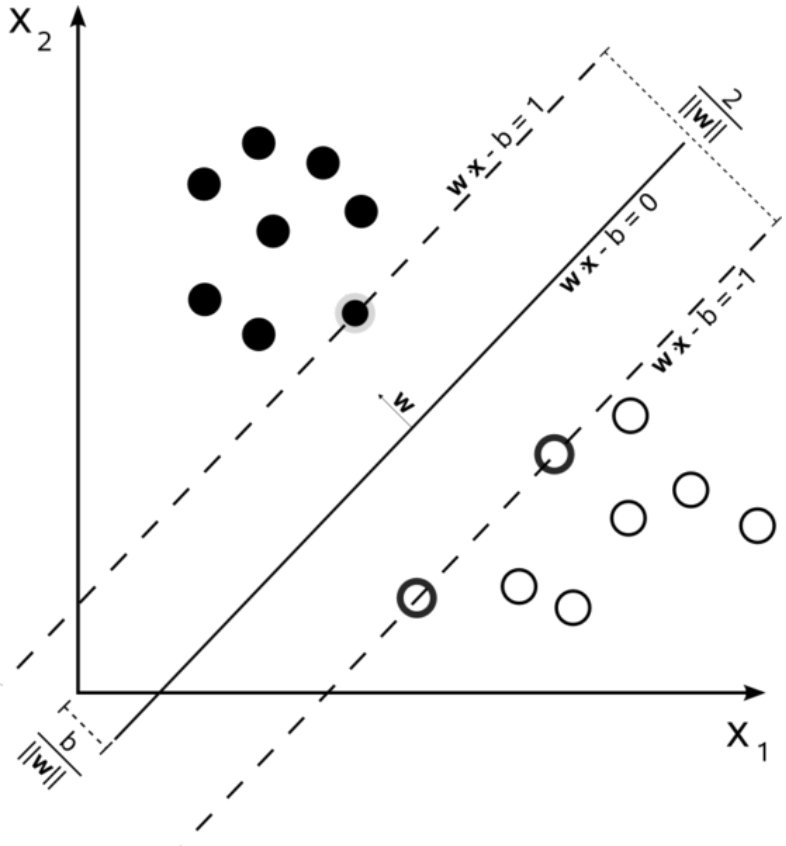
- rozważamy zadanie analizy dyskryminacyjnej w $\mathbb{R}^p$ z $g = 2$ (liczba klas),

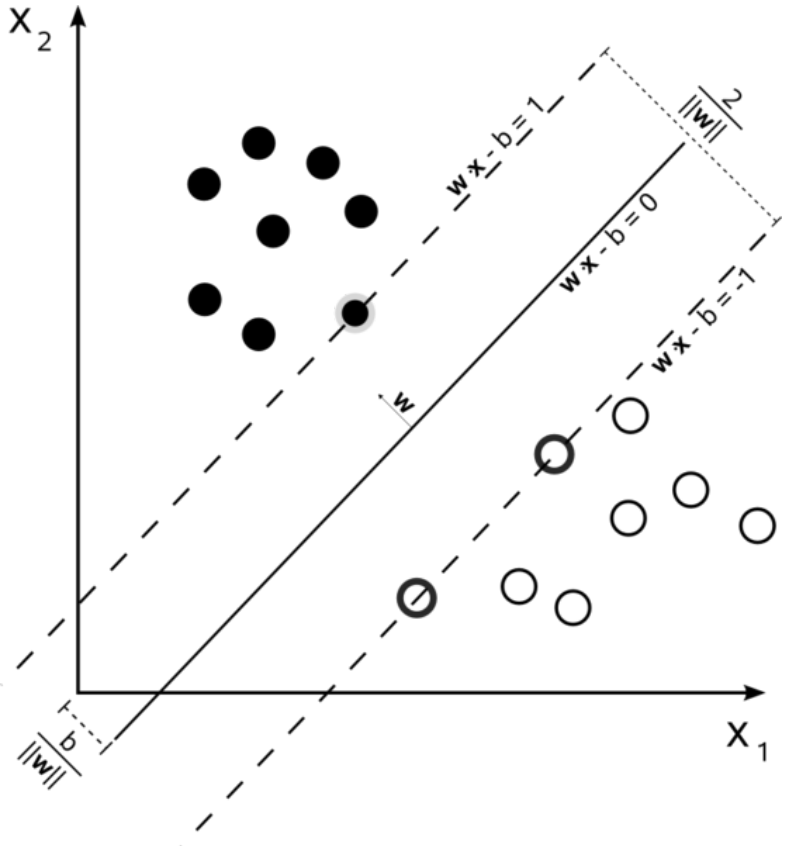


- rozważamy zadanie analizy dyskryminacyjnej w $\mathbb{R}^p$ z $g = 2$ (liczba klas),
- zakładamy, że obie podpróby są **liniowo separowalne**, to znaczy, że można je idealnie rozdzielić hiperpłaszczyzną dyskryminacyjną,

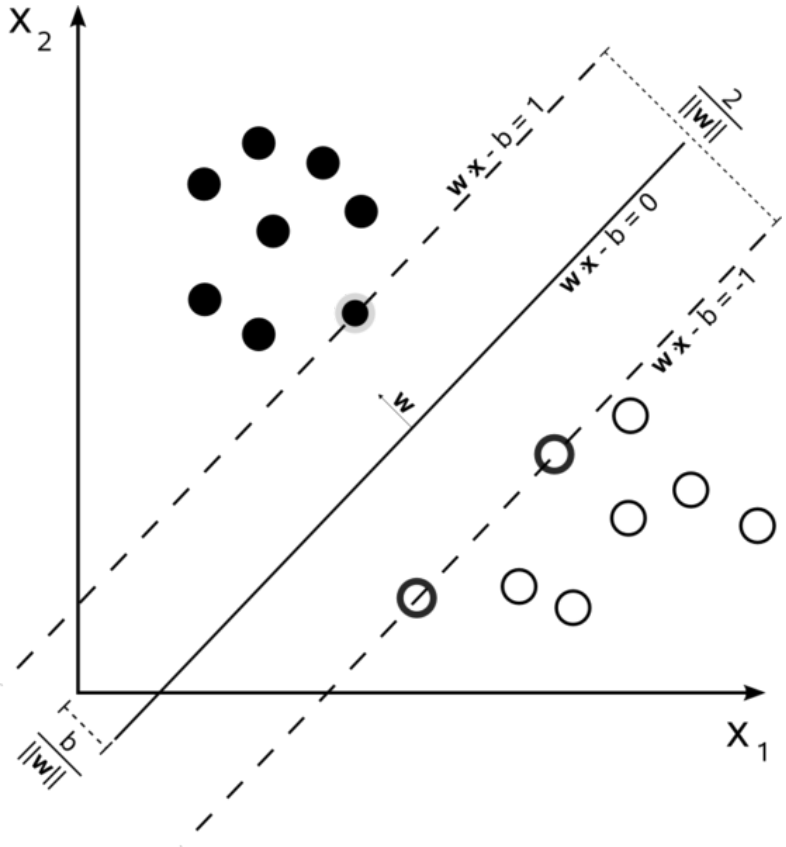


- rozważamy zadanie analizy dyskryminacyjnej w $\mathbb{R}^p$ z $g = 2$ (liczba klas),
- zakładamy, że obie podpróby są **liniowo separowalne**, to znaczy, że można je idealnie rozdzielić hiperpłaszczyzną dyskryminacyjną,
- implikuje to istnienie **marginesów** ograniczonych dwiema równoległymi hiperpłaszczyznami, wewnątrz których nie leży **ani jeden** element próby uczącej,

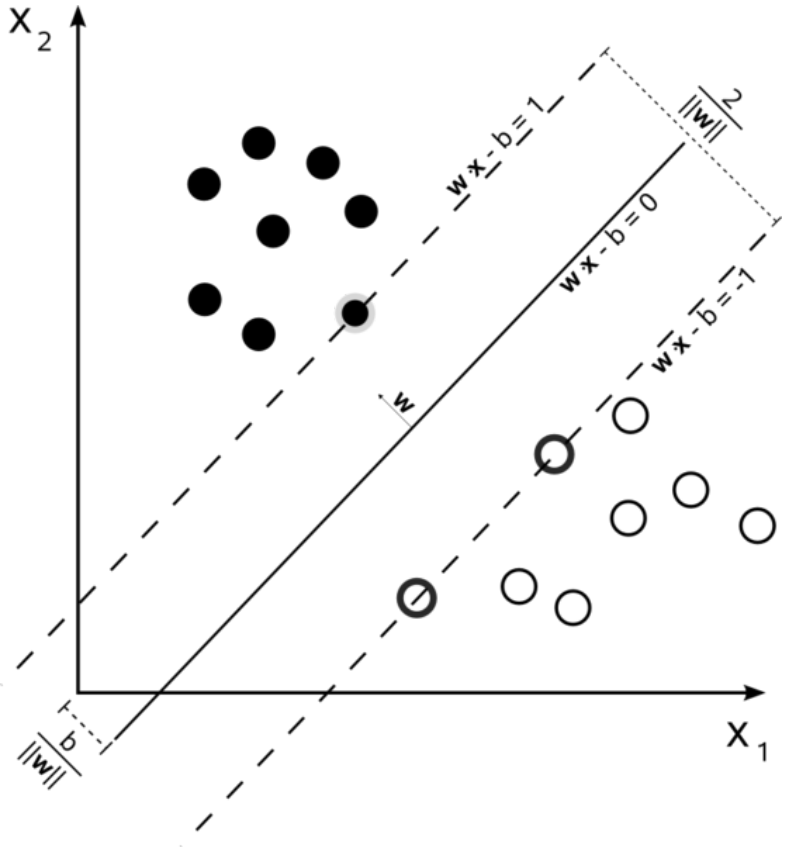

X2
X1
w x - b = 1
w x - b = 0
w x - b = -1
2/||w||
w
b/||w||



- zadanie optymalizacji polega na znalezieniu najszerszego możliwego marginesu,

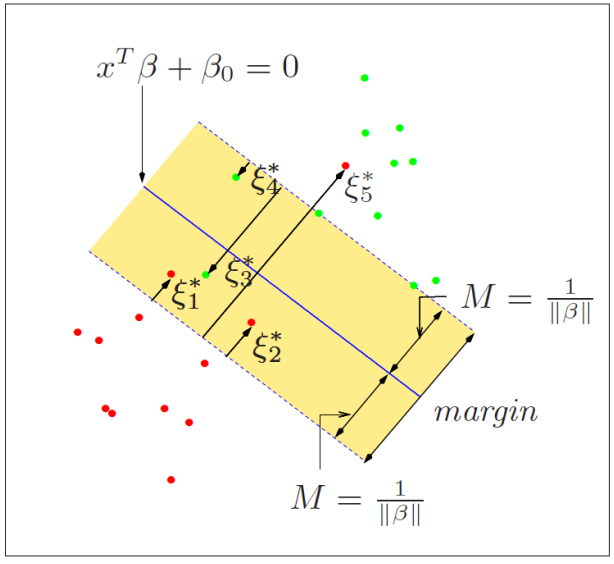


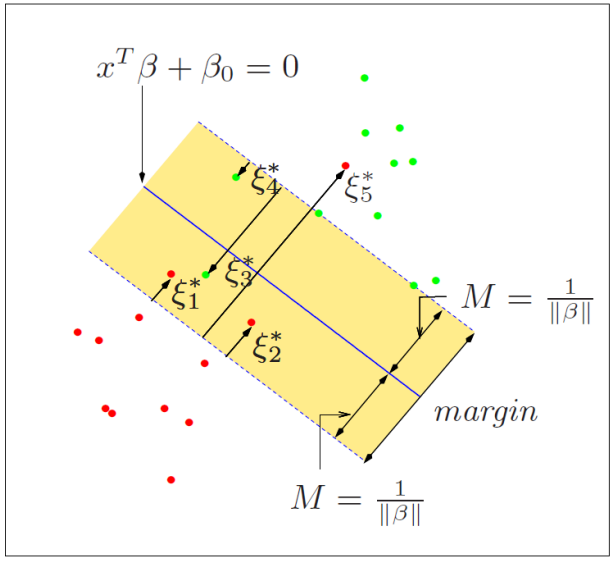
- zadanie optymalizacji polega na znalezieniu najszerszego możliwego marginesu,
- pośrodku marginesu umieszcza się hiperpłaszyznę dyskryminacyjną,

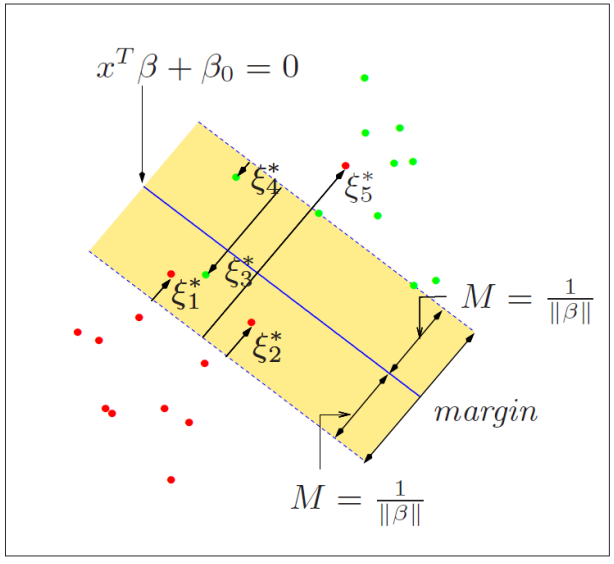


- zadanie optymalizacji polega na znalezieniu najszerszego możliwego marginesu,
- pośrodku marginesu umieszcza się hiperpłaszyznę dyskryminacyjną,
- nazwa SVM ma swoje źródło w tym, że hiperpłaszczyzny marginesów muszą przechodzić przez konkretne elementy prób uczących (inaczej margines można byłoby rozszerzyć) - są to właśnie **wektory podpierające**.

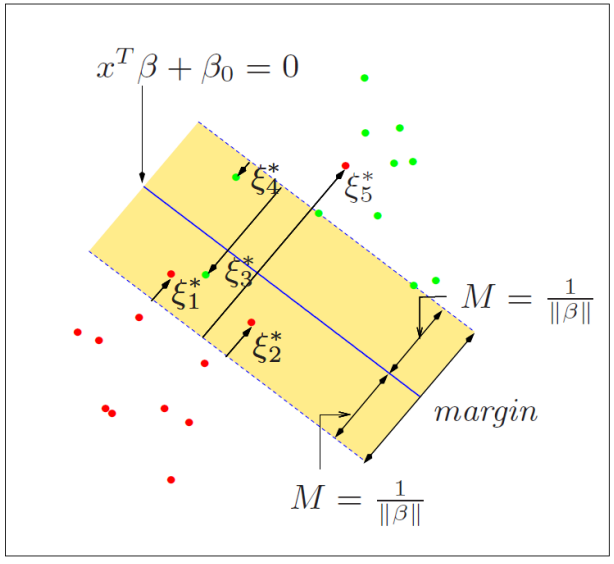
Ogólny opis


$x^T\beta + \beta_0 = 0$
ξ_4^*
ξ_5^*
ξ_3^*
ξ_1^*
ξ_2^*
$M = \frac{1}{\|\beta\|}$
$margin$
$M = \frac{1}{\|\beta\|}$

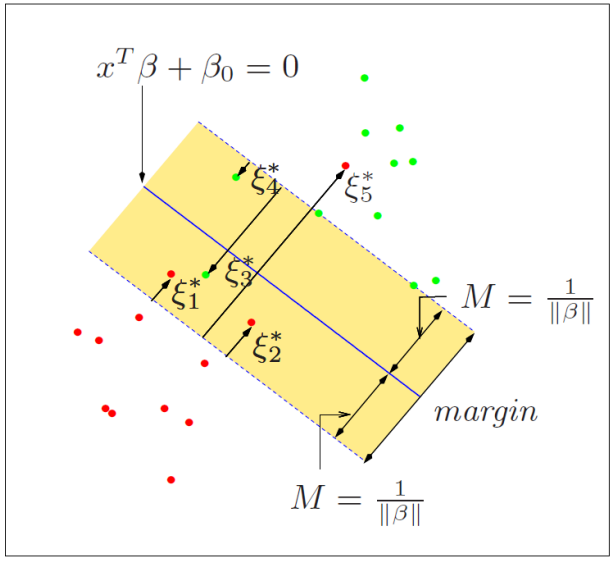

$x^T\beta + \beta_0 = 0$
ξ_4^*
ξ_5^*
ξ_3^*
ξ_1^*
ξ_2^*
$M = \frac{1}{\|\beta\|}$
$margin$
$M = \frac{1}{\|\beta\|}$



W przypadku, gdy próby nie są liniowo separowalne, wprowadza się karę (podobnie jak to jest w przypadku funkcji celu w algorytmach genetycznych) za nieidealne rozdzielenie podprób.

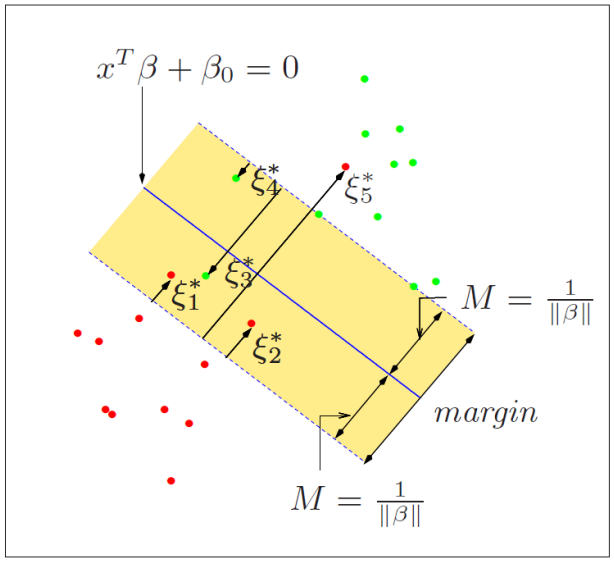


W przypadku, gdy próby nie są liniowo separowalne, wprowadza się karę (podobnie jak to jest w przypadku funkcji celu w algorytmach genetycznych) za nieidealne rozdzielenie podprób.



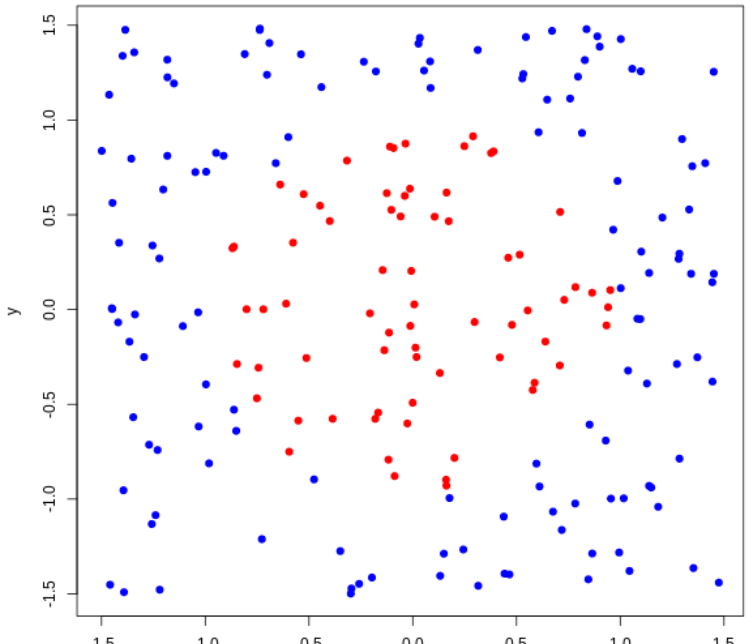
W przypadku, gdy próby nie są liniowo separowalne, wprowadza się karę (podobnie jak to jest w przypadku funkcji celu w algorytmach genetycznych) za nieidealne rozdzielenie podprób.

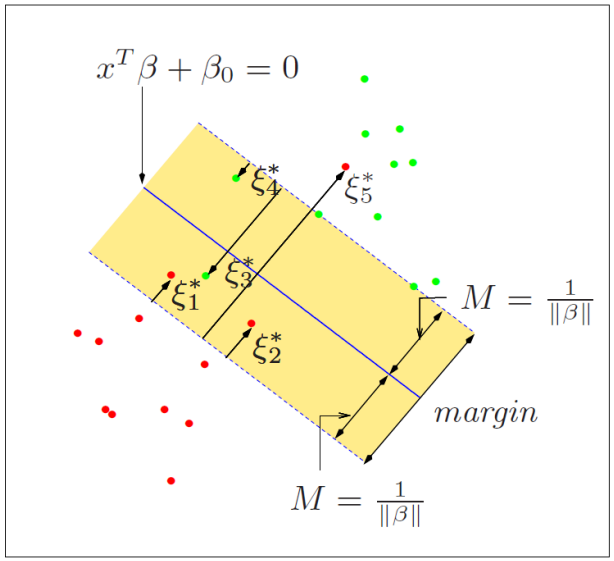
Aby dokonać jak najlepszego rozdzielenia, zadania często rozwiązuje się w przestrzeni o znacznie większym wymiarze niż *p* (ponieważ hiperpłaszczyzny są opisywane iloczynem skalarnym, można wybrać inne **jądro**).



W przypadku, gdy próby nie są liniowo separowalne, wprowadza się karę (podobnie jak to jest w przypadku funkcji celu w algorytmach genetycznych) za nieidealne rozdzielenie podprób.

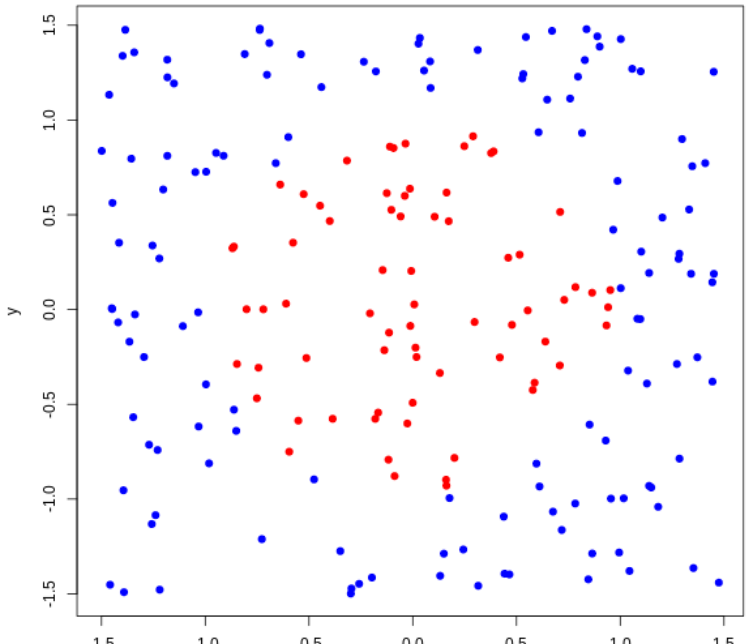
Aby dokonać jak najlepszego rozdzielenia, zadania często rozwiązuje się w przestrzeni o znacznie większym wymiarze niż *p* (ponieważ hiperpłaszczyzny są opisywane iloczynem skalarnym, można wybrać inne **jądro**).





W przypadku, gdy próby nie są liniowo separowalne, wprowadza się karę (podobnie jak to jest w przypadku funkcji celu w algorytmach genetycznych) za nieidealne rozdzielenie podprób.

Aby dokonać jak najlepszego rozdzielenia, zadania często rozwiązuje się w przestrzeni o znacznie większym wymiarze niż *p* (ponieważ hiperpłaszczyzny są opisywane iloczynem skalarnym, można wybrać inne **jądro**).



Rozpoczynamy od rozważań, dotyczących najprostszego przypadku: liniowej separowalności klas.

Rozpoczynamy od rozważań, dotyczących najprostszego przypadku: liniowej separowalności klas. Zakładamy, że mamy do czynienia z problemem dwuklasowym ($g = 2$). Mamy daną próbę uczącą:

Rozpoczynamy od rozważań, dotyczących najprostszego przypadku: liniowej separowalności klas. Zakładamy, że mamy do czynienia z problemem dwuklasowym ($g = 2$). Mamy daną próbę uczącą:

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n),$$

Rozpoczynamy od rozważań, dotyczących najprostszego przypadku: liniowej separowalności klas. Zakładamy, że mamy do czynienia z problemem dwuklasowym ($g = 2$). Mamy daną próbę uczącą:

$$(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n), \quad y_i \in \{-1, 1\},$$

Rozpoczynamy od rozważań, dotyczących najprostszego przypadku: liniowej separowalności klas. Zakładamy, że mamy do czynienia z problemem dwuklasowym ($g = 2$). Mamy daną próbę uczącą:

$$(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n), \quad y_i \in \{-1, 1\}, \quad \mathbf{x}_i \in \mathbb{R}^p,$$

Rozpoczynamy od rozważań, dotyczących najprostszego przypadku: liniowej separowalności klas. Zakładamy, że mamy do czynienia z problemem dwuklasowym ($g = 2$). Mamy daną próbę uczącą:

$$(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n), \quad y_i \in \{-1, 1\}, \quad \mathbf{x}_i \in \mathbb{R}^p,$$

przy czym klasy zostały zakodowane jako 1 oraz -1.

Rozpoczynamy od rozważań, dotyczących najprostszego przypadku: liniowej separowalności klas. Zakładamy, że mamy do czynienia z problemem dwuklasowym ($g = 2$). Mamy daną próbę uczącą:

$$(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n), \quad y_i \in \{-1, 1\}, \quad \mathbf{x}_i \in \mathbb{R}^p,$$

przy czym klasy zostały zakodowane jako 1 oraz -1. Zakładamy, że istnieje hiperpłaszczyzna punktów $\mathbf{x}$ w $\mathbb{R}^p$, rozdzielająca klasy

Rozpoczynamy od rozważań, dotyczących najprostszego przypadku: liniowej separowalności klas. Zakładamy, że mamy do czynienia z problemem dwuklasowym ($g = 2$). Mamy daną próbę uczącą:

$$(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n), \quad y_i \in \{-1, 1\}, \quad \mathbf{x}_i \in \mathbb{R}^p,$$

przy czym klasy zostały zakodowane jako 1 oraz -1. Zakładamy, że istnieje hiperpłaszczyzna punktów $\mathbf{x}$ w $\mathbb{R}^p$, rozdzielająca klasy

$$\mathbf{w}^T\mathbf{x} + b \equiv \mathbf{w} \cdot \mathbf{x} + b = 0$$

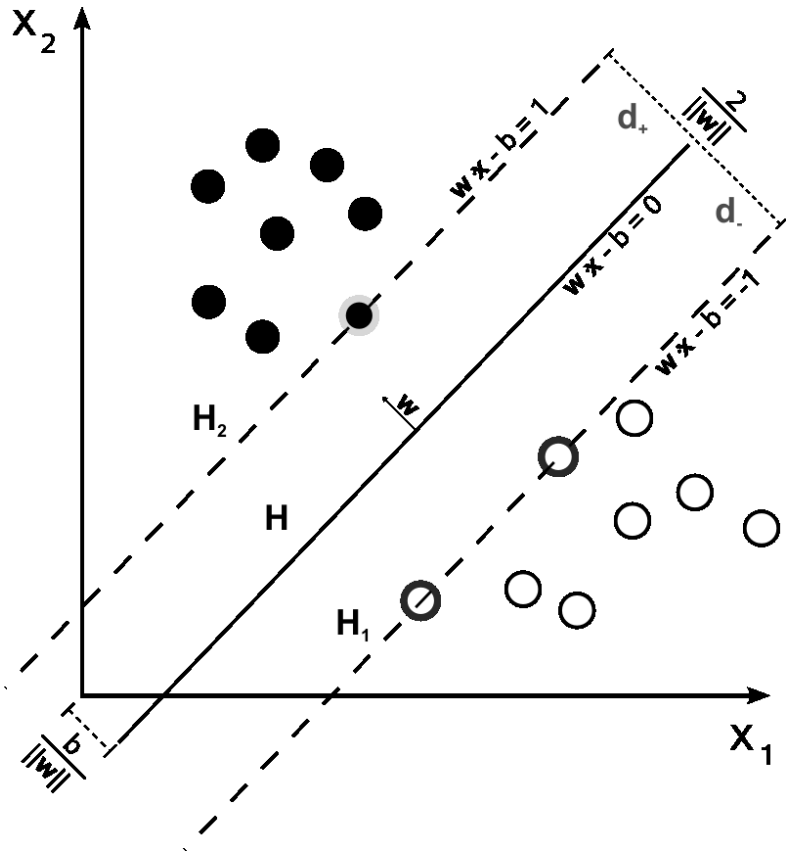
Rozpoczynamy od rozważań, dotyczących najprostszego przypadku: liniowej separowalności klas. Zakładamy, że mamy do czynienia z problemem dwuklasowym ($g = 2$). Mamy daną próbę uczącą:

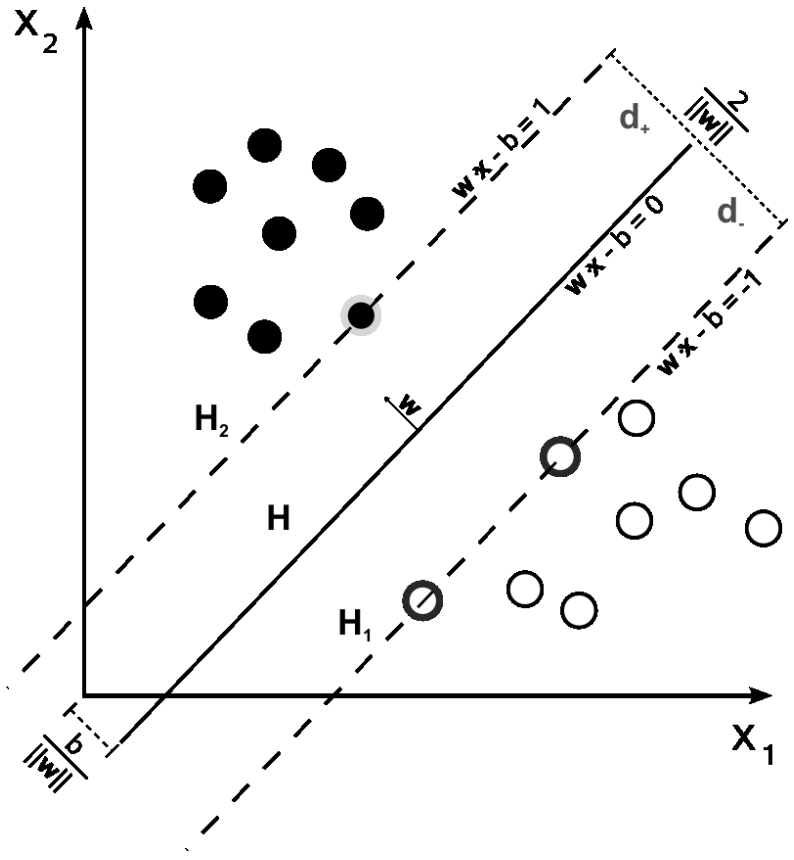
$$(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n), \quad y_i \in \{-1, 1\}, \quad \mathbf{x}_i \in \mathbb{R}^p,$$

przy czym klasy zostały zakodowane jako 1 oraz -1. Zakładamy, że istnieje hiperpłaszczyzna punktów $\mathbf{x}$ w $\mathbb{R}^p$, rozdzielająca klasy

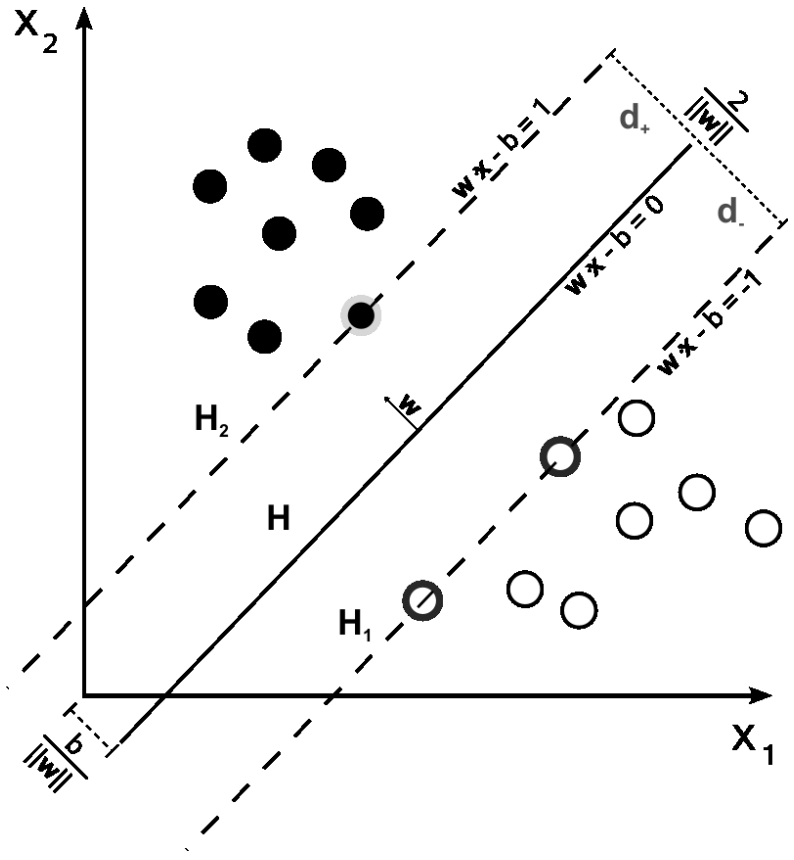
$$\mathbf{w}^T\mathbf{x} + b \equiv \mathbf{w} \cdot \mathbf{x} + b = 0$$

gdzie $\cdot$ oznacza iloczyn skalarny, a wektor $\mathbf{w}$ oraz stała b są odpowiednio dobrane.

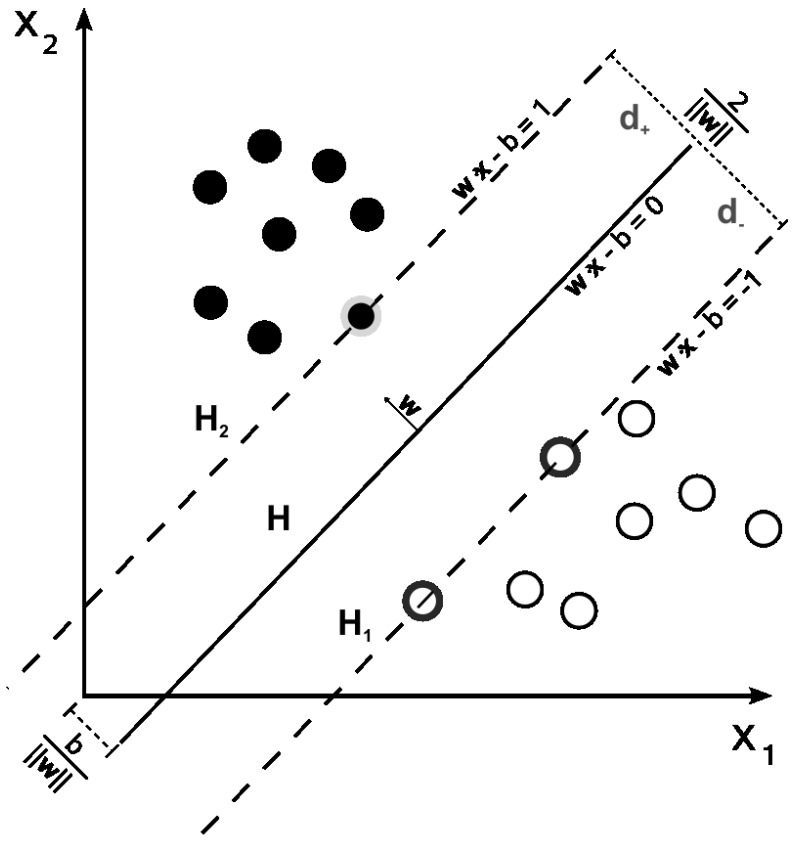

X2
X1
wx - b = 1
wx - b = 0
wx - b = -1
d+
d-
2/||w||
H1
H2
H
w
b/||w||

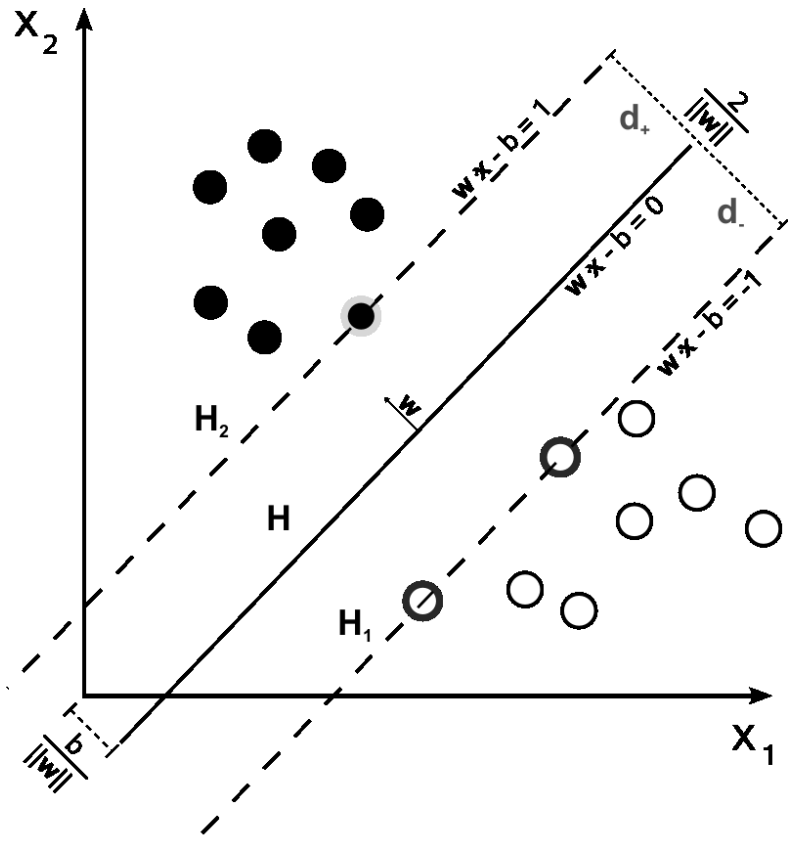


Ze względu na liniową separowalność, hiperpłaszczyzna dyskryminująca (tu: prosta) H leży w pasie ograniczonym dwiema hiperpłaszczyznami H_1 i H_2, wewnątrz którego nie ma żadnego elementu próby uczącej.

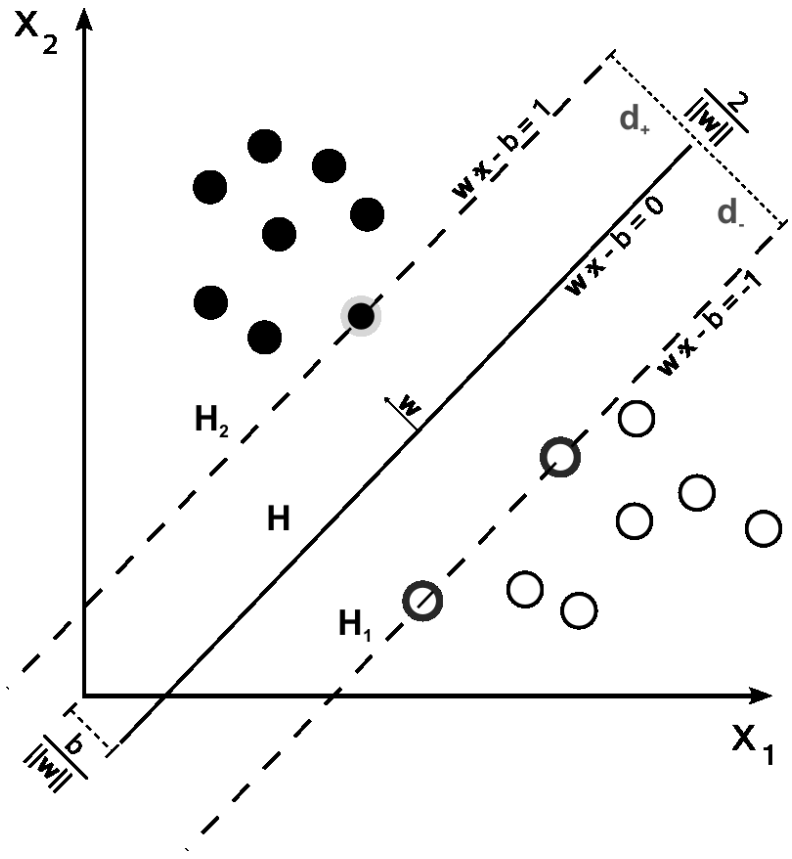


Ze względu na liniową separowalność, hiperpłaszczyzna dyskryminująca (tu: prosta) H leży w pasie ograniczonym dwiema hiperpłaszczyznami H_1 i H_2, wewnątrz którego nie ma żadnego elementu próby uczącej.
Aby uczynić ten pas (marginines) jak największym, trzeba oprzeć jego brzegi o punkty próby uczącej.


X2
X1
wx - b = 1
wx - b = 0
wx - b = -1
d+
d-
2/||w||
H2
H
H1
w
b/||w||

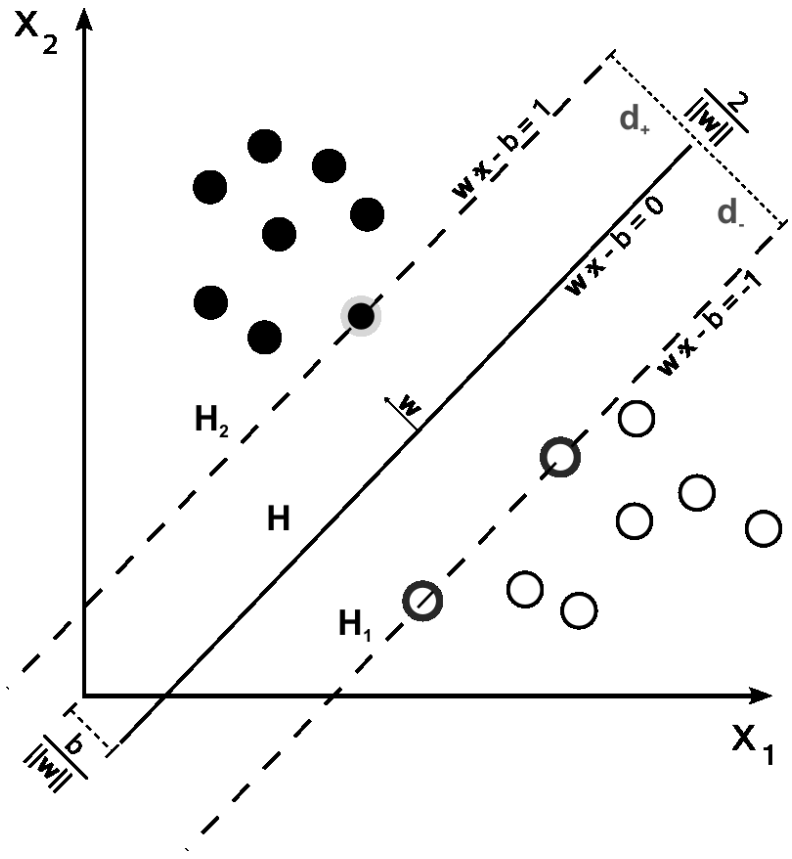


Dla każdej obserwacji z próby uczącej $\mathbf{x}_i$, $i = 1, ..., n$ jest spełniona jedna z nierówności



Dla każdej obserwacji z próby uczącej $\mathbf{x}_i$, $i = 1, ..., n$ jest spełniona jedna z nierówności

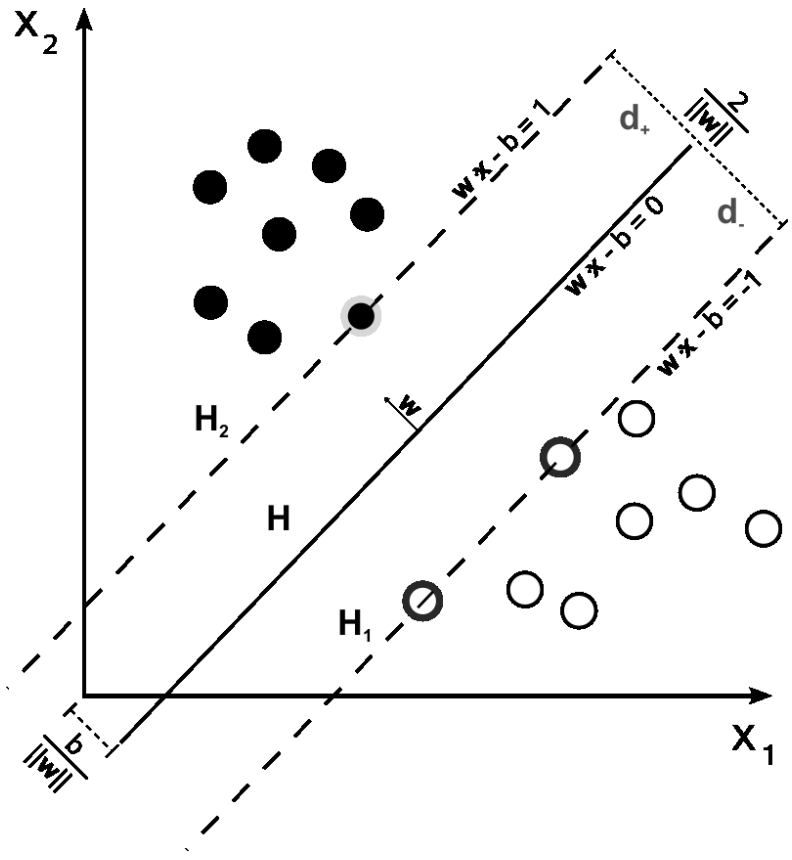
$$\mathbf{x}_i \cdot \mathbf{w} + b \geq +1, \quad \text{gdy} \quad y_i = +1,$$



Dla każdej obserwacji z próby uczącej $\mathbf{x}_i$, $i = 1, ..., n$ jest spełniona jedna z nierówności

$$\mathbf{x}_i \cdot \mathbf{w} + b \geq +1, \quad \text{gdy} \quad y_i = +1,$$

$$\mathbf{x}_i \cdot \mathbf{w} + b \leq -1, \quad \text{gdy} \quad y_i = -1,$$

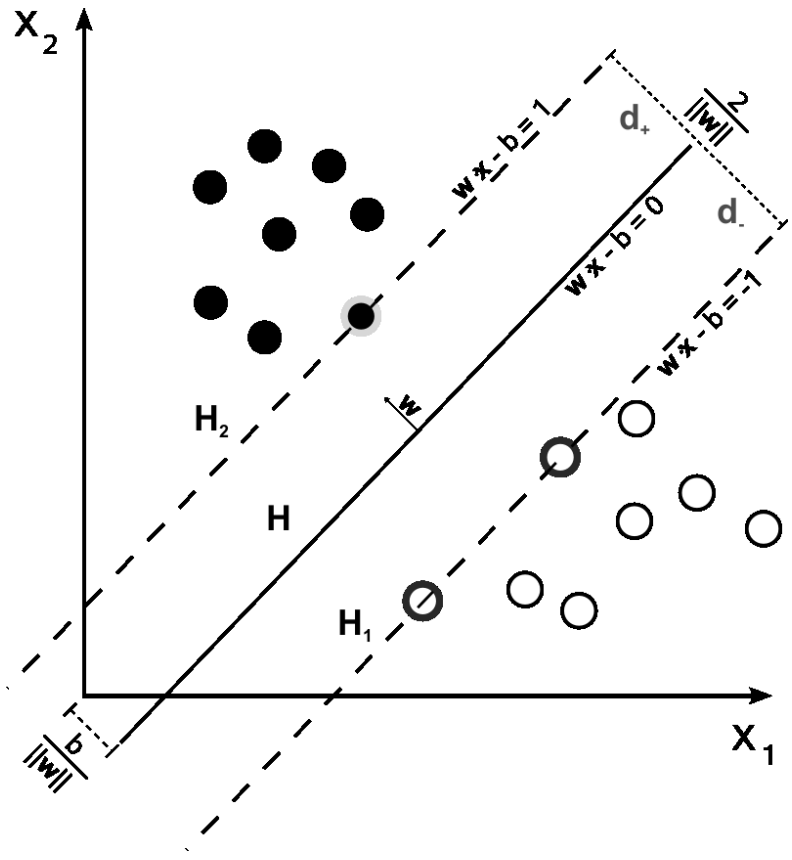


Dla każdej obserwacji z próby uczącej $\mathbf{x}_i$, $i = 1, ..., n$ jest spełniona jedna z nierówności

$$\mathbf{x}_i \cdot \mathbf{w} + b \geq +1, \quad \text{gdy} \quad y_i = +1,$$

$$\mathbf{x}_i \cdot \mathbf{w} + b \leq -1, \quad \text{gdy} \quad y_i = -1,$$

Dla wektorów leżących na H_1 i H_2 przechodzą one w równości.



Dla każdej obserwacji z próby uczącej $\mathbf{x}_i$, $i = 1, ..., n$ jest spełniona jedna z nierówności

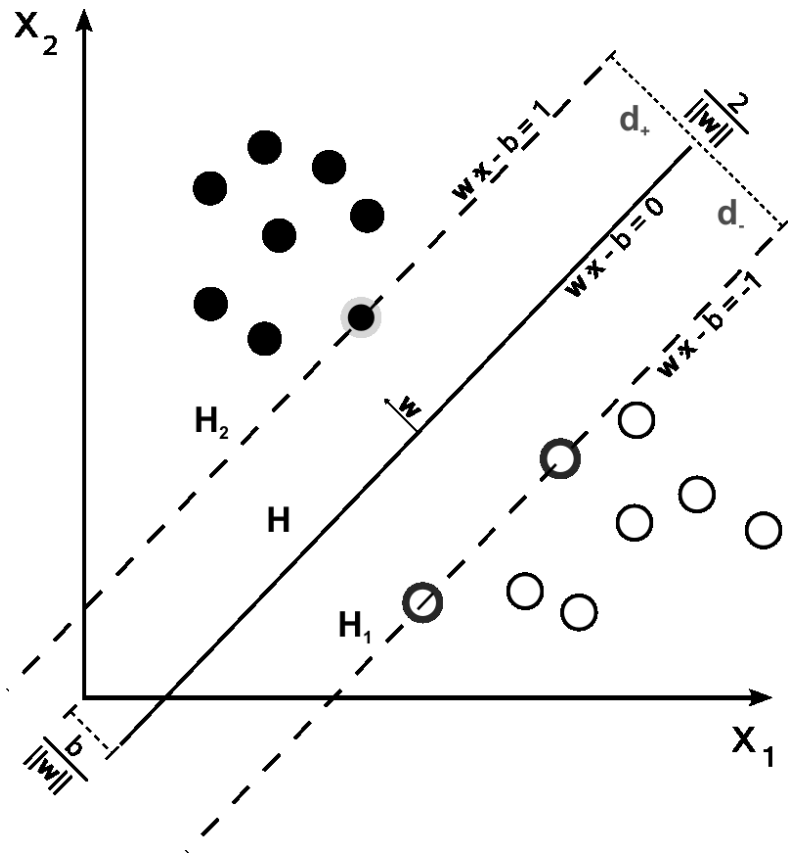
$$\mathbf{x}_i \cdot \mathbf{w} + b \geq +1, \quad \text{gdy} \quad y_i = +1,$$

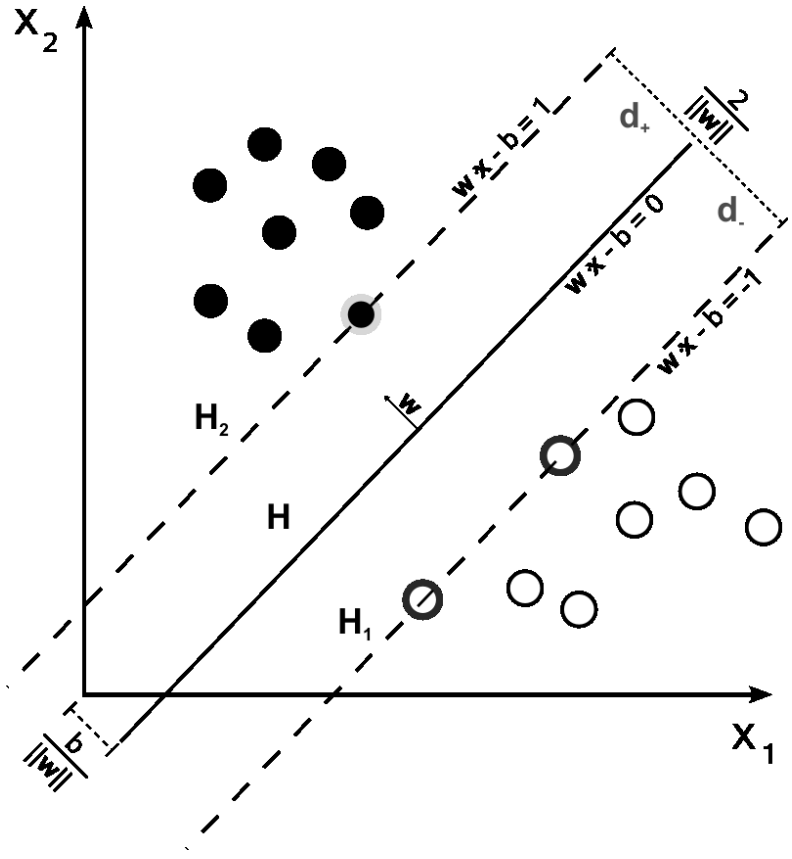
$$\mathbf{x}_i \cdot \mathbf{w} + b \leq -1, \quad \text{gdy} \quad y_i = -1,$$

Dla wektorów leżących na H_1 i H_2 przechodzą one w równości.

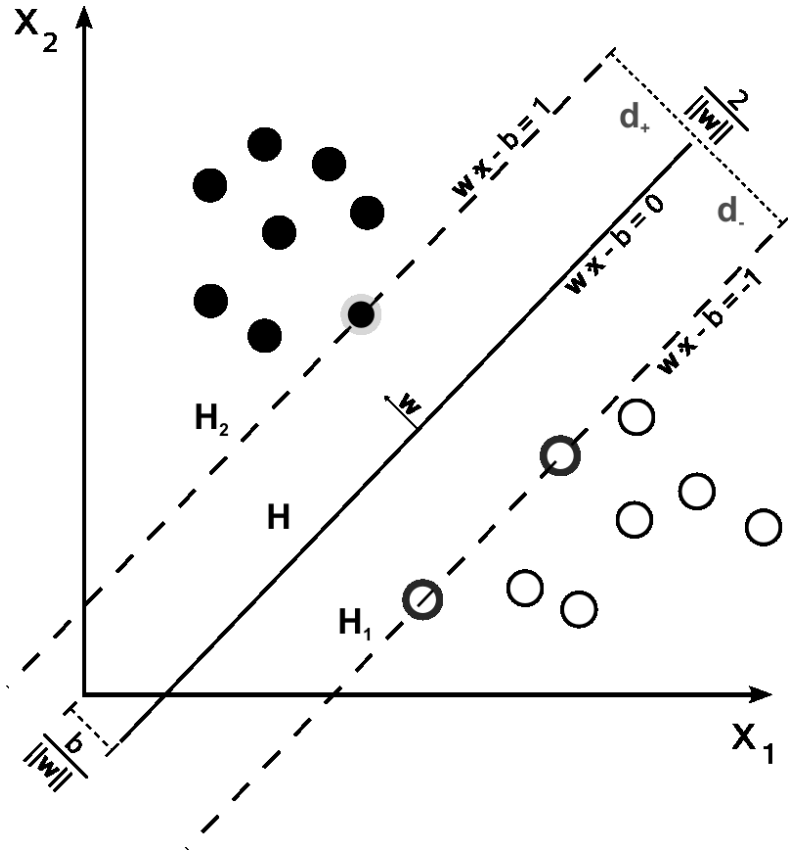
Można je zapisać jako jedną nierówność dla wszystkich $\mathbf{x}_i$ jako

$$y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \geq 0$$


X2
X1
wx - b = 1
wx - b = 0
wx - b = -1
d+
d-
2/||w||
H2
H
H1
w
b/||w||



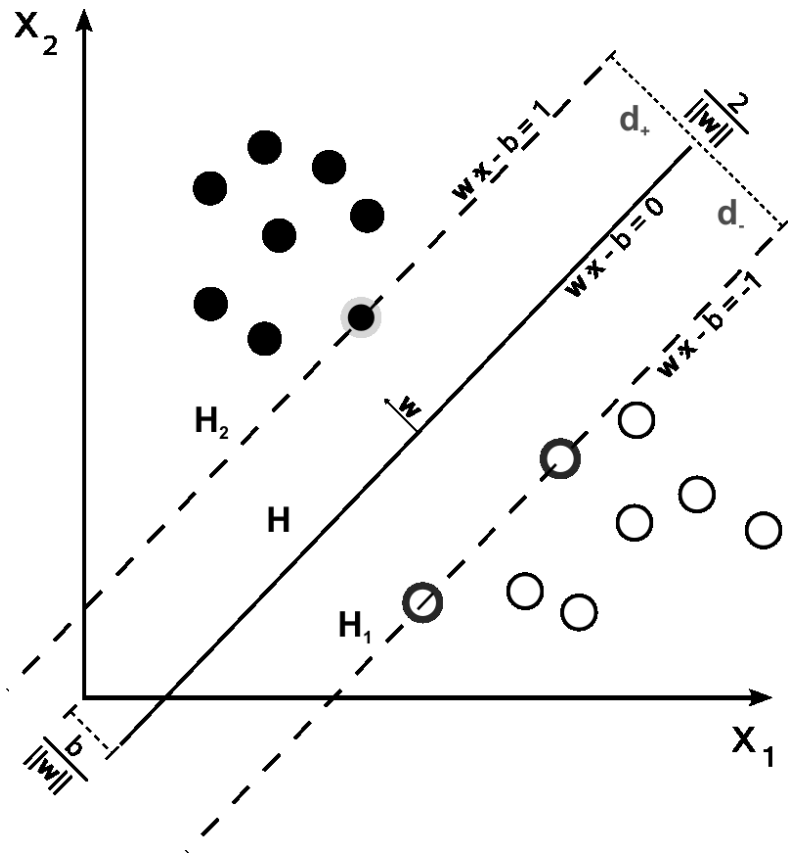
H_1 jest odległa od początku układu współrzędnych o



H_1 jest odległa od początku układu współrzędnych o

$$\frac{|1-b|}{||\mathbf{w}||}$$

natomiast H_2 jest odległa od początku układu współrzędnych o



H_1 jest odległa od początku układu współrzędnych o

$$\frac{|1-b|}{||\mathbf{w}||}$$

natomiast H_2 jest odległa od początku układu współrzędnych o

$$\frac{|-1-b|}{||\mathbf{w}||}$$

Stąd odległość pomiędzy H_1 i H_2 to

$$d_+ + d_- = \frac{2}{||\mathbf{w}||}$$

W efekcie zadanie znalezienia optymalnego położenia hiperpłaszczyzny H sprowadza się do maksymalizacji wyrażenia

W efekcie zadanie znalezienia optymalnego położenia hiperpłaszczyzny H sprowadza się do maksymalizacji wyrażenia

$$d_+ + d_- = \frac{2}{||\mathbf{w}||}$$

W efekcie zadanie znalezienia optymalnego położenia hiperpłaszczyzny H sprowadza się do maksymalizacji wyrażenia

$$d_+ + d_- = \frac{2}{||\mathbf{w}||}$$

lub, co jest równoważne, do minimalizacji

W efekcie zadanie znalezienia optymalnego położenia hiperpłaszczyzny H sprowadza się do maksymalizacji wyrażenia

$$d_+ + d_- = \frac{2}{||\mathbf{w}||}$$

lub, co jest równoważne, do minimalizacji

$$\frac{||\mathbf{w}||^2}{2}$$

W efekcie zadanie znalezienia optymalnego położenia hiperpłaszczyzny H sprowadza się do maksymalizacji wyrażenia

$$d_+ + d_- = \frac{2}{||\mathbf{w}||}$$

lub, co jest równoważne, do minimalizacji

$$\frac{||\mathbf{w}||^2}{2}$$

przy danych ograniczeniach

W efekcie zadanie znalezienia optymalnego położenia hiperpłaszczyzny H sprowadza się do maksymalizacji wyrażenia

$$d_+ + d_- = \frac{2}{||\mathbf{w}||}$$

lub, co jest równoważne, do minimalizacji

$$\frac{||\mathbf{w}||^2}{2}$$

przy danych ograniczeniach

$$y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \geq 0$$

W efekcie zadanie znalezienia optymalnego położenia hiperpłaszczyzny H sprowadza się do maksymalizacji wyrażenia

$$d_+ + d_- = \frac{2}{||\mathbf{w}||}$$

lub, co jest równoważne, do minimalizacji

$$\frac{||\mathbf{w}||^2}{2}$$

przy danych ograniczeniach

$$y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \geq 0$$

Optymalna hiperpłaszczyzna dyskryminacyjna będzie umieszczona w środku, tzn tak, aby $d_+ = d_-$.

Jest to zadanie minimalizacji funkcjonału, które można zapisać w formie funkcji Lagrange'a:

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}(\mathbf{w} \cdot \mathbf{w}) - \sum_{i=1}^{n} \alpha_i \{[\mathbf{x}_i \cdot \mathbf{w} + b] y_i - 1\}$$

Jest to zadanie minimalizacji funkcjonału, które można zapisać w formie funkcji Lagrange'a:

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}(\mathbf{w} \cdot \mathbf{w}) - \sum_{i=1}^{n} \alpha_i \{[\mathbf{x}_i \cdot \mathbf{w} + b] y_i - 1\}$$

gdzie $\boldsymbol{\alpha}$ to wektor nieujemnych współczynników Lagrange'a. Szukamy maksimum funkcji względem α_i i minimum względem $\mathbf{w}$ i b.

Jest to zadanie minimalizacji funkcjonału, które można zapisać w formie funkcji Lagrange'a:

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}(\mathbf{w} \cdot \mathbf{w}) - \sum_{i=1}^{n} \alpha_i \{[\mathbf{x}_i \cdot \mathbf{w} + b] y_i - 1\}$$

gdzie $\boldsymbol{\alpha}$ to wektor nieujemnych współczynników Lagrange'a. Szukamy maksimum funkcji względem α_i i minimum względem $\mathbf{w}$ i b. Żądamy więc, aby pochodne L względem $\mathbf{w}$ oraz b zerowały się oraz, aby został spełniony warunek

$$\sum_{i=1}^{n} \alpha_i \{[\mathbf{x}_i \cdot \mathbf{w_0} + b_0] y_i - 1\} = 0$$

Zerowanie się pochodnej (gradientu) względem **w** daje

$$\mathbf{w} = \sum_{i=1}^{n} y_i \alpha_i \mathbf{x}_i$$

Zerowanie się pochodnej (gradientu) względem **w** daje

$$\mathbf{w} = \sum_{i=1}^{n} y_i \alpha_i \mathbf{x}_i$$

a względem b

$$\sum_{i=1}^{n} y_i \alpha_i = 0$$

Uwzględniając powyższe mamy funkcję Lagrange'a

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

Zerowanie się pochodnej (gradientu) względem **w** daje

$$\mathbf{w} = \sum_{i=1}^{n} y_i \alpha_i \mathbf{x}_i$$

a względem b

$$\sum_{i=1}^{n} y_i \alpha_i = 0$$

Uwzględniając powyższe mamy funkcję Lagrange'a

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

maksymalizowaną przy ograniczeniach

$$\alpha_i \geq 0 \quad i = 1, ..., n \quad \sum_{i=1}^{n} y_i \alpha_i = 0.$$

Bardzo istotnym jest fakt, iż warunki

$$\sum_{i=1}^{n} \alpha_i \{[\mathbf{x}_i \cdot \mathbf{w_0} + b_0] y_i - 1\} = 0$$

oznaczają, że **nie zerują się** tylko te współczynniki α_i, które odpowiadają wektorom podpierającym!

Bardzo istotnym jest fakt, iż warunki

$$\sum_{i=1}^{n} \alpha_i \{[\mathbf{x}_i \cdot \mathbf{w_0} + b_0] y_i - 1\} = 0$$

oznaczają, że **nie zerują się** tylko te współczynniki α_i, które odpowiadają wektorom podpierającym! W efekcie wszystkie sumowania wykonywane są **tylko** po tych i, którym odpowiadają wektory leżące na H_1 i H_2.

Bardzo istotnym jest fakt, iż warunki

$$\sum_{i=1}^{n} \alpha_i \{[\mathbf{x}_i \cdot \mathbf{w_0} + b_0] y_i - 1\} = 0$$

oznaczają, że **nie zerują się** tylko te współczynniki α_i, które odpowiadają wektorom podpierającym! W efekcie wszystkie sumowania wykonywane są **tylko** po tych i, którym odpowiadają wektory leżące na H_1 i H_2. Jako rozwiązanie zadania otrzymujemy optymalne wsp. Lagrange'a $\boldsymbol{\alpha}^0 = (\alpha_1^0, ..., \alpha_n^0)$.

Bardzo istotnym jest fakt, iż warunki

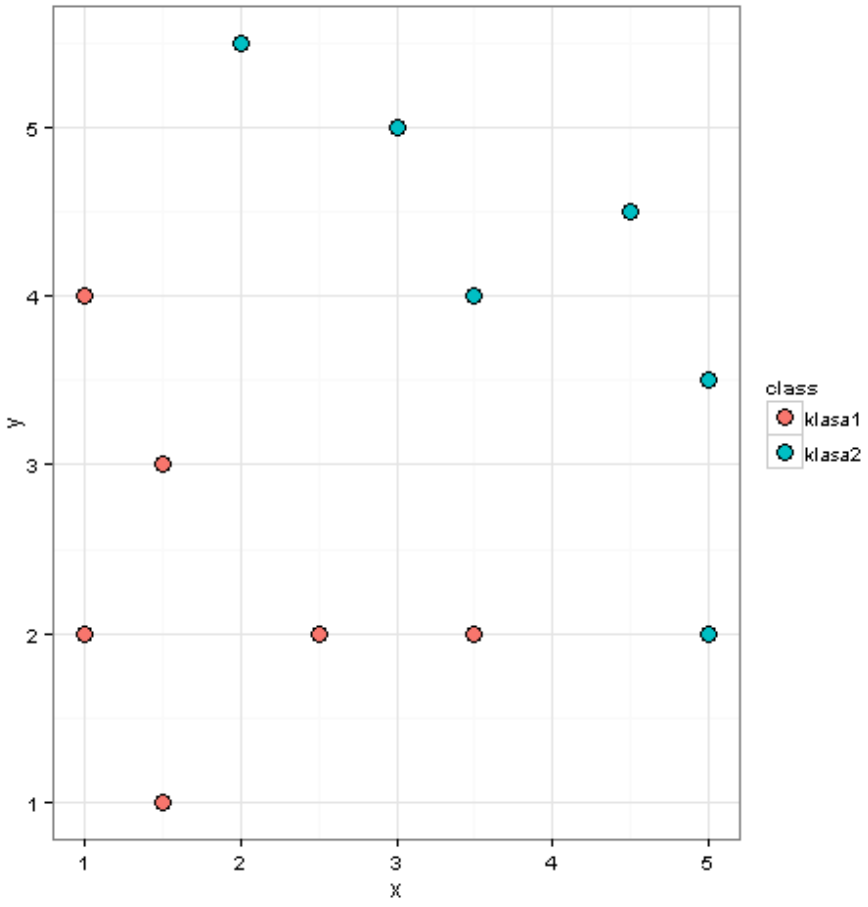
$$\sum_{i=1}^{n} \alpha_i \{[\mathbf{x}_i \cdot \mathbf{w_0} + b_0] y_i - 1\} = 0$$

oznaczają, że **nie zerują się** tylko te współczynniki α_i, które odpowiadają wektorom podpierającym! W efekcie wszystkie sumowania wykonywane są **tylko** po tych i, którym odpowiadają wektory leżące na H_1 i H_2. Jako rozwiązanie zadania otrzymujemy optymalne wsp. Lagrange'a $\boldsymbol{\alpha}^0 = (\alpha_1^0, ..., \alpha_n^0)$. Optymalna hiperpłaszczyzna przyjmuje postać

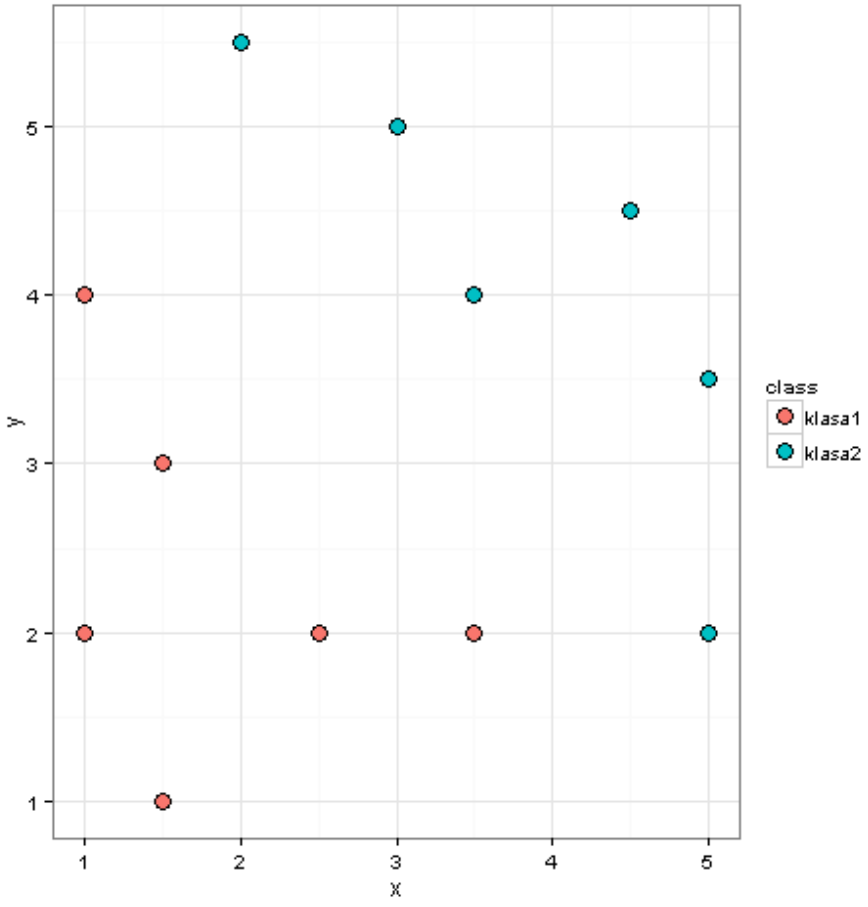
$$\sum_{SV} y_i \alpha_i^0 (\mathbf{x}_i \cdot \mathbf{x}) + b^0 = 0$$

a stałą b^0 można wziąć jako

$$b^0 = \frac{1}{2} \left(\mathbf{w} \cdot \mathbf{x}_1^* + \mathbf{w} \cdot \mathbf{x}_{-1}^* \right)$$

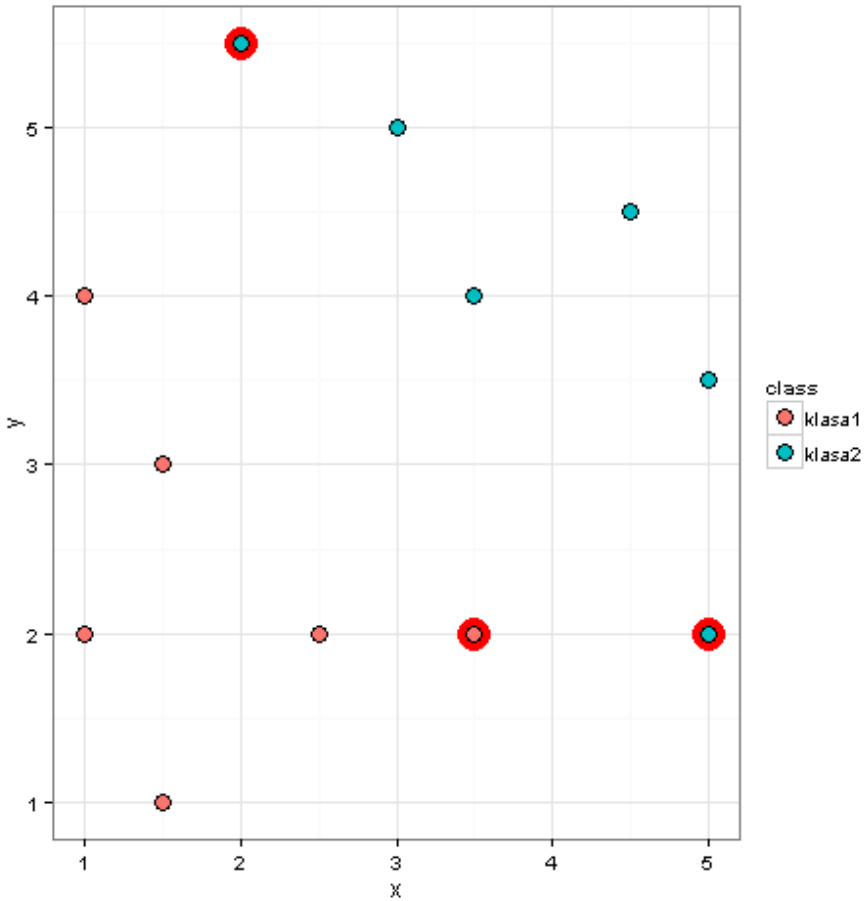


Do próby należy 12 punktów - po 6 z każdej klasy. Wywołujemy funkcję svm() z pakietu R.



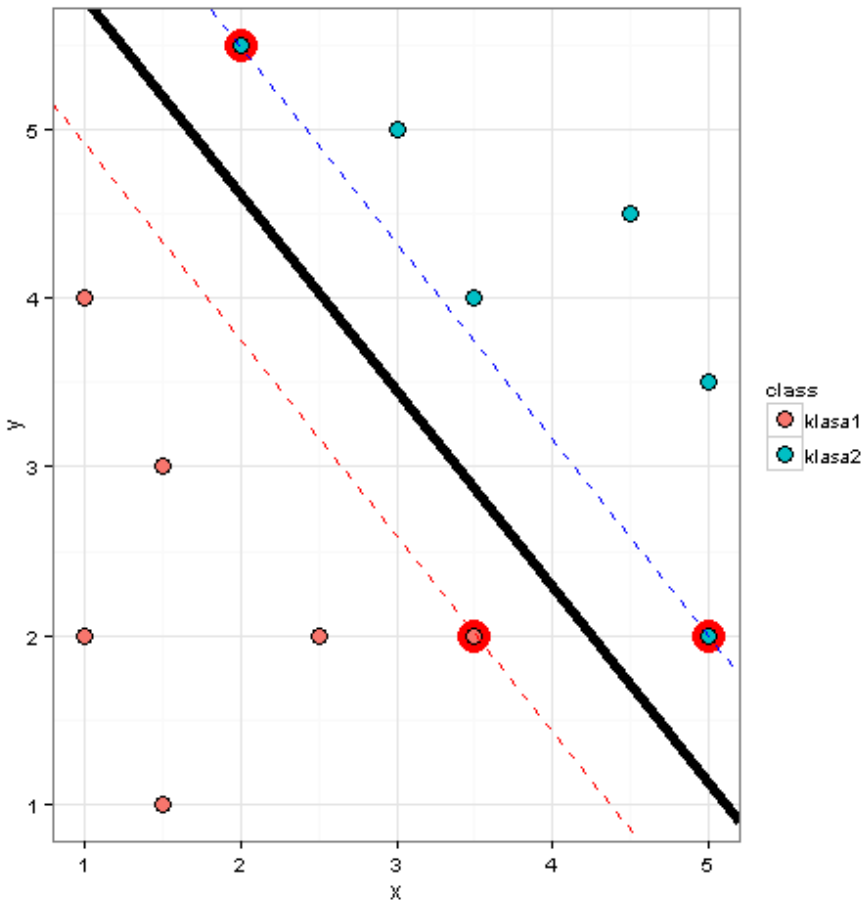
Do próby należy 12 punktów - po 6 z każdej klasy. Wywołujemy funkcję svm() z pakietu R. Otrzymujemy następujące wartości α_i.

x_i	y_i	α_i
1.0	2.0	0.000
1.5	1.0	0.000
1.0	4.0	0.000
1.5	3.0	0.000
2.5	2.0	0.000
3.5	2.0	1.542
2.0	5.5	0.327
3.0	5.0	0.000
3.5	4.0	0.000
4.5	4.5	0.000
5.0	2.0	1.215
5.0	3.5	0.000



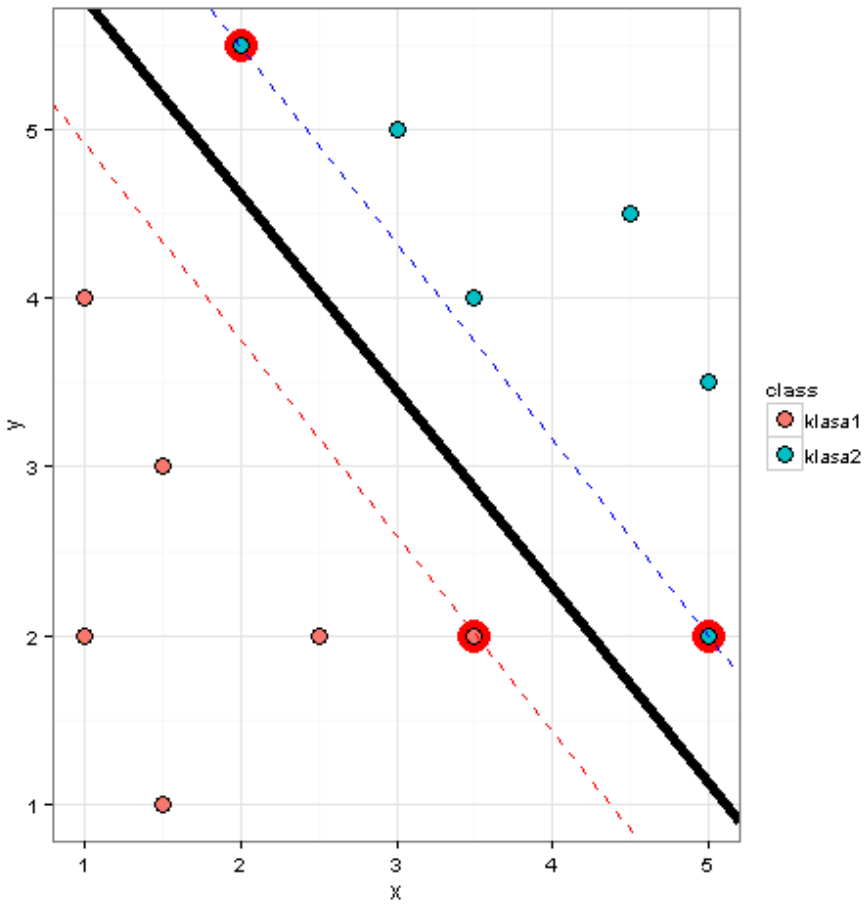
Do próby należy 12 punktów - po 6 z każdej klasy. Wywołujemy funkcję svm() z pakietu R. Otrzymujemy następujące wartości α_i. Te punkty, dla których $\alpha_i > 0$ są wektorami podpierającymi.

x_i	y_i	α_i
1.0	2.0	0.000
1.5	1.0	0.000
1.0	4.0	0.000
1.5	3.0	0.000
2.5	2.0	0.000
3.5	**2.0**	**1.542**
2.0	**5.5**	**0.327**
3.0	5.0	0.000
3.5	4.0	0.000
4.5	4.5	0.000
5.0	**2.0**	**1.215**
5.0	3.5	0.000



Wektor **w**

$$\mathbf{w} = \sum_{i=1}^{n} y_i \alpha_i \mathbf{x}_i = \begin{pmatrix} -1.333 \\ -1.142 \end{pmatrix}$$

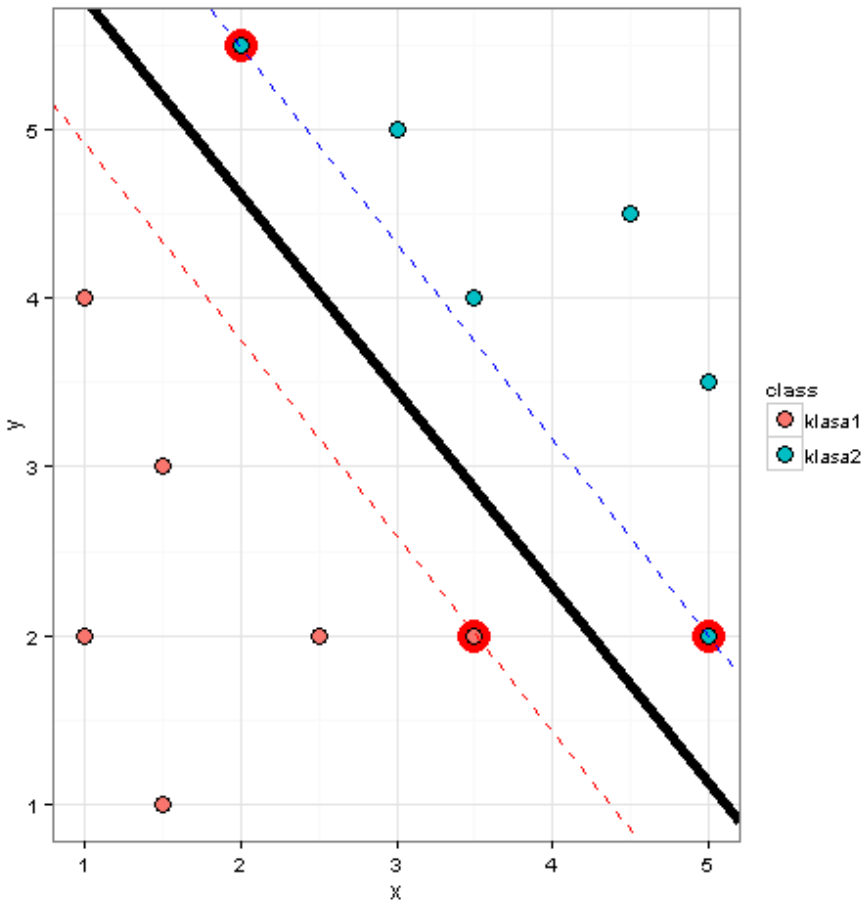


Wektor **w**

$$\mathbf{w} = \sum_{i=1}^{n} y_i \alpha_i \mathbf{x}_i = \begin{pmatrix} -1.333 \\ -1.142 \end{pmatrix}$$

Stała *b*

$$b = \frac{1}{2}\left(\mathbf{w} \cdot \mathbf{x}_1^* + \mathbf{w} \cdot \mathbf{x}_{-1}^*\right) = 7.95$$



Wektor **w**

$$\mathbf{w} = \sum_{i=1}^{n} y_i \alpha_i \mathbf{x}_i = \begin{pmatrix} -1.333 \\ -1.142 \end{pmatrix}$$

Stała *b*

$$b = \frac{1}{2}\left(\mathbf{w} \cdot \mathbf{x}_1^* + \mathbf{w} \cdot \mathbf{x}_{-1}^*\right) = 7.95$$

a interesujące nas proste to

$$H_1 : y = -\frac{w_x}{w_y} x + \frac{1-b}{w_x}$$

$$H_2 : y = -\frac{w_x}{w_y} x - \frac{1+b}{w_x}$$

$$H_0 : y = -\frac{w_x}{w_y} x - \frac{b}{w_x}$$

Oczywiście, bardzo często separowalność klas jest zbyt dużym wymaganiem (np. ze względu na losowośc danych). Na szczęście, ujęcie tego faktu w metodzie SVM nie nastręcza zbyt dużych trudności. Oryginalne nierówności zostają wtedy zastąpione przez

$$\mathbf{x}_i \cdot \mathbf{w} + b \geq 1 - \xi_i, \quad \text{gdy} \quad y_i = +1,$$

Oczywiście, bardzo często separowalność klas jest zbyt dużym wymaganiem (np. ze względu na losowośc danych). Na szczęście, ujęcie tego faktu w metodzie SVM nie nastręcza zbyt dużych trudności. Oryginalne nierówności zostają wtedy zastąpione przez

$$\mathbf{x}_i \cdot \mathbf{w} + b \geq 1 - \xi_i, \quad \text{gdy} \quad y_i = +1,$$

$$\mathbf{x}_i \cdot \mathbf{w} + b \leq -1 + \xi_i, \quad \text{gdy} \quad y_i = -1,$$

Oczywiście, bardzo często separowalność klas jest zbyt dużym wymaganiem (np. ze względu na losowośc danych). Na szczęście, ujęcie tego faktu w metodzie SVM nie nastręcza zbyt dużych trudności. Oryginalne nierówności zostają wtedy zastąpione przez

$$\mathbf{x}_i \cdot \mathbf{w} + b \geq 1 - \xi_i, \quad \text{gdy} \quad y_i = +1,$$

$$\mathbf{x}_i \cdot \mathbf{w} + b \leq -1 + \xi_i, \quad \text{gdy} \quad y_i = -1,$$

przy czym obecność stałych $\xi_i \geq 0$ umożliwia naruszenie oryginalnych ograniczeń.

Oczywiście, bardzo często separowalność klas jest zbyt dużym wymaganiem (np. ze względu na losowośc danych). Na szczęście, ujęcie tego faktu w metodzie SVM nie nastręcza zbyt dużych trudności. Oryginalne nierówności zostają wtedy zastąpione przez

$$\mathbf{x}_i \cdot \mathbf{w} + b \geq 1 - \xi_i, \quad \text{gdy} \quad y_i = +1,$$

$$\mathbf{x}_i \cdot \mathbf{w} + b \leq -1 + \xi_i, \quad \text{gdy} \quad y_i = -1,$$

przy czym obecność stałych $\xi_i \geq 0$ umożliwia naruszenie oryginalnych ograniczeń. W efekcie zadanie sprowadza się do minimalizacji funkcji

$$\frac{||\mathbf{w}||^2}{2} + C \sum_{i=1}^{n} \xi_i$$

gdzie C jest z góry ustalonym współczynnikiem kary za niespełnienie oryginalnych ograniczeń.

Rozwiązaniem problemu jest maksymalizacja funkcji

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

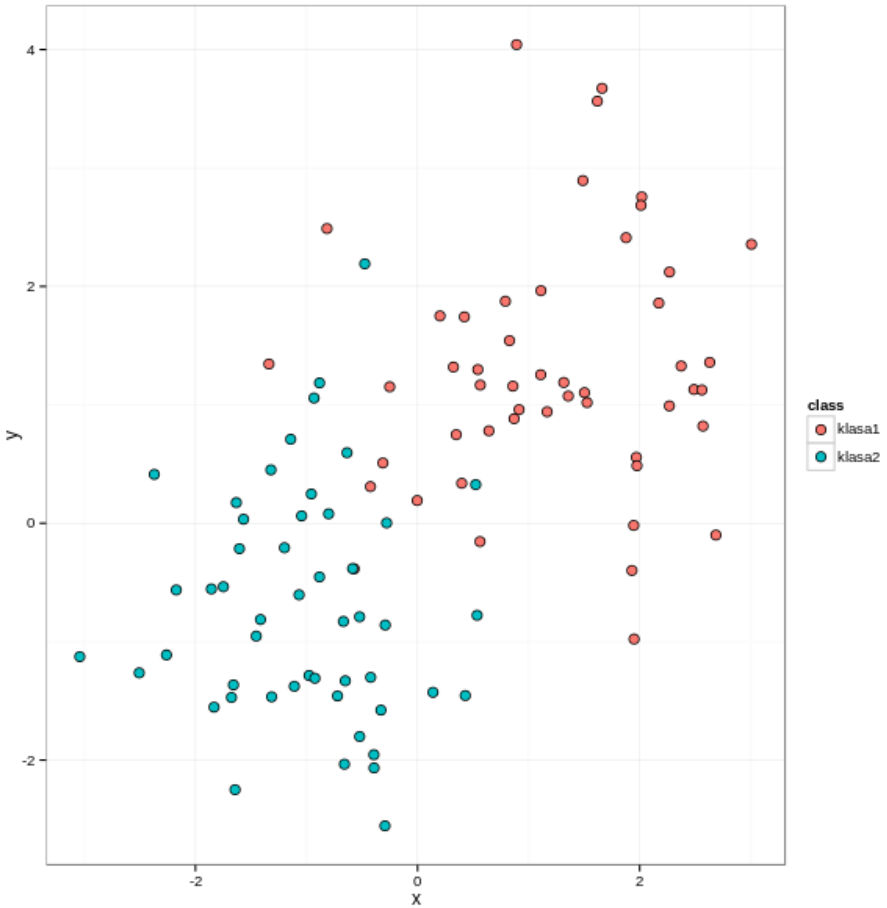
Rozwiązaniem problemu jest maksymalizacja funkcji

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

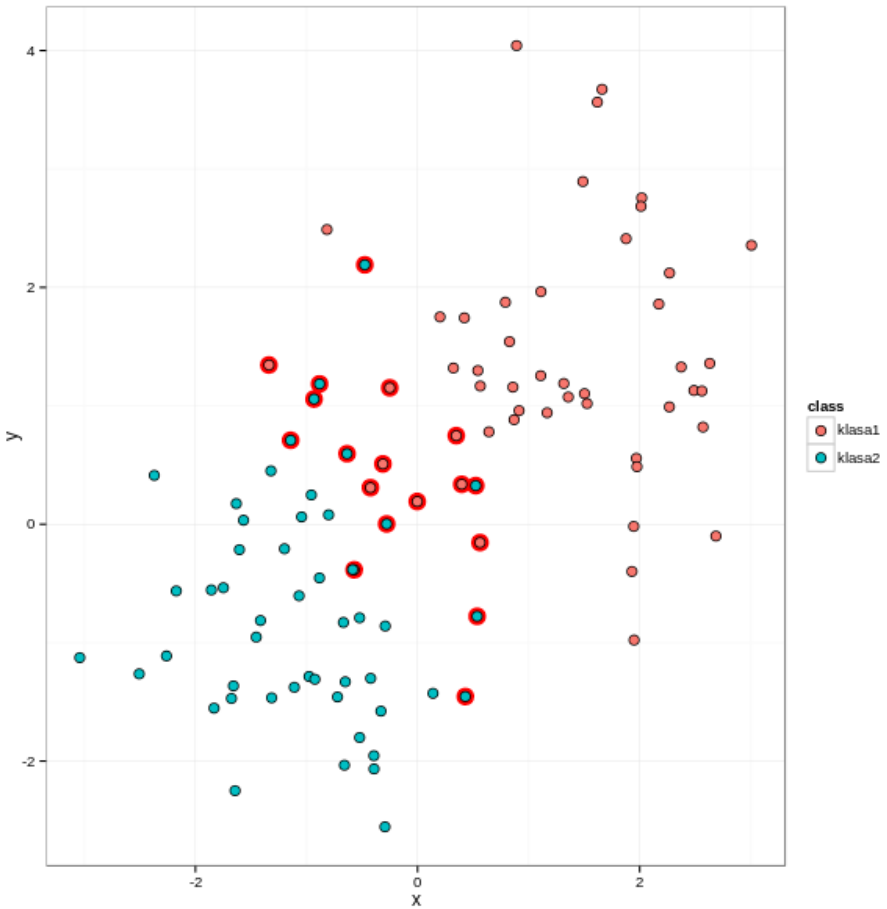
przy ograniczeniach

$$0 \leq \alpha_i \leq C \quad i = 1, ..., n \quad \sum_{i=1}^{n} y_i \alpha_i = 0.$$

Rzecz jasna, stałą C musimy dobrać sami, np. na podstawie próby testowej, kroswalidacji albo oceny prawdopodobieństwa błędnej klasyfikacji. Biorąc pod uwagę losowy charakter danych, zaproponowanie małej wartości C umożliwia przeciwdziałanie nadmiernemu dopasowaniu do próby uczącej.

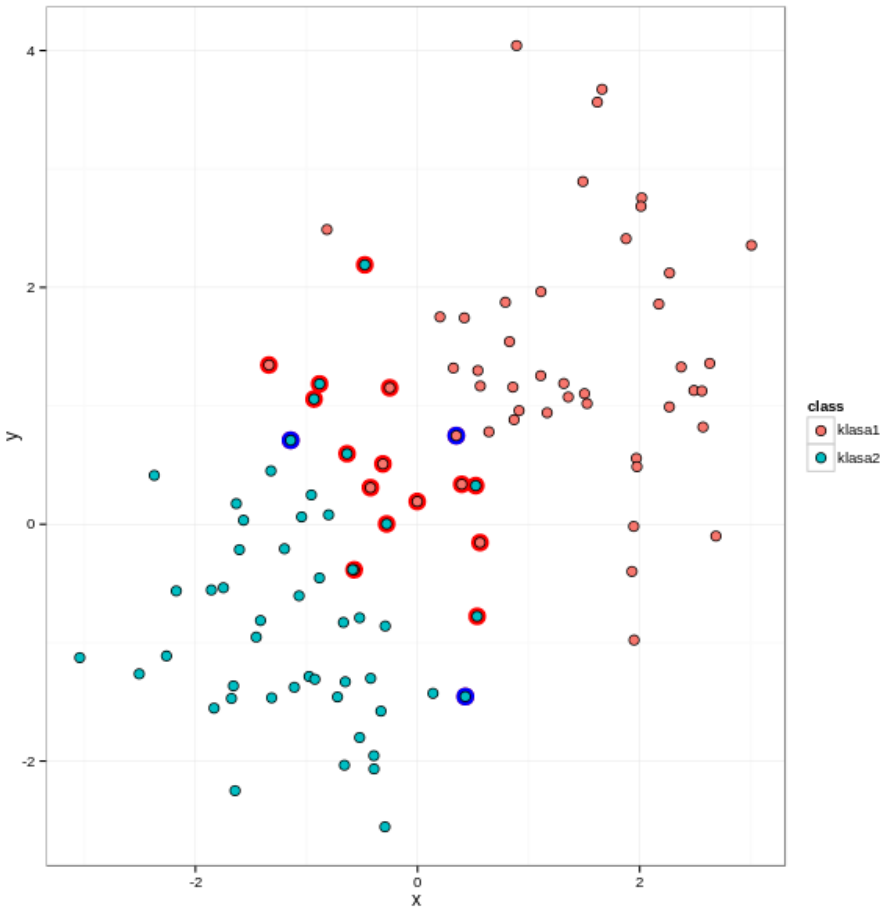


Do próby należy 100 punktów - 50 z klasy 1 i tyle samo z klasy 2. Wywołujemy funkcję svm() z pakietu R z wartością kosztu $C = 10$.



Do próby należy 100 punktów - 50 z klasy 1 i tyle samo z klasy 2. Wywołujemy funkcję svm() z pakietu R z wartością kosztu $C = 10$. Następujące punkty są wektorami podpierającymi.

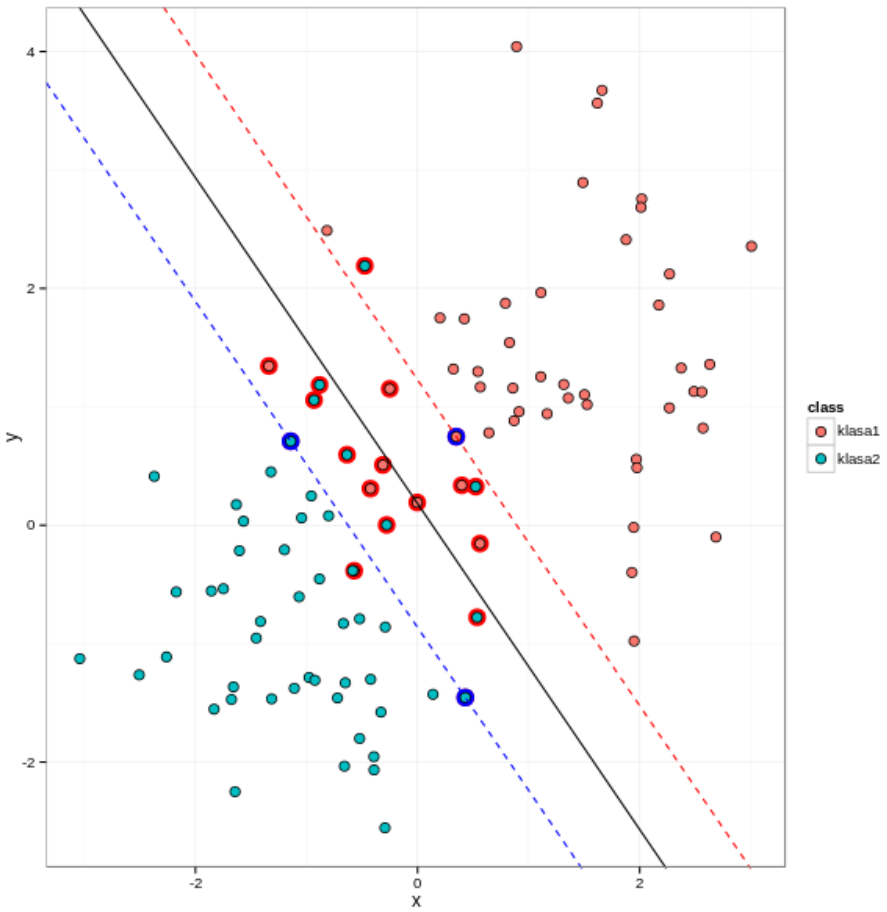
x_i	y_i	α_i
-0.571	-0.384	10.000
-0.312	0.510	10.000
0.562	-0.154	10.000
-0.424	0.310	10.000
-0.250	1.153	10.000
0.398	0.338	10.000
-1.338	1.345	10.000
-0.003	0.192	10.000
0.348	0.749	1.947
-0.882	1.186	10.000
0.536	-0.777	10.000
-1.143	0.710	2.372
0.430	-1.454	9.575
-0.278	0.003	10.000
-0.475	2.192	10.000
-0.932	1.058	10.000
-0.635	0.596	10.000
0.524	0.327	10.000



Do próby należy 100 punktów - 50 z klasy 1 i tyle samo z klasy 2. Wywołujemy funkcję svm() z pakietu R z wartością kosztu $C = 10$. Następujące punkty są wektorami podpierającymi.

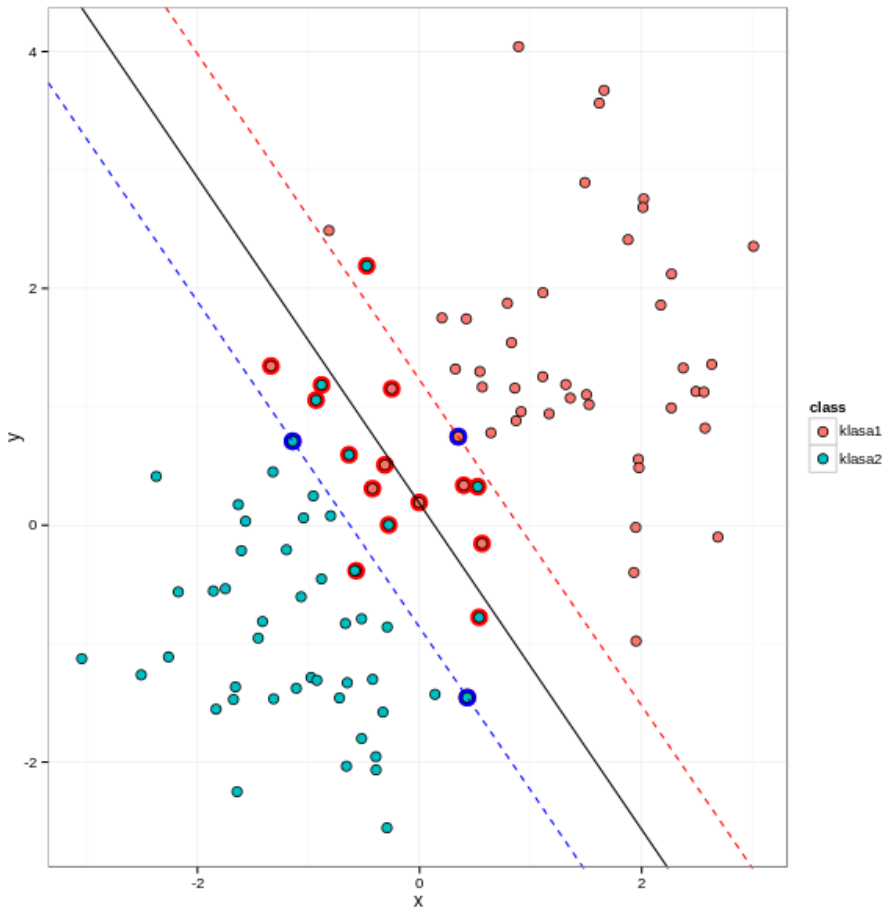
x_i	y_i	α_i
-0.571	-0.384	10.000
-0.312	0.510	10.000
0.562	-0.154	10.000
-0.424	0.310	10.000
-0.250	1.153	10.000
0.398	0.338	10.000
-1.338	1.345	10.000
-0.003	0.192	10.000
0.348	**0.749**	**1.947**
-0.882	1.186	10.000
0.536	-0.777	10.000
-1.143	**0.710**	**2.372**
0.430	**-1.454**	**9.575**
-0.278	0.003	10.000
-0.475	2.192	10.000
-0.932	1.058	-10.000
-0.635	0.596	10.000
0.524	0.327	10.000

Tylko punkty, dla których $\alpha_i < C$ są brane pod uwagę przy wyznaczaniu b.



Wektor **w** liczymy jak poprzednio

$$\mathbf{w} = \sum_{i=1}^{n} y_i \alpha_i \mathbf{x}_i = \begin{pmatrix} 1.316 \\ 0.957 \end{pmatrix}$$

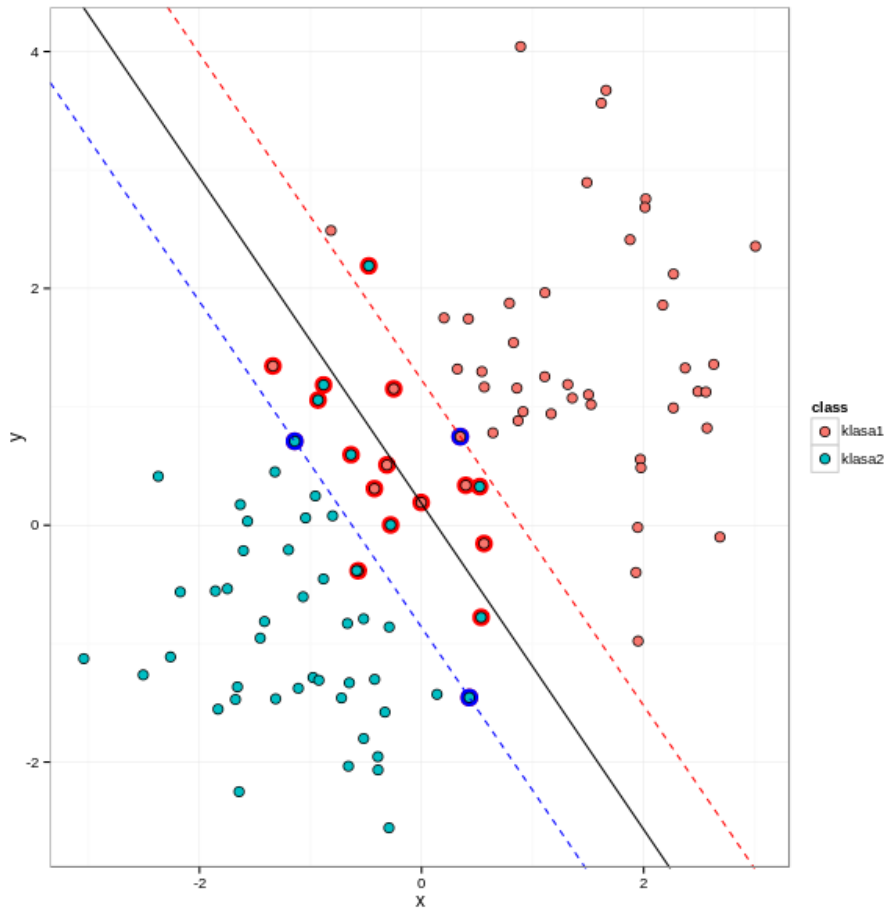


Wektor **w** liczymy jak poprzednio

$$\mathbf{w} = \sum_{i=1}^{n} y_i \alpha_i \mathbf{x}_i = \begin{pmatrix} 1.316 \\ 0.957 \end{pmatrix}$$

stałą b bierzemy z uśrednienia

$$b = \frac{1}{2}\left(\mathbf{w} \cdot \mathbf{x}_1^* + \mathbf{w} \cdot \mathbf{x}_{-1}^*\right) = -0.175$$



Wektor **w** liczymy jak poprzednio

$$\mathbf{w} = \sum_{i=1}^{n} y_i \alpha_i \mathbf{x}_i = \begin{pmatrix} 1.316 \\ 0.957 \end{pmatrix}$$

stałą b bierzemy z uśrednienia

$$b = \frac{1}{2}\left(\mathbf{w} \cdot \mathbf{x}_1^* + \mathbf{w} \cdot \mathbf{x}_{-1}^*\right) = -0.175$$

a interesujące nas proste to

$$H_1 : y = -\frac{w_x}{w_y} x + \frac{1-b}{w_x}$$

$$H_2 : y = -\frac{w_x}{w_y} x - \frac{1+b}{w_x}$$

$$H_0 : y = -\frac{w_x}{w_y} x - \frac{b}{w_x}$$

Z poprzednich wyprowadzeń jest jasne, że zależność od przestrzeni obserwacji $\mathbb{R}^p$ przejawia się **jedynie** przez obliczanie iloczynu skalarnego.

Z poprzednich wyprowadzeń jest jasne, że zależność od przestrzeni obserwacji $\mathbb{R}^p$ przejawia się **jedynie** przez obliczanie iloczynu skalarnego. Z drugiej strony, z algebry liniowej wiadomo, iż przejście od zależności liniowych w przestrzeni $\mathbb{R}^p$ do zależności nieliniowych można opisać jako zależność liniową w **bogatszej przestrzeni**.

Z poprzednich wyprowadzeń jest jasne, że zależność od przestrzeni obserwacji $\mathbb{R}^p$ przejawia się **jedynie** przez obliczanie iloczynu skalarnego. Z drugiej strony, z algebry liniowej wiadomo, iż przejście od zależności liniowych w przestrzeni $\mathbb{R}^p$ do zależności nieliniowych można opisać jako zależność liniową w **bogatszej przestrzeni**. Weźmy przykład z $p = 2$ i policzmy funkcję

$$(1 + \mathbf{x} \cdot \mathbf{y})^2 = (1 + x_1 y_1 + x_2 y_2)^2 = 1 + 2x_1 y_1 + 2x_2 y_2 + x_1^2 y_1^2 + 2x_1 y_1 x_2 y_2 + x_2^2 y_2^2$$

Taka funkcja jest równoważna iloczynowi skalarnemu

$$\mathbf{\Phi}(\mathbf{x})\mathbf{\Phi}(\mathbf{y})$$

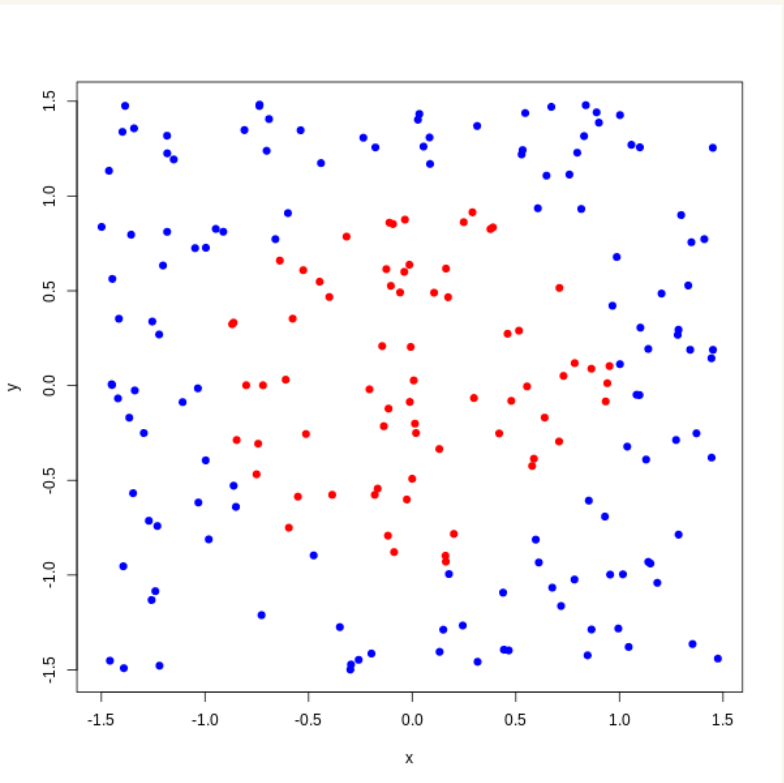
przekształconych zmiennych

$$\mathbf{\Phi}(\mathbf{x}) = \left(1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1 x_2\right)$$
$$\mathbf{\Phi}(\mathbf{y}) = \left(1, \sqrt{2}y_1, \sqrt{2}y_2, y_1^2, y_2^2, \sqrt{2}y_1 y_2\right)$$

Dlaczego jest to takie istotne? Otóż (i) wszystkie takie obliczenia dotyczą **jedynie iloczynu skalarnego**, a nie przekształconych obserwacji, (ii) dokonanie nieliniowej transformacji często umożlwia dokładną klasyfikację:

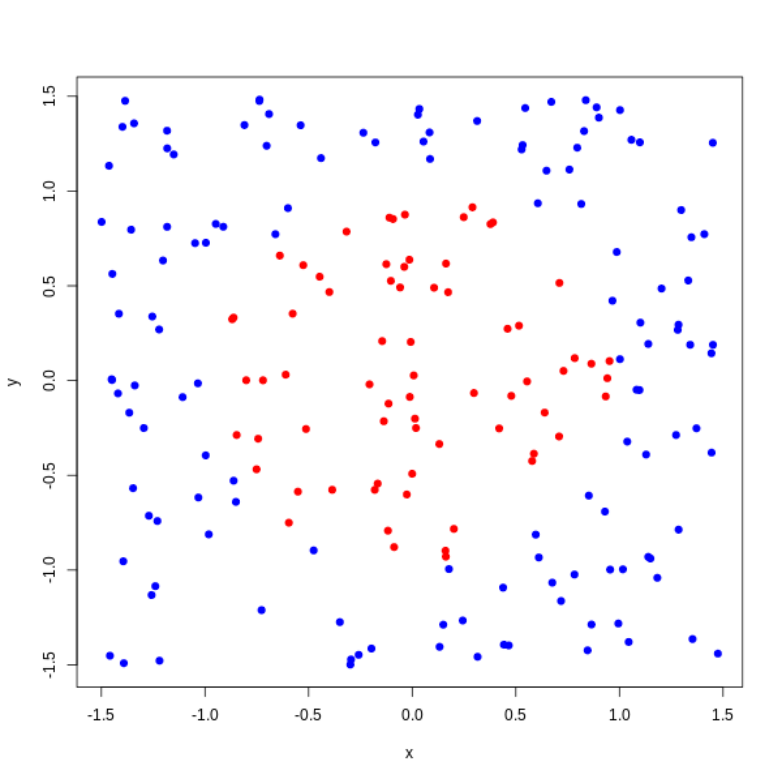
Dlaczego jest to takie istotne? Otóż (i) wszystkie takie obliczenia dotyczą **jedynie iloczynu skalarnego**, a nie przekształconych obserwacji, (ii) dokonanie nieliniowej transformacji często umożlwia dokładną klasyfikację:

Oryginalne dane

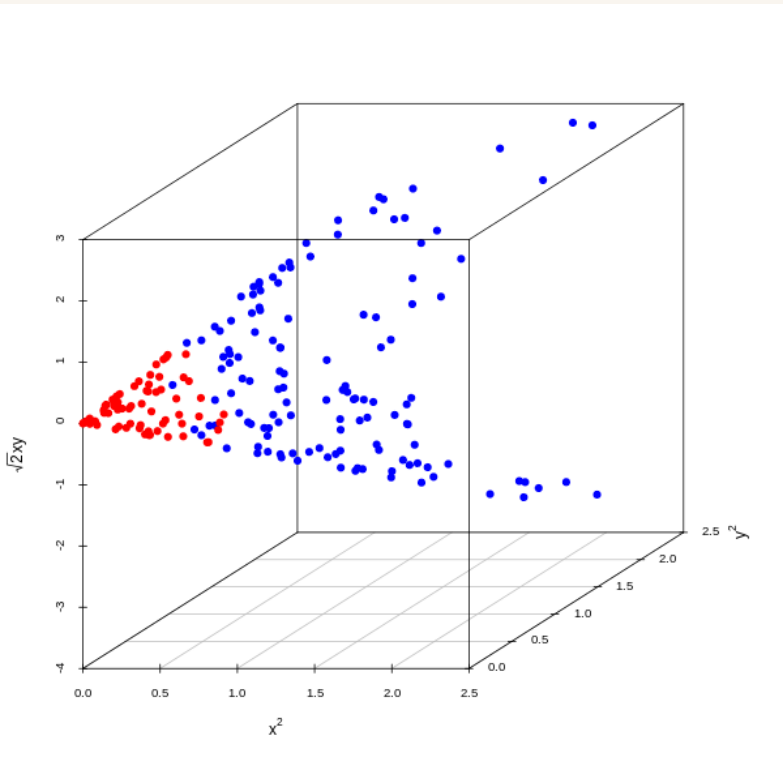


Dlaczego jest to takie istotne? Otóż (i) wszystkie takie obliczenia dotyczą **jedynie iloczynu skalarnego**, a nie przekształconych obserwacji, (ii) dokonanie nieliniowej transformacji często umożlwia dokładną klasyfikację:

Oryginalne dane



Dane przekształcone $\mathbf{z} = (x^2, y^2, \sqrt{2}xy)$



W praktyce takie podejście sprowadza się do zamiany iloczynu skalarnego w funkcji Lagrange'a

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

W praktyce takie podejście sprowadza się do zamiany iloczynu skalarnego w funkcji Lagrange'a

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

na odpowiednie **jądro** $K(\mathbf{x_i}, \mathbf{x_j})$

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)$$

W praktyce takie podejście sprowadza się do zamiany iloczynu skalarnego w funkcji Lagrange'a

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

na odpowiednie **jądro** $K(\mathbf{x_i}, \mathbf{x_j})$

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)$$

Co więcej, nie musimy de facto znać nawet nowej przestrzeni. Wystarczy, że jądro spełnia określone warunki, wynikające z tw. Mercera z analizy funkcjonalnej.

W praktyce takie podejście sprowadza się do zamiany iloczynu skalarnego w funkcji Lagrange'a

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

na odpowiednie **jądro** $K(\mathbf{x_i}, \mathbf{x_j})$

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)$$

Co więcej, nie musimy de facto znać nawet nowej przestrzeni. Wystarczy, że jądro spełnia określone warunki, wynikające z tw. Mercera z analizy funkcjonalnej.Najczęściej stosowanymi jądrami są: wielomianowe, radialne i sigmoidalne

W praktyce takie podejście sprowadza się do zamiany iloczynu skalarnego w funkcji Lagrange'a

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

na odpowiednie **jądro** $K(\mathbf{x_i}, \mathbf{x_j})$

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)$$

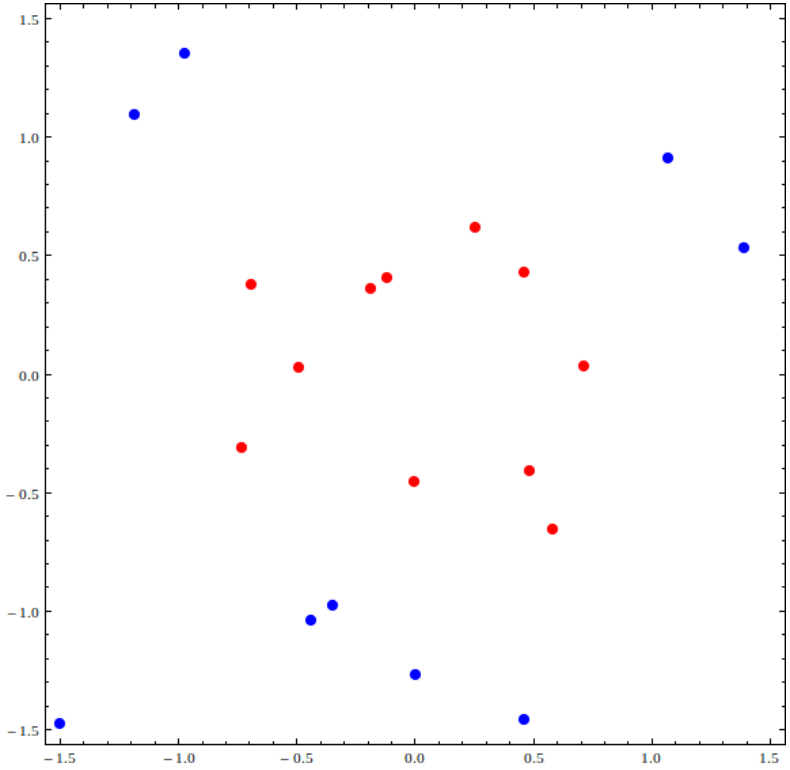
Co więcej, nie musimy de facto znać nawet nowej przestrzeni. Wystarczy, że jądro spełnia określone warunki, wynikające z tw. Mercera z analizy funkcjonalnej.Najczęściej stosowanymi jądrami są: wielomianowe, radialne i sigmoidalne

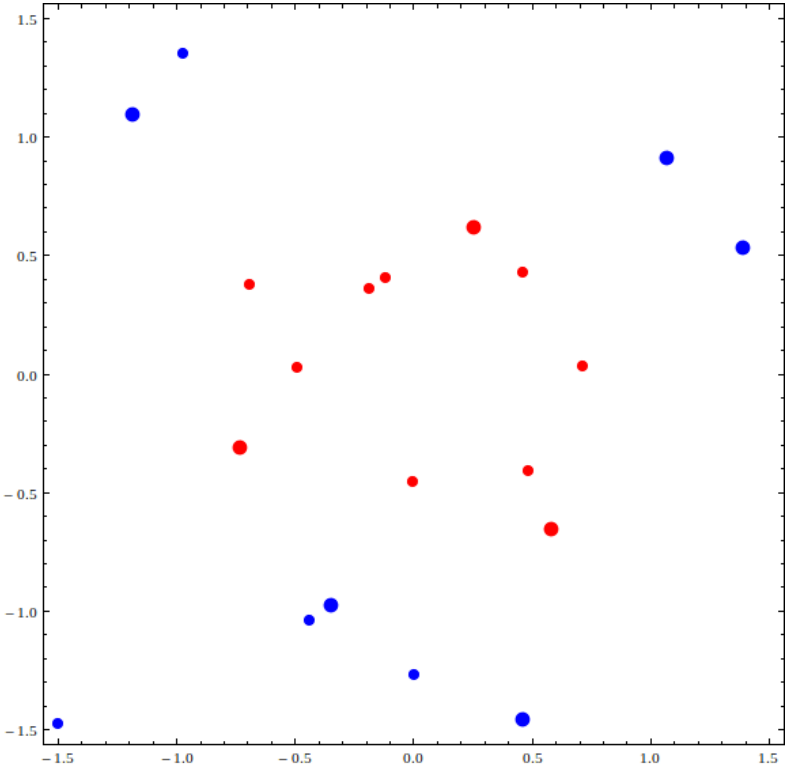
$$K(\mathbf{x}_i, \mathbf{x}_j) = (1 + \mathbf{x}_i \cdot \mathbf{x}_j)^d$$

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma ||\mathbf{x}_i - \mathbf{x}_j||^2)$$

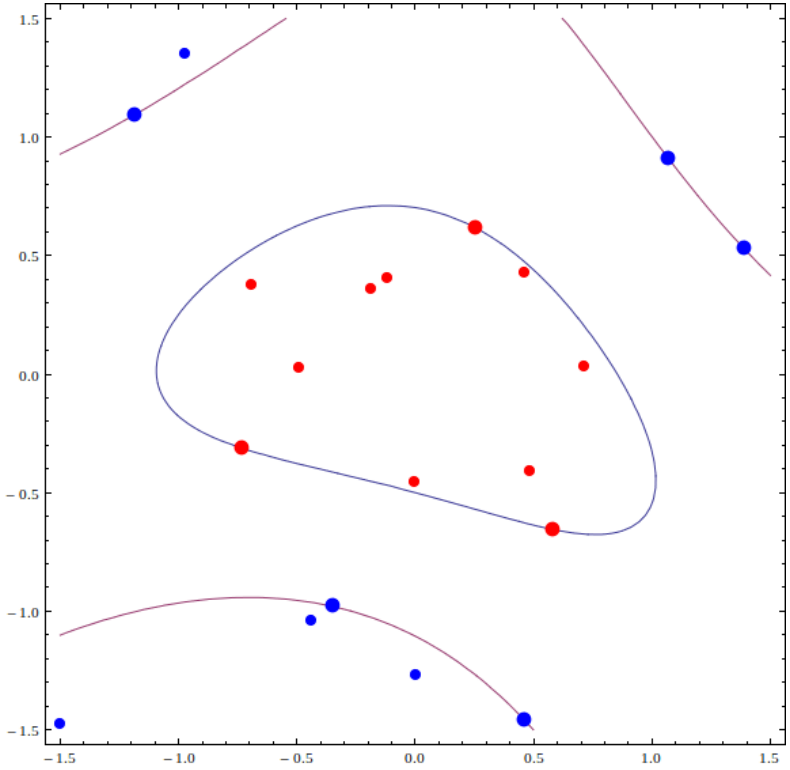
$$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\Psi_1 (\mathbf{x}_i \cdot \mathbf{x}_j) + \Psi_2)$$

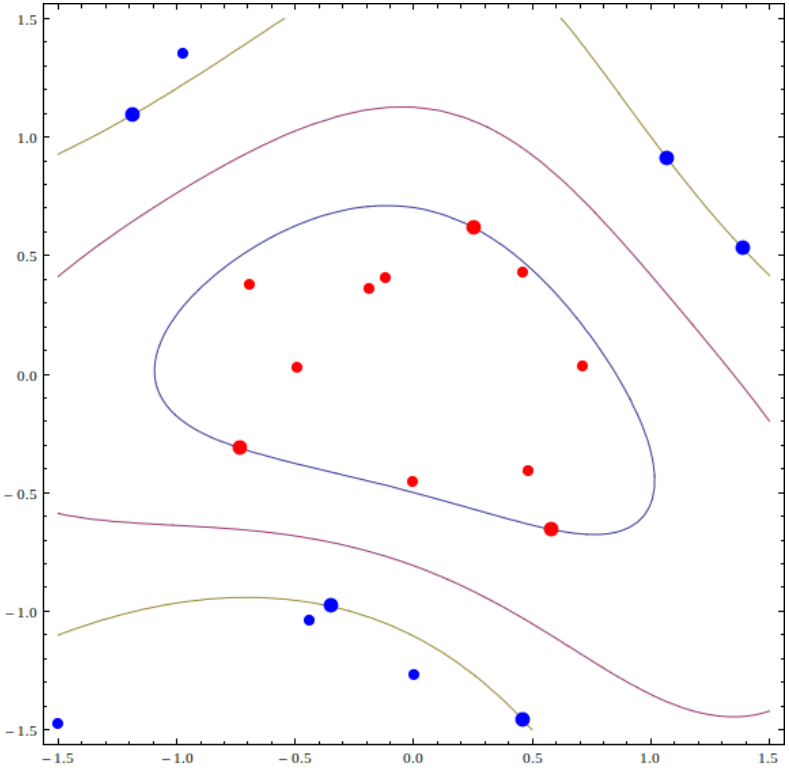
Przykład dla jądra radialnego z $\gamma = 0.5$. Wyróżnione punkty są wektorami podpierającymi.



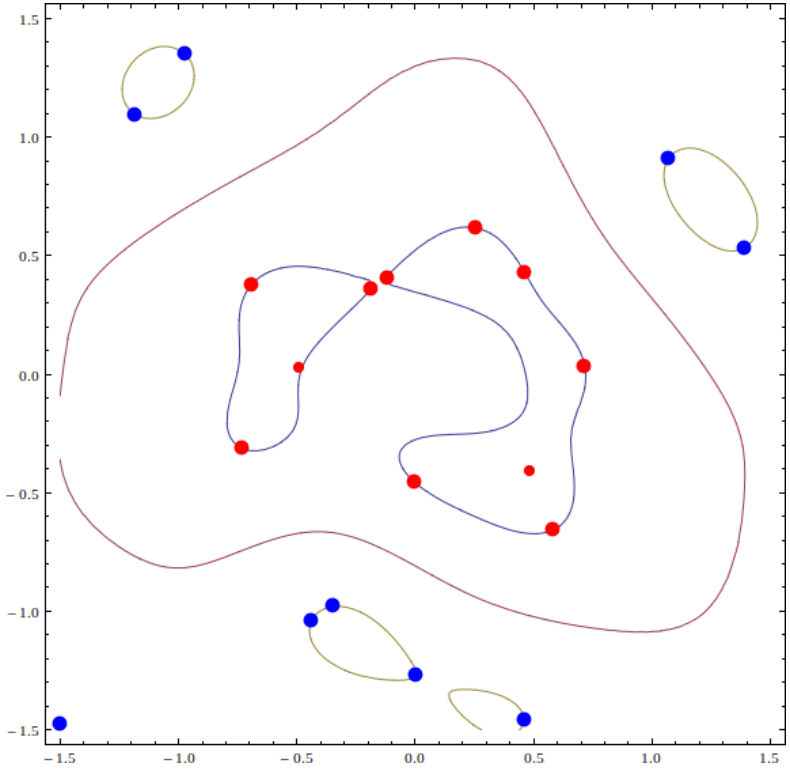


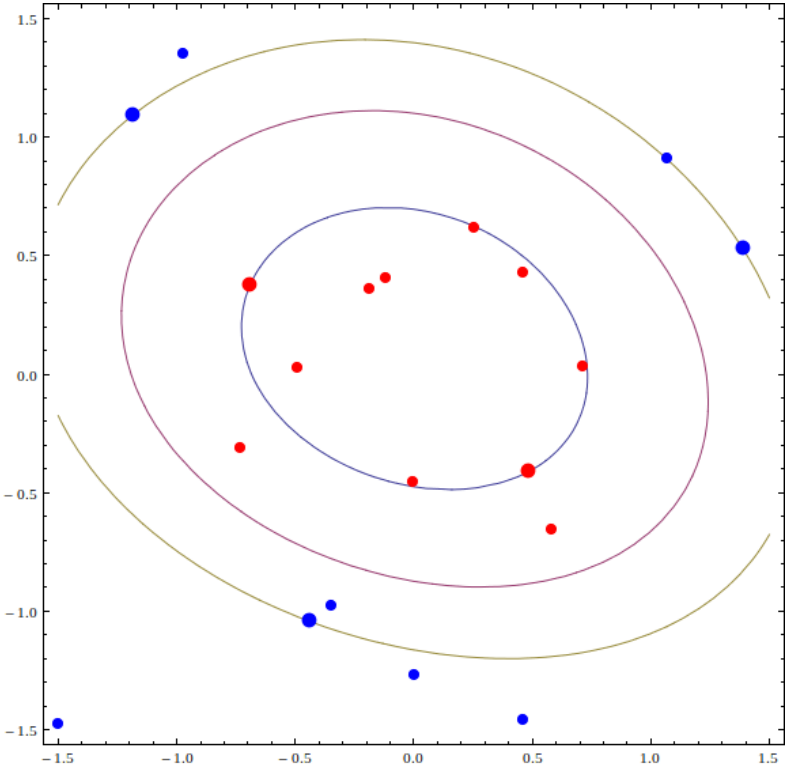
Poniżej hiperpłaszczyzny marginesów oraz hiperpłaszczyzna dyskryminująca.





Ten sam przypadek obliczony dla $\gamma = 5$ (lewa strona) i $\gamma = 0.05$ (prawa strona).





Poniżej przypadek większej ilości punktów obliczony dla $\gamma = 0.5$ (lewa strona - brak separowalności) i $\gamma = 5$ (prawa strona).