LINIOWY KLASYFIKATOR SVM

Rozważamy zadanie klasyfikacji dla dwóch klas

Zbiór uczący:

$$(x_1, d_1), \ldots, (x_N, d_N) \in R^p \times D, \quad D = \{-1, +1\}$$

Regułę decyzyjną g określamy w ten sposób, że obiekt jest zaliczany do klasy 1, gdy $g(x) \geq 0$, a do klasy –1, gdy g(x)<0.

LINIOWY KLASYFIKATOR SVM

Szukamy hiperpłaszczyzny rozdzielającej o równaniu:
$w^T x + b = 0$,

gdzie: x –klasyfikowany wektor
w –zmieniający się wektor wag
b –przesunięcie (bias)

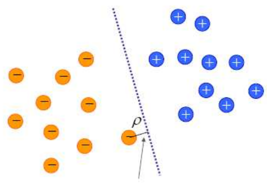
Reguła decyzyjna:

$$g(x) \geq w^T x + b$$

PRZYPADEK LINIOWO SEPAROWALNY

W przypadku gdy próbki są liniowo separowalne istnieje hiperpłaszczyzna rozdzielająca spełniająca warunki:

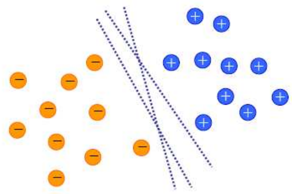
$$w^T x + b \geq 0 \text{ dla } d_i = +1$$
$$w^T x + b < 0 \text{ dla } d_i = -1$$



PRZYPADEK LINIOWO SEPAROWALNY

Która z hiperpłaszczyzn rozdzielających jest optymalna?

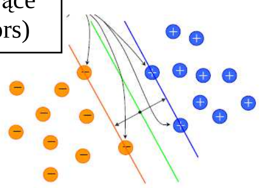
margines – odległość hiperpłaszczyzny rozdzielającej do najbliższej próbki uczącej



KLASYFIKATOR SVM

- Celem SVM jest znalezienie takiej hiperpłaszczyzny rozdzialającej, która maksymalizuje margines.
- Nazywamy ją **optymalną hiperpłaszczyzną rozdzielającą OSH** (optimal separating hyperplane)

Wektory podpierające
(ang. Suport Vectors)



UCZENIE KLASYFIKATORA SVM

Uczenie sprowadza się do znalezienia równania OSH

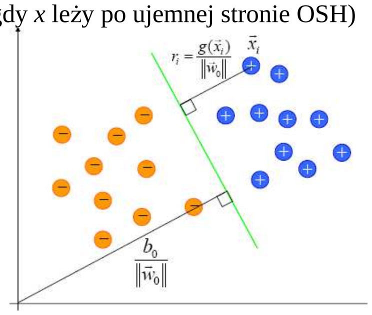
- Niech w_0, oraz b_0 optymalne wartości w i b.
- Równanie OSH:
  $$w_0^T x + b_0 = 0$$
- Funkcja dyskryminacyjna
  $$g(x) = w_0^T x + b_0$$

Uwaga

$g(x)$ wyznacza miarę odległości obiektu x od OSH

UCZENIE KLASYFIKATORA SVM

Można pokazać, że $r = \frac{g(x)}{\|w_0\|}$, gdzie r oznacza odległość próbki od OSH (ujemna gdy x leży po ujemnej stronie OSH)

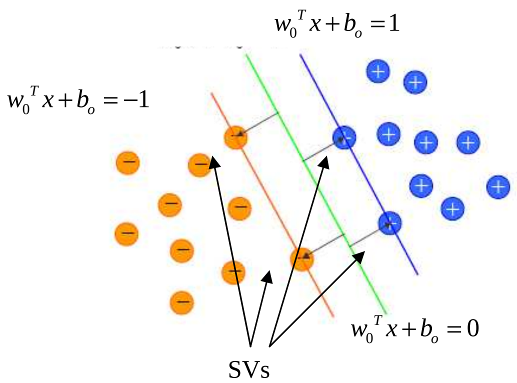


UCZENIE KLASYFIKATORA SVM

1. Spośród nieskończonego zbioru rozwiązań będących OSH (każdą z nich można uzyskać poprzez proste przeskalowanie) wybieramy taką, dla której moduł wartości funkcji dyskryminacyjnej dla najbliższego wektora wynosi 1.
2. Jest to tzw. kanoniczna OSH.
3. Z 1. wynika zatem, że dla wszystkich wektorów zbioru uczącego zachodzi:

$$d_i \left(w_0^T x + b_o \right) \geq 1$$

UCZENIE KLASYFIKATORA SVM



UCZENIE KLASYFIKATORA SVM

- Dla każdego wektora podpierającego $x^{(s)}$ zachodzi $g\left(x^{(s)}\right) = w_0^T x^{(s)} + b_0 = \pm 1$ dla $d^{(s)} = \pm 1$.
- Odległość wektora podpierającego od OSH

$$r = \frac{g\left(x^{(s)}\right)}{\|w_0\|} = \begin{cases} \frac{1}{\|w_0\|} & \text{gdy } d^{(s)} = +1 \\ -\frac{1}{\|w_0\|} & \text{gdy } d^{(s)} = -1 \end{cases}$$

UCZENIE KLASYFIKATORA SVM

- Niech ρ oznacza optymalną wartość marginesu, między dwoma klasami w zbiorze uczącym V. Wtedy:

$$\rho = 2r = \frac{2}{\|w_0\|}.$$

- Z powyższej postaci wynika, że maksymalizowanie marginesu jest równoważne minimalizowaniu normy euklidesowej $\|w\|$.

UCZENIE KLASYFIKATORA SVM

Poszukujemy wektora w minimalizującego funkcję kosztu:

$$\Phi(w) = \frac{1}{2} w^T w$$

(½ występuje tutaj dla wygody przedstawienia)

Przy ograniczeniach:

$$d_i \left(w^T x_i + b \right) \geq 1, \text{ dla } i=1,\ldots,N.$$

Jest to zadanie **programowania kwadratowego**

- Funkcja kosztu $\Phi(w)$ jest wypukła
- Ograniczenia są liniowe ze względu na w.

OPTYMALIZACJA KWADRATOWA

- Zadanie rozwiązujemy metodą mnożników Lagrange'a.
- Wprowadzamy N nieujemnych mnożników Lagrange'a, po jednym dla każdego z ograniczeń i budujemy funkcję Lagrange'a dla zadania pierwotnego:

$$L(w,b,\alpha)=\frac{1}{2}w^T w-\sum_{i=1}^{N}\alpha_i\left[d_i\left(w^T x_i+b\right)-1\right],$$

gdzie α_i są mnożnikami Lagrange'a.

- Rozwiązaniem zadania jest punkt siodłowy funkcji $L(w,b,\alpha)$, którą minimalizujemy ze względu na w i b i maksymalizujemy ze względu na α.

OPTYMALIZACJA KWADRATOWA

- Z warunków Kuhna-Tuckera dla zadania pierwotnego (że gradient funkcji $L(w,b,\alpha)$ zeruje się ze względu na w i b) otrzymujemy warunki:

$$\begin{cases}\dfrac{\partial L(w,b,\alpha)}{\partial w}=0\\[2ex]\dfrac{\partial L(w,b,\alpha)}{\partial b}=0\end{cases}$$

OPTYMALIZACJA KWADRATOWA

$$\begin{cases}\dfrac{\partial L(w,b,\alpha)}{\partial w}=\dfrac{1}{2}\cdot 2w-\dfrac{\partial\sum_{i=1}^{N}\left[\alpha_i d_i w^T x_i+\alpha_i d_i b-\alpha_i\right]}{\partial w}=w-\sum_{i=1}^{N}\alpha_i d_i x_i=0\\[3ex]\dfrac{\partial L(w,b,\alpha)}{\partial b}=0-\sum_{i=1}^{N}\alpha_i d_i=0\end{cases}$$

Skąd:

$$\begin{cases}w=\sum_{i=1}^{N}\alpha_i d_i x_i\\[2ex]\sum_{i=1}^{N}\alpha_i d_i=0\end{cases}\qquad(*)$$

OPTYMALIZACJA KWADRATOWA

Warunki Kuhna-Tuckera dla zadania pierwotnego:

- $\dfrac{\partial L(w,b,\alpha)}{\partial w}=w$
- $d_i\left(w^T x_i+b\right)\geq 1$
- $\alpha_i>0$, dla i=1,...,m
- $\alpha_i\left(d_i\left(w^T x_i+b\right)-1\right)=0$, dla i=1,...,N

OPTYMALIZACJA KWADRATOWA

- Zadanie pierwotne jest z wypukłą funkcja kosztu i liniowymi ograniczeniami. Możemy zatem skonstruować zadanie dualne, które ma to samo rozwiązanie optymalne, lecz z mnożnikami Lagrange'a zapewniającymi to rozwiązanie.
- Z tw. Wolfa o dualności zadań programowania, wnioskujemy, że maksimum funkcji $L(w,b,\alpha)$ przy ograniczeniach α_i >0 znajduje się w tym samym miejscu co minimum tej funkcji przy ograniczeniach zadania pierwotnego.

OPTYMALIZACJA KWADRATOWA

Przekształcamy do zadania dualnego.

$$L(w,b,\alpha)=\frac{1}{2}w^T w-\sum_{i=1}^{N}\alpha_i\left[d_i\left(w^T x_i+b\right)-1\right]=\frac{1}{2}w^T w-\sum_{i=1}^{N}\alpha_i d_i w^T x_i-$$

z warunku optymalności $\sum_{i=1}^{N}\alpha_i d_i=0$

Z (*) $w=\sum_{i=1}^{N}\alpha_i d_i x_i$. Mnożąc przez w^T z lewej strony:

$$w^T w=w^T\sum_{i=1}^{N}\alpha_i d_i x_i=\sum_{i=1}^{N}\alpha_i d_i w^T x_i=\sum_{i=1}^{N}\alpha_i d_i\sum_{j=1}^{N}\alpha_j d_j x_j^T x_i=$$

$$=\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j d_i d_j x_i^T x_j$$

OPTYMALIZACJA KWADRATOWA

Oznaczając $L(w,b,\alpha)=Q(\alpha)$ i wstawiając powyższe do równania (6.15) otrzymujemy:

$$Q(\alpha)=\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j d_i d_j x_i^T x_j-\sum_{i=1}^{N}\alpha_i d_i w^T x_i+\sum_{i=1}^{N}\alpha_i=$$
$$=\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j d_i d_j x_i^T x_j-\sum_{i=1}^{N}\alpha_i d_i\sum_{j=1}^{N}\alpha_j d_j x_j^T x_i+\sum_{i=1}^{N}\alpha_i=$$
$$=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j d_i d_j x_i^T x_j$$

gdzie $\alpha_i \geq 0$.

UCZENIE KLASYFIKATORA SVM
(przypadek liniowo separowany)

Przy danym zbiorze uczącym V znajdź mnożniki Lagrange'a α_i, i=1,...,N, które maksymalizują funkcję celu

$$Q(\alpha)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j d_i d_j x_i^T x_j$$

przy ograniczeniach:

1. $\sum_{i=1}^{N}\alpha_i d_i=0$
2. $\alpha_i \geq 0$, i=1,...,N.

UCZENIE KLASYFIKATORA SVM
(przypadek liniowo separowany)

Optymalny wektor wag: $w_0=\sum_{i=1}^{N}\alpha_{0,i}d_i x_i$

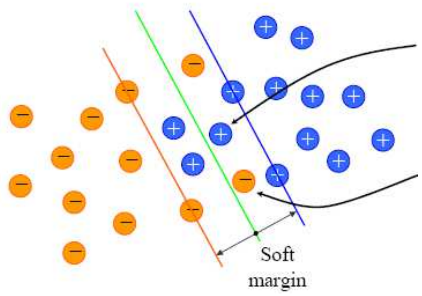
Przesunięcie: $b_0=1-w_0^T x^{(s)}$ dla $d^{(s)}=1$,

$x^{(s)}$ - wektor podpierający

Uwaga:

- Wartości współczynników Lagrange'a są niezerowe tylko dla wektorów podpierających

UCZENIE KLASYFIKATORA SVM
(przypadek liniowo nieseparowalny)



Wektor uczący znajduje się w obszarze marginesu (po dobrej bądź złej stronie)

UCZENIE KLASYFIKATORA SVM
(przypadek liniowo nieseparowalny)

- Poszukujemy hiperpłaszczyzny, która minimalizuje prawdopodobieństwo błędnej klasyfikacji.
- Tzw miękki margines
- Jeżeli próbka (x_i, d_i) narusza warunek $d_i\left(w^T x_i+b\right)\geq 1$, może dziać się tak w dwóch przypadkach:

1. Próbka wpada w region oddzielenia ale jest po dobrej stronie hiperpłaszczyzny decyzyjnej.
2. Próbka jest po drugiej stronie hiperpłaszczyzny decyzyjnej.

W przypadku 1 klasyfikacja będzie poprawna, natomiast w przypadku 2 błędna.

UCZENIE KLASYFIKATORA SVM
(przypadek liniowo nieseparowalny)

- Wprowadzamy nowy zbiór nieujemnych zmiennych $\{\xi_i\}_{i=1}^{N}$ do definicji hiperpłaszczyzny rozdzielającej otrzymując $d_i\left(w^T x_i+b\right)\geq 1-\xi_i$, i=1,...,N.
- Są to tak zwane zmienne osłabiające
- Dla $0\leq\xi_i\leq 1$ punkt wpadnie w region rozdzielający ale będzie po odpowiedniej stronie płaszczyzny decyzyjnej, natomiast dla $\xi_i\geq 1$ wpadnie na złą stronę.

Uczenie SVM -przypadek liniowo nieseparowalny

- Celem jest znalezienie hiperpłaszczyzny rozdzielającej, dla której błąd błędnej klasyfikacji jest minimalny w zbiorze uczącym.

ZADANIE PIERWOTNE:

- Uczenie odbywa się poprzez minimalizację funkcji:

$$\Phi(w,\xi)=\frac{1}{2}w^T w+C\sum_{i=1}^{N}(\xi_i),$$

Przy ograniczeniach:

$d_i\left(w^T x_i+b\right)\geq 1-\xi_i$, i=1,...,N.

$\xi_i\geq 0$, i=1,...,N

C – ustalony parametr

Uczenie SVM -przypadek liniowo nieseparowalny

Przy użyciu mnożników Lagrange'a postępując w sposób jak wcześniej formuujemy **zadanie dualne** maksymalizacji funkcji celu: $Q(\alpha)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j d_i d_j x_i^T x_j$

przy ograniczeniach:

1. $\sum_{i=1}^{N}\alpha_i d_i=0$
2. $C\geq\alpha_i\geq 0$, i=1,...,N

Rozwiązanie dla wektora wag jest takie jak poprzednio, b_0 uzyskujemy z warunku $\alpha_i\left(d_i\left(w^T x+b_0\right)-1+\zeta_i\right)=0$, dla wekora podpierającego $\left(0\leq\alpha_i\leq C\right)$, $\zeta_i=0$

Liniowy klasyfikator SVM - przykład

Przykład.

Utożsamiamy klasę 1 z klasą –1 a klasę 2 z klasą 1. Uczenie klasyfikatora polega na rozwiązaniu zadania programowania kwadratowego. Dla C=10 niezerowe wartości mnożników Lagrange'a α_i rozwiązania optymalnego, współrzędne wektorów podpierających i ich klasy znajdują się w tabeli:

α_i	x_{i1}	x_{i2}	d_i
1,3661	-0,565	-2,653	-1
1,0893	1,241	-4,208	1
0,2768	-2,024	5,388	1

Wynik uczenia liniowego klasyfikatora SVM

Liniowy klasyfikator SVM - przykład

Wektor wag optymalnej hiperpłaszczyzny decyzyjnej $w_0^T x_i+b_0=0$ otrzymujemy z równania $w_0=\sum_{i=1}^{N}\alpha_{0,i}d_i x_i$ zatem:

$$w_0=-1{,}3661\cdot\begin{bmatrix}-0{,}565\\-2{,}653\end{bmatrix}+1{,}0893\cdot\begin{bmatrix}1{,}241\\-4{,}208\end{bmatrix}+0{,}2768\cdot\begin{bmatrix}1{,}241\\-4{,}208\end{bmatrix}=$$

$$=\begin{bmatrix}0{,}7724\\3{,}6243\end{bmatrix}+\begin{bmatrix}1{,}3523\\-4{,}5836\end{bmatrix}+\begin{bmatrix}-0{,}5602\\1{,}4916\end{bmatrix}=\begin{bmatrix}1{,}5645\\0{,}5323\end{bmatrix}$$

Liniowy klasyfikator SVM - przykład

Wartość b_0 otrzymujemy warunku, że dla każdego wektora podpierającego zachodzi równanie $b_0=d^{(s)}-w_0^T x^{(s)}$, biorąc jeden z wektorów podpierających (taki dla którego odpowiadający mu mnożnik Lagrange'a jest niezerowy) dostajemy:

$$b_0=\begin{bmatrix}-1{,}5645\\-0{,}5323\end{bmatrix}^T\cdot\begin{bmatrix}1{,}241\\-4{,}208\end{bmatrix}+1=1{,}2984$$

Liniowy klasyfikator SVM - przykład

Ostatecznie otrzymujemy zatem regułę decyzyjną liniowego klasyfikatora SVM postaci:

$$\psi^{linSVM}(x)=\begin{cases}-1 & gdy\ \left[1{,}5645\quad 0{,}5323\right]\cdot x+1{,}2984\leq 0\\ 1 & gdy\ \left[1{,}5645\quad 0{,}5323\right]\cdot x+1{,}2984>0\end{cases}.$$

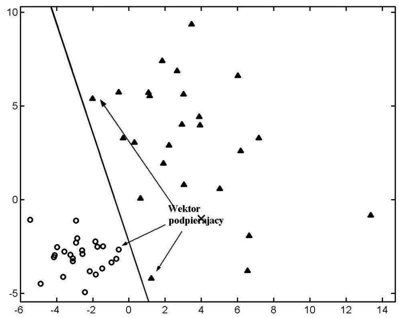
Dla przykładu zaklasyfikujmy obiekt opisany wektorem cech równym $\left[4\quad -1\right]^T$:

$$\left[1{,}5645\quad 0{,}5323\right]\cdot\begin{bmatrix}4\\-1\end{bmatrix}+1{,}2984=6{,}258-0{,}5323+1{,}2984=$$

$$=7{,}0241$$

czyli rozpatrywany obiekt należy do klasy2.

Liniowy klasyfikator SVM – przykład



Prosta $1,5645x_1 + 0,5323x_2 + 1,2984 = 0$ jest powierzchnią decyzyjną klasyfikatora – optymalną hiperpłaszczyzną rozdzielającą

NIELINIOWY KLASYFIKATOR SVM

Idea:

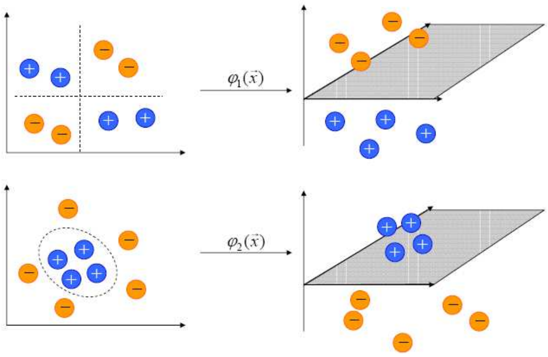
- Wyznaczenie optymalnej hiperpłaszczyzny rozdzielającej, jak w klasyfikatorze liniowym, tyle że w przetransformowanej za pomocą pewnego przekształcenia φ przestrzeni wejściowej
- W nowej wysokowymiarowej przestrzeni próbki są już liniowo separowane

Problem

- Znalezienie przekształcenia φ.

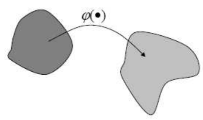
NIELINIOWY KLASYFIKATOR SVM

LINIOWA SEPAROWALNOŚĆ W PRZ. WYŻSZEGO WYMIARU



NIELINIOWY KLASYFIKATOR SVM

$$\varphi(x_i) = \varphi\left(\begin{bmatrix} x_{i1} \\ \vdots \\ x_{id} \end{bmatrix}\right) = \tilde{x} = \begin{bmatrix} \tilde{x}_{i1} \\ \vdots \\ \tilde{x}_{ie} \end{bmatrix}$$



Funkcja jądra:

$$K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$$

NIELINIOWY KLASYFIKATOR SVM

- Dzięki funkcji jądra nie jest konieczne operowanie w sposób jawny na obrazach $\varphi(x_i)$
- Nie ma potrzeby jawnego podawania postaci odwzorowania φ
- Warunki, które musi spełniać funkcja jądra określa tw. Melcera

NIELINIOWY KLASYFIKATOR SVM

PRZYKŁADY FUNKCJI JĄDRA:

- Jądro wielomianowe

$$K(x_i, x_j) = (x_i^T x_j + \beta)^p$$

- Jądro Gaussa

$$K(x_i, x_j) = \exp\left(-\frac{1}{2\sigma^2}\|x_i - x_j\|^2\right)$$

gdzie β, p, σ są stałymi

UCZENIE NIELINIOWEGO SVM

(przypadek liniowo nieseparowalny)

Zadanie dualne optymalizacji

Wyznacz mnożniki Lagrange'a α_i dla i=1,…,N, tak by maksymalizować funkcję celu:

$$Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j K(x_i, x_j)$$

przy ograniczeniach:

1. $\sum_{i=1}^{N} \alpha_i d_i = 0$
2. $C \geq \alpha_i \geq 0$, i=1,...,N

UCZENIE NIELINIOWEGO SVM

- Wektor wag określających OSH w przestrzeni wtórnej:

$$w = \sum_{i=1}^{N_s} \alpha_i d_i \varphi(x_i),$$

gdzie N_s -liczba wektorów podpierających.

- Funkcja dyskryminacyjna:

$$g(x) = w^T \varphi(x) + b_0$$

Podstawiając:

$$g(x) = \left(\sum_{i=1}^{N_s} \alpha_i d_i \varphi(x_i) \right)^T \varphi(x) + b_0 = \sum_{i=1}^{N_s} \alpha_i d_i \varphi(x_i)^T \varphi(x) + b_0$$

UCZENIE NIELINIOWEGO SVM

I dalej:

$$g(x) = \sum_{i=1}^{N_s} \alpha_i d_i K(x_i, x) + b_0$$

- Wartość b_0 uzyskujemy z odpowiedniego warunku Kuhna-Tuckera poprzez uśrednienei po wektorach podpierających

$$b_0 = \frac{1}{N_s} \sum_{j=1}^{N_s} \left(d_j - \sum_{j=1}^{N_s} d_i \alpha_i K(x_i, x_j) \right)$$

KLASYFIKATOR SVM – UWAGI

ZALETY:

- Maksymalizacja marginesu zapewnia dobre zdolności generalizacyjne
- Dokonywanie transformacji w sposób niejawny
- Rozwiązaniem problemu optymalizacji jest rozwiązanie globalne
- Rozwiązanie uzyskuje się tylko w oparciu o mały podzbiór zbioru uczącego najbardziej znaczących wektorów (leżących najbiżej OSH).

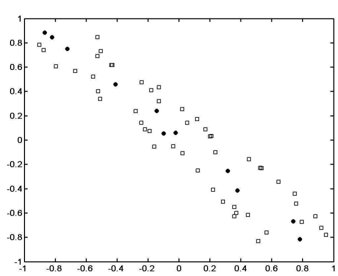
PROBLEM: Wybór odpowiedniej postaci funkcji jądra.

PORÓWNANIE SVM z kNN

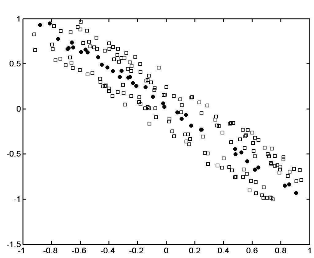
- W tym przykładzie porównamy sprawność gaussowskiego klasyfikatora SVM w odniesieniu do klasyfikatora k-NN.

- Zbiór danych do tego przykładu liczący 500 próbek został podzielony losowo na trzy rozłączne podzbiory. Walidacyjny liczący 200 punktów, który będzie służył do optymalnego doboru parametrów używanych klasyfikatorów, drugi uczący oraz zbiór testowy

PORÓWNANIE SVM z kNN



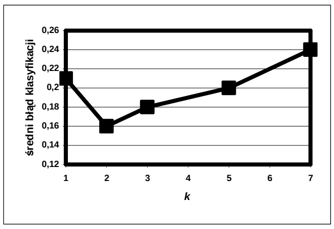
Zbiór uczący



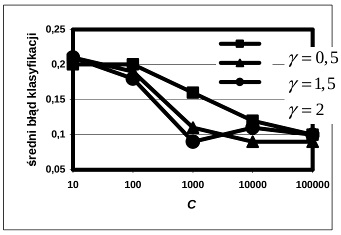
Zbiór testowy

PORÓWNANIE SVM z kNN

Najwyższą sprawność kNN osiąga on dla $k = 2$
Sprawność Gaussowskiego klasyfikatora SVM najwyższa dla $\gamma = 1,5$ oraz $C = 10000$.

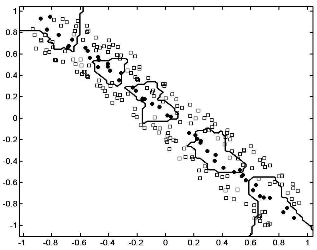


Wykres zależności błędu klasyfikacji klasyfikatora *k*-NN od parametru *k*.

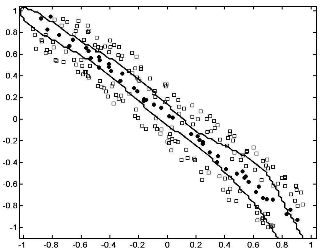


Wykres zależności błędu klasyfikacji Gaussowskiego klasyfikatora SVM od parametru *C* dla γ =0,5; 1,5; 2.

PORÓWNANIE SVM z kNN



Rys.24. Powierzchnie decyzyjne klasyfikatora 2-NN na tle zbioru testowego



Powierzchnie decyzyjne gaussowskiego klasyfikatora SVM dla $\gamma = 1,5$ oraz C=10000 na tle zbioru testowego.
.