Metoda wektorów wspierających

- wstęp -

Jacek Kluska

Politechnika Rzeszowska

Wprowadzenie

1. V. Vapnik, *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995
2. Statystyczne metody uczenia maszynowego:
    1. Sieci neuronowe wykorzystują *minimalizację ryzyka empirycznego (ERM)*, co prowadzi do minimalizacji błędu uczenia.
    2. SVM wykorzystuje *minimalizację ryzyka strukturalnego (SRM)*, co prowadzi do minimalizacji błędu uogólniania (cel uczenia).
3. SVM wykorzystuje się do

    - **Klasyfikacji**
    - **Regresji.**

Problem klasyfikacji

Dla danych uczących

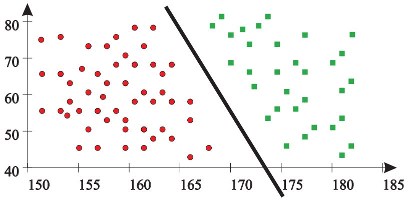
$$\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_Q, y_Q)\} \subset \mathbb{X} \times \{-1, 1\}, \quad \mathbb{X} \subset \mathbb{R}^n,$$

znaleźć hiperpłaszczyznę

$$< \mathbf{x}, \mathbf{w} > +b = x_1 w_1 + \ldots + x_n w_n + b = 0$$

dzieląca zbiór $\mathbb{X}$ na 2 podzbiory

$$\mathbb{X} = \{\mathbf{x}_i \in \mathbb{X} \mid y_i = -1\} \cup \{\mathbf{x}_i \in \mathbb{X} \mid y_i = +1\}$$



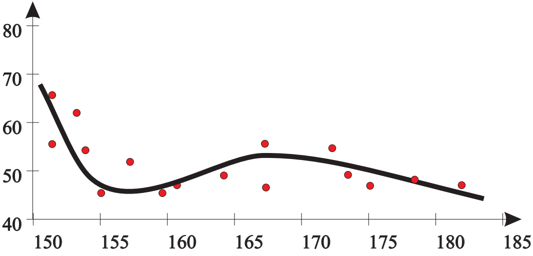
Rysunek: Klasyfikacja.

Problem regresji

Dla danych uczących

$$\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_Q, y_Q)\} \subset \mathbb{R}^n \times \mathbb{R},$$

znaleźć $f(\mathbf{x})$ z danej klasy dopuszczalnych, która najlepiej aproksymuje dane.



Rysunek: Regresja.

Optymalna hiperpłaszczyzna separująca (OSH)

Dla danych uczących

$$\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_Q, y_Q)\} \subset \mathbb{X} \times \{-1, 1\}, \quad \mathbb{X} \subset \mathbb{R}^n,$$

znaleźć taką hiperpłaszczyznę

$$< \mathbf{x}, \mathbf{w} > + b = 0,$$

która **bezbłędnie separuje** dane (perceptron) i jednocześnie **maksymalizuje** najmniejszą odległość każdego punktu $\mathbf{x}_i$, $(i = 1, \ldots, Q)$ od hiperpłaszczyzny.

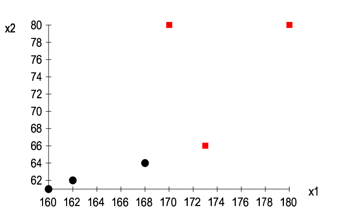
- Perpceptron separuje dane – nie robi tego jednak optymalnie (dla danych liniowo separowalnych).

Przykład - dane liniowo separowalne

Example

$$\mathbf{X} = \begin{bmatrix} 180 & 173 & 170 & 160 & 162 & 168 \\ 80 & 66 & 80 & 61 & 62 & 64 \end{bmatrix}^T \begin{array}{r} \leftarrow \text{wysokość} \\ \leftarrow \text{waga} \end{array}$$

$$\mathbf{y} = \begin{bmatrix} -1 & -1 & -1 & +1 & +1 & +1 \end{bmatrix}^T.$$



Rysunek: Jak znaleźć OSH ?

Dane liniowo separowalne

Założenie:

dane są liniowo separowalne.

Fakt 1.

Odległość $\rho\left(\mathbf{x}_i, \mathbf{w}, b\right)$ punktu $\mathbf{x}_i \in \mathbb{R}^n$ od hiperpłaszczyzny $\langle \mathbf{x}_i, \mathbf{w} \rangle + b = 0$ jest równa

$$\rho\left(\mathbf{x}_i, \mathbf{w}, b\right) = \frac{\left|\langle \mathbf{x}_i, \mathbf{w} \rangle + b\right|}{\|\mathbf{w}\|}.$$

Fakt 2.

Jeżeli istnieje (jakakolwiek) hiperpłaszczyzna separująca, to

$$y_i\left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b\right) - 1 \geqslant 0, \qquad i = 1, \ldots, Q.$$

Dane liniowo separowalne - Dow. Faktu 2

Dowód.

$$\exists\ \mathbf{v}, r: \qquad y = \operatorname{sign}(n) = \operatorname{sign}(\langle \mathbf{v}, \mathbf{x}_i \rangle + r).$$

Stąd, $\exists\ \varepsilon > 0:$

$$n = \langle \mathbf{v}, \mathbf{x}_i \rangle + r \geqslant \varepsilon \quad \Leftrightarrow \quad y_i = +1,$$
$$n = \langle \mathbf{v}, \mathbf{x}_i \rangle + r \leqslant -\varepsilon \quad \Leftrightarrow \quad y_i = -1.$$

Niech $\mathbf{w} := \mathbf{v}/\varepsilon$, $b = r/\varepsilon$

$$\langle \mathbf{w}, \mathbf{x}_i \rangle + b \geqslant +1 \quad \Leftrightarrow \quad y_i = +1$$
$$\langle \mathbf{w}, \mathbf{x}_i \rangle + b \leqslant -1 \quad \Leftrightarrow \quad y_i = -1.$$

Ostatecznie

$$y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 \geqslant 0\,, \quad \forall\ i = 1, \ldots, Q.$$

Dane liniowo separowalne - c.d.

Fakt 3. Margines

$$M = \min_{\mathbf{x}_i : y_i = -1} \rho\left(\mathbf{x}_i, \mathbf{w}, b\right) + \min_{\mathbf{x}_i : y_i = +1} \rho\left(\mathbf{x}_i, \mathbf{w}, b\right) = \frac{2}{\|\mathbf{w}\|}$$

Musimy znaleźć OSH, która maksymalizuje minimalną odległość między hiperpłaszczyzną i zadanymi punktami:

$$\max_{h_j} \min_{\mathbf{x}_i \in \mathbb{X}} \rho\left(\mathbf{x}_i, \mathbf{w}, b\right) \; = \; \max_{h_j} \frac{1}{\|\mathbf{w}\|}.$$

Postawienie problemu

$$
\begin{cases}
\min \quad \Phi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 \\
y_i\left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b\right) - 1 \geqslant 0, \quad i = 1, \ldots, Q.
\end{cases}
$$

Funkcja Lagrange'a

1. Wektor mnożników Lagrange'a

$$\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_Q]^T \geqslant \mathbf{0}$$

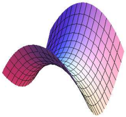
2. Wektor ograniczeń

$$\mathbf{g}(\mathbf{w}, b) \geqslant \mathbf{0}$$
$$\Updownarrow$$
$$[y_1 (\langle \mathbf{w}, \mathbf{x}_1 \rangle + b) - 1, \ldots, y_Q (\langle \mathbf{w}, \mathbf{x}_Q \rangle + b) - 1]^T \geqslant \mathbf{0}$$

3. Funkcja Lagrange'a

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2} \langle \mathbf{w}, \mathbf{w} \rangle - \langle \boldsymbol{\alpha}, \mathbf{g}(\mathbf{w}, b) \rangle$$

Punkt siodłowy funkcji Lagrange'a



Rysunek: Punkt siodłowy.

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) \underset{(\mathbf{w},b)}{\longrightarrow} \min, \qquad L(\mathbf{w}, b, \boldsymbol{\alpha}) \underset{\boldsymbol{\alpha} \geqslant \mathbf{0}}{\longrightarrow} \max$$

$$\frac{\partial}{\partial b} L(\mathbf{w}, b, \boldsymbol{\alpha}) = 0, \qquad \frac{\partial}{\partial \mathbf{w}} L(\mathbf{w}, b, \boldsymbol{\alpha}) = \mathbf{0}$$

Stąd,

$$\sum_{i=1}^{Q} \alpha_i y_i = 0, \qquad \mathbf{w}^0 = \sum_{i=1}^{Q} \alpha_i y_i \mathbf{x}_i$$

- **Istnieje jedyne minimum globalne** funkcji $\frac{1}{2}\langle \mathbf{w}, \mathbf{w} \rangle$.

Dlaczego wektory wspierające są tak ważne ?

- Wektory wspierające to takie punkty $\mathbf{x}_i$ dla których $\alpha_i > 0$.
  Stąd

$$\mathbf{w}^0 = \textstyle\sum_{i=1}^{Q} \alpha_i^0 y_i \mathbf{x}_i = \displaystyle\sum_{i \in SV} \alpha_i^0 y_i \mathbf{x}_i$$

$$SV = \{i \mid \alpha_i > 0\}$$

- Można wykazać, że

$$b^0 = -\frac{1}{2} \left\langle \mathbf{w}^0, \mathbf{x}_r + \mathbf{x}_s \right\rangle$$

$$r, s \in SV, \quad y_r = -1, \quad y_s = +1$$

Example

Niech

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \\ \mathbf{x}_4 \end{bmatrix} = \begin{bmatrix} 4 & 3 \\ 2 & 1 \\ 6 & 4 \\ 1 & 1 \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ -1 \\ 1 \end{bmatrix}.$$

Hessjan

$$\mathbf{H} = \begin{bmatrix} 25 & -11 & 36 & -7 \\ -11 & 5 & -16 & 3 \\ 36 & -16 & 52 & -10 \\ -7 & 3 & -10 & 2 \end{bmatrix}.$$

Prosty przykład - c.d.

Example

Problem dualny

$$\max_{\alpha_1,\ldots,\alpha_4}\left\{-\frac{1}{2}\begin{bmatrix}\alpha_1\\ \alpha_2\\ \alpha_3\\ \alpha_4\end{bmatrix}^T\begin{bmatrix}25 & -11 & 36 & -7\\ -11 & 5 & -16 & 3\\ 36 & -16 & 52 & -10\\ -7 & 3 & -10 & 2\end{bmatrix}\begin{bmatrix}\alpha_1\\ \alpha_2\\ \alpha_3\\ \alpha_4\end{bmatrix}+\sum_{i=1}^{4}\alpha_i\right\}$$

ograniczenia

$$-\alpha_1+\alpha_2-\alpha_3+\alpha_4=0,$$

$$\alpha_1\geqslant 0,\ \alpha_2\geqslant 0,\ \alpha_3\geqslant 0,\ \alpha_4\geqslant 0.$$

Prosty przykład - c.d.

Example

$$\begin{bmatrix} \alpha_1^0 \\ \alpha_2^0 \\ \alpha_3^0 \\ \alpha_4^0 \end{bmatrix} = \begin{bmatrix} 0.25 \\ 0.25 \\ 0 \\ 0 \end{bmatrix} \begin{array}{l} \Rightarrow \mathbf{x}_1 \in SVs \\ \Rightarrow \mathbf{x}_2 \in SVs \\ \Rightarrow \mathbf{x}_3 \notin SVs \\ \Rightarrow \mathbf{x}_4 \notin SVs \end{array} \Rightarrow \begin{bmatrix} w_1^0 \\ w_2^0 \end{bmatrix} = \begin{bmatrix} -0.5 \\ -0.5 \end{bmatrix}, \quad b^0 = 2.5.$$

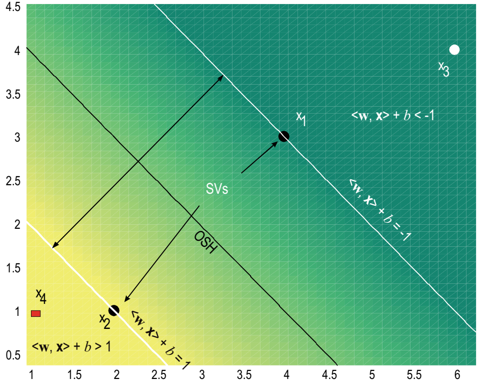
minimalizowana funkcja osiąga minimum globalne, oraz

$$OSH : \langle \mathbf{w}^0, \mathbf{x} \rangle + b^0 = [-0.5, -0.5]\,\mathbf{x} + 2.5 = 0.$$

$$M = 2/\sqrt{(-0.5)^2 + (-0.5)^2} \approx 2.8284$$

Prosty przykład - c.d.

Example



Rysunek: Punkty: x_1, x_2 są SVs, ale x_3, x_4 nie są.

Dane liniowo separowalne - wnioski

1. $\mathbf{w}^0$ jest kombinacją wypukłą wektorów SV mnożonych przez "wartości etykiet" y_i.
2. **Jedynie $\mathbf{x}_i$ będące SV decydują o rozwiązaniu.**
3. Preocedurę SVM można nazwać **perceptronem optymalnym:**

$$\begin{aligned} output &= \operatorname{sign}\left(\left\langle \mathbf{w}^0, \mathbf{x} \right\rangle + b^0\right) \\ &= \operatorname{sign}\left(\sum_{i \in SV} \alpha_i^0 \left\langle y_i \mathbf{x}_i, \mathbf{x} \right\rangle + b^0\right) \end{aligned}$$

Dążymy do sformułowania problemu dualnego

Niech $\mathbf{w} = \mathbf{w}^0$ i $b = b^0$. Opuszczając indeksy $(\cdot)^0$ mamy

$$L = \frac{1}{2}\langle \mathbf{w}, \mathbf{w} \rangle - \langle \boldsymbol{\alpha}, \mathbf{g}(\mathbf{w}, b) \rangle$$

Fakty

1. $$\frac{1}{2}\langle \mathbf{w}, \mathbf{w} \rangle = \frac{1}{2}\left\langle \sum_{i=1}^{Q} \alpha_i y_i \mathbf{x}_i, \sum_{i=1}^{Q} \alpha_i y_i \mathbf{x}_i \right\rangle$$
2. $$\sum_{i=1}^{Q} \alpha_i y_i = 0$$
3. $$\sum_{i=1}^{Q} \alpha_i y_i \langle \mathbf{w}, \mathbf{x}_i \rangle = \left\langle \mathbf{w}, \sum_{i=1}^{Q} \alpha_i y_i \mathbf{x}_i \right\rangle = \langle \mathbf{w}, \mathbf{w} \rangle$$
4. $$\langle \boldsymbol{\alpha}, \mathbf{g}(\mathbf{w}, b) \rangle = \sum_{i=1}^{Q} \alpha_i y_i \langle \mathbf{w}, \mathbf{x}_i \rangle - \sum_{i=1}^{Q} \alpha_i = \langle \mathbf{w}, \mathbf{w} \rangle - \sum_{i=1}^{Q} \alpha_i$$

Postawienie problemu dualnego

$$L\left(\mathbf{w},b,\boldsymbol{\alpha}\right)=-\frac{1}{2}\sum_{i=1}^{Q}\sum_{j=1}^{Q}\alpha_i\alpha_j y_i y_j\left\langle \mathbf{x}_i,\mathbf{x}_j\right\rangle+\sum_{i=1}^{Q}\alpha_i$$

Niech

$$W\left(\boldsymbol{\alpha}\right)=-\frac{1}{2}\left\langle \boldsymbol{\alpha},\mathbf{H}\boldsymbol{\alpha}\right\rangle+\sum_{i=1}^{Q}\alpha_i$$

gdzie macierz Hessjanu

$$\mathbf{H}=\left\{\left\langle y_i\mathbf{x}_i,y_j\mathbf{x}_j\right\rangle\right\}_{Q\times Q}$$

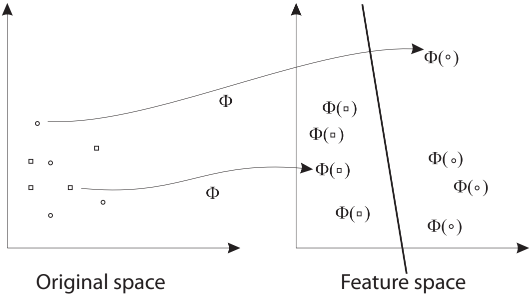
Problem dualny

$$\begin{cases}\max \quad W\left(\boldsymbol{\alpha}\right)=-\frac{1}{2}\left\langle \boldsymbol{\alpha},\mathbf{H}\boldsymbol{\alpha}\right\rangle+\sum_{i=1}^{Q}\alpha_i\\ \left\langle \boldsymbol{\alpha},\mathbf{y}\right\rangle=0\\ \boldsymbol{\alpha}\geqslant\mathbf{0}\end{cases}$$

- **Istnieje jedyne maksimum globalne** funkcji $W\left(\boldsymbol{\alpha}\right)$.

Problem liniowo nieseparowalny

Praktyka: dane nie są liniowo separowalne.
Wprowadzamy **klasyfikator nieliniowy**. Może on być otrzymany poprzez odwzorowanie oryginalnej przestrzeni n-wymiarowej wektorów wejściowych $[x_1, \ldots, x_n]$ do **przestrzeni cech o dużym wymiarze**. Nowy klasyfikator będzie utożsamiony z **uogólnioną hiperpłaszczyzną separującą (GSH)**.



Rysunek: Odwzorowanie przestrzeni oryginalnej do przestrzeni cech.

Example

Niech $\mathbf{x} = [x_1, x_2]^T \in \mathbb{R}^2, \quad \boldsymbol{\phi} : \mathbb{R}^2 \to \mathbb{R}^6$

$$\boldsymbol{\phi}(\mathbf{x}) = \begin{bmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \\ \sqrt{2}x_1 \\ \sqrt{2}x_2 \\ 1 \end{bmatrix}$$

Przykłady odwzorowań nieliniowych - c.d.

Example

Niech $\mathbf{x} = [x_1, x_2, x_3]^T \in \mathbb{R}^3, \quad \boldsymbol{\phi} : \mathbb{R}^3 \to \mathbb{R}^9$

$$\boldsymbol{\phi}(\mathbf{x}) = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_1^2 \\ x_2^2 \\ x_3^2 \\ \sqrt{2}x_1x_2 \\ \sqrt{2}x_1x_3 \\ \sqrt{2}x_2x_3 \end{bmatrix}$$

Iloczyn skalarny w przestrzeni cech - przykłady

Example

$$\boldsymbol{\phi}(\mathbf{x}) = \left[x_1^2,\ x_2^2,\ \sqrt{2}x_1x_2,\ \sqrt{2}x_1,\ \sqrt{2}x_2, 1\right]^T,$$

$$\langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{y}) \rangle = (1 + \langle \mathbf{x}, \mathbf{y} \rangle)^2 .$$

Example

$$\boldsymbol{\phi}(\mathbf{x}) = \left[x_1,\ x_2,\ x_3,\ x_1^2,\ x_2^2,\ x_3^2,\ \sqrt{2}x_1x_2,\ \sqrt{2}x_1x_3,\ \sqrt{2}x_2x_3\right]^T,$$

$$\langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{y}) \rangle = (1 + \langle \mathbf{x}, \mathbf{y} \rangle) \langle \mathbf{x}, \mathbf{y} \rangle .$$

- Klasyfikator SVM będzie wykorzystywał iloczyn skalarny zamiast $\boldsymbol{\phi}(\mathbf{x})$.

Uwagi

- Jeżeli $\dim \mathbf{x}$ jest duży, to $\dim \boldsymbol{\phi}$ jest olbrzymi, np. wielomian 2-go st. 256-zmiennych zawiera 35245 współrzędnych.
- **Interpretacja.**
  - Zamiast
  $$\langle \mathbf{x}, \mathbf{y} \rangle$$
  wykorzystamy
  $$\langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{y}) \rangle$$
  - W przestrzeni oryginalnej $\mathbf{x} = [x_1, x_2]^T$ chcieliśmy otrzymać separatrysę
  $$w_1 x_1 + w_2 x_2 + b = 0$$
  W nowej przestrzeni będziemy poszukiwać hiperpowierzchni
  $$w_1 x_1^2 + w_2 x_2^2 + w_3 x_1 x_2 + w_4 x_1 + w_5 x_2 + b = 0$$

Przykłady funkcji jądra

$$K(\mathbf{u},\mathbf{v}) = \langle \boldsymbol{\phi}(\mathbf{u}), \boldsymbol{\phi}(\mathbf{v}) \rangle$$

1. 'linear':
$$K(\mathbf{u},\mathbf{v}) = \langle \mathbf{u},\mathbf{v} \rangle$$
2. 'polynomial':
$$K(\mathbf{u},\mathbf{v}) = (1 + \langle \mathbf{u},\mathbf{v} \rangle)^p \quad , \qquad p = 2, 3, \ldots$$
3. 'rbf':
$$K(\mathbf{u},\mathbf{v}) = \exp\left(-\frac{\|\mathbf{u}-\mathbf{v}\|^2}{2\sigma^2}\right)$$
4. 'erbf':
$$K(\mathbf{u},\mathbf{v}) = \exp\left(-\frac{\|\mathbf{u}-\mathbf{v}\|}{2\sigma^2}\right)$$
5. 'sigmoid':
$$K(\mathbf{u},\mathbf{v}) = \operatorname{tgh}(p \cdot \langle \mathbf{u},\mathbf{v} \rangle + q)$$

Przykłady funkcji jądra - c.d.

6. 'spline' ($n = \dim \mathbf{x}$)

$$K(\mathbf{u}, \mathbf{v}) = \prod_{i=1}^{n} \left(1 + u_i v_i + \frac{1}{2} u_i v_i q_i - \frac{1}{6} q_i^3 \right)$$

   gdzie

$$q_i = \min(u_i, v_i), \qquad i = 1, 2, \ldots, n$$

7. 'bspline' ($n = \dim \mathbf{x}$)

$$K(\mathbf{u}, \mathbf{v}) = \prod_{i=1}^{n} z_i$$

   gdzie

$$z_i = \sum_{r=0}^{2(N+1)} (-1)^r \binom{2N+2}{r} \times \left[\max(0, u_i - v_i + N + 1 - r) \right]^{2N+1}$$

   N - parametr, np. $N = 1$.

Uogólniona hiperpłaszczyzna separująca (GSH)

- Zmienne swobodne ξ_i:

$$\boldsymbol{\xi} = [\xi_1, \ldots, \xi_Q]^T \geqslant \mathbf{0}.$$

- Można wykazać, że problem znajdowania GSH sprowadza się do rozwiązania zadania

$$\begin{cases} \min \quad \Phi(\mathbf{w}, \boldsymbol{\xi}) = \frac{1}{2}\langle \mathbf{w}, \mathbf{w} \rangle + C \sum_{i=1}^{Q} \xi_i \\ y_i \left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b \right) \geqslant 1 - \xi_i \;, \quad i = 1, \ldots, Q \\ \boldsymbol{\xi} \geqslant \mathbf{0} \end{cases}$$

- $C > 0$ jest zadane przez projektanta **a'priori**.
- Dla danych liniowo separowalnych otrzymujemy

$$\mathbf{w}^0 = \sum_{i=1}^{Q} \alpha_i^0 y_i \mathbf{x}_i = \sum_{i \in SV} \alpha_i^0 y_i \mathbf{x}_i$$

oraz b^0 wynika z warunków Kuhna-Tuckera.

Postawienie problemu dualnego dla danych liniowo nieseparowalnych

- Hessjan

$$\mathbf{H} = \left\{ y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \right\}_{Q \times Q}$$

- **Problem dualny**

$$\begin{cases} \max \quad W(\boldsymbol{\alpha}) = -\dfrac{1}{2} \langle \boldsymbol{\alpha}, \mathbf{H}\boldsymbol{\alpha} \rangle + \sum_{i=1}^{Q} \alpha_i \\ \langle \boldsymbol{\alpha}, \mathbf{y} \rangle = 0 \\ 0 \leqslant \alpha_i \leqslant C \,, \quad i = 1, \ldots, Q \end{cases}$$

- **Istnieje jedyne maksimum globalne funkcji** $W(\boldsymbol{\alpha})$.

Rozwiązanie problemu dualnego dla danych liniowo nieseparowalnych

OSH:

$$\underline{\underline{f(\mathbf{x}) = \operatorname{sign}\left(\sum_{i \in SV} \alpha_i^0 K(\mathbf{x}_i, \mathbf{x}) + b^0\right)}}$$

gdzie

$$b^0 = -\frac{1}{2} \sum_{i \in SV} \alpha_i^0 y_i \left[K(\mathbf{x}_i, \mathbf{x}_r) + K(\mathbf{x}_i, \mathbf{x}_s)\right]$$

$$r, s \in SV, \quad y_r = -1, \quad y_s = +1$$

Obliczanie funkcji jądra - przykłady

Example

'linear'

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \ \mathbf{x}_2 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \ \mathbf{x}_3 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\mathbf{y} = [-1, +1, +1]$$

$$K(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle \Rightarrow \mathbf{H} = \left\{ y_i y_j \left\langle \mathbf{x}_i, \mathbf{x}_j \right\rangle \right\}_{Q \times Q}$$

$$\mathbf{H} = \begin{bmatrix} \langle \mathbf{x}_1, \mathbf{x}_1 \rangle & -\langle \mathbf{x}_1, \mathbf{x}_2 \rangle & -\langle \mathbf{x}_1, \mathbf{x}_3 \rangle \\ -\langle \mathbf{x}_2, \mathbf{x}_1 \rangle & \langle \mathbf{x}_2, \mathbf{x}_2 \rangle & \langle \mathbf{x}_2, \mathbf{x}_3 \rangle \\ -\langle \mathbf{x}_3, \mathbf{x}_1 \rangle & \langle \mathbf{x}_3, \mathbf{x}_2 \rangle & \langle \mathbf{x}_3, \mathbf{x}_3 \rangle \end{bmatrix} = \begin{bmatrix} 2 & -3 & -1 \\ -3 & 5 & 2 \\ -1 & 2 & 1 \end{bmatrix}$$

Obliczanie funkcji jądra - przykłady

Example

'poly'

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix},\ \mathbf{x}_2 = \begin{bmatrix} 2 \\ 1 \end{bmatrix},\ \mathbf{x}_3 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\mathbf{y} = [-1, +1, +1]$$

$$K(\mathbf{u}, \mathbf{v}) = (1 + \langle \mathbf{u}, \mathbf{v} \rangle)^p,\ p = 2, 3, \ldots$$

$$\mathbf{H} = \left\{ y_i y_j \left(1 + \langle \mathbf{x}_i, \mathbf{x}_j \rangle\right)^p \right\}_{Q \times Q}$$

$$\mathbf{H} = \begin{bmatrix} 3^p & -4^p & -2^p \\ -4^p & 6^p & 3^p \\ -2^p & 3^p & 2^p \end{bmatrix},\ p = 2, 3, \ldots$$

Obliczanie funkcji jądra - przykłady

Example

'rbf'

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix},\ \mathbf{x}_2 = \begin{bmatrix} 2 \\ 1 \end{bmatrix},\ \mathbf{x}_3 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\mathbf{y} = [-1, +1, +1]$$

$$K(\mathbf{u}, \mathbf{v}) = \exp\left(-\frac{1}{2\sigma^2}\left\|\mathbf{u}-\mathbf{v}\right\|^2\right) = \exp\left(-\frac{1}{2\sigma^2}\left\langle \mathbf{u}-\mathbf{v},\ \mathbf{u}-\mathbf{v}\right\rangle\right)$$

$$\mathbf{H} = \left\{ y_i y_j \exp\left(-\frac{1}{2\sigma^2}\left\langle \mathbf{x}_i - \mathbf{x}_j,\ \mathbf{x}_i - \mathbf{x}_j \right\rangle\right)\right\}_{Q\times Q}$$

Dla $\sigma = 1$:

$$\mathbf{H} = \begin{bmatrix} 1 & -e^{-0.5} & -e^{-0.5} \\ -e^{-0.5} & 1 & e^{-1} \\ -e^{-0.5} & e^{-1} & 1 \end{bmatrix}.$$

Obliczanie funkcji jądra - przykłady w Matlabie

% kernel='rbf'

```
X=[1, 1;2,1;1,0]; y=[-1; 1; 1];
Q = size(X,1); % liczność zbioru danych
sigma=1;
for i=1:Q
    for j=1:Q
        u=X(i,:); v=X(j,:);
        H(i,j)=y(i)*y(j)*exp(-(u-v)*(u-v)'/(2*sigma^2));
    end
end
%=============
```

Obliczanie funkcji jądra - przykłady w Matlabie - c.d.

```
% Inne funkcje jądra:    H(i,j)=y(i)*y(j)*k
% 'linear'
k = u*v';
% 'poly'
k = (u*v' + 1)^p1;
% 'rbf'
k = exp(-(u-v)*(u-v)'/(2*p1^2));
% 'erbf'
k = exp(-sqrt((u-v)*(u-v)')/(2*p1^2));
% 'sigmoid'
k = tanh(p1*u*v'/length(u) + p2);
%===========
```

- S. R. Gunn, *Support Vector Machines for Classification and Regression*, University of Southampton, 1998.

Normalizacja danych

Dane

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_Q \end{bmatrix} = \begin{bmatrix} (x_{1,1} & \cdots & x_{1,n}) \\ (x_{2,1} & \cdots & x_{2,n}) \\ \vdots & \vdots & \vdots \\ (x_{Q,1} & \cdots & x_{Q,n}) \end{bmatrix} \in \mathbb{R}^{Q \times n}, \ \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_Q \end{bmatrix} \in \{-1, +1\}$$

$x_{i,k}$ powinno być znormalizowane:

$$L \leqslant x_{i,k} \leqslant U, \qquad i = 1, 2, \ldots, Q, \quad k = 1, 2, \ldots, n,$$

gdzie:

- $[L, U] = [-1, 1]$ – 'linear', 'poly', 'rbf', 'erbf', 'sigmoid'
- $[L, U] = [0, 1]$ – 'spline', 'b-spline'.

Normalizacja danych - c.d.

Niech

$$\mathbf{M} = [M_1, \cdots, M_n], \quad \mathbf{m} = [m_1, \cdots, m_n], \quad \mathbf{R} = \mathbf{M} - \mathbf{m} = [r_1, \cdots, r_n],$$

gdzie

- M_k jest maksymalną wartością k-tej współrzędnej po wszystkich wektorach $\mathbf{x}_i$:

$$M_k = \max_{i=1,\ldots,Q} \{x_{i,k}\}, \quad \text{for} \quad k = 1, 2, \ldots, n,$$

- m_k jest minimalną wartością k-tej współrzędnej po wszystkich wektorach $\mathbf{x}_i$:

$$m_k = \min_{i=1,\ldots,Q} \{x_{i,k}\}, \quad \text{for} \quad k = 1, 2, \ldots, n.$$

Normalizacja homogeniczna i niehomogeniczna

$$r_k = \begin{cases} M_k - m_k & \text{normalizacja homogeniczna} \\ \\ \max\limits_{k=1,\dots,n} \{M_k - m_k\} & \text{normalizacja niehomogeniczna} \end{cases}$$

gdzie $k = 1, 2, \dots, n$. Dla k-tej współrzędnej $\mathbf{x}$ wprowadzamy

$$a_k = \frac{U - L}{r_k}, \quad b_k = L - \frac{U - L}{r_k} m_k, \qquad k = 1, 2, \dots, n.$$

Stąd

$$\mathbf{X}^{norm} = \begin{bmatrix} a_1 x_{1,1} + b_1, & \cdots, & a_n x_{1,n} + b_n \\ a_1 x_{2,1} + b_1, & \cdots, & a_n x_{2,n} + b_n \\ \vdots & \vdots & \vdots \\ a_1 x_{Q,1} + b_1, & \cdots, & a_n x_{Q,n} + b_n \end{bmatrix}.$$

Normalizacja homogeniczna i niehomogeniczna - przykłady

Example

Niech $[L, U] = [-1, 1]$,

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \\ \mathbf{x}_4 \\ \mathbf{x}_5 \\ \mathbf{x}_6 \end{bmatrix} = \begin{bmatrix} 180 & 80 \\ 173 & 66 \\ 170 & 80 \\ 160 & 61 \\ 162 & 62 \\ 168 & 64 \end{bmatrix} \to \mathbf{X}^{norm} = \begin{bmatrix} 1.00 & 1.00 \\ 0.30 & -0.47 \\ 0.00 & 1.00 \\ -1.00 & -1.00 \\ -0.88 & -0.89 \\ -0.20 & -0.68 \end{bmatrix}.$$

$a_1 = 1/10, b_1 = -17$ - dla pierwszej kolumny $\mathbf{X}^{norm}$,
$a_2 = 2/19, b_2 = -141/19$ - dla drugiej kolumny $\mathbf{X}^{norm}$.

Normalizacja niehomogeniczna i niehomogeniczna - przykłady

Example

Niech $[L, U] = [-1, 1]$,

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \\ \mathbf{x}_4 \\ \mathbf{x}_5 \\ \mathbf{x}_6 \end{bmatrix} = \begin{bmatrix} 180 & 80 \\ 173 & 66 \\ 170 & 80 \\ 160 & 61 \\ 162 & 62 \\ 168 & 64 \end{bmatrix} \to \mathbf{X}^{norm} = \begin{bmatrix} 1.0 & 0.9 \\ 0.3 & -0.5 \\ 0.0 & 0.9 \\ -1.0 & -1.0 \\ -0.8 & -0.9 \\ -0.2 & -0.7 \end{bmatrix}.$$

$a_1 = 1/10, b_1 = -17$ - dla pierwszej kolumny $\mathbf{X}^{norm}$,
$a_2 = 1/10, b_2 = -71/10$ - dla drugiej kolumny $\mathbf{X}^{norm}$.

```
QUADPROG(H,f,A,b,Aeq,beq,LB,UB,X0,OPTIONS,P1,P2,...)
```

rozwiązuje:

`min(alpha'*H*alpha+f'*alpha)` % $\min \left(\alpha^T \mathbf{H}\alpha + \mathbf{f}^T\alpha\right)$

`A*alpha<= b,` % $\mathbf{A}\alpha \leqslant b$

`Aeq*alpha= beq,` % $\mathbf{A}_{eq}\alpha = \mathbf{b}_{eq}$

`LB <= alpha<= UB,` % $\mathbf{L}_B \leqslant \alpha \leqslant \mathbf{U}_B$

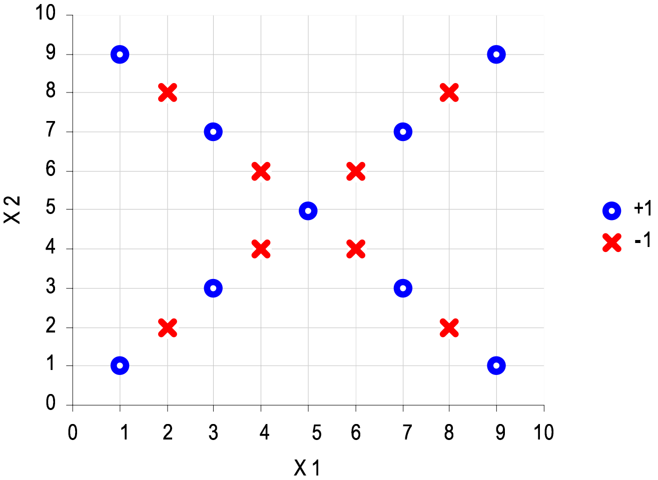
1. Normalizacja danych wejściowych $\mathbf{X}$;
2. Wyznaczenie Hessjanu $\mathbf{H}$ dla $\mathbf{X}$;
3. `n = size(X,1); f = -ones(n,1);`
   `LB=zeros(n,1);UB=C*ones(n,1);` %$0 \leqslant \alpha_i \leqslant C$
   %Kernel: 'poly', 'rbf', 'erbf', 'spline' or 'bspline':
   `A=[];b=[];Aeq=[];beq=[];` %$\mathbf{A}\alpha \leqslant \mathbf{b}$, $\mathbf{A}_{eq}\alpha = \mathbf{b}_{eq}$ - brakt
   %Kernel: 'lin' or 'sigmoid':
   `A=y';b=0;Aeq=y';beq=0;` %$\mathbf{A}\alpha \leqslant \mathbf{b}$, $\mathbf{A}_{eq}\alpha = \mathbf{b}_{eq}$
4. `alpha=QUADPROG(H,f,A,b,Aeq,beq,LB,UB);`

Sekwencyjna minimalizacja optymalna (SMO)

Jak rozwiązać problem QP dla bardzo dużych zbiorów danych ?
SMO: John Platt

1. Metoda analityczna rozwiązująca problem QP jedynie dla dwóch mnożników Lagrange'a. Przechowuje się tylko macierze 2×2.
2. Do optymalizacji stosuje się heurystyczny algorytm wyboru mnożników Lagrange'a.
3. Złożoność obliczeniowa jest między liniową i kwadratową.

Przykłady numeryczne: SVM i SMO



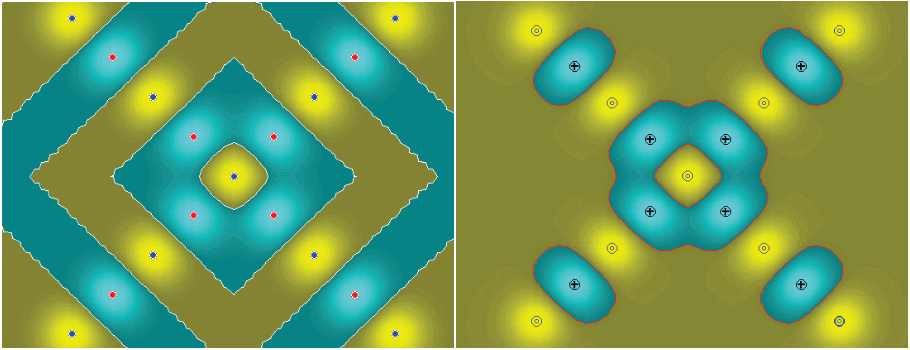
Rysunek: Zbiór "X".

Zbiór X – sigma=0.5

'rbf', $C = 10^4$.

a) $\sigma = 0.5$

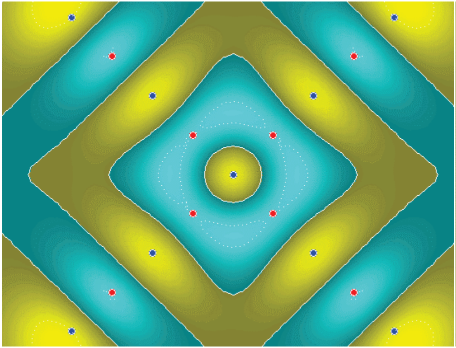
SVM SMO



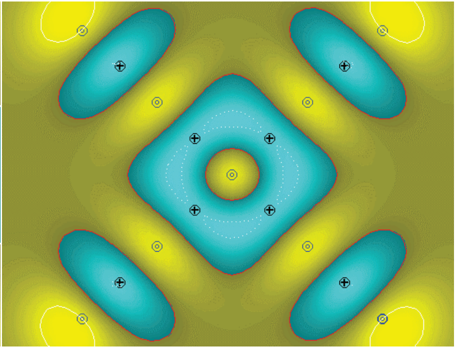
Zbiór X – sigma=1.0

'rbf', $C = 10^4$.
a) $\sigma = 1.0$

SVM



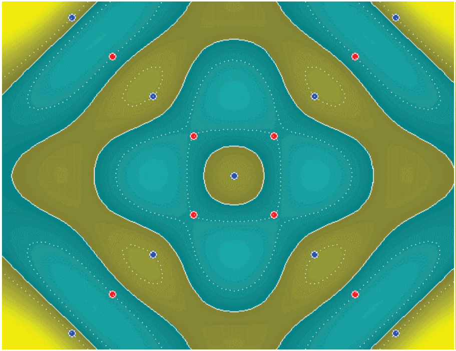
SMO



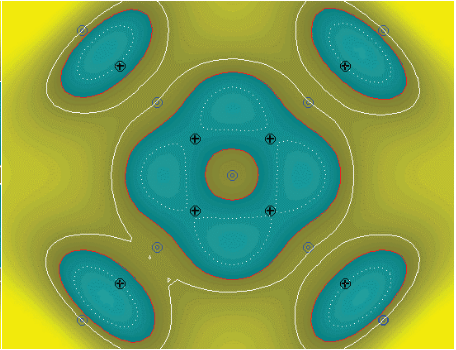
'rbf', $C = 10^4$.

a) $\sigma = 2.0$

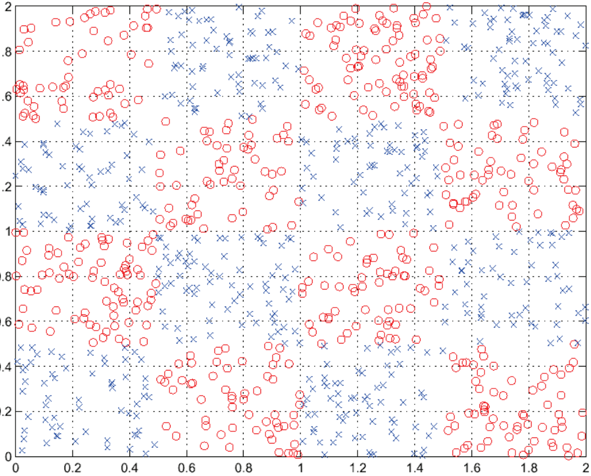
SVM



SMO



Przykłady numeryczne: SVM i SMO - c.d.

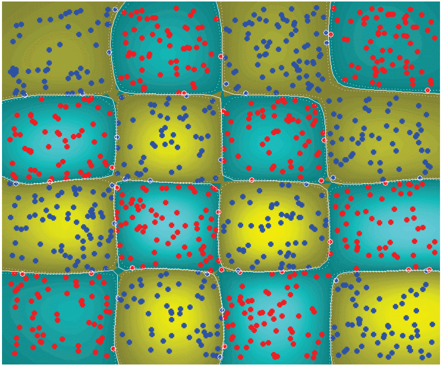


Rysunek: Szachownica.

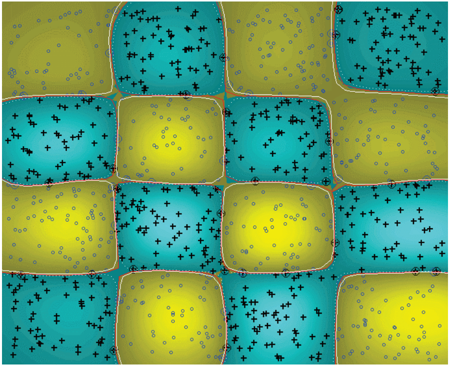
Szachownica – sigma=0.25

'rbf', $C = 10^4$.
$\sigma = 0.25$

SVM



SMO



Kwiaty irysa

http://www.badbear.com/signa/

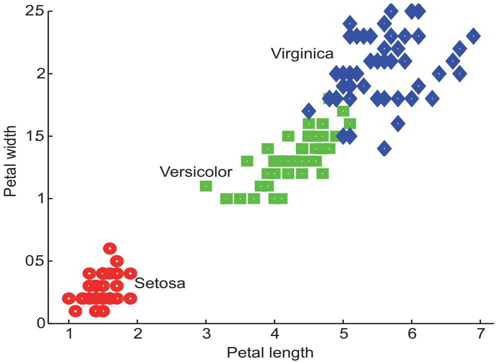
Rysunek: Odmiany: Virginica, Versicolor and Setosa;
http://www.badbear.com/signa/

Kwiaty irysa - c.d.

Rysunek: Atrybuty (Versicolor): W - płatek (petal), Z - działka (sepal).

Kwiaty irysa - c.d.

Ograniczamy się do dwóch atrybutów (dł. płatka, szer. płatka).



Rysunek: Dane kwiatów irysa.

Kwiaty irysa - c.d.

Dane: `http://www.badbear.com/signa/ iris_data.txt`

- 150 rekordów:

  Sepal length, Sepal width, Petal length, Petal width [cm]

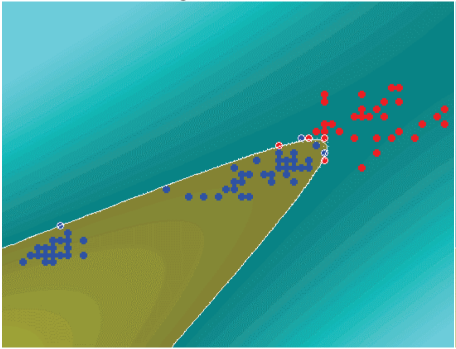
001 – Iris Setosa, 002 – Iris Versicolour, 003 – Iris Virginica

1.	5.1 3.5 1.4 0.2 001	7.0 3.2 4.7 1.4 002	6.3 3.3 6.0 2.5 003
2.	4.9 3.0 1.4 0.2 001	6.4 3.2 4.5 1.5 002	5.8 2.7 5.1 1.9 003
3.	4.7 3.2 1.3 0.2 001	6.9 3.1 4.9 1.5 002	7.1 3.0 5.9 2.1 003
4.	4.6 3.1 1.5 0.2 001	5.5 2.3 4.0 1.3 002	6.3 2.9 5.6 1.8 003
⋮			
49.	5.3 3.7 1.5 0.2 001	5.1 2.5 3.0 1.1 002	6.2 3.4 5.4 2.3 003
50.	5.0 3.3 1.4 0.2 001	5.7 2.8 4.1 1.3 002	5.9 3.0 5.1 1.8 003

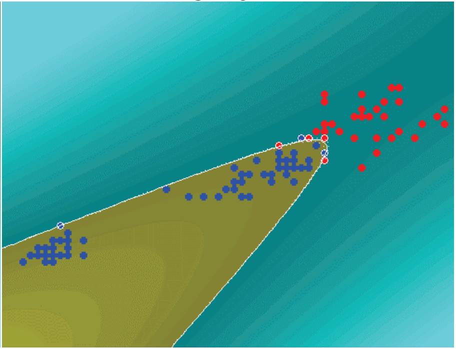
Kwiaty irysa - c.d.

'poly' 2-go rzędu (p1 = 2, p2 = 1, p3 = 1), $C = 10^8$.

SVM

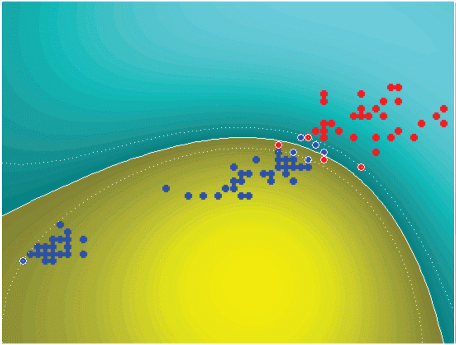


SMO

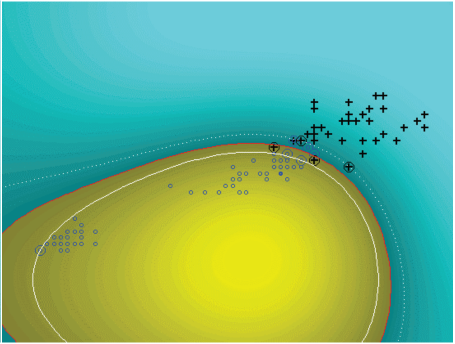


'rbf', $C = 100$, $\sigma = 2$.

SVM



SMO



Regresja za pomocą SVM

Dla

$$\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_Q, y_Q)\} \in \mathbb{R}^n \times \mathbb{R}$$

znaleźć "najlepszą" funkcję regresyjną

$$f(\mathbf{x}) = \langle \mathbf{x}, \mathbf{w} \rangle + b.$$

Znalezienie "najlepszej" funkcji regresyjnej polega na rozwiązaniu problemu:

$$\Phi(\mathbf{w}, \boldsymbol{\xi}) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \left(\xi_i^- + \xi_i^+\right)$$

gdzie C jest zadane a'priori oraz ξ_i^-, ξ_i^+ są zmiennymi swobodnymi.

Funkcje strat

1. Kwadratowa
$$L(z) = z^2$$
2. Liniowa
$$L(z) = |z|$$
3. Liniowa ze strefą nieczułości ε:
$$L(z) = \max(|z| - \varepsilon, 0),$$
4. Kwadratowo-liniowa (funkcja Hubera)
$$L(z) = \begin{cases} \frac{1}{2}z^2 & \text{dla} \quad |z| \leqslant \mu \\ \mu|z| - \frac{\mu^2}{2} & \text{dla} \quad |z| > \mu \end{cases}$$

Niech

$$L(y) = (f(\mathbf{x}) - y)^2$$

Problem dualny

$$\begin{cases} \min\limits_{\boldsymbol{\beta}} \quad W(\boldsymbol{\beta}) = \dfrac{1}{2}\sum\limits_{i=1}^{Q}\sum\limits_{j=1}^{Q}\beta_i\beta_j\left\langle \mathbf{x}_i, \mathbf{x}_j \right\rangle - \sum\limits_{i=1}^{Q}\beta_i y_i + \dfrac{1}{2C}\sum\limits_{i=1}^{Q}\beta_i^2 \\ -C \leqslant \beta_i \leqslant C, \qquad i = 1, \dots, Q \\ \sum\limits_{i=1}^{Q}\beta_i = 0 \end{cases}$$

Rozwiązanie: $f(\mathbf{x}) = \left\langle \mathbf{x}, \mathbf{w}^0 \right\rangle + b^0$:

$$w^0 = \sum_{i=1}^{Q}\beta_i \mathbf{x}_i, \qquad b^0 = -\frac{1}{2}\left\langle \mathbf{w}^0, \mathbf{x}_r + \mathbf{x}_s \right\rangle.$$

Podsumowanie

Procedura SVM jest metodą uczenia nadzorowanego, minimalizującą ryzyko strukturalne.

1. Hiperpowierzchnia separująca maksymalizuje margines między dwoma zbiorami danych; **czym szerszy jest margines, tym mniejszy jest błąd generalizacji** klasyfikatora.
2. SVC dostarcza wektorów wspierających.
3. SVC można wykorzystać do rozwiązania wielu problemów, np. rozpoznawania (twarzy, linii papilarnych, ...), wykrywania błędów, ... Powinniśmy dążyć do minimalizacji błędu testowania i uwzględniając inne wskaźniki.
4. Otwartymi problemami - w sensie rozwiązań analitycznych, są:
   - wybór **stałej** C,
   - wybór **funkcji jądra**,
   - wybór **parametrów funkcji jądra**.
5. Warto normalizować dane wejściowe, aby uniknąć złego uwarunkowania macierzy Hessjanu.

Dodatek - podstawowe pojęcia teorii optymalizacji

1. **Gradient** funkcji skalarnej $f(\mathbf{x}) = f(x_1, \ldots, x_n) : \mathbb{R}^n \to \mathbb{R}$

$$\nabla f(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \left[\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}, \ \ldots, \ \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}\right]^T$$

2. **Hessjan** funkcji skalarnej $f(\mathbf{x}) = f(x_1, \ldots, x_n) : \mathbb{R}^n \to \mathbb{R}$

$$\mathbf{H} = \nabla^2 f(\mathbf{x}) = \left\{\frac{\partial^2 f}{\partial x_i \partial x_j}\right\}$$

3. Zbiór $\mathbb{X} \in \mathbb{R}^n$ jest **wypukły**, jeżeli

$$\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{X}, \ 0 \leqslant \alpha \leqslant 1 \qquad \Longrightarrow \qquad \alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2 \in X$$

Dodatek - podstawowe pojęcia teorii optymalizacji - c.d.

1. $f(\mathbf{x})$ jest **wypukły** na $\mathbb{X}$, jeżeli dla $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{X}$ oraz $0 \leqslant \alpha \leqslant 1$ zachodzi

$$f(\alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leqslant \alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2)$$

   Jeżeli zamienimy $\leqslant$ ze znakiem $\geqslant$, to funkcję nazywamy **wklęsłą**. Jeżeli nierówność jest spełniona dla $0 < \alpha < 1$, to funkcja jest **ściśle wypukła** (lub **ściśle wklęsła**).

2. f jest wklęsła, jeżeli $-f$ jest wypukła. Niech $f(x) = x^2$.

$$\begin{aligned} r &= f(\alpha x_1 + (1-\alpha) x_2) - (\alpha f(x_1) + (1-\alpha) f(x_2)) \\ &= (\alpha x_1 + (1-\alpha) x_2)^2 - \left(\alpha x_1^2 + (1-\alpha) x_2^2\right) \\ &= \alpha (x_2 - x_1)^2 (\alpha - 1) < 0 \end{aligned}$$

   dla $0 < \alpha < 1$. Funkcja ta jest **ściśle wypukła**.

Dodatek - podstawowe pojęcia teorii optymalizacji - c.d.

1. Jeżeli $f(\mathbf{x})$ jest **wypukła** na zamkniętym zbiorze wypukłym zawartym w $\mathbb{R}^n$, to każde minimum lokalne jest jednocześnie minimum globalnym.
2. Jeżeli $f(\mathbf{x})$ jest **ściśle wypukła** na zamkniętym zbiorze wypukłym zawartym w $\mathbb{R}^n$, to minimum **globalne** jest osiągane tylko w jednym punkcie.
3. Jeżeli $f_1, \ldots, f_k : \mathbb{R}^n \to \mathbb{R}$ są wypukłe, to
$$f(\mathbf{x}) = \alpha_1 f_1(\mathbf{x}) + \ldots + \alpha_k f_k(\mathbf{x}) \quad , \quad \alpha_1, \ldots, \alpha_k \geqslant 0$$
jest wypukła.
4. $f : \mathbb{R}^n \to \mathbb{R}$ oraz f jest różniczkowalna na $\mathbb{R}^n$, to f jest **wypukła** w.t.w., gdy dla dowolnych dwóch punktów $\mathbf{x}^0$, $\mathbf{x} \in \mathbb{R}^n$ zachodzi
$$f(\mathbf{x}) \geqslant f\left(\mathbf{x}^0\right) + \nabla_{\mathbf{x}} f(\mathbf{x})^T (\mathbf{x} - \mathbf{x}_0)$$

Dodatek - podstawowe pojęcia teorii optymalizacji - c.d.

1. Jeżeli $f : \mathbb{R}^n \to \mathbb{R}$ oraz f jest z klasy C^2 (dwukrotnie różniczkowalna w każdym punkcie $\mathbb{R}^n$ i jej Hessjan $\mathbf{H}$ jest ciągły), to jest ona
   1. **ściśle wypukła** w.t.w. $\mathbf{H} > 0$,
   2. **wypukła** w.t.w. $\mathbf{H} \geqslant 0$.
2. Wartości własne macierzy rzeczywistej i symetrycznej są rzeczywiste.
3. Symetryczna macierz $\mathbf{A}$ jest **dodatnio określona**, jeżeli zachodzi jeden z warunków równoważnych:
   1. $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$ dla każdego $\mathbf{x} \neq \mathbf{0}$,
   2. $\lambda\left[\mathbf{A}\right] > 0$,
   3. istnieje nieosobliwa macierz $\mathbf{W} : \mathbf{A} = \mathbf{W}^T\mathbf{W}$,
   4. wyznaczniki minorów głównych są dodatnie: $\det \mathbf{A}_k > 0$ dla $k = 1, \ldots, n$ (tw. Sylvestera).

Dodatek - podstawowe pojęcia teorii optymalizacji - c.d.

1. Symetryczna macierz $\mathbf{A}$ jest **dodatnio półokreślona**, jeżeli $\lambda\left[\mathbf{A}\right] \geqslant 0$.
2. Symetryczna macierz $\mathbf{A}$ jest **ujemnie określona**, jeżeli
    1. $\lambda\left[\mathbf{A}\right] < 0$.
    2. minory główne $\mathbf{A}_k$:

$$(-1)^k \det \mathbf{A}_k > 0 \quad ,k = 1, \ldots, n$$

3. Symetryczna macierz $\mathbf{A}$ jest **ujemnie półokreślona**, jeżeli $\lambda\left[\mathbf{A}\right] \leqslant 0$.

Dodatek - podstawowe pojęcia teorii optymalizacji - c.d.

- Niech $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{b} \in \mathbb{R}^n$ oraz

$$f(\mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x}$$

Stąd,

$$\nabla_{\mathbf{x}} \left(\mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x} \right) = \frac{1}{2} \left(\mathbf{A} + \mathbf{A}^T \right) \mathbf{x} + \mathbf{b}$$
$$\nabla^2_{\mathbf{x}} \left(\mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x} \right) = \frac{1}{2} \left(\mathbf{A} + \mathbf{A}^T \right).$$

Jeżeli $\mathbf{A}$ jest symetryczna i $\mathbf{A} > 0$, to funkcja f jest wypukła i osiąga minimum dla:

$$\frac{1}{2} \left(\mathbf{A} + \mathbf{A}^T \right) \mathbf{x} + \mathbf{b} = \mathbf{0}$$

czyli

$$\mathbf{x}^0 = \mathbf{A}^{-1} \mathbf{b} \ .$$

Dodatek - podstawowe pojęcia teorii optymalizacji - c.d.

1. Jeżeli $f : \mathbb{R}^n \to \mathbb{R}$ jest z klasy C^1 (różniczkowalna i jej gradient jest ciągły), to koniecznym warunkiem, aby w punkcie $\hat{\mathbf{x}} \in \mathbb{R}^n$ istniało minimum, jest
$$\nabla_{\mathbf{x}} f\left(\hat{\mathbf{x}}\right) = \mathbf{0}$$
Jeżeli dodatkowo f jest **wypukła**, to warunek ten jest koniecznym i dostatecznym na to, aby punkt $\hat{\mathbf{x}}$ był **minimum globalnym**.
2. Niech $f : \mathbb{R}^n \to \mathbb{R}$ będzie z klasy C^2. Jeżeli punkt $\hat{\mathbf{x}} \in \mathbb{R}^n$ jest **minimum lokalnym**, to
$$\begin{aligned} \nabla_{\mathbf{x}} f\left(\hat{\mathbf{x}}\right) &= \mathbf{0} \\ \mathbf{H} = \nabla_{\mathbf{x}}^2 f\left(\hat{\mathbf{x}}\right) &\geqslant \mathbf{0} \end{aligned}$$
Jeżeli
$$\begin{aligned} \nabla_{\mathbf{x}} f\left(\hat{\mathbf{x}}\right) &= \mathbf{0} \\ \mathbf{H} = \nabla_{\mathbf{x}}^2 f\left(\hat{\mathbf{x}}\right) &> \mathbf{0} \end{aligned}$$
to $\hat{\mathbf{x}}$ stanowi **minimum globalne**.

Dodatek - podstawowe pojęcia teorii optymalizacji - c.d.

- Zadanie programowania nieliniowego (ZPN)

$$\begin{cases} \min f(\mathbf{x}) \\ g_i(\mathbf{x}) \leqslant 0, & i = 1, \ldots, m_g \\ h_i(\mathbf{x}) = 0, & i = 1, \ldots, m_h \end{cases}$$

  gdzie wszystkie funkcje są zdefiniowane na $\mathbb{R}^n$.

- **Tw. Weierstrassa**. Jeżeli f jest ciągła i zbiór możliwych rozwiązań jest ograniczony i domknięty, to istnieje co najmniej jedno minimum globalne funkcji f.
- Zbiór ograniczeń aktywnych

$$A = A(\mathbf{x}) = \{i : g_i(\mathbf{x}) = 0\}$$

- Zbiór kierunków

$$D(\mathbf{x}) = \{\mathbf{d} \in \mathbb{R}^n : i \in A \Rightarrow \langle \nabla g_i, \mathbf{d} \rangle \leqslant 0, \langle \nabla h_i, \mathbf{d} \rangle = 0, i = 1, \ldots, m_h\}$$

- Punkt $\mathbf{x} \in \mathbb{X}$ jest **regularnym**, jeżeli istnieje funkcja $\mathbf{e} : [0, \infty) \to \mathbb{R}^n$, oraz $\mathbf{e}(t) \in \mathbb{X}$ przy $0 \leqslant t \leqslant \bar{t}$ taka, że

$$\mathbf{e}(0) = \mathbf{x}$$
$$\lim_{t \to 0+} (\mathbf{e}(t) - \mathbf{e}(0) / t) = \mathbf{d}$$

gdzie $\mathbf{d} \in D(\mathbf{x})$.

Dodatek - podstawowe pojęcia teorii optymalizacji - c.d.

- Niech

$$\lambda = \left[\lambda_1, \ldots, \lambda_{m_g}\right]^T$$
$$\mu = \left[\mu_1, \ldots, \mu_{m_h}\right]^T$$
$$\mathbf{g}(\mathbf{x}) = \left[g_1(\mathbf{x}), \ldots, g_{m_g}(\mathbf{x})\right]^T$$
$$\mathbf{h}(\mathbf{x}) = \left[h_1(\mathbf{x}), \ldots, h_{m_h}(\mathbf{x})\right]^T$$

- **Warunki konieczne Kuhna-Tuckera**. Jeżeli $\hat{\mathbf{x}} \in \mathbb{X}$ jest minimum lokalnym dla ZPN i jest punktem regularnym, to istnieją takie liczby zwane **mnożnikami Lagrange'a** $\hat{\lambda}_i$ dla $i = 1, \ldots, m_g$ oraz $\hat{\mu}_i$ dla $i = 1, \ldots, m_h$, że

$$\frac{\partial f(\hat{\mathbf{x}})}{\partial \mathbf{x}} + \left\langle \hat{\lambda}, \frac{\partial g(\hat{\mathbf{x}})}{\partial \mathbf{x}} \right\rangle + \left\langle \hat{\mu}, \frac{\partial h(\hat{\mathbf{x}})}{\partial \mathbf{x}} \right\rangle = \mathbf{0}$$
$$\hat{\lambda}_i \nabla g_i(\hat{\mathbf{x}}) = 0, \quad i = 1, \ldots, m_g$$
$$\hat{\lambda}_i \geqslant 0, \quad i = 1, \ldots, m_g$$

Warunki wystarczające regularności (Slater). Jeżeli g_i są wypukłe i h_i są liniowe i istnieje $\mathbf{x}^0$ takie, że

$$\begin{aligned} g_i\left(\mathbf{x}^0\right) &< 0, \quad i = 1, \ldots, m_g \\ h_i\left(\mathbf{x}^0\right) &= 0, \quad i = 1, \ldots, m_h \end{aligned}$$

to każdy punkt $\mathbf{x} \in \mathbb{X}$ jest regularny.

Dodatek - podstawowe pojęcia teorii optymalizacji - c.d.

- **Warunki wystarczające regularności** (dotyczące niezależności liniowej). Jeżeli w punkcie $\mathbf{x}$ gradienty ∇g_i dla $i \in A$ oraz ∇h_i dla $i = 1, \ldots, m_h$ są liniowo niezależne, to $\mathbf{x}$ jest regularny.
- **Warunki wystarczające Kuhna-Tuckera**. Jeżeli

1. f jest wypukła,
2. g_i są wypukłe, $i = 1, \ldots, m_g$,
3. h_i są liniowe, $i = 1, \ldots, m_h$,

to każdy punkt spełniający warunki Kuhna-Tuckera stanowi minimum globalne.

- **Funkcja Lagrange'a**, która związana jest z ZPN jest funkcją skalarną $L : \mathbb{R}^n \times \mathbb{R}^{m_g} \times \mathbb{R}^{m_h} \to \mathbb{R}$:

$$L(\mathbf{x}, \lambda, \mu) = f(\mathbf{x}) + \langle \lambda, \mathbf{g}(\mathbf{x}) \rangle + \langle \mu, \mathbf{h}(\mathbf{x}) \rangle$$

Dodatek - podstawowe pojęcia teorii optymalizacji - c.d.

- **Warunki konieczne Kuhna-Tuckera** bazujące na funkcji Lagrange'a. Istnieją wektory $\hat{\lambda}$ i $\hat{\mu}$ takie, że

$$\begin{aligned} \frac{\partial}{\partial \mathbf{x}} L\left(\hat{\mathbf{x}}, \hat{\lambda}, \hat{\mu}\right) &= \mathbf{0} \\ \left\langle \hat{\lambda},\ \frac{\partial}{\partial \lambda} L\left(\hat{\mathbf{x}}, \hat{\lambda}, \hat{\mu}\right) \right\rangle &= \mathbf{0} \\ \hat{\lambda} &\geqslant \mathbf{0} \end{aligned}$$