Tartóvektorgépek

Ispány Márton és Jeszenszky Péter

2017. október 10.

Tartalom

Hipersík

Ha $\mathbf{w} \in \mathbb{R}^n$ és $b \in \mathbb{R}$ adottak, akkor a $\mathbf{w}^\top \mathbf{x} + b = 0$ egyenlet $\mathbf{x} \in \mathbb{R}^n$ megoldásai egy hipersíkot határoznak meg, ahol $\mathbf{w}$ a hipersík normálvektora.

A hipersík a teret két féltérre bontja:

- $\left\{ \mathbf{x} \mid \mathbf{w}^\top \mathbf{x} + b \geq 0 \right\}$ (a hipersík feletti féltér)
- $\left\{ \mathbf{x} \mid \mathbf{w}^\top \mathbf{x} + b < 0 \right\}$ (a hipersík alatti féltér)

Megjegyzés: a hipersíkot az egyenlőséggel a hipersík feletti féltérhez csatoltuk, ugyanúgy csatolhatnánk a hipersík alatti féltérhez is.

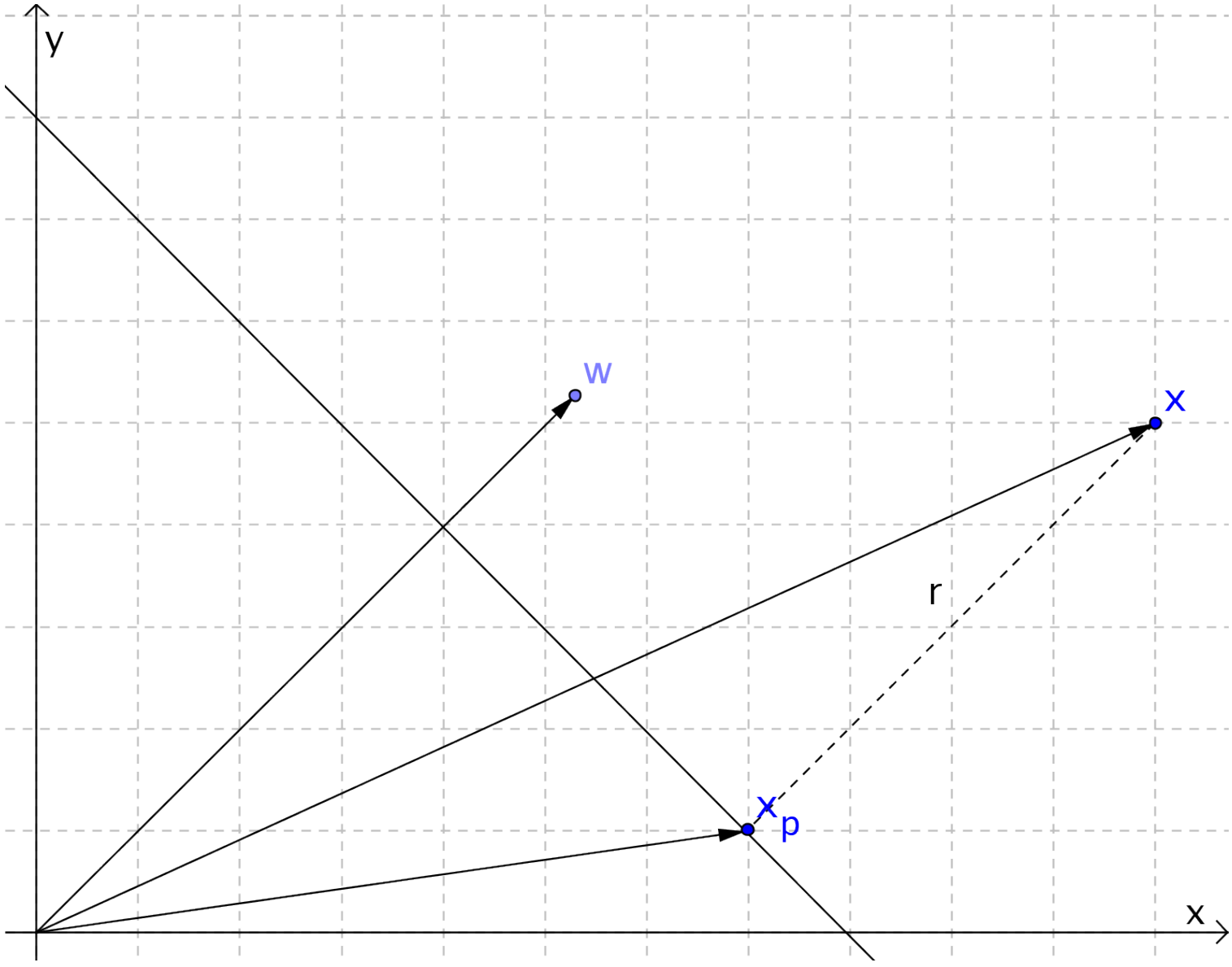
Távolság a hipersíktól (1)

Tekintsük a $\mathbf{w}^\top \mathbf{x} + b = 0$ egyenlettel meghatározott hipersíkot. Bármely $\mathbf{x}$ vektor kifejezhető

$$\mathbf{x} = \mathbf{x}_p + r \frac{\mathbf{w}}{\|\mathbf{w}\|}$$

alakban, ahol $\mathbf{x}_p$ a vektor merőleges vetülete a hipersíkra, r pedig $\mathbf{x}$-nek a hipersíktól mért távolsága.

Távolság a hipersíktól (2)



Távolság a hipersíktól (3)

A $g(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ mennyiség az $\mathbf{x}$ vektor távolságát méri a hipersíktól. Lásd:

$$g(\mathbf{x}) = \mathbf{w}^\top \left(\mathbf{x}_p + r \frac{\mathbf{w}}{\|\mathbf{w}\|} \right) + b = (\mathbf{w}^\top \mathbf{x}_p + b) + r \frac{\mathbf{w}^\top \mathbf{w}}{\|\mathbf{w}\|}$$

$$= \underbrace{g(\mathbf{x}_p)}_{=0} + r \frac{\mathbf{w}^\top \mathbf{w}}{\|\mathbf{w}\|} = r\|\mathbf{w}\|$$

Azaz $g(\mathbf{x})$ az $\mathbf{x}$ vektor a hipersíktól mért távolságának a konstansszorosa. Ekkor az $\mathbf{x}$ vektor a hipersíktól mért távolsága:

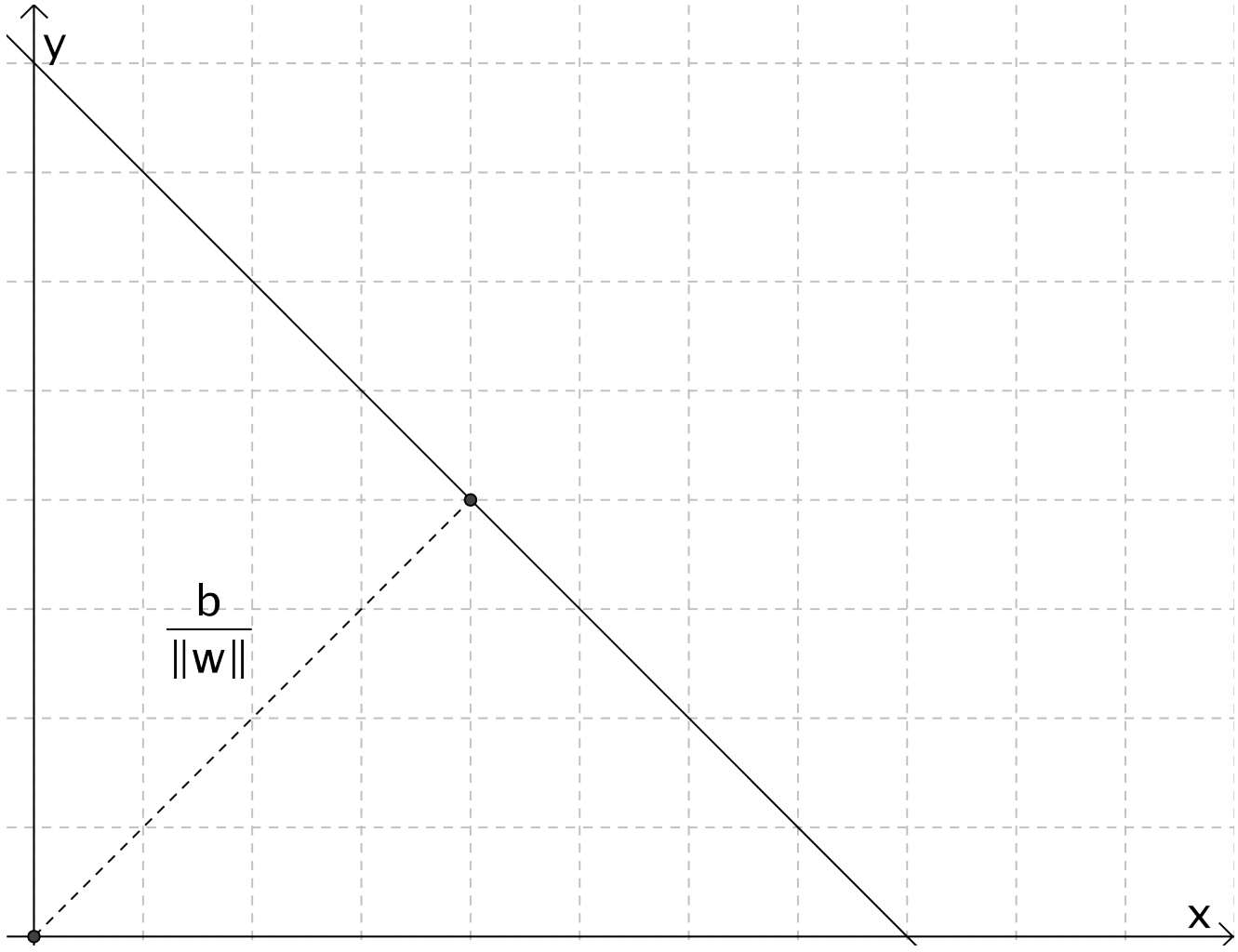
$$r = \frac{g(\mathbf{x})}{||\mathbf{w}||}.$$

Távolság a hipersíktól (4)

Az előbbiek alapján a hipersík távolsága az origótól:

$$r = \frac{g(\mathbf{0})}{\|\mathbf{w}\|} = \frac{b}{\|\mathbf{w}\|}.$$

Távolság a hipersíktól (5)



Lineáris szeparálási feladat (1)

Tekintsünk egy olyan bináris osztályozási feladatot, ahol a két osztály lineárisan szeparálható, azaz szétválasztható egy hipersíkkal.

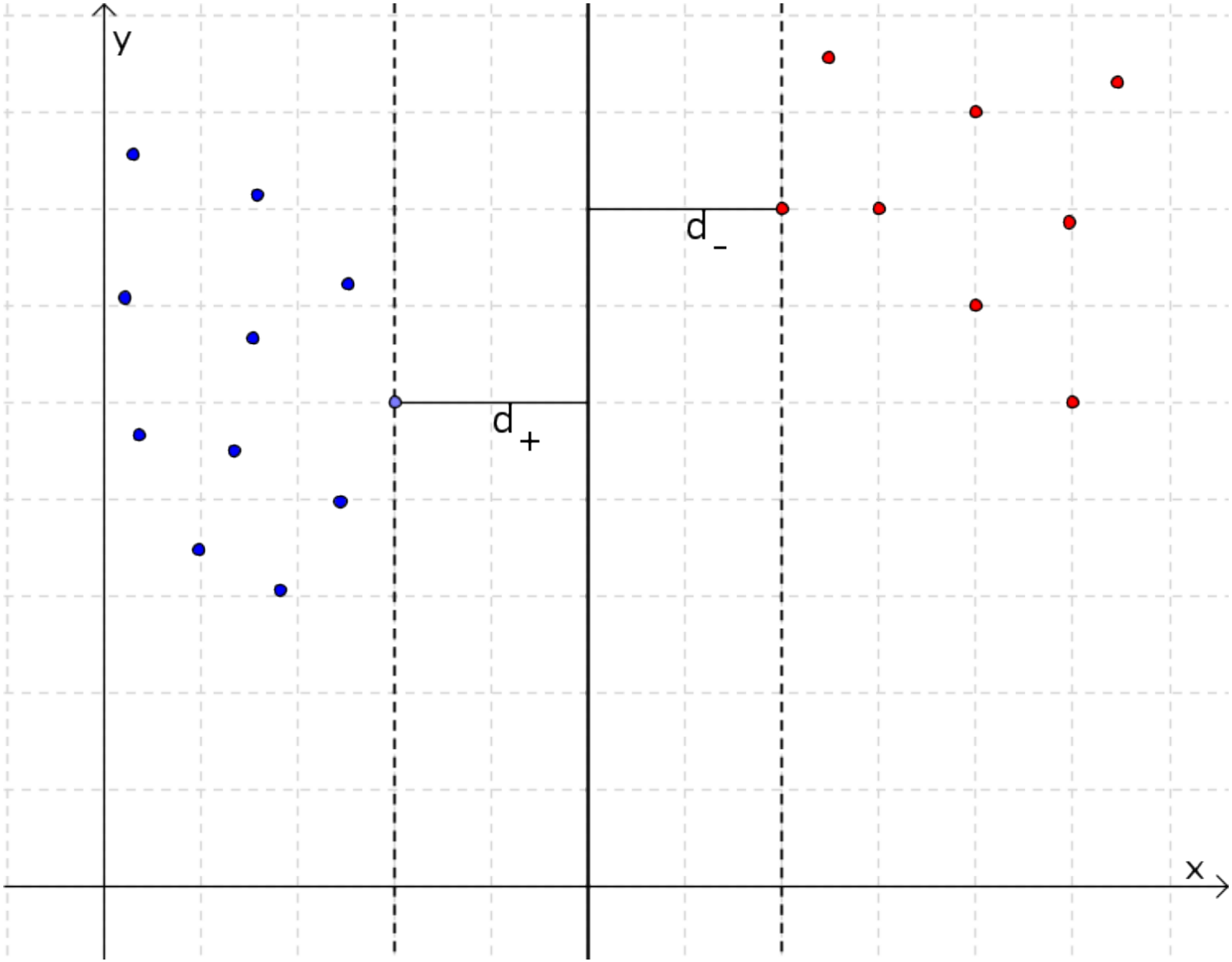
Ehhez legyenek adottak $(\mathbf{x}_i, y_i)$ elempárok, ahol $\mathbf{x}_i \in \mathbb{R}^n$, $y_i \in \{-1, +1\}$ pedig egy osztálycímke $(i = 1, \ldots, N)$.

Ha a két osztály lineárisan szeperálható, akkor alkalmas $\mathbf{w}$ és b mellett teljesül, hogy

$$\begin{aligned} \mathbf{w}^\top \mathbf{x}_i + b &\geq 0, \quad \text{ha } y_i = +1, \\ \mathbf{w}^\top \mathbf{x}_i + b &< 0, \quad \text{ha } y_i = -1. \end{aligned}$$

Rögzített adatok esetén sok ilyen elválasztó hipersík adható meg, keressük meg az optimálisat!

Lineáris szeparálási feladat (2)



Margó (1)

Jelölje d_+ a hipersíkhoz legközelebbi olyan vektor távolságát a hipersíktól, amelyre $y_i = +1$, d_- pedig a hipersíkhoz legközelebbi olyan vektor távolságát, amelyre $y_i = -1$. A hipersík margója a

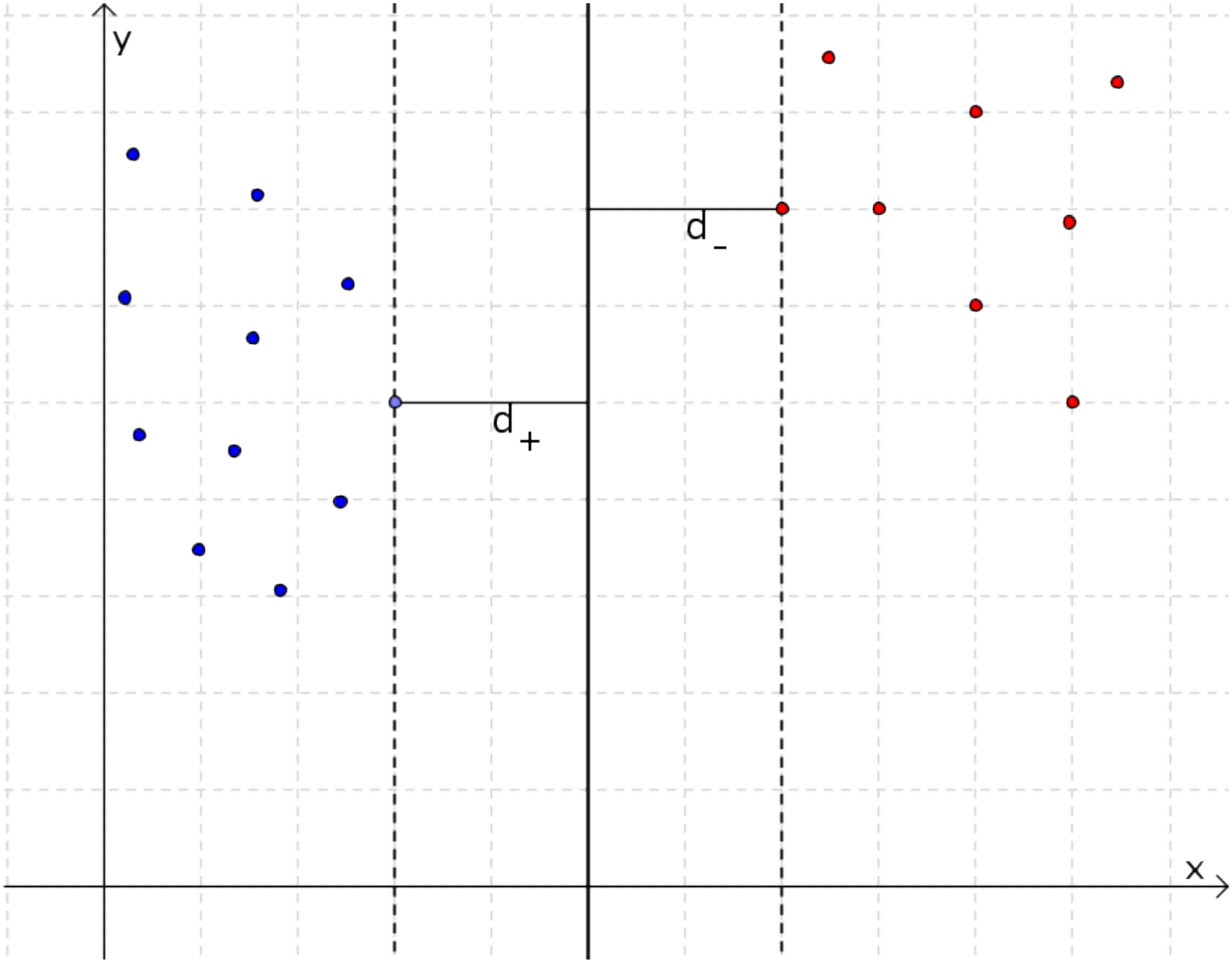
$$\rho = d_+ + d_-$$

távolság.

Megjegyzés

Az elválasztó hipersíkot úgy szokták a megfelelő irányba párhuzamosan eltolni, hogy a két osztály legközelebbi pontjaitól azonos távolságra legyen, azaz $d_+ = d_-$. A továbbiakban ilyen elválasztó hipersíkot feltételezünk.

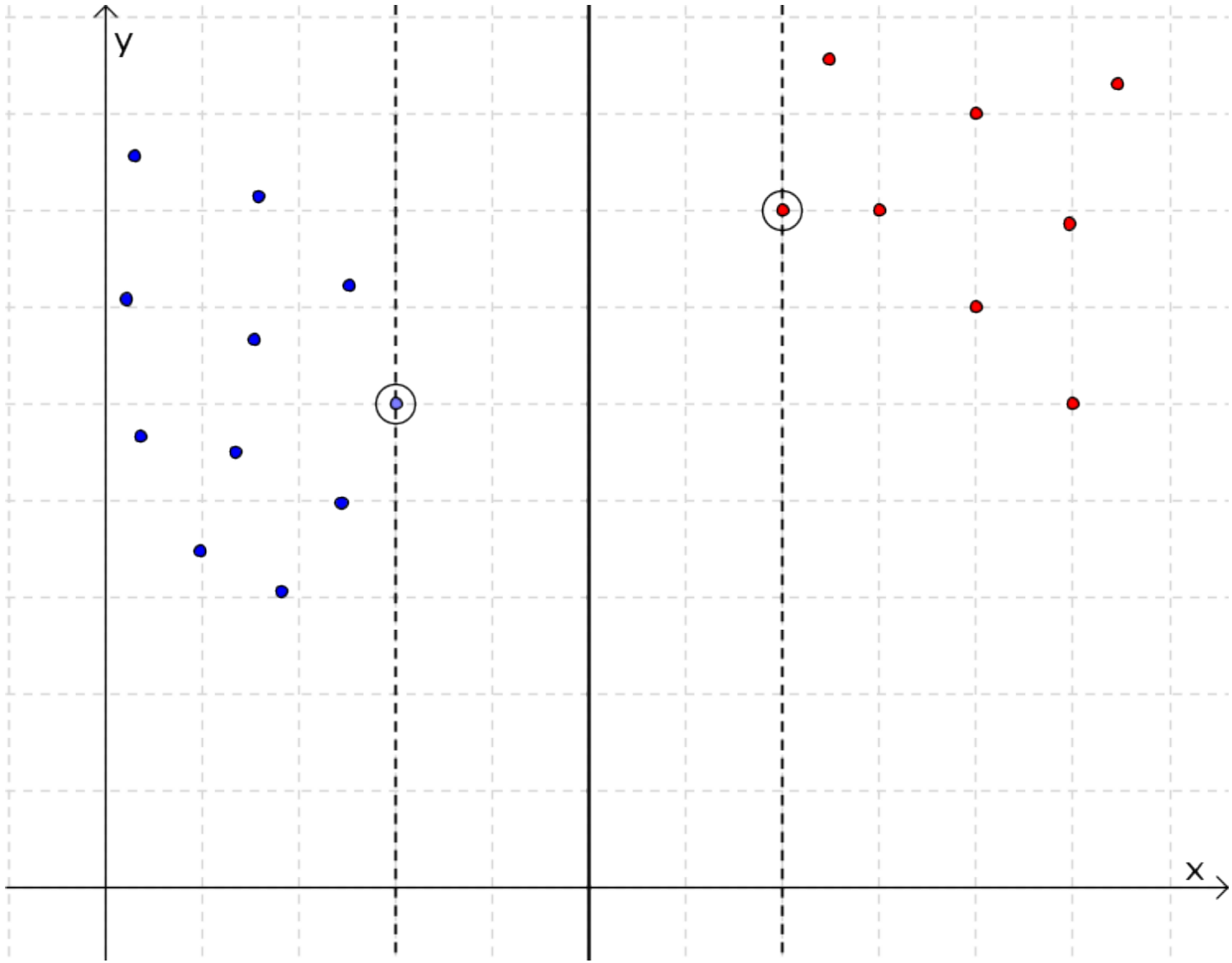
Margó (2)



Tartóvektorok (1)

Tartóvektoroknak (**support vectors**) nevezzük az elválasztó hipersíkhoz legközelebbi vektorokat. Ezek nagyon fontosak, mert ezeket a legnehezebb osztályozni.

Tartóvektorok (2)



Optimális elválasztó hipersík (1)

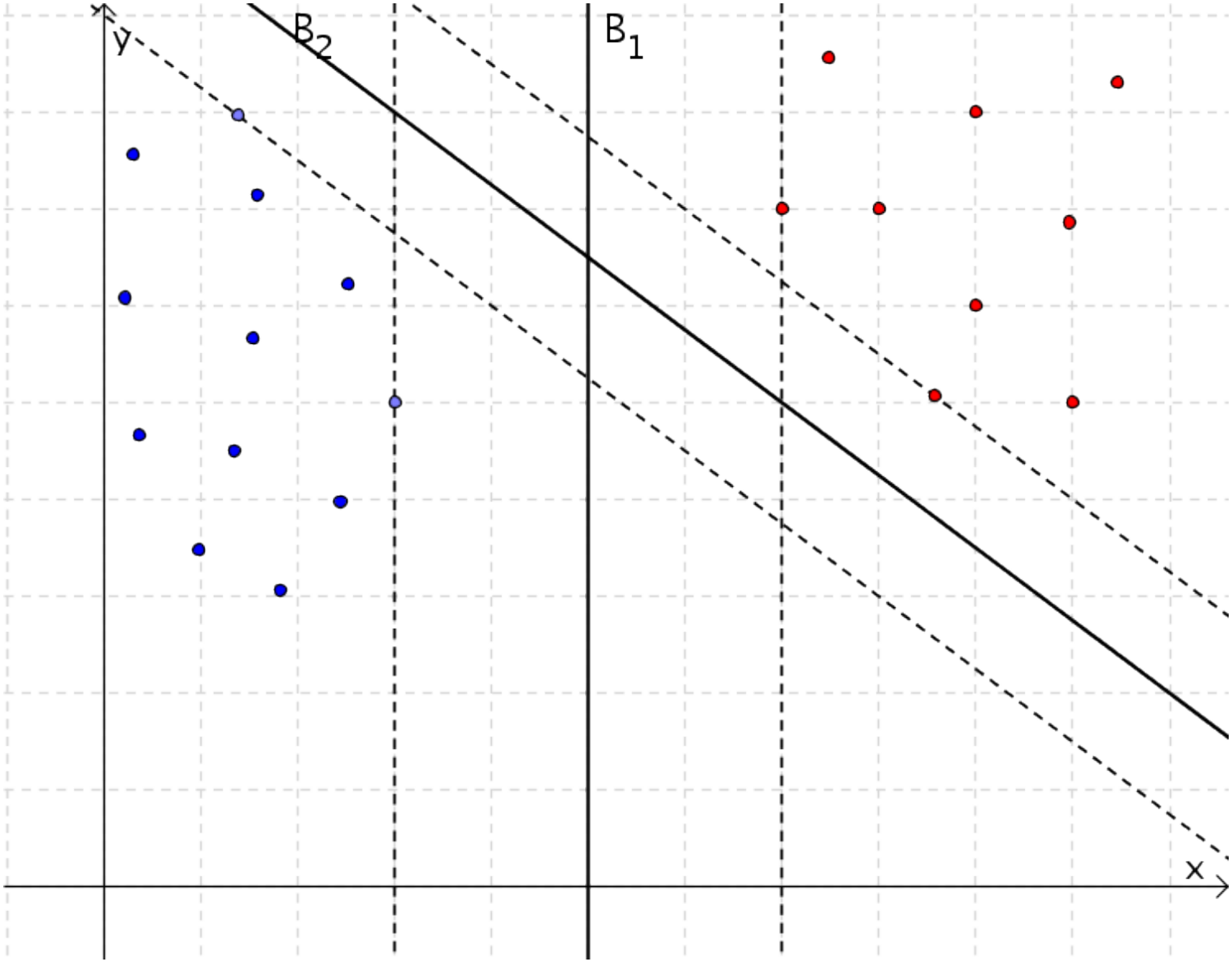
Válasszuk a maximális margójú elválasztó hipersíkot!

Ez egy jó stratégia, mert nagy margó esetén általában jobb az általánosítási hiba, mint kis margó esetén. (Intuitívan: ha a margó kicsi, akkor a döntési határ bármilyen kis perturbációjának elég jelentős hatása lehet az osztályozásra.)

Kis margó esetén nagyobb eséllyel léphet fel modell túlillesztés.

A fentiekre egy formálisabb magyarázatot ad a strukturális kockázat minimalizálás (SRM – Structural Risk Minimization) módszere.

Optimális elválasztó hipersík (2)



Lineáris szeparálási feladat (újra)

Az alábbi módon módosítjuk a feladat megfogalmazását.

Lineáris szeparálhatóság esetén $\mathbf{w}$ és b megfelelő skálázásával mindig elérhető, hogy az alábbi teljesüljön:

$$\begin{aligned} \mathbf{w}^\top \mathbf{x}_i + b &\geq 1, \quad \text{ha } y_i = +1, \\ \mathbf{w}^\top \mathbf{x}_i + b &\leq -1, \quad \text{ha } y_i = -1. \end{aligned}$$

Ez tömören a következő alakba írható[1]:

$$y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \qquad (i = 1, \ldots, N).$$

[1] Itt használjuk ki, hogy az osztályokat a {-1, +1} halmazból vett értékekkel cimkéztük fel.

Optimális elválasztó hipersík (újra) (1)

A $\mathbf{w}^\top \mathbf{x} + b = \pm 1$ egyenletek egy-egy, az elválasztó hipersíkkal párhuzamos hipersíkot határoznak meg, egyenlőség a tartóvektorok esetén teljesül.

Egy $\mathbf{x}_s$ tartóvektor távolsága az elválasztó hipersíktól a korábbiak szerint:

$$r = \frac{g(\mathbf{x}_s)}{\|\mathbf{w}\|} = \frac{\mathbf{w}^\top \mathbf{x}_s + b}{\|\mathbf{w}\|} = \begin{cases} \frac{1}{\|\mathbf{w}\|}, & \text{ha } y_s = +1, \\ -\frac{1}{\|\mathbf{w}\|}, & \text{ha } y_s = -1. \end{cases}$$

Ezért a margó

$$\rho = \frac{2}{\|\mathbf{w}\|}.$$

Optimális elválasztó hipersík (újra) (2)

Azt az elválasztó hipersíkot válasszuk, amelyre a $\rho = \frac{2}{\|\mathbf{w}\|}$ margó maximális.

Ez ekvivalens annak a hipersíknak a választásával, amelyre $\|\mathbf{w}\|$ minimális, amivel ekvivalens $\frac{\|\mathbf{w}\|^2}{2}$ minimalizálása. Azért ezt a célfüggvényt érdemes választani mert négyzetösszegként egy kvadratikus optimalizálási feladatot kapunk.

A feladat megfogalmazása optimalizálási problémaként

Feltételes optimalizálási probléma:

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

$$\text{feltéve, hogy} \qquad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, \quad i = 1, \ldots, N.$$

Mivel a célfüggvény kvadratikus és a korlátozások lineárisak a $\mathbf{w}$ és b paraméterekben, ezt **konvex optimalizálás**nak nevezzük, amely a szokásos **Lagrange-multiplikátor** módszerrel oldható meg.

Az optimalizálási probléma megoldása (1)

Írjuk fel az alábbi, ún. Lagrange-függvényt:

$$L_P = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \lambda_i \Big(y_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1 \Big),$$

ahol a $\lambda_i \geq 0$ paramétereket **Lagrange-multiplikátorok**nak nevezzük.

A feltételes optimalizálási probléma megoldását úgy kapjuk, hogy meghatározzuk az L_P Lagrange-függvény nyeregpontját. (A $\mathbf{w}$ és b paraméterek szerint minimalizálni, a λ_i paraméterek szerint pedig maximalizálni kell L_P értékét.)

A fenti Lagrange-függvényt primál Lagrange-függvénynek nevezik, a megoldandó optimalizálási feladatot pedig **primál feladat**nak.

Az optimalizálási probléma megoldása (2)

A Lagrange-függvény minimalizálásához vennünk kell és nullává kell tegyük L_P $\mathbf{w}$ és b szerinti deriváltját:

$$\frac{\partial L_P}{\partial \mathbf{w}} = 0 \Longrightarrow \mathbf{w} = \sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_i, \tag{1}$$

$$\frac{\partial L_P}{\partial b} = 0 \Longrightarrow \sum_{i=1}^{N} \lambda_i y_i = 0. \tag{2}$$

A Lagrange-multiplikátorok azonban továbbra is ismeretlenek!

A (1) képletben $\mathbf{w}$-re kapott eredmény azt jelenti, hogy az elválasztó hipersík az inputvektorok lineáris kombinációjaként adódik, ahol az együtthatók a Lagrange-multiplikátorok az osztálycimke szerinti előjellel véve!
A (2) képlet azt mutatja, hogy a két osztályhoz tartozó Lagrange-multiplikátorok összege megegyezik!

Az optimalizálási probléma megoldása (3)

Teljesülnek az alábbi, ún. Karush-Kuhn-Tucker (KKT) feltételek:

$$\lambda_i \left[y_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1 \right] = 0, \qquad i = 1, \ldots, N.$$

Ez azt jelenti, hogy az $y_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1$ egyenletet kielégítő tartóvektorok kivételével minden Lagrange-multiplikátor 0.

Ezért a $\mathbf{w}$ vektor számolásakor csak a tartóvektorokat kell figyelembe venni! Az elválasztó hipersík paraméterei csak a tartóvektorokhoz tartozó tanítópontoktól fognak függeni, újabb nem tartóvektor tanító pont hozzáadása az adatállományhoz az elválasztó hipersíkot és így az osztályozást változatlanul hagyja.

Az optimalizálási probléma átfogalmazása (1)

Úgy fogalmazzuk át az optimalizálási problémát, hogy ha az eredeti feladatnak van optimális megoldása, akkor az átfogalmazott problémának is, és az optimális értékeik megegyeznek.

Az optimalizálási probléma átfogalmazása (2)

Legyen

$$L_D = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{N} \lambda_i \lambda_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j,$$

ahol $\lambda_i \geq 0$. A feladat L_D maximalizálása, feltéve hogy teljesülnek az alábbiak:

1. $\sum_{i=1}^{N} \lambda_i y_i = 0$,
2. $\lambda_i \geq 0$ minden $i = 1, \ldots, N$ esetén.

A fenti L_D függvényt duális Lagrange-függvénynek nevezik, a megoldandó optimalizálási feladatot pedig **duális feladat**nak.[2]

[2] Az L_D függvényt úgy kapjuk, hogy az L_P primál Lagrange-függvénybe $\mathbf{w}$ helyére behelyettesítjük az (1) képlet eredményét, valamint egyszerűsítést végzünk a (2) képlet eredménye alapján.

Az optimalizálási probléma átfogalmazása (3)

A számolás részletezése, felhasználva az (1) és (2) formulákat:

$$
\begin{aligned}
L_P =& \frac{1}{2}\|\mathbf{w}\|^2 - \mathbf{w}^\top \underbrace{\sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_i}_{=\mathbf{w}} - b \underbrace{\sum_{i=1}^{N} \lambda_i y_i}_{=0} + \sum_{i=1}^{N} \lambda_i \\
=& \sum_{i=1}^{N} \lambda_i - \frac{1}{2}\mathbf{w}^\top \mathbf{w} = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \left(\sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_i \right)^\top \left(\sum_{j=1}^{N} \lambda_j y_j \mathbf{x}_j \right) \\
=& \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{N} \lambda_i \lambda_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j = L_D
\end{aligned}
$$

Az optimalizálási probléma átfogalmazása (4)

Az L_D duális Lagrange-függvény már csak a Lagrange-multiplikátorokat és a tanulóadatokat tartalmazza!

Optimális elválasztó hipersík

Jelölje $\mathcal{S}$ a tartóvektorok indexhalmazát az $\{1, 2, \ldots, N\}$ halmazban. Ha meghatároztuk az optimális λ_i értékeket, akkor az alábbi lesz az optimális elválasztó hipersík $\mathbf{w}$ paraméterének értéke (ld. (1)):

$$\mathbf{w}_{opt} = \sum_{i=1}^{N} \lambda_i^{opt} y_i \mathbf{x}_i = \sum_{s \in \mathcal{S}} \lambda_s^{opt} y_s \mathbf{x}_s.$$

Úgy kapjuk meg az optimális elválasztó hipersík b paraméterének értékét, hogy véve egy $\mathbf{x}_s$ tartóvektort amelyre $y_s = \pm 1$,:

$$b_{opt} = y_s - \mathbf{w}_{opt}^{\top} \mathbf{x}_s.$$

Robusztus becsléshez a legszerencsésebb ha az összes tartóvektorra ezen paraméterértékeknek az átlagát vesszük:

$$b_{opt} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} (y_s - \mathbf{w}_{opt}^{\top} \mathbf{x}_s) = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \left(y_s - \sum_{s' \in \mathcal{S}} \lambda_{s'}^{opt} y_{s'} \mathbf{x}_{s'}^{\top} \mathbf{x}_s \right)$$

Osztályozás

Az optimális elválasztó hipersík birtokában

$$f(\mathbf{z}) = \text{sgn}\left(\mathbf{w}_{opt}^{\top}\mathbf{z} + b_{opt}\right)$$

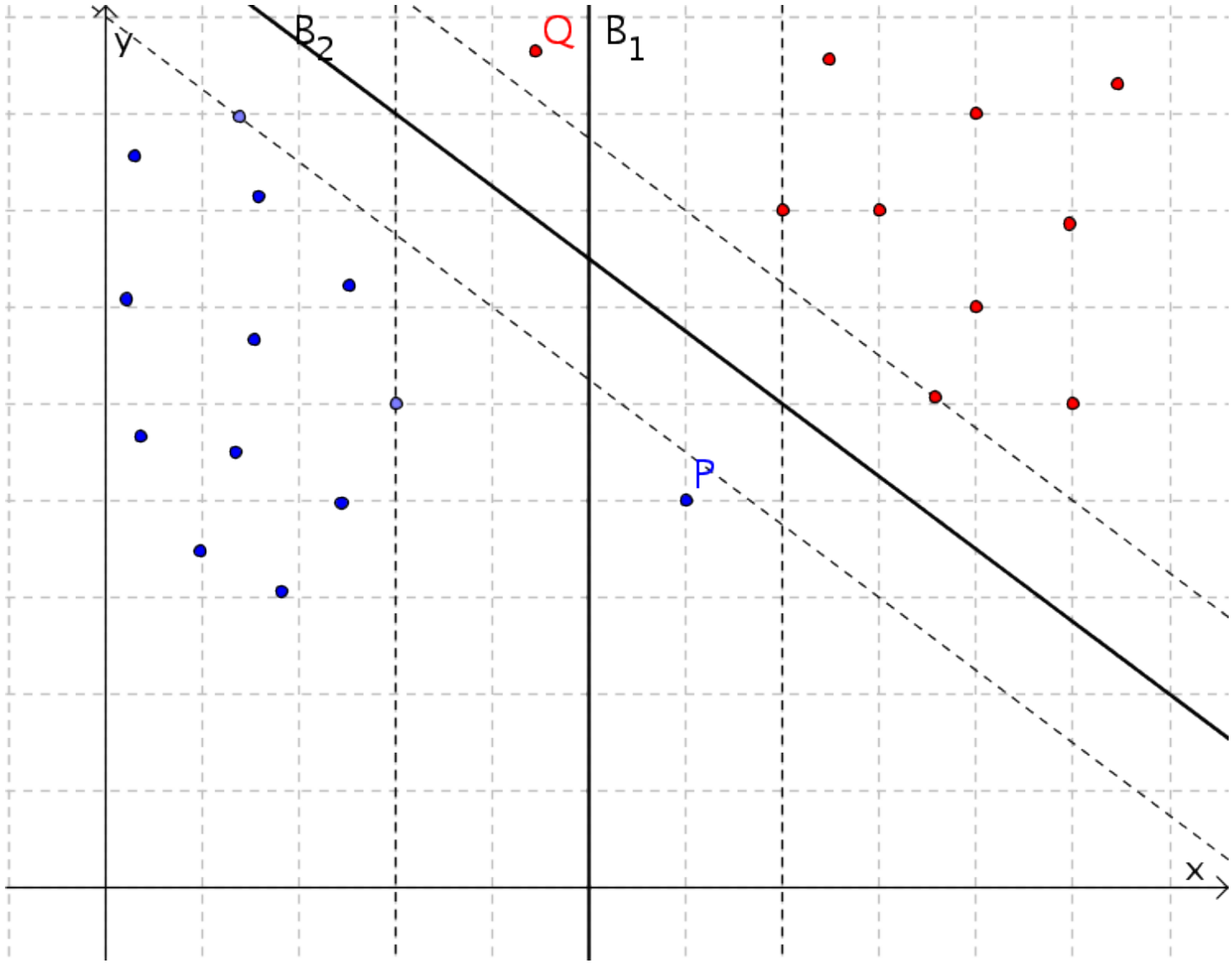
prediktálja egy $\mathbf{z} \in \mathbb{R}^n$ vektor osztálycímkéjét, azaz $\mathbf{z}$ egy lineáris függvényének előjele prediktálja az osztálycimkéjét.

Valójában az osztály predikcióhoz sincs másra szükségünk csak a tartóvektorokra és a hozzátartozó Lagrange-multiplikátorokra:

$$f(\mathbf{z}) = \text{sgn}\left(\sum_{s\in\mathcal{S}} \lambda_s^{opt} y_s \mathbf{x}_s^{\top}\mathbf{z} + b_{opt}\right)$$

és mint láttuk b_{opt} is csak a tartóvektoroktól függ. Így az osztály predikciója is csak akkor változik meg egy $\mathbf{z}$ vektornak amennyiben megváltozik a tanító adatállomány, ha ez a változás hatással van a tartóvektorokra is. Látható, hogy $f(\mathbf{z})$ csak belsőszorzatokon keresztül függ az input tanítóvektoroktól.

Lineáris szeparálás hibával (1)



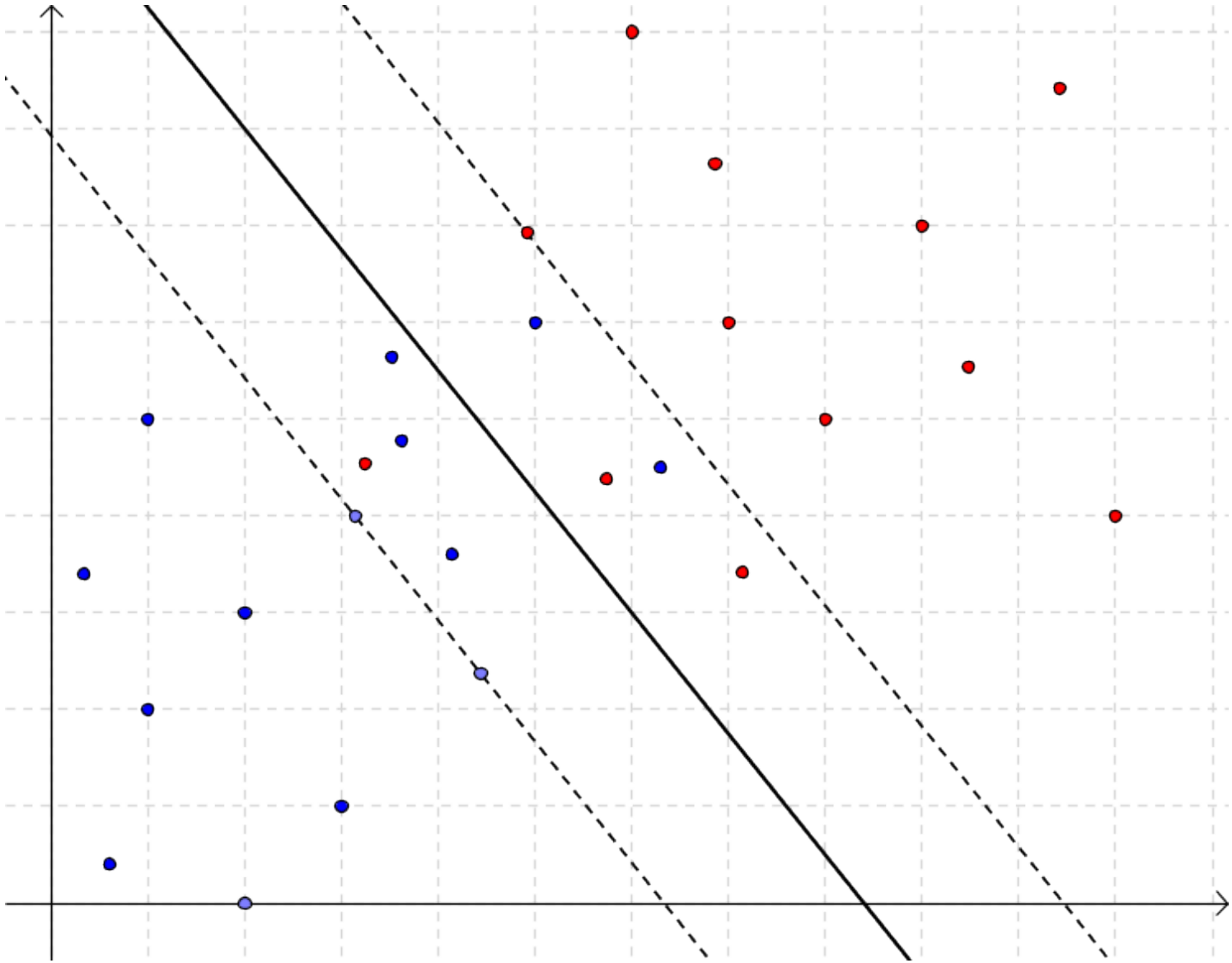
Lineáris szeparálás hibával (2)

Az előző ábrán a B_2 hipersík tökéletesen szétválasztja a két osztályt, a B_1 hipersík azonban nem.

Ez azonban nem feltétlenül jelenti azt, hogy B_2 jobb B_1-nél, mivel a B_1 által hibásan osztályozott P és Q pont zajnak felelhet meg.

B_1 a jobb elválasztó hipersík, mivel a nagyobb margó következtében kevésbé hajlamos a modell túlillesztésre.

Lineárisan nem szeparálható eset (1)



Lineárisan nem szeparálható eset (2)

A linerisan szeparálható osztályokra kidolgozott konstrukciót kiterjeszthetjük olyan esetekre, amelyekben nem lehetséges az osztályok tökéletes szétválasztása egy hipersíkkal.

Egy olyan elválasztó hipersíkot keresünk, amely egy kompromisszumot jelent az elkövetett osztályozási hibák és a margó szélessége között.

A probléma következőként bemutatásra kerülő megfogalmazása **puha margó**ként (**soft margin**) ismert.

Lineárisan nem szeparálható eset (3)

A lineárisan szeparálható eset egyenlőtlenség alakú korlátozásait a lineárisan nem szeparálható esetre gyengítjük az alábbi módon:

$$\begin{aligned} \mathbf{w}^\top \mathbf{x}_i + b &\geq 1 - \xi_i, \quad && \text{ha } y_i = +1, \\ \mathbf{w}^\top \mathbf{x}_i + b &\leq -1 + \xi_i, && \text{ha } y_i = -1, \end{aligned}$$

ahol ξ_i-k nemnegatív értékű, úgynevezett **kiegészítő változók** (**slack variables**).

A fenti egyenlőtlenségeket tömören

$$y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \qquad i = 1, \ldots, N, \tag{3}$$

alakba írhatjuk.

Lineárisan nem szeparálható eset (3)

Egy ξ_i kiegészítő változó az elválasztó hipersík hibáját jelenti az $\mathbf{x}_i$ vektor esetén, amikor a hipersíkot használjuk a két osztály szétválasztására:

- Ha $0 \leq \xi_i \leq 1$, akkor az $\mathbf{x}_i$ vektor a hipersík megfelelő oldalára esik, azonban vagy a margóra ($\xi_i = 0$) vagy a margón belülre.
- Ha $\xi_i > 1$, akkor az $\mathbf{x}_i$ vektor a hipersík nem megfelelő oldalára esik.

A tartóvektorok azok a vektorok, amelyekre egyenlőség áll fenn a (3) képletben. Nem szükséges azonban, hogy egy $\mathbf{x}_s$ tartóvektorra $\xi_s = 0$ teljesüljön!

A feladat megfogalmazása optimalizálási problémaként

Úgy módosítjuk a lineárisan szeparálható eset célfüggvényét, hogy szerepeljenek benne a hibát képviselő kiegészítő változók.

Tekintsük az alábbi feltételes optimalizálási problémát:

$$\min_{\mathbf{w}} \left(\frac{\|\mathbf{w}\|^2}{2} + C \sum_{i=1}^{N} \xi_i \right)$$

feltéve, hogy $\quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, N.$

A célfüggvényben szereplő C egy olyan pozitív értékű paraméter, melyet a felhasználó kell hogy meghatározzon, és amelyet **regularizációs paraméter**nek is neveznek.

Az optimalizálási probléma megoldása (1)

A probléma primál Lagrange-függvénye:

$$L_P = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\lambda_i\Big(y_i(\mathbf{w}^\top\mathbf{x}_i + b) - 1 + \xi_i\Big) - \sum_{i=1}^{N}\mu_i\xi_i,$$

ahol λ_i és μ_i paraméterek, úgynevezett Lagrange-multiplikátorok, amelyekre $\lambda_i \geq 0$ és $\mu_i \geq 0$ minden $i = 1, \ldots, N$ esetén.

Teljesülnek az alábbi, ún. Karush-Kuhn-Tucker feltételek:

$$\lambda_i\big[y_i(\mathbf{w}^\top\mathbf{x}_i + b) - 1 + \xi_i\big] = 0, \qquad \mu_i\xi_i = 0, \qquad i = 1, \ldots, N.$$

Az optimalizálási probléma megoldása (2)

Az L_P Lagrange-függvény $\mathbf{w}$, b és ξ_i szerinti elsőrendű deriváltjait nullává téve a következő egyenleteket kapjuk:

$$\frac{\partial L_P}{\partial \mathbf{w}} = 0 \Longrightarrow \mathbf{w} = \sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_i$$

$$\frac{\partial L_P}{\partial b} = 0 \Longrightarrow \sum_{i=1}^{N} \lambda_i y_i = 0$$

$$\frac{\partial L_P}{\partial \xi_i} = 0 \Longrightarrow \lambda_i + \mu_i = C, \qquad i = 1, \ldots, N$$

Az optimalizálási probléma megoldása (3)

Az előbbiek felhasználásával a lineárisan szeparálható eset mintájára az alábbi duális Lagrange-függvény alkotható:

$$L_D = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j,$$

amely azonos a lineárisan szeparálható adatok duális Lagrange-függvényével.

A korábbiakhoz hasonlóan L_D maximalizálása a feladat, feltéve hogy teljesülnek az alábbiak:

1. $\sum_{i=1}^{N} \lambda_i y_i = 0$,
2. $0 \leq \lambda_i \leq C$ minden $i = 1, \ldots, N$ esetén (ez az ún. doboz-feltétel).[3]

[3] A korábbi, $\lambda_i \geq 0$ feltételeket azért cseréltük $0 \leq \lambda_i \leq C$ feltételekre, mert a λ_i és μ_i multiplikátorok nemnegativitása miatt $\lambda_i + \mu_i = C$ azt jelenti, hogy egyik λ_i sem haladhatja meg C-t.

Az optimalizálási probléma megoldása (4)

A számolás részletezése:

$$
\begin{aligned}
L_P =& \frac{1}{2}\|\mathbf{w}\|^2 - \mathbf{w}^\top \underbrace{\sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_i}_{=\mathbf{w}} - b \underbrace{\sum_{i=1}^{N} \lambda_i y_i}_{=0} + \sum_{i=1}^{N} \lambda_i + \sum_{i=1}^{N} \underbrace{(C - \lambda_i - \mu_i)}_{=0} \xi_i \\
=& \sum_{i=1}^{N} \lambda_i - \frac{1}{2}\mathbf{w}^\top \mathbf{w} = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \left(\sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_i \right)^\top \left(\sum_{j=1}^{N} \lambda_j y_j \mathbf{x}_j \right) \\
=& \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{N} \lambda_i \lambda_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j = L_D
\end{aligned}
$$

Az optimalizálási probléma megoldása (5)

A λ_i Lagrange-multiplikátorok a lineárisan szeparálható esethez hasonlóan csak a tartóvektorok esetén nemnulla értékűek.

A $\lambda_i + \mu_i = C$ és $\mu_i \xi_i = 0$ egyenletekből következik azonban, hogy ha $\lambda_i < C$, akkor $\xi_i = 0$.

Ebből következik, hogy kétféle tartóvektor van:

- Ha $0 < \lambda_s < C$, akkor $\mathbf{x}_s$ olyan tartóvektor, amelyre $\xi_s = 0$. (Ezek a tartóvektorok a jóoldali margóra esnek.)
- Ha $\lambda_s = C$, akkor $\mathbf{x}_s$ olyan tartóvektor, amelyre $\xi_s > 0$. (Ezek a tartóvektorok vagy a margón belül vagy a hipersík másik oldalán vannak.)

Optimális elválasztó hipersík

Ha meghatároztuk az optimális λ_i értékeket, akkor az alábbi lesz az optimális elválasztó hipersík $\mathbf{w}$ paraméterének értéke:

$$\mathbf{w}_{opt} = \sum_{i=1}^{N} \lambda_i^{opt} y_i \mathbf{x}_i = \sum_{s \in \mathcal{S}} \lambda_s^{opt} y_s \mathbf{x}_s.$$

Jelölje $\mathcal{S}' \subset \mathcal{S}$ azon tartóvektorok s indexét melyre $\xi_s = 0$. Úgy kapjuk meg az optimális elválasztó hipersík b paraméterének értékét, hogy veszünk egy $\mathbf{x}_{s'}$ tartóvektort, amelyre $s' \in \mathcal{S}'$. Ekkor

$$b_{opt} = y_{s'} - \mathbf{w}_{opt}^{\top} \mathbf{x}_{s'},$$

ahol a robusztus becslés érdekében szintén érdemes az összes szóbajöhető esetre átlagolni:

$$b_{opt} = \frac{1}{|\mathcal{S}'|} \sum_{s' \in \mathcal{S}'} (y_{s'} - \mathbf{w}_{opt}^{\top} \mathbf{x}_{s'}) = \frac{1}{|\mathcal{S}'|} \sum_{s' \in \mathcal{S}'} \left(y_{s'} - \sum_{s \in \mathcal{S}} \lambda_s^{opt} y_s \mathbf{x}_s^{\top} \mathbf{x}_{s'} \right)$$

Osztályozás

Egy $\mathbf{z} \in \mathbb{R}^n$ vektor osztálycímkéjének prediktálása az optimális elválasztó hipersík birtokában pontosan úgy történik, mint a lineárisan szeparálható esetben:

$$f(\mathbf{z}) = \operatorname{sgn}\left(\sum_{s \in \mathcal{S}} \lambda_s^{opt} y_s \mathbf{x}_s^\top \mathbf{z} + b_{opt}\right)$$

A predikció csak az $\mathbf{x}_s^\top \mathbf{z}$ és $\mathbf{x}_s^\top \mathbf{x}_{s'}$ belsőszorzatokon keresztül függ az input tanítóvektoroktól, amelyek gyorsan számolhatóak.

A C paraméter hatásának szemléltetése (1)

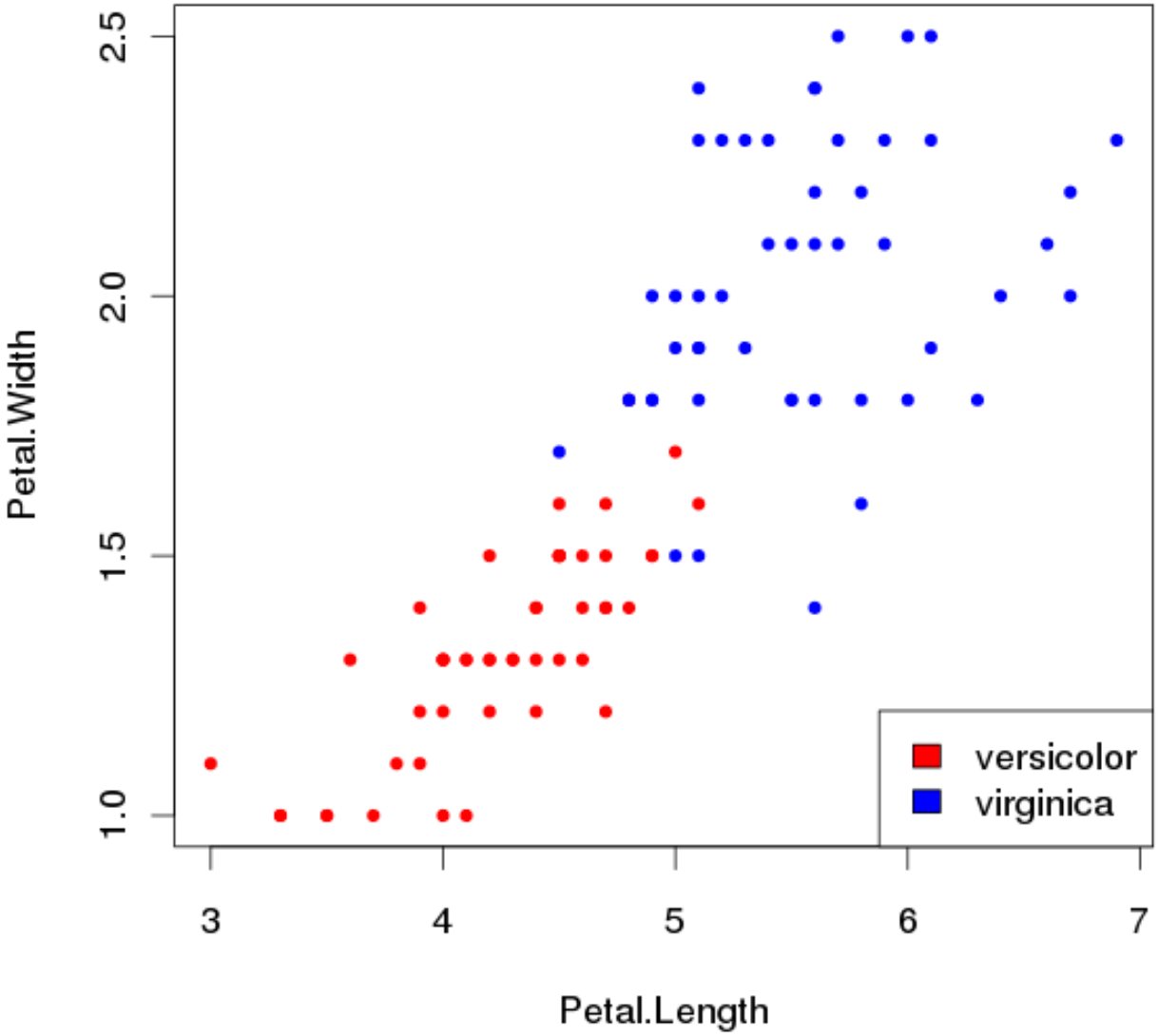
Az alábbi osztályozási feladat segítségével szemléltetjük a C paraméter hatását az optimális elválasztó hipersíkra.

Példa

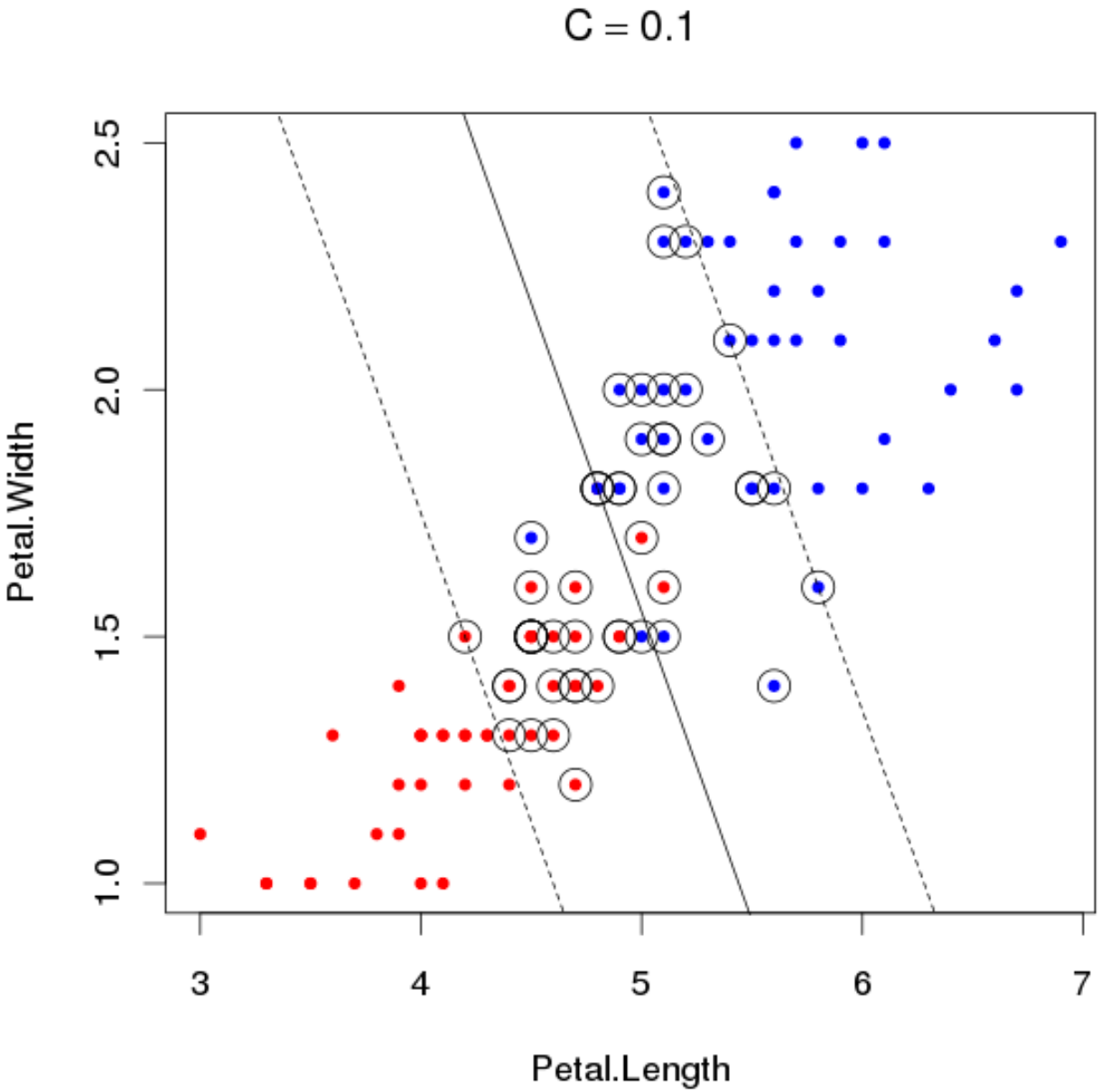
Tekintsük az írisz virágok adatállományából csak a foltos nőszirom (Iris versicolor) és a virginiai nőszirom (Iris virginica) virágok adatait. Osztályozzuk a virágokat a sziromlevél hossza és szélessége alapján a bemutatott konstrukció segítségével.

A következő ábrán jól látható, hogy a két osztály lineárisan nem szeparálható. Mindkét osztályba 50-50 virág tartozik.

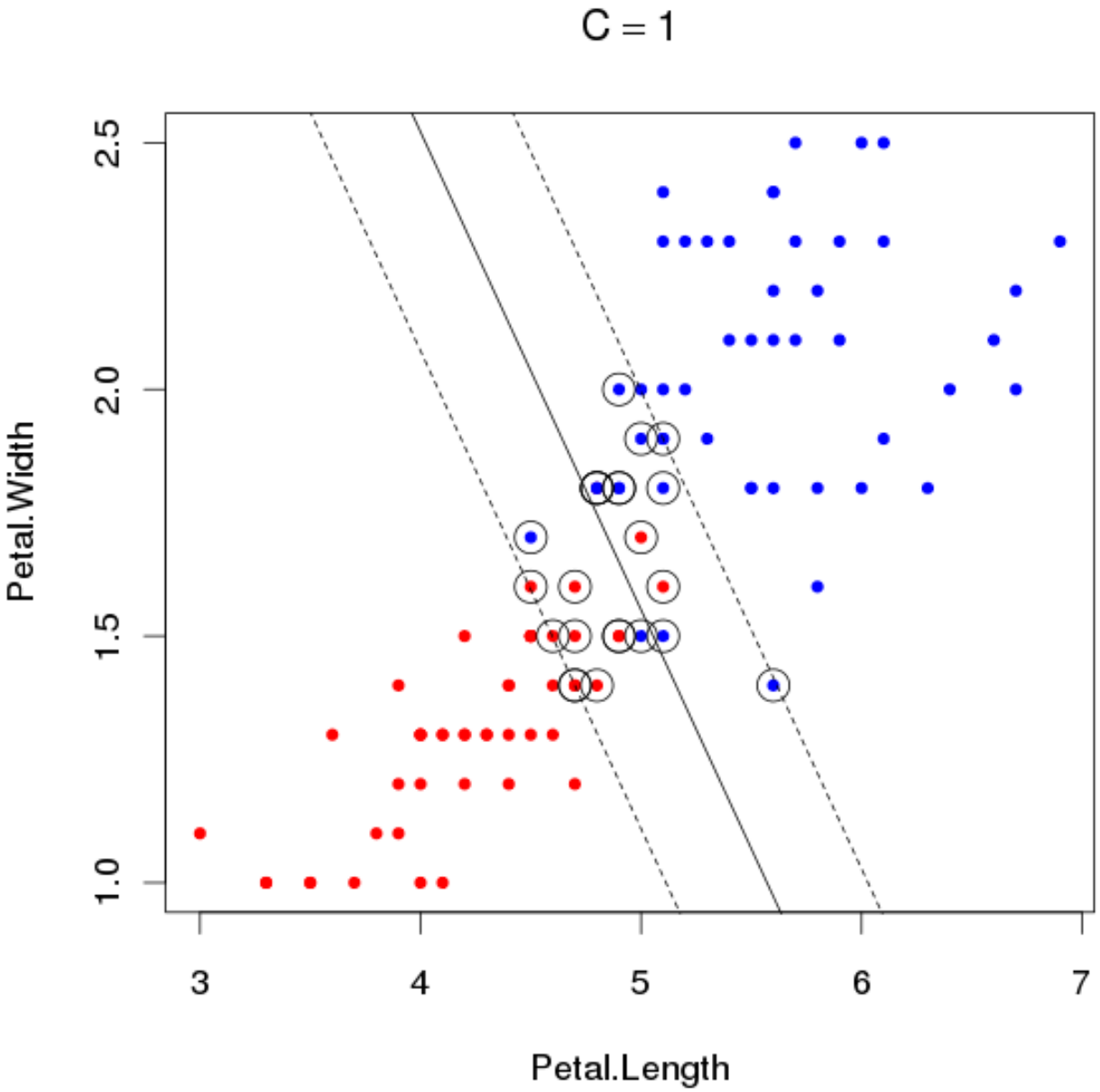
A C paraméter hatásának szemléltetése (2)



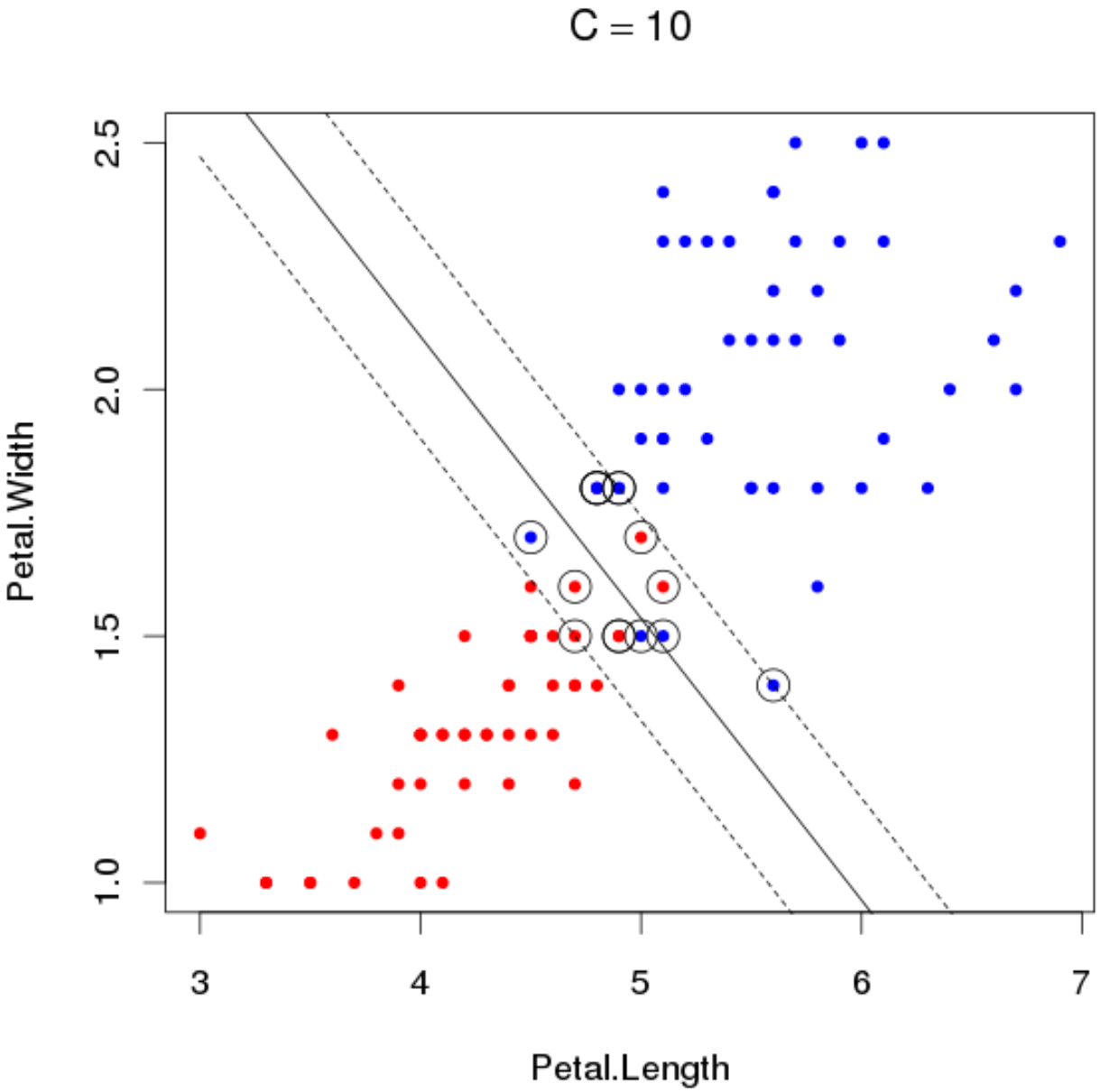
A C paraméter hatásának szemléltetése (3)



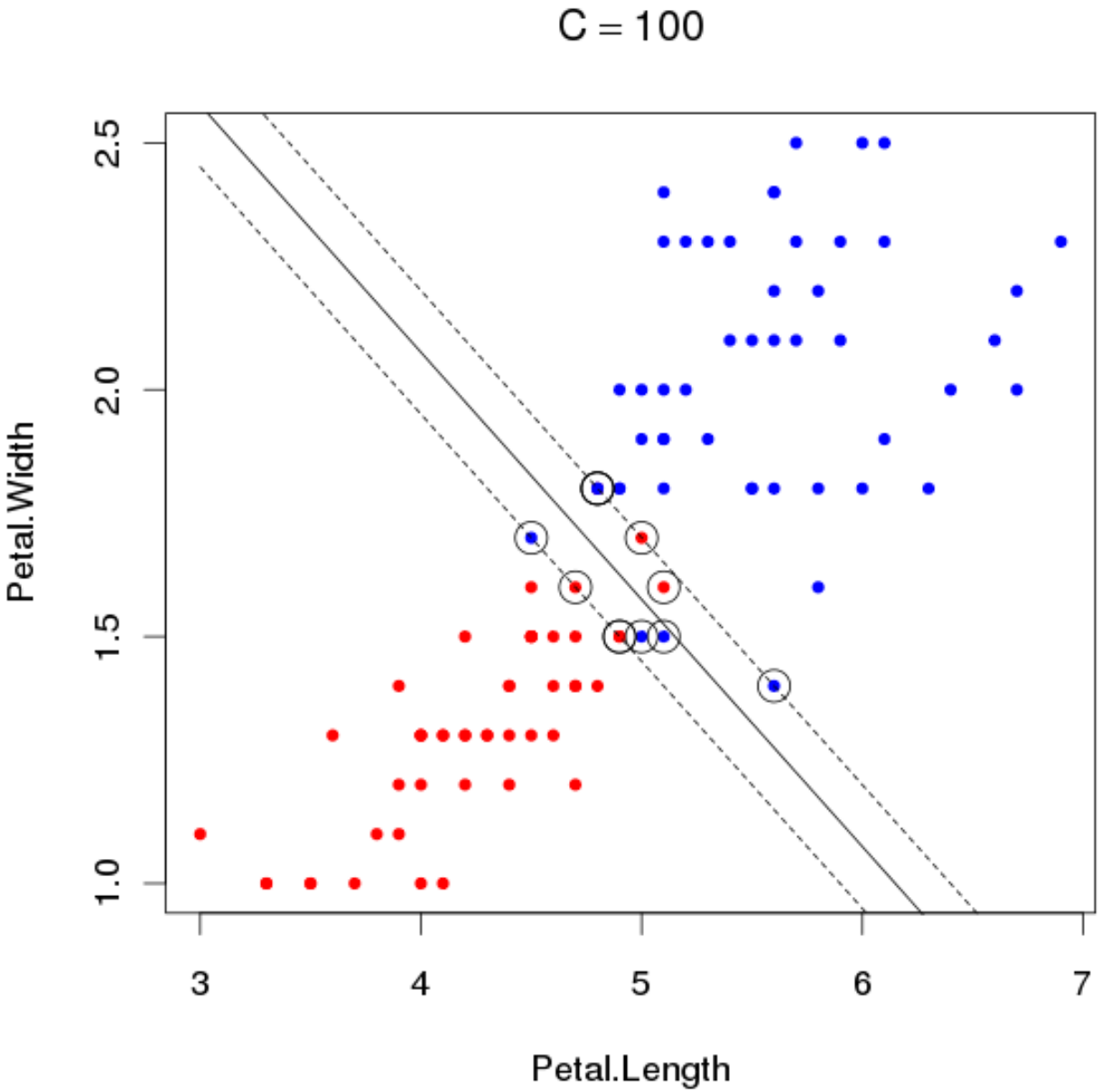
A C paraméter hatásának szemléltetése (4)



A C paraméter hatásának szemléltetése (5)



A C paraméter hatásának szemléltetése (6)



A C paraméter hatásának szemléltetése (7)

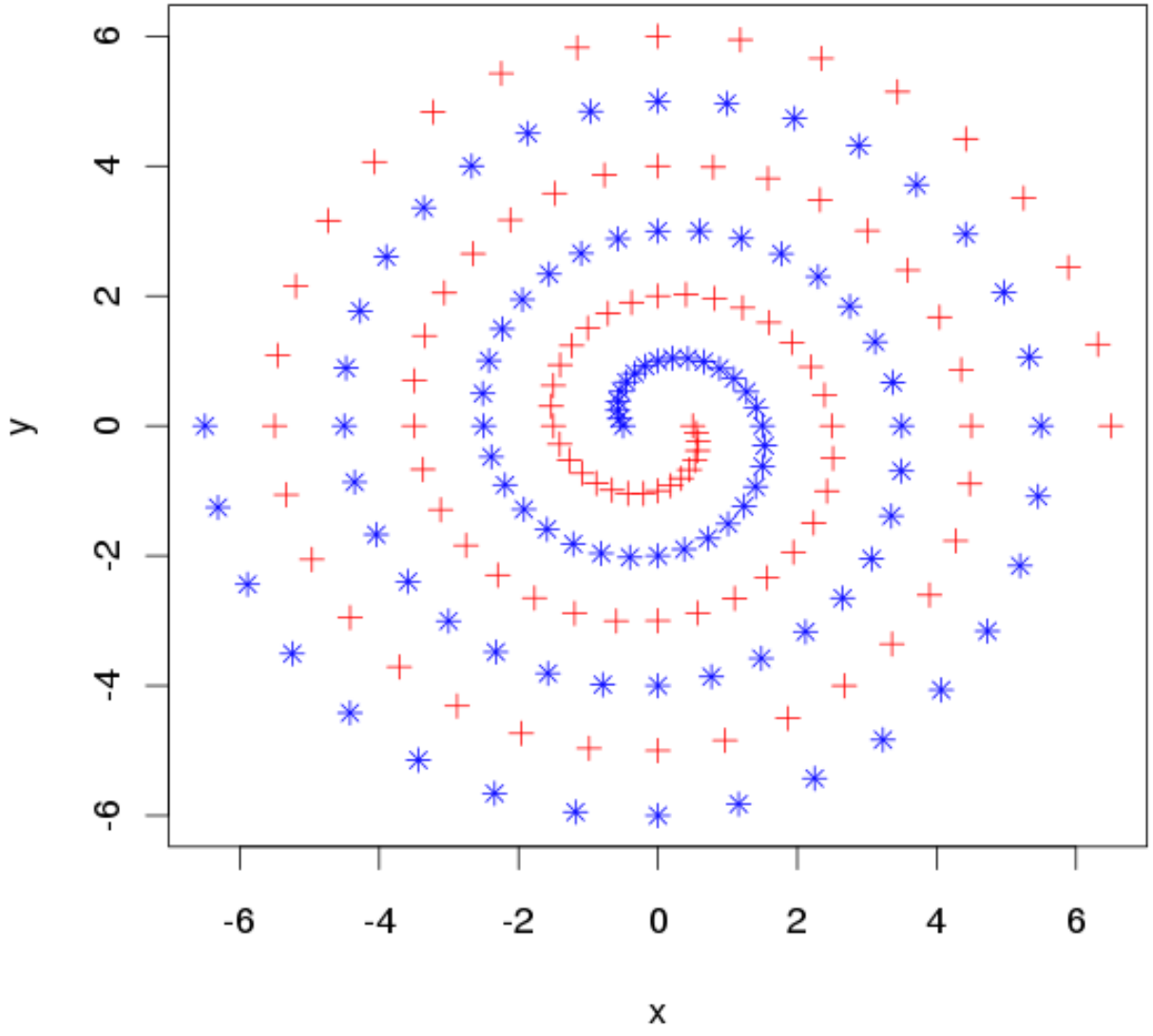
C	Tartóvektorok száma	Tanító halmazon mért pontosság
0,1	50	95%
1	24	95%
10	15	95%
100	12	94%

Nemlineáris SVM

Az előzőleg bemutatott mindkét konstrukció annak megfelelően osztályoz egy vektort, hogy az az optimális elválasztó hipersík melyik oldalára esik.

Egy módszertant adunk az SVM olyan adatokra történő alkalmazására, amelyek egy hipersík segítségével nem választhatók szét megfelelően két osztályra.

Példa lineárisan nem szeparálható adatokra



Alapötlet

Transzformáljuk az adatokat egy olyan nagy dimenziószámú térbe, amelyben azokat már megfelelően szét lehet választani két osztályra egy hipersík segítségével. A transzformáció nemlineáris függvények felhasználásával történik.

Implementációs kérdések

Az alapötlet kapcsán felmerülő implementációs kérdések:

- Nem világos, hogy adott adatok esetén pontosan milyen leképezést használjunk.
- Ha ismernénk is a megfelelő leképezést, számításigényes lehet a feltételes optimalizálási feladat megoldása a sokdimenziós térben.

Egy olyan módszert adunk, amely a fenti problémákat megfelelően kezeli. Látni fogjuk, hogy nem szükséges a leképezés explicit megadása, és a felmerülő feltételes optimalizálási feladat is hatékonyan megoldható.

Lineáris szeparálás a transzformált térben (1)

Tegyük fel, hogy ismerjük azt a $\mathbf{\Phi}$ transzformációt, amely az adatokat egy olyan m-dimenziós térbe transzformálja, amelyben már lehetséges azok lineáris szeparálása. Legyen

$$\mathbf{\Phi}(\mathbf{x}) = (\phi_1(\mathbf{x}), \ldots, \phi_m(\mathbf{x})),$$

ahol minden ϕ_i egy nemlineáris függvény ($i = 1, \ldots, m$).

A $\mathbf{\Phi}$ leképezés által meghatározott transzformált teret **tulajdonságtér**nek (**feature space**) nevezik.

Lineáris szeparálás a transzformált térben (2)

A transzfomált térben a

$$\sum_{i=1}^{m} w_i \phi_i(\mathbf{x}) + b = 0$$

egyenlettel adható meg egy elválasztó hipersík, amely tömören

$$\mathbf{w}^\top \mathbf{\Phi}(\mathbf{x}) + b = 0$$

alakba írható.

Megjegyzés

Ha a $\mathbf{w}$ vektort kibővítjük a $w_0 = b$ elemmel, a $\mathbf{\Phi}$ vektort pedig a $\phi_0(\mathbf{x}) \equiv 1$ leképezéssel, akkor a hipersíkot meghatározó képlet

$$\mathbf{w}^\top \mathbf{\Phi}(\mathbf{x}) = 0$$

alakot ölt.

A feladat megfogalmazása optimalizálási problémaként

A korábbiak mintájára az alábbi feltételes optimalizálási probléma fogalmazható meg:

$$\min_{\mathbf{w}} \left(\frac{\|\mathbf{w}\|^2}{2} + C \sum_{i=1}^{N} \xi_i \right)$$

feltéve, hogy $\quad y_i(\mathbf{w}^\top \mathbf{\Phi}(\mathbf{x}_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, N,$

ahol C a felhasználó által meghatározott nemnegatív értékű paraméter.

A problémát pontosan úgy oldhatjuk meg a Lagrange-multiplikátor módszerrel, mint a korábbi hasonló problémát.

Az optimalizálási probléma megoldása (1)

Az alábbi duális Lagrange-függvény alkotható:

$$L_D = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j \mathbf{\Phi}(\mathbf{x}_i)^\top \mathbf{\Phi}(\mathbf{x}_j).$$

A korábbiakhoz hasonlóan L_D maximalizálása a feladat, feltéve hogy teljesülnek az alábbiak:

1. $\sum_{i=1}^{N} \lambda_i y_i = 0$,
2. $0 \leq \lambda_i \leq C$ minden $i = 1, \ldots, N$ esetén.

Megjegyzés

Mindössze annyi változott a duális Lagrange-függvényben, hogy benne az $\mathbf{x}_i^\top \mathbf{x}_j$ belső szorzatok helyett a transzformált térbeli $\mathbf{\Phi}(\mathbf{x}_i)^\top \mathbf{\Phi}(\mathbf{x}_j)$ belső szorzatok szerepelnek.

Az optimalizálási probléma megoldása (2)

A feltételes optimalizálási feladat megoldására a korábbiakhoz hasonlóan teljesül

$$\mathbf{w} = \sum_{i=1}^{N} \lambda_i y_i \mathbf{\Phi}(\mathbf{x}_i).$$

Ha meghatároztuk az optimális λ_i értékeket, akkor ez alapján az alábbi lesz az optimális elválasztó hipersík $\mathbf{w}$ paraméterének értéke:

$$\mathbf{w}_{opt} = \sum_{i=1}^{N} \lambda_i^{opt} y_i \mathbf{\Phi}(\mathbf{x}_i) = \sum_{s \in \mathcal{S}} \lambda_s^{opt} y_s \mathbf{\Phi}(\mathbf{x}_s).$$

A $\mathbf{w}_{opt}$ paraméter továbbra is csak a tartóvektorokon keresztül függ a tanító adatállománytól.

Az optimalizálási probléma megoldása (3)

Jelölje továbbra is $\mathcal{S}'$ a jóoldali margóra eső tartóvektorok indexeinek halmazát, azaz olyan $s \in \mathcal{S}$ melyekre $\xi_s = 0$. Az optimális elválasztó hipersík b paraméterének becslése:

$$\begin{aligned} b_{opt} &= \frac{1}{|\mathcal{S}'|} \sum_{s' \in \mathcal{S}'} (y_{s'} - \mathbf{w}_{opt}^\top \mathbf{\Phi}(\mathbf{x}_{s'})) \\ &= \frac{1}{|\mathcal{S}'|} \sum_{s' \in \mathcal{S}'} \left(y_{s'} - \sum_{s \in \mathcal{S}} \lambda_s^{opt} y_s \mathbf{\Phi}(\mathbf{x}_s)^\top \mathbf{\Phi}(\mathbf{x}_{s'}) \right) \end{aligned}$$

A b_{opt} paraméter is csak a tartóvektorokon keresztül függ a tanító adatállománytól.

Osztályozás

Az optimális elválasztó hipersík birtokában

$$
\begin{aligned}
f(\mathbf{z}) &= \operatorname{sgn}\left(\mathbf{w}_{opt}^{\top}\mathbf{\Phi}(\mathbf{z}) + b_{opt}\right) \\
&= \operatorname{sgn}\left(\sum_{i=1}^{N} \lambda_i^{opt} y_i \mathbf{\Phi}(\mathbf{x}_i)^{\top}\mathbf{\Phi}(\mathbf{z}) + b_{opt}\right) \\
&= \operatorname{sgn}\left(\sum_{s\in\mathcal{S}} \lambda_s^{opt} y_s \mathbf{\Phi}(\mathbf{x}_s)^{\top}\mathbf{\Phi}(\mathbf{z}) + b_{opt}\right)
\end{aligned}
$$

prediktálja egy $\mathbf{z} \in \mathbb{R}^n$ vektor osztálycímkéjét.

Kernel-trükk (1)

Az L_D duális Lagrange-függvényben és az osztálycímkét prediktáló $f(\mathbf{z})$ függvényben szereplő $\mathbf{\Phi}(\mathbf{x}_i)^\top \mathbf{\Phi}(\mathbf{x}_j)$ és $\mathbf{\Phi}(\mathbf{x}_i)^\top \mathbf{\Phi}(\mathbf{z})$ kifejezések az adatvektorok transzformált térbeli belső szorzatai.

A transzformált térbeli belső szorzat kiszámításához nem feltétlenül szükséges a $\mathbf{\Phi}$ leképezés ismerete! A belső szorzat ráadásul gyorsan és olcsón számolható az eredeti adatvektorokból.

Kernek-trükknek (**kernel trick**) nevezik a bemutatásra kerülő módszert.

Kernel-trükk (2)

Kernelfüggvénynek nevezünk egy olyan függvényt, amely vektorok egy transzformált térbeli belső szorzatát az eredeti adattérben számítja ki.

Egy K kernelfüggvényre tehát

$$K(\mathbf{u}, \mathbf{v}) = \mathbf{\Phi}(\mathbf{u})^\top \mathbf{\Phi}(\mathbf{v}) = \sum_{i=1}^{m} \phi_i(\mathbf{u}) \phi_i(\mathbf{v})$$

teljesül valamilyen $\mathbf{\Phi}$ leképezés esetén.

Egy kernelfüggvény szimmetrikus a változóiban, azaz

$$K(\mathbf{u}, \mathbf{v}) = K(\mathbf{v}, \mathbf{u}).$$

Kernel-trükk (3)

Példa

Szemléltetésül tekintsük a

$$\boldsymbol{\Phi}(x_1, x_2) = \left(1, x_1^2, \sqrt{2}x_1x_2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2\right),$$

leképezést, ahol $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$. Ekkor az $\mathbf{u} = (u_1, u_2) \in \mathbb{R}^2$ és $\mathbf{v} = (v_1, v_2) \in \mathbb{R}^2$ vektorok belső szorzata a $\boldsymbol{\Phi}$ leképezés által meghatározott transzformált térben

$$\begin{aligned}
&\boldsymbol{\Phi}(\mathbf{u})^\top \boldsymbol{\Phi}(\mathbf{v}) = \\
&\quad = \left(1, u_1^2, \sqrt{2}u_1u_2, u_2^2, \sqrt{2}u_1, \sqrt{2}u_2\right)^\top \left(1, v_1^2, \sqrt{2}v_1v_2, v_2^2, \sqrt{2}v_1, \sqrt{2}v_2\right) = \\
&\qquad = u_1^2v_1^2 + 2u_1v_1u_2v_2 + u_2^2v_2^2 + 2u_1v_1 + 2u_2v_2 + 1 = (\mathbf{u}^\top\mathbf{v} + 1)^2.
\end{aligned}$$

Tehát a transzformált térbeli belső szorzatot ki tudjuk fejezni az eredeti adatvektorok belső szorzata alapján.

Kernelfüggvény tehát a

$$K(\mathbf{u}, \mathbf{v}) = (\mathbf{u}^\top\mathbf{v} + 1)^2.$$

leképezés.

Kernel-trükk (4)

Csak bizonyos függvényeket lehet kifejezni

$$K(\mathbf{u}, \mathbf{v}) = \mathbf{\Phi}(\mathbf{u})^\top \mathbf{\Phi}(\mathbf{v})$$

alakban. Ehhez egy szükséges és elégséges feltételt a Mercer-tétel fogalmaz meg.

A gyakorlatban nem kell tehát explicit módon megadni a $\mathbf{\Phi}$ leképezést, hanem egy alkalmas kernelfüggvényt kell csupán választani.

Kernel-trükk (5)

Az optimalizálási feladat megoldásánál is a kernelfüggvényt használjuk a transzformált térbeli belső szorzat kiszámításához. Így az optimalizálási feladat az alábbi alakot ölti:

Legyen

$$L_D = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j).$$

A feladat L_D maximalizálása a feladat, feltéve hogy teljesülnek az alábbiak:

1. $\sum_{i=1}^{N} \lambda_i y_i = 0$,
2. $0 \leq \lambda_i \leq C$ minden $i = 1, \ldots, N$ esetén.

Kernel-trükk (6)

Az alábbi módon fejezhető ki az optimális elválasztó (nemlineáris) hipersík a kernelfüggvény segítségével:

$$\sum_{s \in \mathcal{S}} \lambda_s^{opt} y_s K(\mathbf{x}_s, \mathbf{x}) + b_{opt} = 0,$$

ahol

$$b_{opt} = \frac{1}{\mathcal{S}'} \sum_{s' \in \mathcal{S}'} \left(y_{s'} - \sum_{s \in \mathcal{S}} \lambda_s^{opt} y_s K(\mathbf{x}_{s'}, \mathbf{x}_s) \right)$$

Egy $\mathbf{z}$ teszteset osztálycímkéjének predikciója pedig:

$$f(\mathbf{z}) = \text{sgn} \left(\sum_{s \in \mathcal{S}} \lambda_s^{opt} y_s K(\mathbf{x}_s, \mathbf{z}) + b_{opt} \right)$$

Azaz mind az elválasztó hipersík, mind a predikciós függvény csak a tartóvektoroktól függ, az input adatoktól való függés pedig a kernel függvényen keresztül valósul meg.

Mercer-tétel

A kernelfüggvényként szóba jöhető függvényekre fogalmaz meg feltételt a következő tétel.

Tétel (Mercer-tétel)

Egy változóiban szimmetrikus folytonos $K(\mathbf{u}, \mathbf{v})$ *függvény akkor, és csak akkor fejezhető ki*

$$K(\mathbf{u}, \mathbf{v}) = \mathbf{\Phi}(\mathbf{u})^\top \mathbf{\Phi}(\mathbf{v})$$

alakban, ha minden olyan $g(\mathbf{x})$ *függvény esetén, amelyre* $\int g(\mathbf{x})^2 \, d\mathbf{x}$ *véges, teljesül*

$$\iint K(\mathbf{x}, \mathbf{y}) g(\mathbf{x}) g(\mathbf{y}) \, d\mathbf{x} \, d\mathbf{y} \geq 0.$$

Elterjedten használt kernelfüggvények

Néhány a Mercer-tételt kielégítő és elterjedten használt kernelfüggvény:

- **Polinomiális kernel:**

$$K(\mathbf{x}, \mathbf{y}) = \left(\mathbf{x}^\top \mathbf{y} + 1\right)^p,$$

ahol p a felhasználó által meghatározott paraméter.

- **Radiális bázisfüggvény (RBF) kernel:**

$$K(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x} - \mathbf{y}\|^2\right),$$

ahol σ a felhasználó által meghatározott paraméter.

- **Szigmoid kernel:**

$$K(\mathbf{x}, \mathbf{y}) = \tanh(\beta_0 \mathbf{x}^\top \mathbf{y} + \beta_1),$$

ahol β_0 és β_1 a felhasználó által meghatározott paraméter. A Mercer-tétel csak bizonyos β_0 és β_1 értékek mellett teljesül!

Kernel-mátrix

Kernel-mátrixnak nevezzük a

$$\mathbf{K} = \left[K(\mathbf{x}_i, \mathbf{x}_j)\right]_{N \times N}$$

mátrixot.

A kernel-mátrix tulajdonságai:

- Szimmetrikus, azaz $\mathbf{K}^\top = \mathbf{K}$.
- Pozitív definit mátrix, azaz minden olyan $\mathbf{x} \in \mathbb{R}^N$ esetén, hogy $\mathbf{x} \neq \mathbf{0}$, teljesül $\mathbf{x}^\mathbf{T}\mathbf{K}\mathbf{x} > 0$.

A tulajdonságtér (1)

Adott kernel esetén nem egyértelmű a $\mathbf{\Phi}$ transzformáció és az általa meghatározott tulajdonságtér.

Példa

Tekintsük a $K(\mathbf{x}, \mathbf{y}) = \left(\mathbf{x}^\top \mathbf{y}\right)^2$ kernelfüggvényt és legyen $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$ esetén

$$\mathbf{\Phi}_1(x_1, x_2) = \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right)$$

valamint

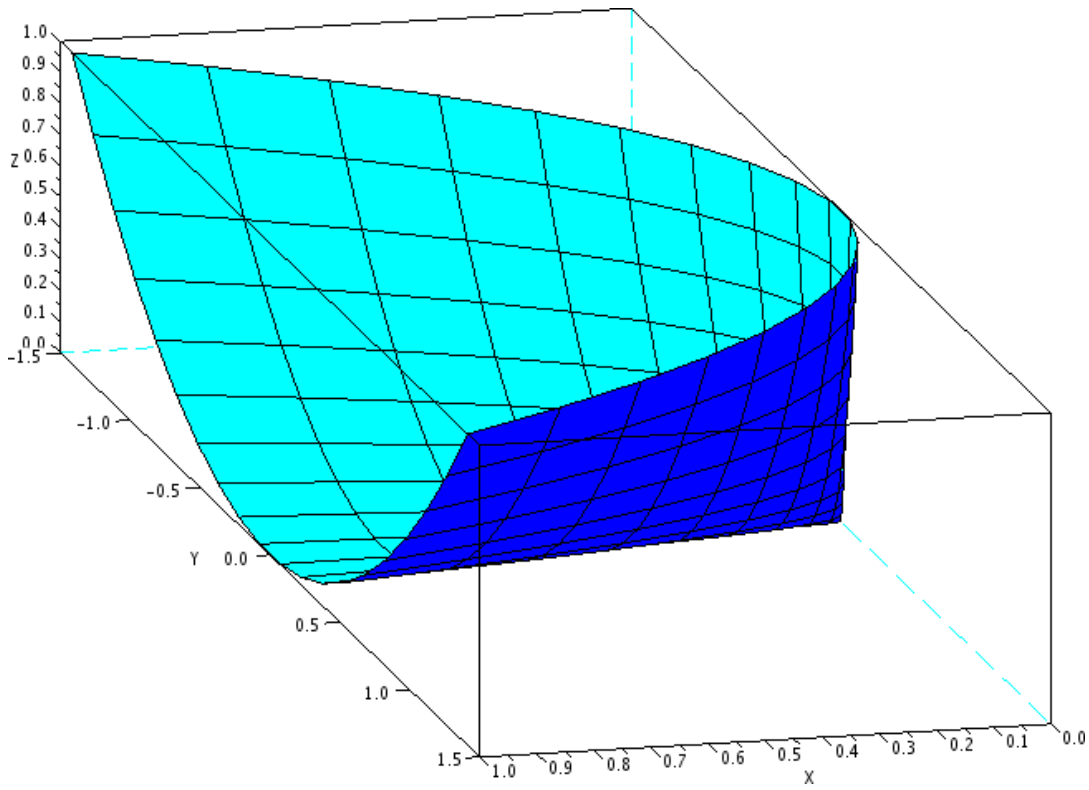
$$\mathbf{\Phi}_2(x_1, x_2) = \frac{1}{\sqrt{2}}\left(x_1^2 + x_2^2, 2x_1x_2, x_1^2 - x_2^2\right).$$

Ekkor könnyen látható, hogy

$$K(\mathbf{u}, \mathbf{v}) = \mathbf{\Phi}_1(\mathbf{u})^\top \mathbf{\Phi}_1(\mathbf{v}) \quad \text{és} \quad K(\mathbf{u}, \mathbf{v}) = \mathbf{\Phi}_2(\mathbf{u})^\top \mathbf{\Phi}_2(\mathbf{v}),$$

ahol $\mathbf{u}, \mathbf{v} \in \mathbb{R}^2$.

A tulajdonságtér (2)



A $[-1,1] \times [-1,1] \subset \mathbb{R}^2$ négyzet képe a $\mathbf{\Phi}_1$ leképezés által meghatározott tulajdonságtérben.

Példa nemlineáris tartóvektorgépre (1)

Példa

Tekintsük az alábbi táblázatban látható adatokat:

	x_1	x_2	y
1.	-1	-1	-1
2.	-1	$+1$	$+1$
3.	$+1$	-1	$+1$
4.	$+1$	$+1$	-1

Ez a közismert XOR adatállomány. Osztályozzuk az adatokat nemlináris tartóvektorgép segítségével.

Példa nemlineáris tartóvektorgépre (2)

Használjuk a

$$K(\mathbf{u}, \mathbf{v}) = (\mathbf{u}^\top \mathbf{v} + 1)^2$$

kernelfüggvényt. Korábban láttuk, hogy $\mathbf{u} = (u_1, u_2) \in \mathbb{R}^2$ és $\mathbf{v} = (v_1, v_2) \in \mathbb{R}^2$ esetén ez a kernelfüggvény az $\mathbf{u}$ és $\mathbf{v}$ vektorok a

$$\mathbf{\Phi}(x_1, x_2) = (1, x_1^2, \sqrt{2}x_1x_2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2)$$

leképezés által meghatározott transzformált térbeli képének belső szorzatát szolgáltatja.

Példa nemlineáris tartóvektorgépre (3)

A kernel-mátrix:

$$\mathbf{K} = \begin{bmatrix} 9 & 1 & 1 & 1 \\ 1 & 9 & 1 & 1 \\ 1 & 1 & 9 & 1 \\ 1 & 1 & 1 & 9 \end{bmatrix}$$

A duális Lagrange-függvény:

$$L_D = \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 - \frac{1}{2}\Big(9\lambda_1^2 - 2\lambda_1\lambda_2 - 2\lambda_1\lambda_3 + 2\lambda_1\lambda_4 + 9\lambda_2^2 + 2\lambda_2\lambda_3 - 2\lambda_2\lambda_4 + 9\lambda_3^2 - 2\lambda_3\lambda_4 + 9\lambda_4^2\Big)$$

Példa nemlineáris tartóvektorgépre (4)

Végezzünk az L_D duális Lagrange-függvényen parciális deriválást a Lagrange-multiplikátorok szerint, az eredményt pedig tegyük egyenlővé nullával. Ekkor az alábbi lineáris egyenletrendszer adódik:

$$\begin{aligned} 9\lambda_1 - \lambda_2 - \lambda_3 + \lambda_4 &= 1 \\ -\lambda_1 + 9\lambda_2 + \lambda_3 - \lambda_4 &= 1 \\ -\lambda_1 + \lambda_2 + 9\lambda_3 - \lambda_4 &= 1 \\ \lambda_1 - \lambda_2 - \lambda_3 + 9\lambda_4 &= 1 \end{aligned}$$

A lineáris egyenletrenszer megoldása szolgáltatja a Lagrange-multiplikátorok optimális értékét.

Példa nemlineáris tartóvektorgépre (5)

A Langrange-multiplikátorok optimális értékei:

$$\lambda_1^{opt} = \lambda_2^{opt} = \lambda_3^{opt} = \lambda_4^{opt} = \frac{1}{8},$$

tehát mind a négy vektor tartóvektor. A duális Lagrange-függvény optimális értéke:

$$L_D^{opt} = L_D(\boldsymbol{\lambda}_{opt}) = \frac{1}{4}.$$

Ez az optimális érték megegyezik az eredeti optimalizálási feladat célfüggvényének optimális értékével, tehát

$$\frac{\|\mathbf{w}_{opt}\|^2}{2} = \frac{1}{4},$$

azaz

$$\|\mathbf{w}_{opt}\| = \frac{1}{\sqrt{2}}.$$

Példa nemlineáris tartóvektorgépre (6)

Az optimális elválasztó hipersík $\mathbf{w}$ paraméterének értéke:

$$\mathbf{w}_{opt} = \frac{1}{8}\left(-\mathbf{\Phi}(\mathbf{x}_1) + \mathbf{\Phi}(\mathbf{x}_2) + \mathbf{\Phi}(\mathbf{x}_3) - \mathbf{\Phi}(\mathbf{x}_4)\right)$$

$$= \frac{1}{8}\left(- \begin{bmatrix} 1 \\ 1 \\ \sqrt{2} \\ 1 \\ -\sqrt{2} \\ -\sqrt{2} \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ -\sqrt{2} \\ 1 \\ -\sqrt{2} \\ \sqrt{2} \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ -\sqrt{2} \\ 1 \\ \sqrt{2} \\ -\sqrt{2} \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ \sqrt{2} \\ 1 \\ \sqrt{2} \\ \sqrt{2} \end{bmatrix} \right) = \begin{bmatrix} 0 \\ 0 \\ -\frac{1}{\sqrt{2}} \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

A $\mathbf{w}_{opt}$ vektor első eleme az optimális elválasztó hipersík b paraméterének értéke, amely tehát 0.

Példa nemlineáris tartóvektorgépre (7)

Az optimális elválasztó hipersík

$$\mathbf{w}_{opt}^{\top}\mathbf{\Phi}(\mathbf{x}) = 0,$$

azaz

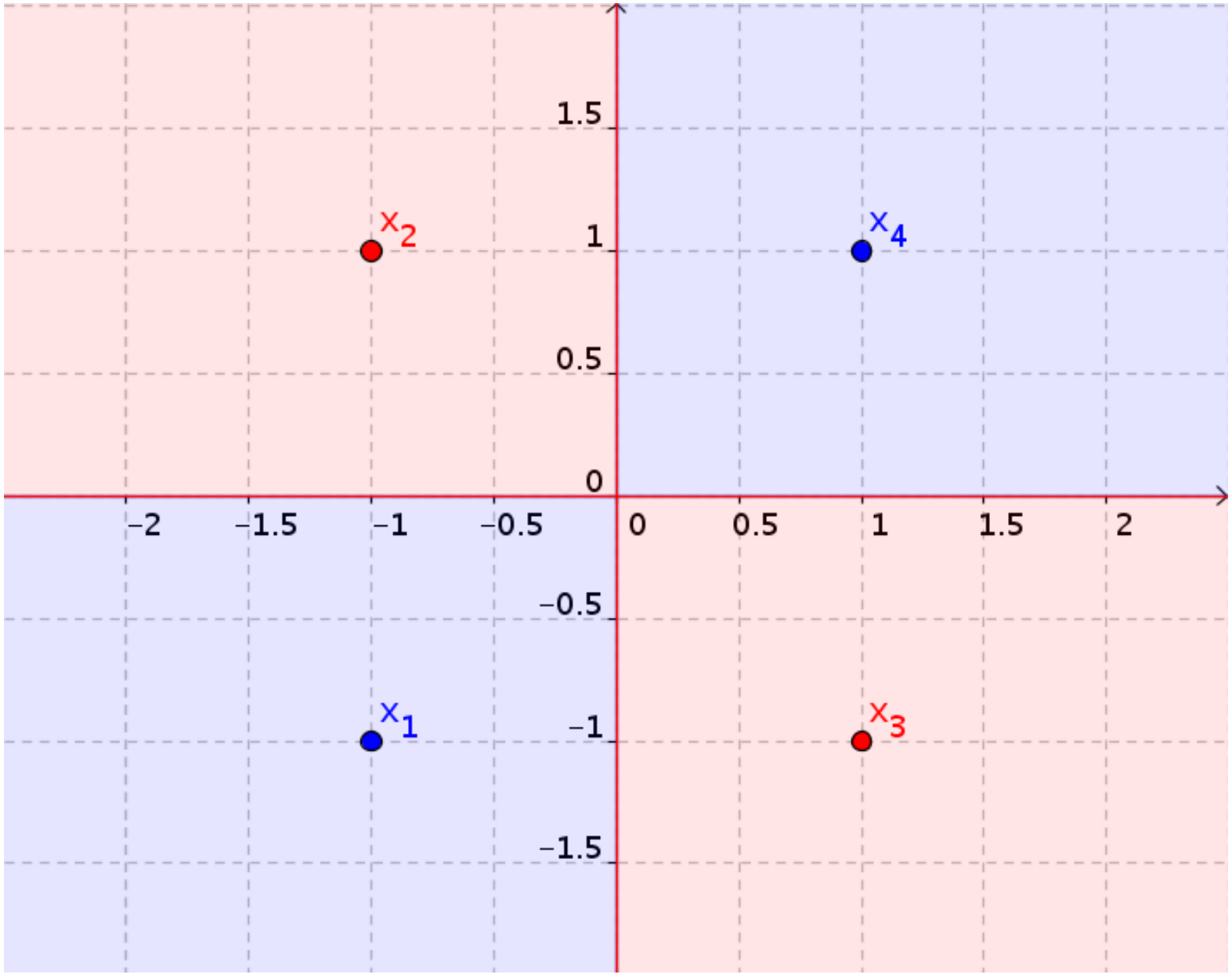
$$\left(0, 0, -1/\sqrt{2}, 0, 0, 0\right)^{\top}\left(1, x_1^2, \sqrt{2}x_1x_2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2\right) = 0,$$

amely röviden a

$$-x_1x_2 = 0$$

alakba írható.

Példa nemlineáris tartóvektorgépre (8)



Példa nemlineáris tartóvektorgépre (1)

Példa

Tekintsük az alábbi táblázatban látható adatokat:

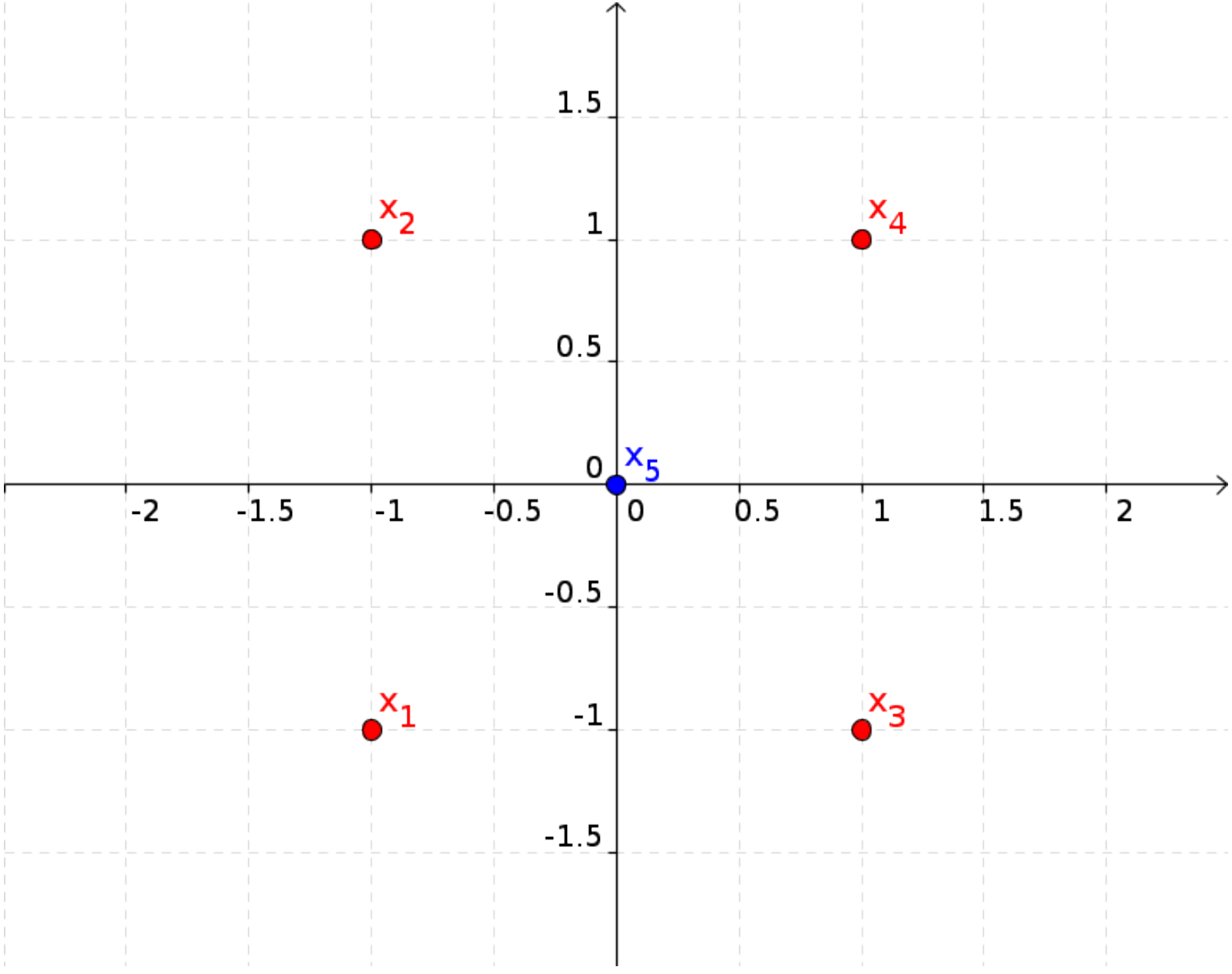
	x_1	x_2	y
1.	-1	-1	$+1$
2.	-1	$+1$	$+1$
3.	$+1$	-1	$+1$
4.	$+1$	$+1$	$+1$
5.	0	0	-1

Osztályozzuk az adatokat nemlináris tartóvektorgép segítségével, amelyhez használjuk a

$$K(\mathbf{u}, \mathbf{v}) = (\mathbf{u}^\top \mathbf{v} + 1)^2$$

kernelfüggvényt.

Példa nemlineáris tartóvektorgépre (2)



Példa nemlineáris tartóvektorgépre (3)

A Langrange-multiplikátorok optimális értékei:

$$\lambda_1^{opt} = \lambda_2^{opt} = \lambda_3^{opt} = \lambda_4^{opt} = \frac{1}{4} \quad \text{és} \quad \lambda_5^{opt} = 1.$$

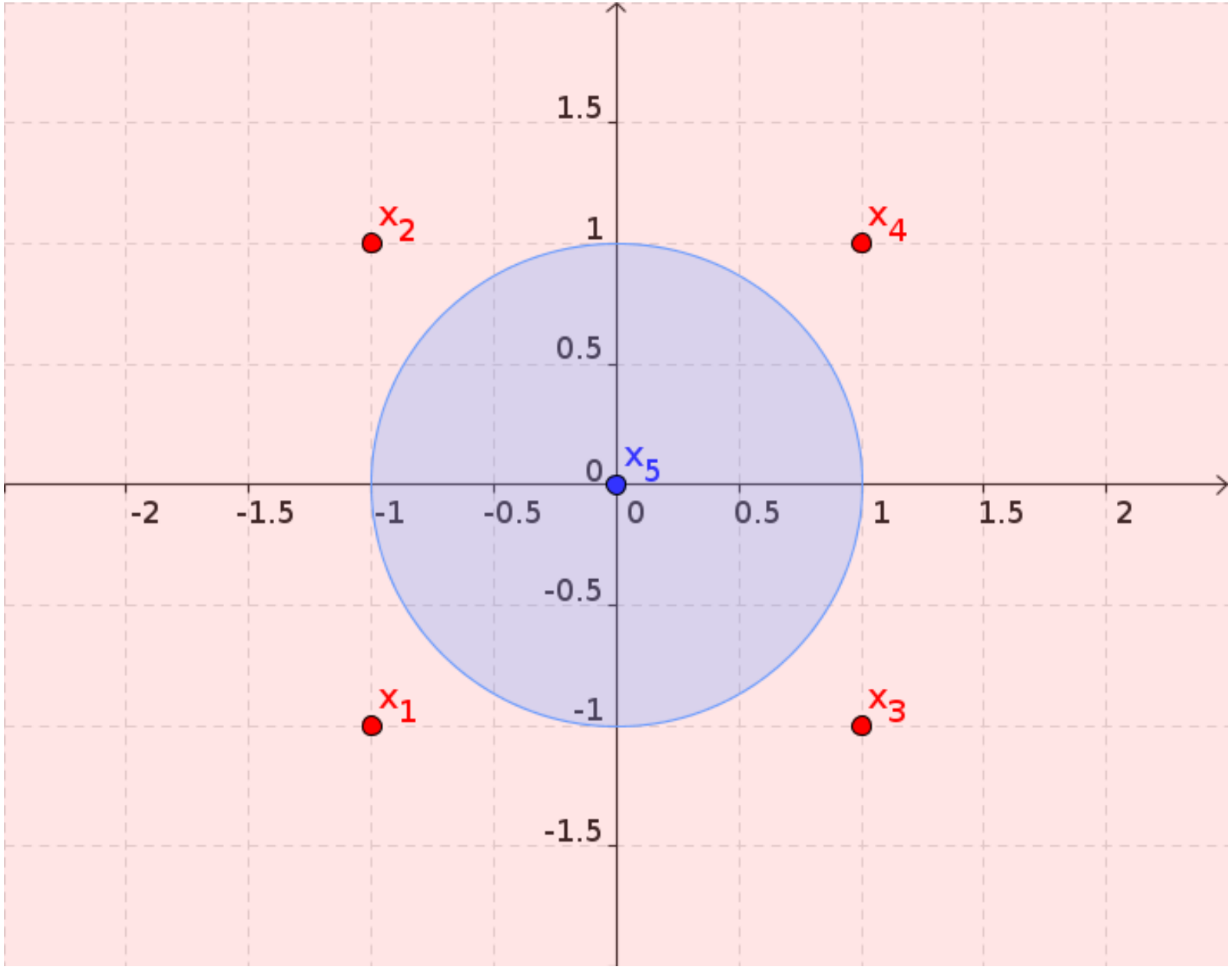
Az optimális elválasztó hipersík b paraméterének értéke:

$$b_{opt} = -1.$$

Az optimális elválasztó hipersík röviden az alábbi alakba írható:

$$x_1^2 + x_2^2 - 1 = 0.$$

Példa nemlineáris tartóvektorgépre (4)



Többosztályos osztályozás

A tartóvektorgépeket eredetileg bináris osztályozási feladatokhoz dolgozták ki.

Olyan osztályozási feladatok kezelésére használható módszerek, amelyekben kettőnél több osztály van:

- Többosztályos osztályozási feladat felbontása több bináris osztályozási feladattá.
- Többosztályos osztályozási feladat megfogalmazása feltételes szélsőértékszámítási feladatként.

Az első csoportba tartozó módszereket nem csupán tartóvektorgépekhez lehet használni, hanem bináris osztályozási feladatokat kezelő tetszőleges osztályozási modellekhez.

Osztályozási feladat felbontása több bináris osztályozási feladattá (1)

Egy vektor osztályozásához a részfeladatokhoz alkotott bináris osztályozók előrejelzéseit kombináljuk.

Ehhez tipikusan **többségi szavazás**t használunk, amelynek során minden egyes bináris osztályozó előrejelzését egy adott osztályra leadott szavazatnak tekintjük, végül a legtöbb szavazatot kapott osztályt választjuk.

Osztályozási feladat felbontása több bináris osztályozási feladattá (2)

A gyakorlatban alkalmazott megoldások:

- „egy az egy ellen” (one-against-one)
- „egy a többi ellen” (one-against-rest)

Osztályozási feladat felbontása több bináris osztályozási feladattá: „egy az egy ellen” (1)

$k > 2$ számú osztály esetén $k(k-1)/2$ bináris osztályozási feladat megoldása.

Az összes különböző osztálypár esetén egy-egy bináris osztályozó építése. Ennek során minden osztályozóhoz csupán a megfelelő két osztályba tartozó vektorokat használjuk fel.

Ezt a megoldást alkalmazza például a LIBSVM.[4]

[4] A LIBSVM egy népszerű tartóvektorgép programkönyvtár, amelyet sok alkalmazásban (például R, RapidMiner) használnak tartóvektorgépek megvalósításához.

Osztályozási feladat felbontása több bináris osztályozási feladattá: „egy az egy ellen” (2)

Példa

Tekintsünk egy olyan osztályozási feladatot, ahol 4 osztály van, és jelölje $Y = \{1, 2, 3, 4\}$ az osztálycímkék halmazát.
Tételezzük fel, hogy egy tesztvektort a következőképpen osztályozunk:

Bináris osztálypár	$+$: 1 $-$: 2	$+$: 1 $-$: 3	$+$: 1 $-$: 4	$+$: 2 $-$: 3	$+$: 2 $-$: 4	$+$: 3 $-$: 4
Osztályozás	$+$	$+$	$-$	$+$	$-$	$+$

A predikciók kombinálása után az 1. és 4. osztály két szavazatot kap, míg a 2. és 3. osztály csak egy szavazatot. A tesztvektort ezért az 1. vagy 4. osztályhoz tartozóként osztályozzuk.

Osztályozási feladat felbontása több bináris osztályozási feladattá: „egy a többi ellen” (1)

$k > 2$ számú osztály esetén k bináris osztályozási feladat megoldása.

Olyan bináris osztályozók építése, amelyek mindegyike az egyik osztályt tekinti pozitív osztályként, az összes többi osztályt pedig negatív osztályként.

Ha egy vektort negatívként osztályozunk, akkor a pozitív osztály kivételével minden osztály egy szavazatot kap.

Osztályozási feladat felbontása több bináris osztályozási feladattá: „egy a többi ellen” (2)

Példa

Tekintsünk egy olyan osztályozási feladatot, ahol 4 osztály van, és jelölje $Y = \{ 1, 2, 3, 4 \}$ az osztálycímkék halmazát.
Tételezzük fel, hogy egy tesztvektort a következőképpen osztályozunk:

Bináris osztálypár	$+$: 1 $-$: $\{2, 3, 4\}$	$+$: 2 $-$: $\{1, 3, 4\}$	$+$: 3 $-$: $\{1, 2, 4\}$	$+$: 4 $-$: $\{1, 2, 3\}$
Osztályozás	$+$	$-$	$-$	$-$

A predikciók kombinálása után az 1. osztály négy szavazatot kap, míg az összes többi osztály csak két szavazatot. A tesztvektort ezért az 1. osztályhoz tartozóként osztályozzuk.

Többosztályos osztályozási feladat megfogalmazása feltételes szélsőértékszámítási feladatként

Koby Crammer and Yoram Singer. **?**On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines**?** In: *Journal of Machine Learning Research* 2 (2001), pp. 265–292. URL: http://www.ai.mit.edu/projects/jmlr/papers/v2/crammer01a.html

Regressziós feladat

Legyenek adottak $(\mathbf{x}_i, y_i)$ elempárok, ahol $\mathbf{x}_i \in \mathbb{R}^n$ és $y_i \in \mathbb{R}$ $(i = 1, \ldots, N)$.

Az adatok egy vagy több független változóra $(\mathbf{x})$ és egy függő változóra (y) vonatkozó megfigyelések.

Feladatunk az $\mathbf{x}$ és y közötti összefüggés modellezése egy, az adatok alapján alkotott $f(\mathbf{x})$ regressziós függvénnyel.

Nemlineáris regresszió tartóvektorgéppel

Használjuk az alábbi regressziós modellt:

$$\begin{aligned} y &= f(\mathbf{x}) \\ &= \sum_{i=1}^{m} w_i \phi_i(\mathbf{x}) + b = \mathbf{w}^\top \mathbf{\Phi}(\mathbf{x}) + b, \end{aligned}$$

ahol

$$\mathbf{\Phi}(\mathbf{x}) = (\phi_1(\mathbf{x}), \ldots, \phi_m(\mathbf{x}))$$

és

$$\mathbf{w} = (w_1, \ldots, w_m)$$

b pedig a konstans vagy torzítás paraméter, melyet beolvaszthatunk $\mathbf{w}$-be ha bevezetjük a $\phi_0(\mathbf{x}) = 0$ bázisfüggvényt.

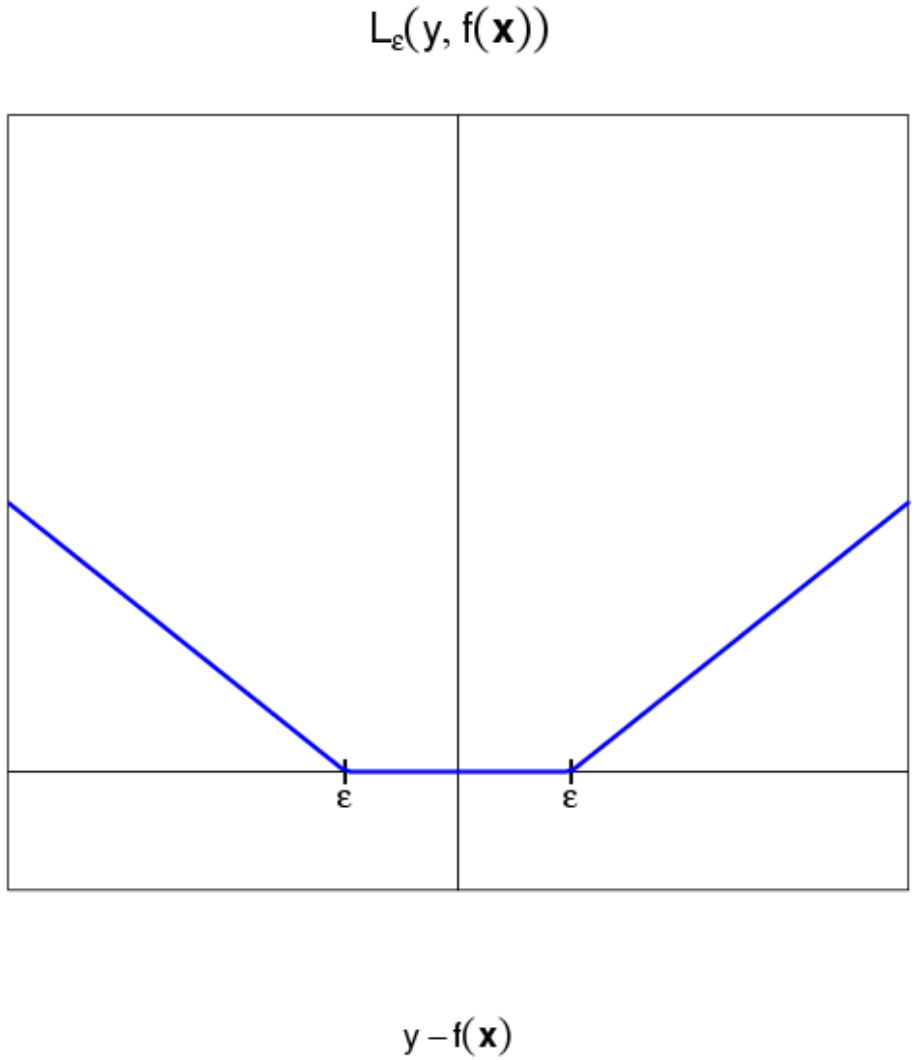
ϵ-érzéketlen veszteségfüggvény (1)

Használjuk az alábbi, úgynevezett ϵ-érzéketlen veszteségfüggvényt a hiba méréséhez:

$$L_\epsilon(y, f(\mathbf{x})) = \begin{cases} |y - f(\mathbf{x})| - \epsilon, & \text{ha } |y - f(\mathbf{x})| \geq \epsilon, \\ 0, & \text{egyébként,} \end{cases}$$

ahol ϵ egy, a felhasználó által rögzített nemnegatív értékű paraméter.

ϵ-érzéketlen veszteségfüggvény (2)



A feladat megfogalmazása optimalizálási problémaként

Feltételes optimalizálási probléma:

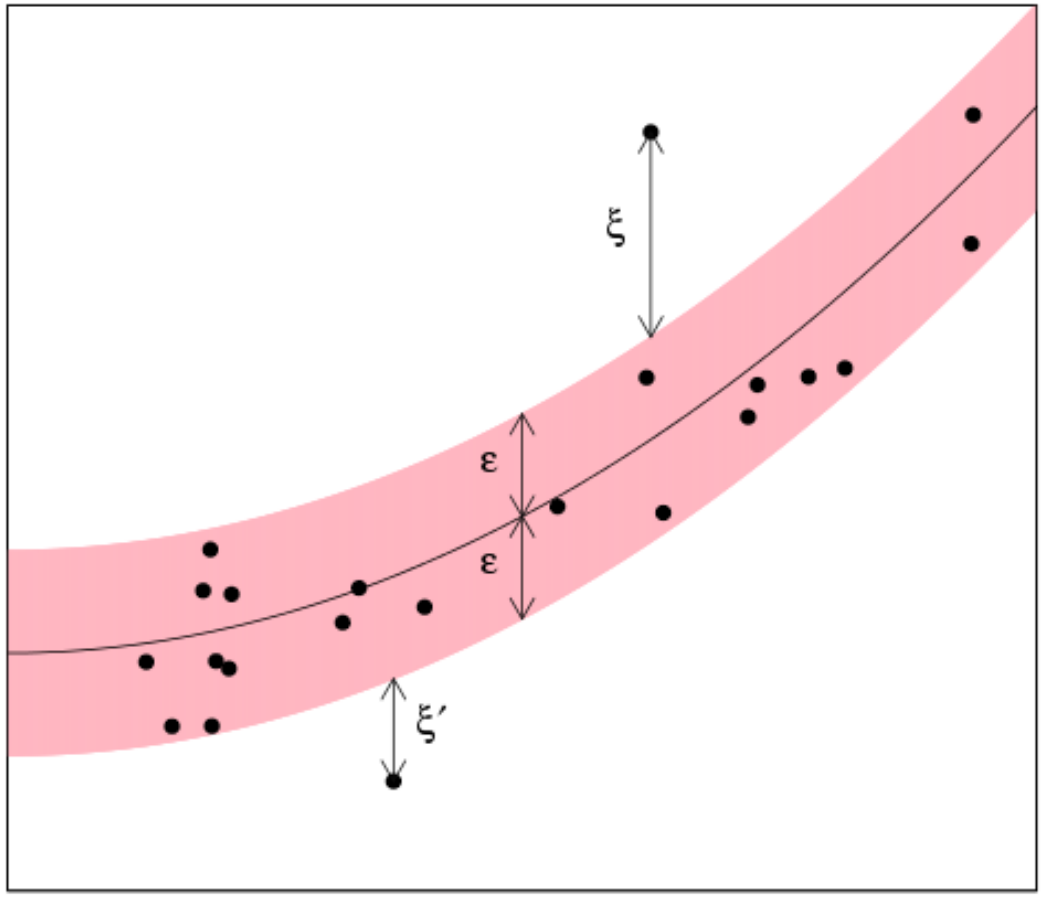
$$\min_{\mathbf{w}} \left(\frac{\|\mathbf{w}\|^2}{2} + C \sum_{i=1}^{N} (\xi_i + \xi_i') \right),$$

feltéve, hogy

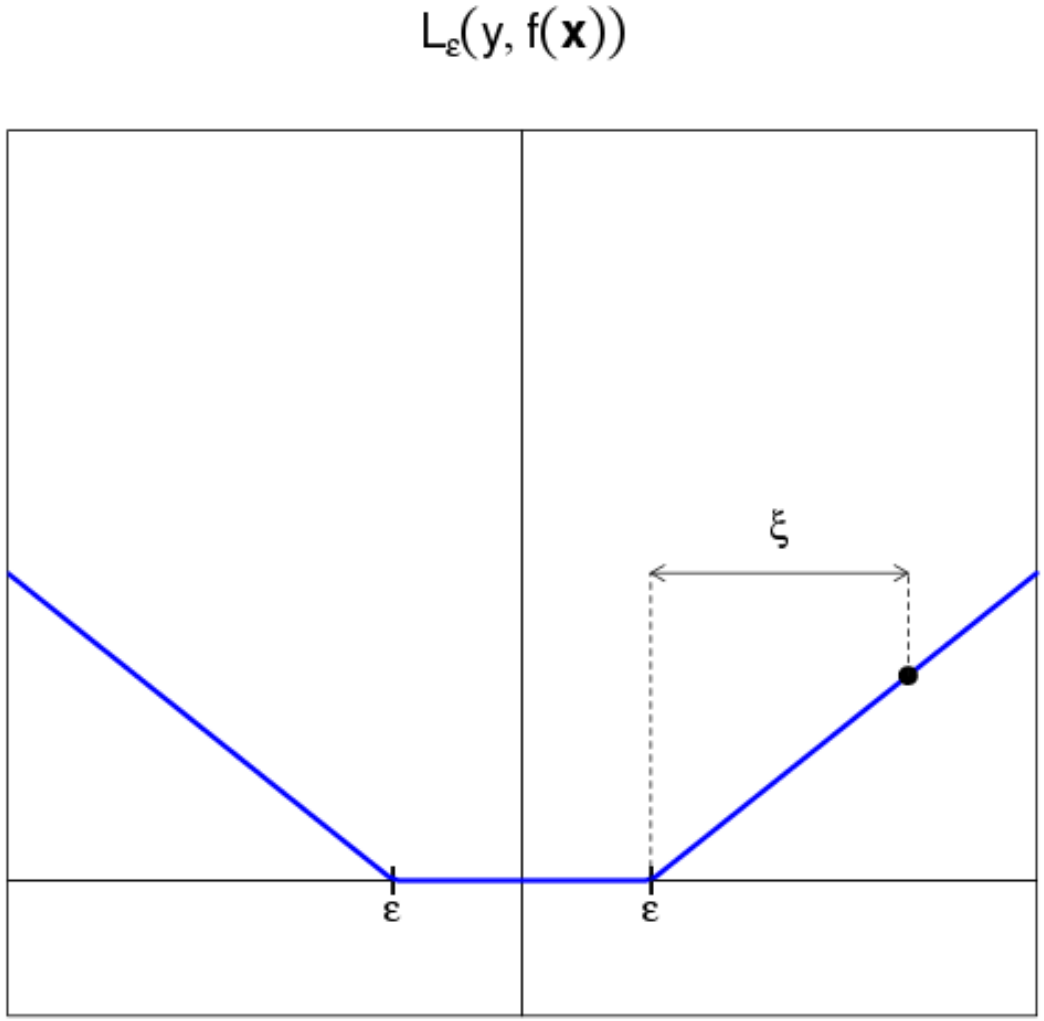
$$\begin{aligned} y_i - \mathbf{w}^\top \mathbf{\Phi}(\mathbf{x}_i) - b &\leq \epsilon + \xi_i, \qquad i = 1, \ldots, N, \\ \mathbf{w}^\top \mathbf{\Phi}(\mathbf{x}_i) + b - y_i &\leq \epsilon + \xi_i', \qquad i = 1, \ldots, N, \\ \xi_i &\geq 0, \qquad i = 1, \ldots, N, \\ \xi_i' &\geq 0, \qquad i = 1, \ldots, N, \end{aligned}$$

ahol a ξ_i-k és ξ_i'-k kiegészítő változók, C pedig egy, a felhasználó által meghatározott paraméter.

A kiegészítő változók jelentése (1)



A kiegészítő változók jelentése (2)



Az optimalizálási probléma megoldása (1)

A probléma primál Lagrange-függvénye:

$$
\begin{aligned}
L_P = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i') \\
- \sum_{i=1}^{N}\lambda_i\Big(\mathbf{w}^\top\boldsymbol{\Phi}(\mathbf{x}_i) + b - y_i + \epsilon + \xi_i\Big) \\
- \sum_{i=1}^{N}\lambda_i'\Big(y_i - \mathbf{w}^\top\boldsymbol{\Phi}(\mathbf{x}_i) - b + \epsilon + \xi_i'\Big) \\
- \sum_{i=1}^{N}(\gamma_i\xi_i + \gamma_i'\xi_i'),
\end{aligned}
$$

ahol a λ_i, λ_i', γ_i és γ_i' paraméterek Lagrange-multiplikátorok, amelyekre $\lambda_i \geq 0$, $\lambda_i' \geq 0$, $\gamma_i \geq 0$ és $\gamma_i' \geq 0$ minden $i = 1, \ldots, N$ esetén.

Az optimalizálási probléma megoldása (2)

A cél az L_P primál Lagrange-függvény értékét minimalizálni a $\mathbf{w}$ és b paraméterek, valamint a ξ_i és ξ_i' kiegészítő változók szerint, ugyanakkor maximalizálni a λ_i, λ_i', γ és γ_i' Lagrange-multiplikátorok szerint.

Az optimalizálási probléma megoldása (3)

A Lagrange-függvény minimalizálásához vennünk kell és nullává kell tegyük L_P $\mathbf{w}$, b, ξ_i és ξ_i' szerinti deriváltját:

$$\frac{\partial L_P}{\partial \mathbf{w}} = 0 \Longrightarrow \mathbf{w} = \sum_{i=1}^{N} (\lambda_i - \lambda_i') \mathbf{\Phi}(\mathbf{x}_i)$$

$$\frac{\partial L_P}{\partial b} = 0 \Longrightarrow \sum_{i=1}^{N} (\lambda_i - \lambda_i') = 0$$

$$\frac{\partial L_P}{\partial \xi_i} = 0 \Longrightarrow \gamma_i = C - \lambda_i$$

$$\frac{\partial L_P}{\partial \xi_i'} = 0 \Longrightarrow \gamma_i' = C - \lambda_i'$$

Az optimalizálási probléma megoldása (4)

A duális Lagrange-függvény:

$$L_D = \sum_{i=1}^{N} y_i(\lambda_i - \lambda_i') - \epsilon \sum_{i=1}^{N}(\lambda_i + \lambda_i') - \frac{1}{2}\sum_{i,j}(\lambda_i - \lambda_i')(\lambda_j - \lambda_j')K(\mathbf{x}_i, \mathbf{x}_j)$$

A korábbiakhoz hasonlóan L_D maximalizálása a feladat, feltéve, hogy teljesülnek az alábbi feltételek:

1. $\sum_{i=1}^{N}(\lambda_i - \lambda_i') = 0$,
2. $0 \leq \lambda_i \leq C$ minden $i = 1, \ldots N$ esetén,
   $0 \leq \lambda_i' \leq C$ minden $i = 1, \ldots N$ esetén,

ahol C egy, a felhasználó által meghatározott paraméter.

Az optimalizálási probléma megoldása (5)

A számolás részletezése:

$$
\begin{aligned}
L_P =& \frac{1}{2}\|\mathbf{w}\|^2 - \mathbf{w}^\top \underbrace{\sum_{i=1}^{N}(\lambda_i - \lambda_i')\mathbf{\Phi}(\mathbf{x}_i)}_{=\mathbf{w}} - b\underbrace{\sum_{i=1}^{N}(\lambda_i - \lambda_i')}_{=0} \\
&+ \sum_{i=1}^{N}\underbrace{(C - \lambda_i - \gamma_i)}_{=0}\xi_i + \sum_{i=1}^{N}\underbrace{(C - \lambda_i' - \gamma_i')}_{=0}\xi_i' \\
&+ \sum_{i=1}^{N} y_i(\lambda_i - \lambda_i') - \epsilon\sum_{i=1}^{N}(\lambda_i + \lambda_i') \\
=& \sum_{i=1}^{N} y_i(\lambda_i - \lambda_i') - \epsilon\sum_{i=1}^{N}(\lambda_i + \lambda_i') - \frac{1}{2}\|\mathbf{w}\|^2 = L_D
\end{aligned}
$$

Az optimalizálási probléma megoldása (6)

Teljesülnek az alábbi KKT feltételek:

$$\begin{aligned}
\lambda_i\left(\mathbf{w}^\top \mathbf{\Phi}(\mathbf{x}_i) + b - y_i + \epsilon + \xi_i\right) &= 0, \qquad i = 1, \ldots, N, \\
\lambda_i'\left(y_i - \mathbf{w}^\top \mathbf{\Phi}(\mathbf{x}_i) - b + \epsilon + \xi_i'\right) &= 0, \qquad i = 1, \ldots, N, \\
(C - \lambda_i)\xi_i &= 0, \qquad i = 1, \ldots, N, \\
(C - \lambda_i')\xi_i' &= 0, \qquad i = 1, \ldots, N.
\end{aligned}$$

Az első és harmadik egyenlet alapján ha $0 < \lambda_i < C$ akkor $\xi_i = 0$ miatt

$$y_i - \mathbf{w}^\top \mathbf{\Phi}(\mathbf{x}_i) - b = \epsilon$$

Hasonlóan ha $0 < \lambda_i' < C$ akkor $\xi_i' = 0$ miatt

$$y_i - \mathbf{w}^\top \mathbf{\Phi}(\mathbf{x}_i) - b = -\epsilon$$

A két egyenlet egyszerre nem állhat fenn. Így $\lambda_i, \lambda_i' > 0$ esetén valamelyik, esetleg mindkettő Lagrange-multiplikátor C-vel egyenlő.

Az optimális regressziós paraméterek (1)

Azokat az adatvektorokat nevezzük tartóvektoroknak, amelyekre $\lambda_i \neq \lambda_i'$ teljesül, így egy tartóvektorra pontosan az egyik Lagrange-multiplikátor egyenló C-vel.
Jelöljük az tartóvektorokhoz tartozó indexek halmazát ismét $\mathcal{S}$-sel.
Az optimális regressziós függvény paraméterei:

$$\mathbf{w}_{opt} = \sum_{i=1}^{N} (\lambda_i^{opt} - \lambda_i'^{opt}) \mathbf{\Phi}(\mathbf{x}_i) = \sum_{s \in \mathcal{S}} (\lambda_s^{opt} - \lambda_s'^{opt}) \mathbf{\Phi}(\mathbf{x}_s)$$

míg a konstans becslését az alábbi formulából kapjuk véve egy $s \in \mathcal{S}$ tartóvektort

$$b_{opt} = y_s - \mathbf{w}_{opt}^{\top} \mathbf{\Phi}(\mathbf{x}_s) - \delta_s \epsilon$$

ahol $\delta_s = +1$ ha $\lambda_i' = C$ $(\xi_i = 0)$ és $\delta_s = -1$ ha $\lambda_i = C$ $(\xi_i' = 0)$.

Az optimális regressziós paraméterek (2)

Most is érdemes robusztusabb becslésért átlagot venni:

$$
\begin{aligned}
b_{opt} =& \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} (y_s - \mathbf{w}_{opt}^\top \mathbf{\Phi}(\mathbf{x}_s) - \delta_s \epsilon) \\
=& \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \left(y_s - \sum_{s' \in \mathcal{S}} (\lambda_{s'}^{opt} - \lambda_{s'}'^{opt}) \mathbf{\Phi}(\mathbf{x}_{s'})^\top \mathbf{\Phi}(\mathbf{x}_s) - \delta_s \epsilon \right) \\
=& \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \left(y_s - \sum_{s' \in \mathcal{S}} (\lambda_{s'}^{opt} - \lambda_{s'}'^{opt}) K(\mathbf{x}_{s'}, \mathbf{x}_s) - \delta_s \epsilon \right)
\end{aligned}
$$

Prediktálás

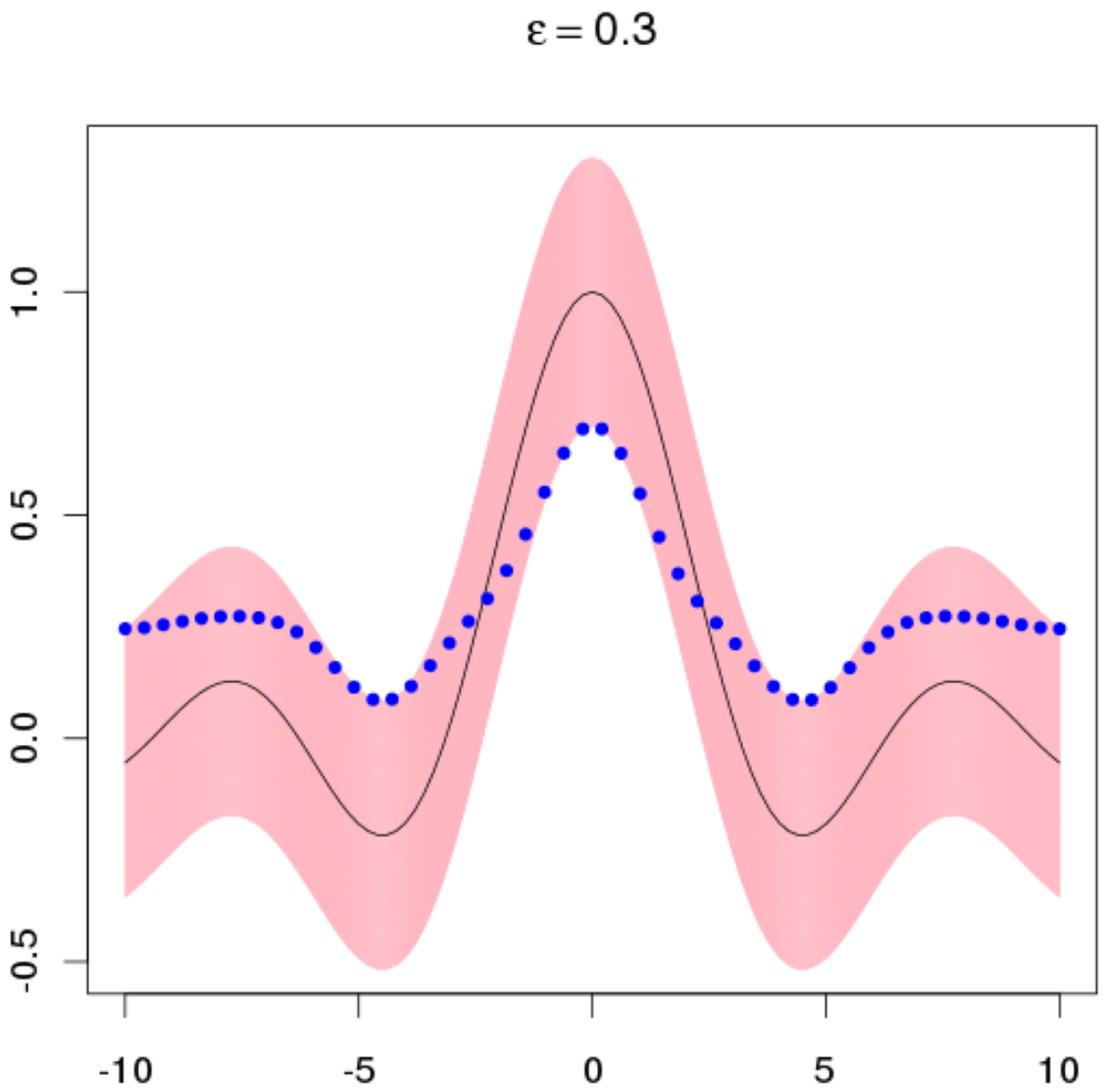
Egy $\mathbf{x} \in \mathbb{R}^n$ vektorhoz tartozó regressziós függvényérték

$$\begin{aligned} f(\mathbf{x}) &= \mathbf{w}_{opt}^{\top} \mathbf{\Phi}(\mathbf{x}) + b_{opt} \\ &= \sum_{s \in \mathcal{S}} (\lambda_s^{opt} - \lambda_s'^{opt}) K(\mathbf{x}, \mathbf{x}_s) + b_{opt} \end{aligned}$$

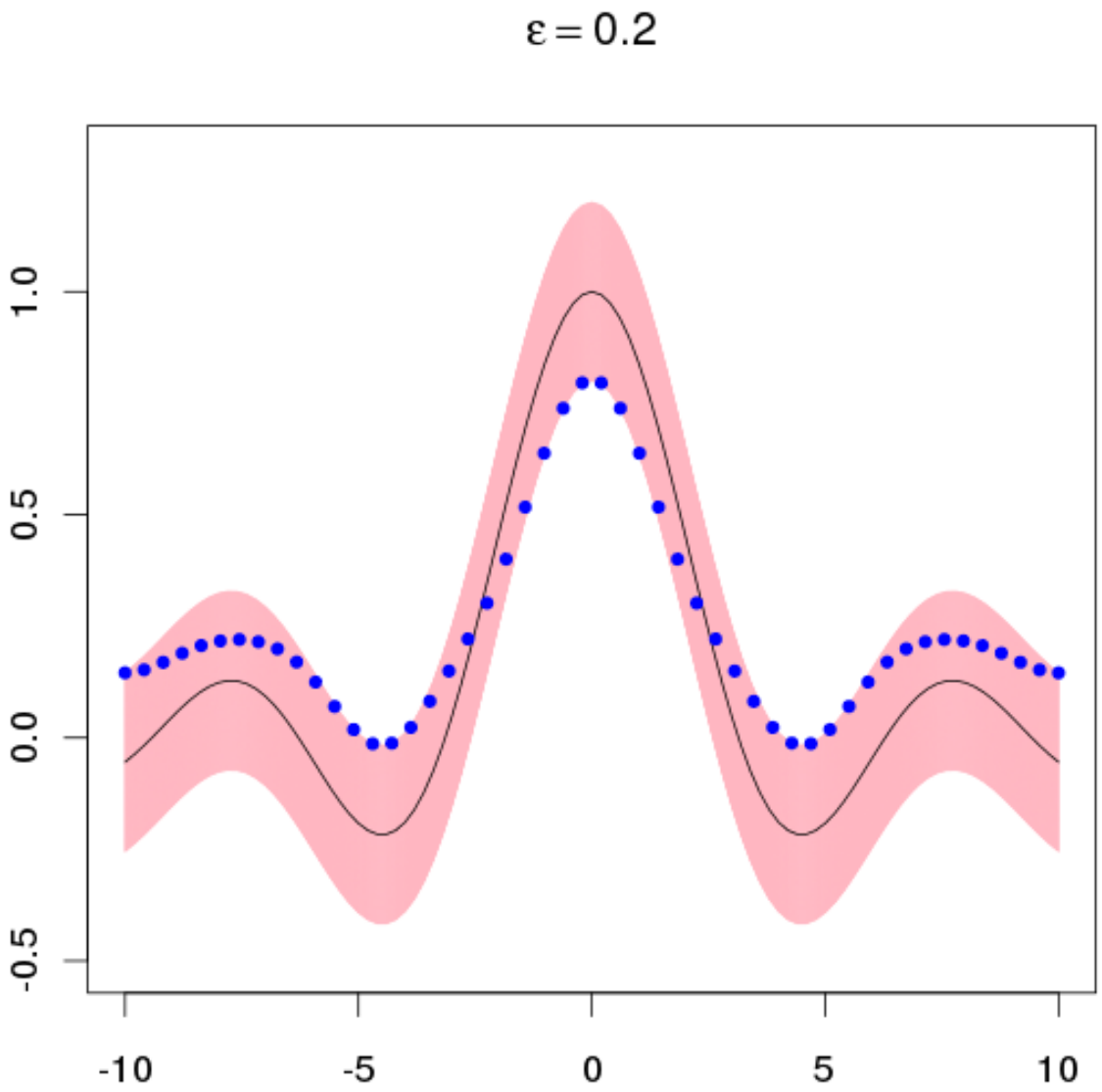
Megjegyzés

A predikció most is csak a kernel függvényen keresztül függ a tartóvektoroktól!

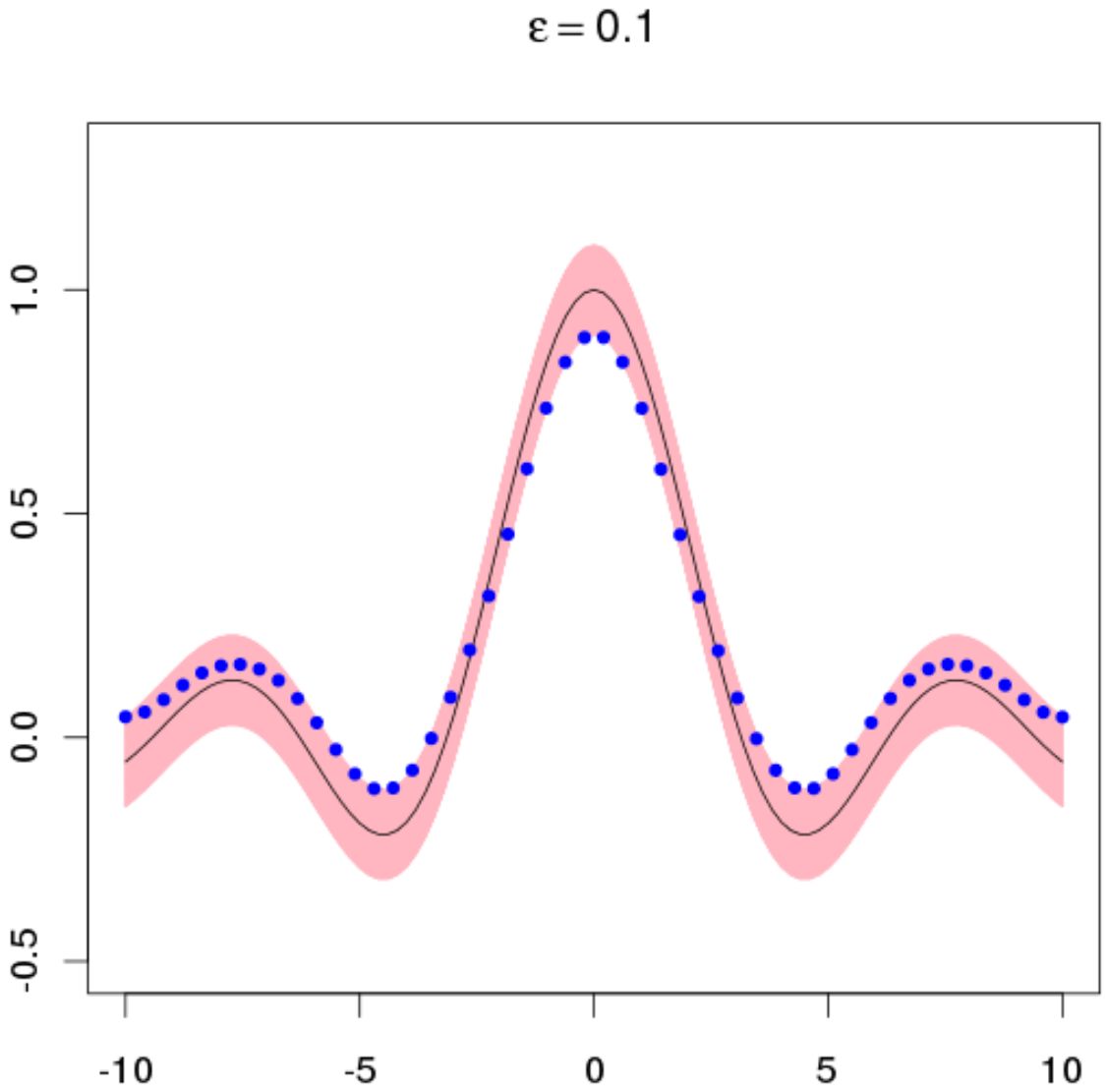
Az ϵ paraméter hatásának szemléltetése (1)



Az ϵ paraméter hatásának szemléltetése (2)



Az ϵ paraméter hatásának szemléltetése (3)



Felhasznált irodalom

- Simon Haykin. *Neural Networks and Learning Machines*. 3rd ed. Prentice Hall, 2008. ISBN: 9780131471399
- Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Addison Wesley, 2005. ISBN: 9780321321367. URL: http://www-users.cs.umn.edu/~kumar/dmbook/
- Alex J. Smola and Bernhard Schölkopf. **?**A Tutorial on Support Vector Regression**?** In: *Statistics and Computing* 14.3 (2004), pp. 199–222. URL: http://alex.smola.org/papers/2004/SmoSch04.pdf