

Üzleti Elemzések Módszertana

6. Előadás: Tartó vektor gépek

Kuknyó Dániel
Budapesti Gazdasági Egyetem

2023/24
2.félév

1 SVM modellek

2 Nemlineáris SVM

3 SVM regresszió

4 Az SVM matematikai alapjai

1 SVM modellek

2 Nemlineáris SVM

3 SVM regresszió

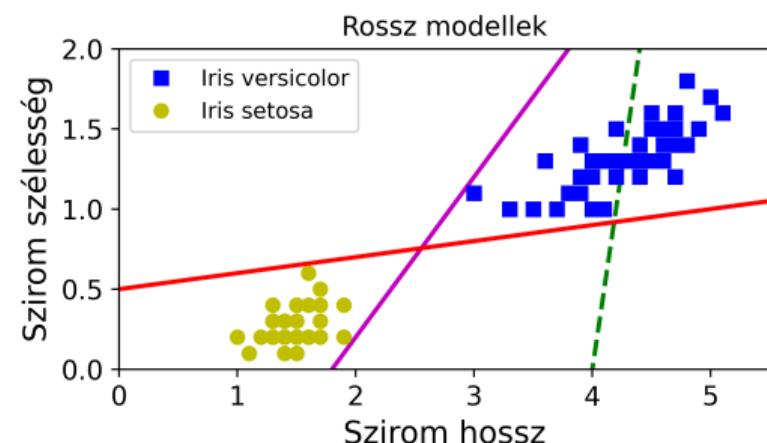
4 Az SVM matematikai alapjai

A vektorgépek mögötti intuíció

Az SVM modellek minden esetben a **legszélesebb utat** keresik egy adathalmaz két osztálya között. Az út közepén húzódik a **döntési határ** és az út két oldalán fekszenek a **margók**.

Az úton kívül hozzáadott mintaegyedek nem befolyásolják a döntési határt, tehát az utat a csoportok szélén lévő mintaegyedek tartják.

A margó egyik oldalán A , a másikon pedig B osztály lesz a predikció.

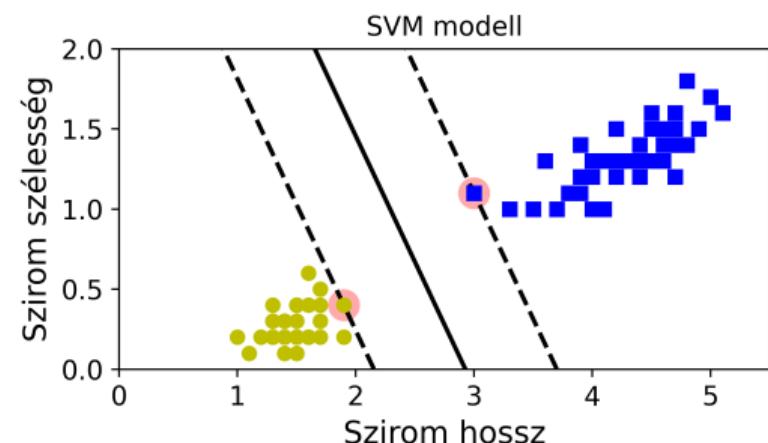


A vektorgépek mögötti intuíció

Az SVM modellek minden esetben a **legszélesebb utat** keresik egy adathalmaz két osztálya között. Az út közepén húzódik a **döntési határ** és az út két oldalán fekszenek a **margók**.

Az úton kívül hozzáadott mintaegyedek nem befolyásolják a döntési határt, tehát az utat a csoportok szélén lévő mintaegyedek tartják.

A margó egyik oldalán A , a másikon pedig B osztály lesz a predikció.



Adat előkészítés

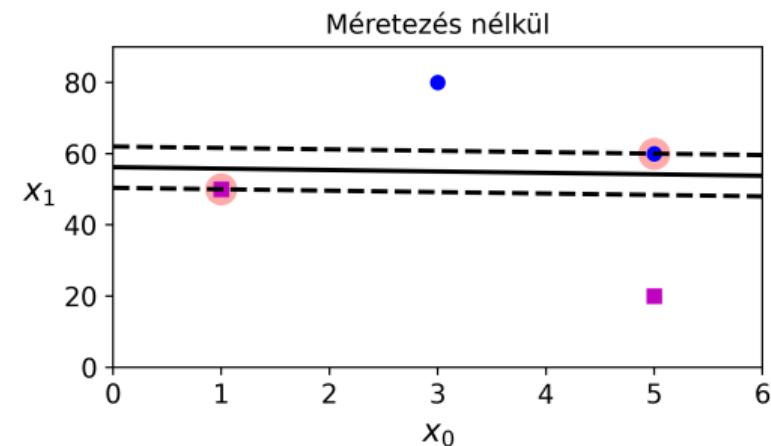
A tartó vektor gépek nagyon szenzitívek az adatok méretezésére.

Normalizálás

A normalizálás (vagy méretezés) célja, hogy az adatkészlet különböző változót egységes mértékrendszerbe hozza.

Ennek következménye, hogy a predikcióhoz minden változó egyenlő súlyjal járul hozzá.

Érdemes megfigyelni az út változását méretezés előtt és után.



Adat előkészítés

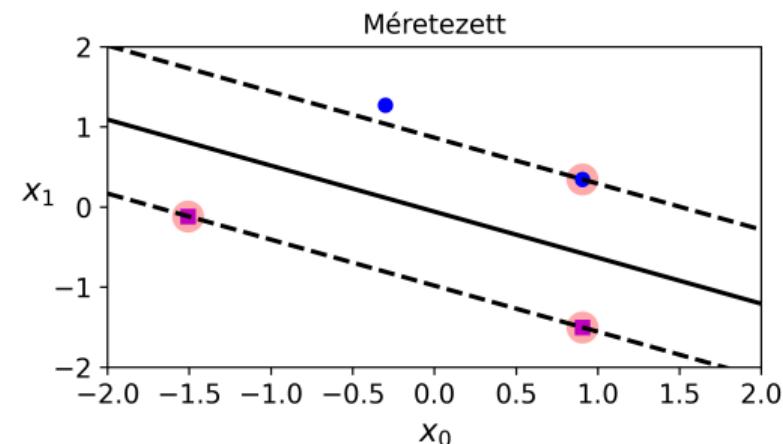
A tartó vektor gépek nagyon szenzitívek az adatok méretezésére.

Normalizálás

A normalizálás (vagy méretezés) célja, hogy az adatkészlet különböző változót egységes mértékrendszerbe hozza.

Ennek következménye, hogy a predikcióhoz minden változó egyenlő súlyjal járul hozzá.

Érdemes megfigyelni az út változását méretezés előtt és után.



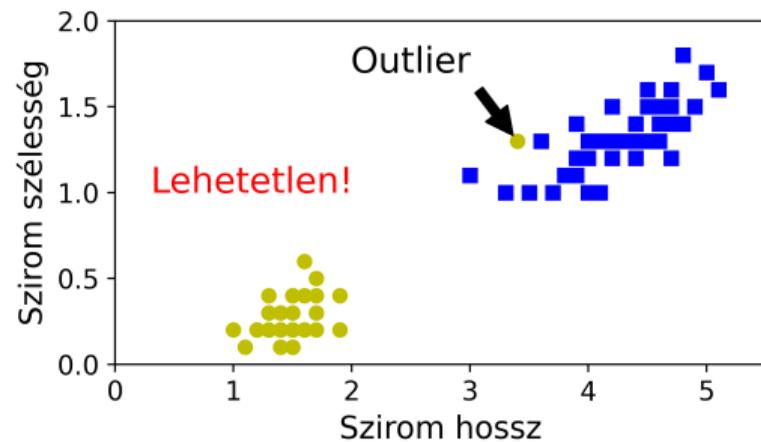
Keménymargós osztályozás

Keménymargós osztályozás

Amennyiben áll az a feltétel, hogy minden mintaegyednek az úton kívül kell esnie, az osztályozás **keménymargós**.

Ez csak akkor lehetséges, amikor az adatok lineárisan szeparálhatóak. Ezenkívül a kiugró adatpontok képesek irracionálisan torzítani a modell határait.

Ennek kiküszöbölésére olyan modellre van szükség, amelyik engedélyezi a **margósértéseket**.



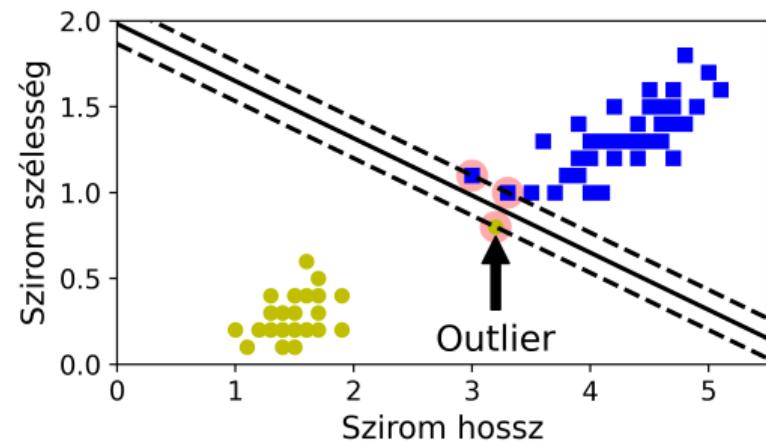
Keménymargós osztályozás

Keménymargós osztályozás

Amennyiben áll az a feltétel, hogy minden mintaegyednek az úton kívül kell esnie, az osztályozás **keménymargós**.

Ez csak akkor lehetséges, amikor az adatok lineárisan szeparálhatóak. Ezenkívül a kiugró adatpontok képesek irracionálisan torzítani a modell határait.

Ennek kiküszöbölésére olyan modellre van szükség, amelyik engedélyezi a margósértéseket.



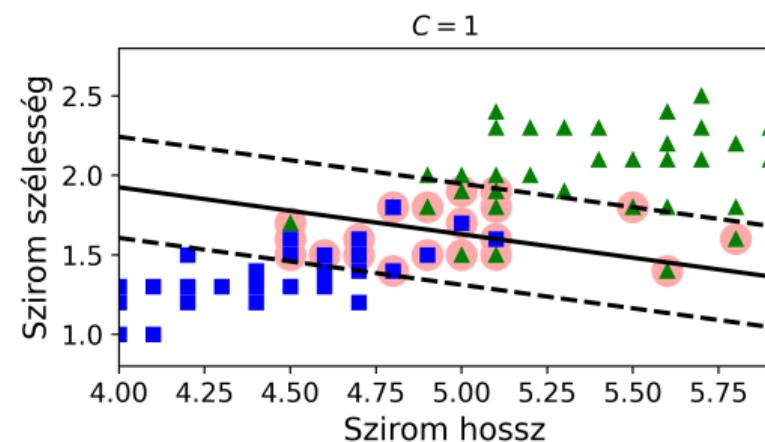
Lágymargós osztályozás

A margósértések és a rosszul generalizáló modellek közötti optimum keresésére hivatottak a lágymargós SVM modellek.

Lágymargós osztályozás

Olyan SVM modell, amely engedélyezi a margósértéseket. A margó keménységét a C hiperparaméter szabályozza.

A kisebb C érték szélesebb úthoz, de több margósértéshez vezet. A nagyobb C érték pedig pedig kevesebb margósértést, de rosszabbul generalizáló modellt eredményez.



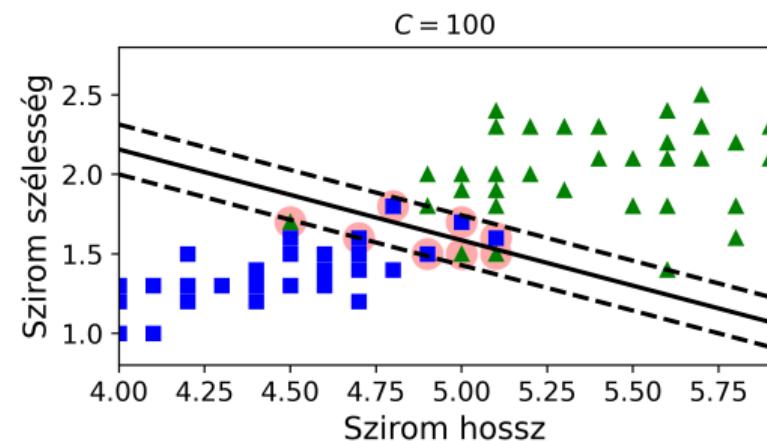
Lágymargós osztályozás

A margósértések és a rosszul generalizáló modellek közötti optimum keresésére hivatottak a lágymargós SVM modellek.

Lágymargós osztályozás

Olyan SVM modell, amely engedélyezi a margósértéseket. A margó keménységét a C hiperparaméter szabályozza.

A kisebb C érték szélesebb úthoz, de több margósértéshez vezet. A nagyobb C érték pedig pedig kevesebb margósértést, de rosszabbul generalizáló modellt eredményez.



1 SVM modellek

2 Nemlineáris SVM

3 SVM regresszió

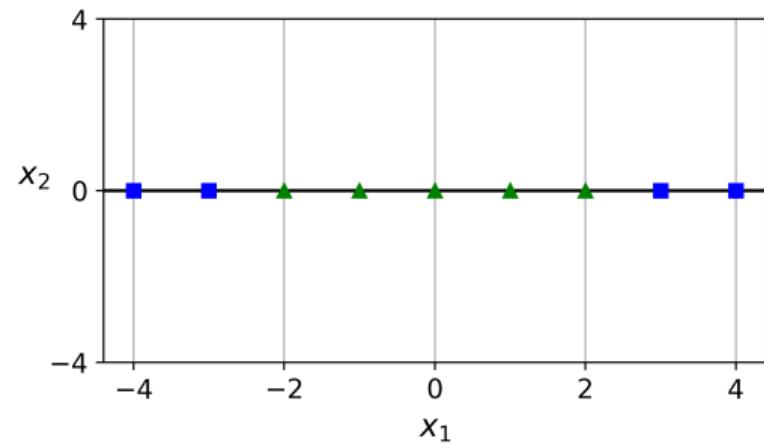
4 Az SVM matematikai alapjai

Nemlineáris SVM osztályozás

Habár az SVM modellek jól teljesítenek lineárisan szeparálható adatok esetén, a valóságban ezek az adathalmazok nagyon ritkának számítanak.

Egy módja a nemlineáris SVM osztályozásnak, ha a meglévő változókra **egy magasabb dimenziójú térben** történik az osztályozás.

Ebben az esetben a magasabb dimenziós tér a **négyzetes transzformációja** az x_1 változónak.

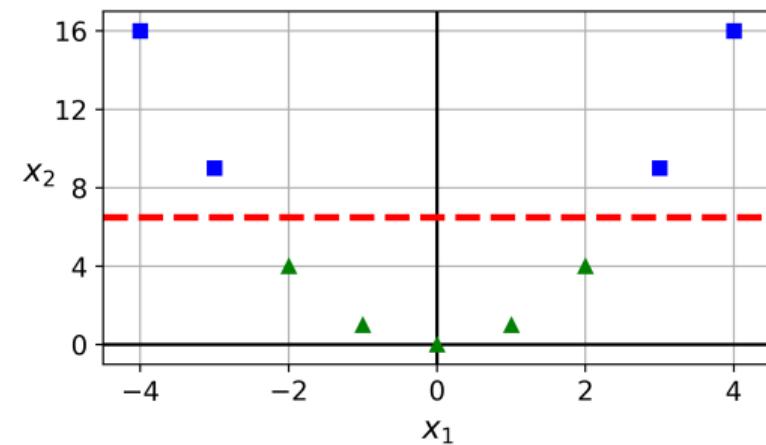


Nemlineáris SVM osztályozás

Habár az SVM modellek jól teljesítenek lineárisan szeparálható adatok esetén, a valóságban ezek az adathalmazok nagyon ritkának számítanak.

Egy módja a nemlineáris SVM osztályozásnak, ha a meglévő változókra **egy magasabb dimenziójú térben** történik az osztályozás.

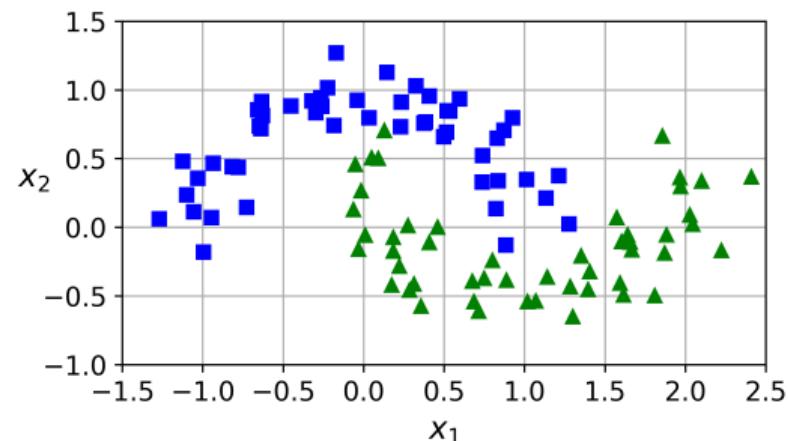
Ebben az esetben a magasabb dimenziós tér a **négyzetes transzformációja** az x_1 változónak.



Nemlineáris SVM a make-moons adathalmazon

A transzformációs csővezeték a következő:

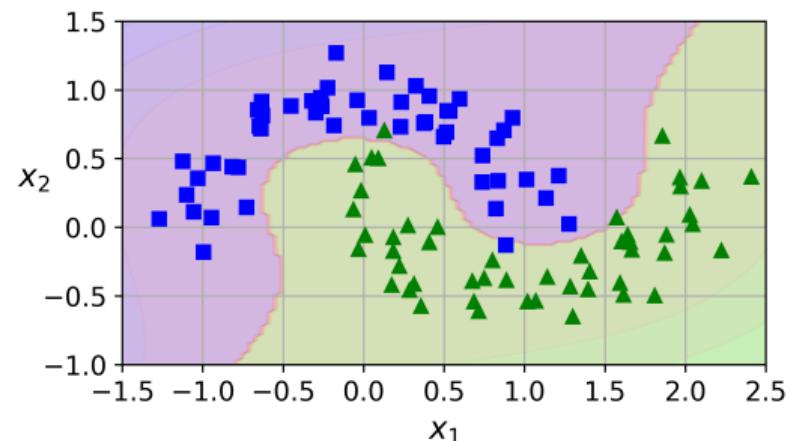
- ① Polinomiális jellemzők felvétele
- ② Jellemzők normalizálása
- ③ SVM osztályozó futtatása



Nemlineáris SVM a make-moons adathalmazon

A transzformációs csővezeték a következő:

- ① Polinomiális jellemzők felvétele
- ② Jellemzők normalizálása
- ③ SVM osztályozó futtatása



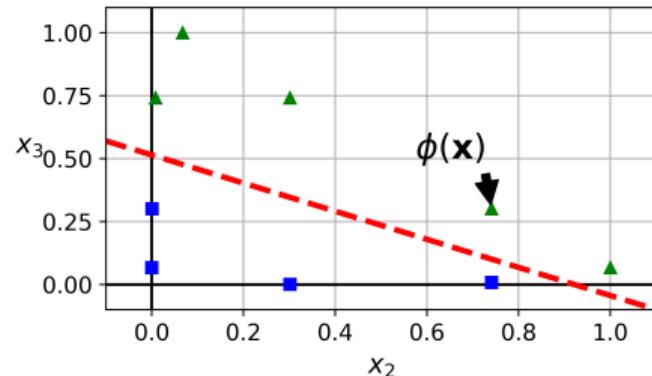
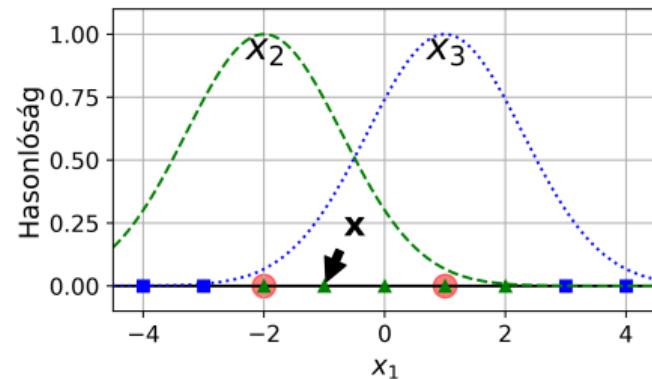
Hasonlósági függvények kernelként

Hasonlósági függvény

Azt reprezentálja, hogy egy adott lekérdezési pont mennyire hasonlít egy előre meghatározott **tájékozódási ponthoz**.

A példában az előző, 1D adathalmazhoz választott két tájékozódási pont $x_1 = -2$ és $x_2 = 1$. A hasonlósági függvény pedig a Gauss-i radiális bázis függvény $\gamma = 0.3$ paraméterrel:

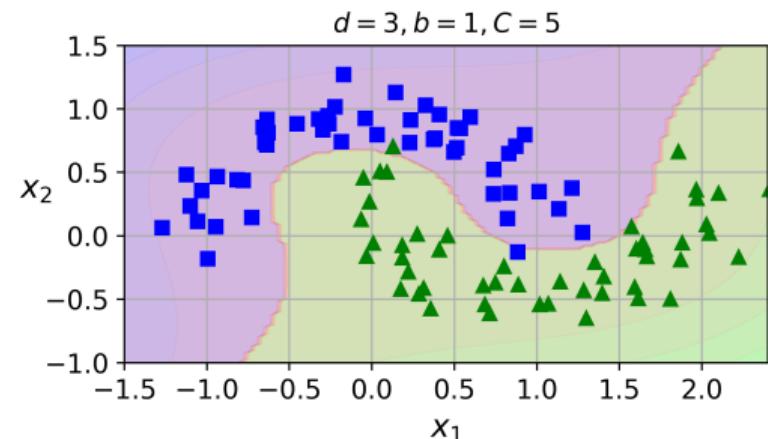
$$\phi_\gamma(x, \ell) = \exp(-\gamma |x - \ell|^2)$$



Polinomiális függvények kernelként

Polinomiális jellemzők hozzáadása alacsony polinomiális szinten nem képes komplex adathalmazokkal dolgozni.

Magas szinten viszont nagyon lassúvá teheti a modellt, mivel minden polinomiális szinten egy külön transzformációt jelent a jellemzőkre.



Polinomiális függvények kernelként

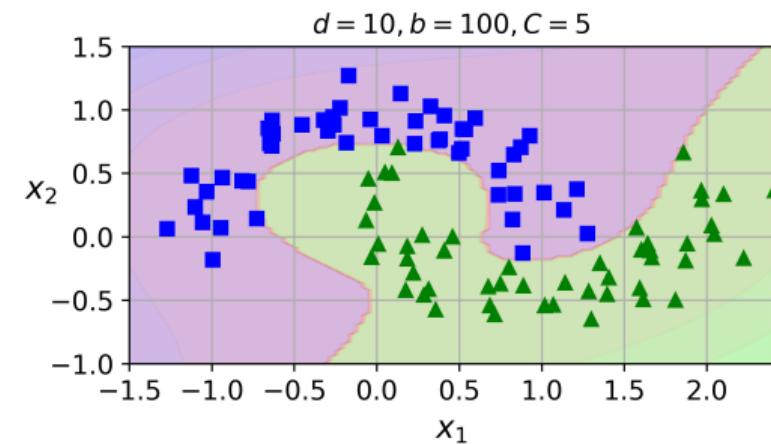
Polinomiális kernel

A polinomiális kernel transzformáció lineárisan nem szeparálható adathalmazokat transzformál olyan magasabb dimenziós térbe, ahol már lineárisan szeparálhatóvá válnak:

$$K(x, y) = ((x \cdot y) + b)^d$$

Ahol:

- b : A konstans torzítás
- d : A transzformáció polinomiális foka



A kernel trükk

Egy φ másodrendű polinomiális leképezés 2D tanító halmazra 3D eredményt ad:

$$\varphi(x) = \varphi\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right) = \begin{pmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{pmatrix}$$

Ezt a másodrendű leképezést alkalmazva, majd a vektorok belső szorzatait kiszámolva:

$$\begin{pmatrix} a_1^2 \\ \sqrt{2}a_1a_2 \\ a_2^2 \end{pmatrix}^T \begin{pmatrix} b_1^2 \\ \sqrt{2}b_1b_2 \\ b_2^2 \end{pmatrix} = a_1^2b_1^2 + 2a_1b_1a_2b_2 + a_2^2b_2^2 = \left(\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}^T \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \right)^2 = (a^T b)^2$$

Látható, hogy a vektorok belső szorzata megegyezik az eredeti vektorok belső szorzatának négyzetével:

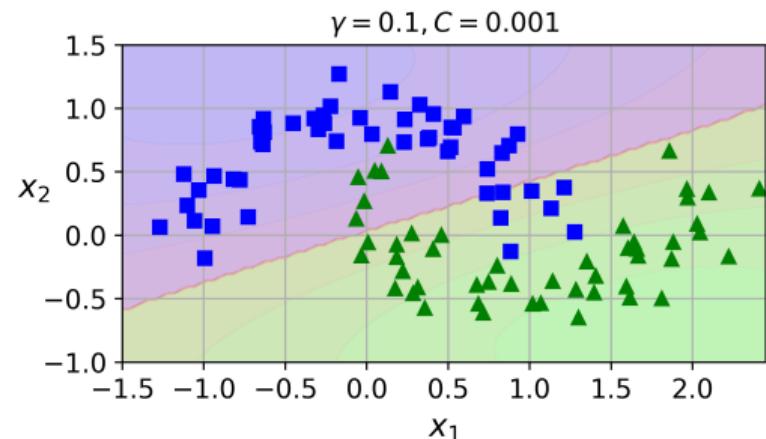
$$\varphi(a)^T \varphi(b) = (a^T b)^2$$

Radiális bázis függvények kernelként

Ahogy a polinomiális jellemzők esetében, úgy a **hasonlósági jellemzők** esetén is működik a kernel trükk.

A diagramon különböző γ és C értékekkel tanított modellek láthatóak.

A γ növelése szűkebb haranggörbékkel végzi az illesztést, ezért minden egyed hasonlósági tartománya kisebb. A γ csökkentése nagyobb befolyási teret ad a pontoknak.

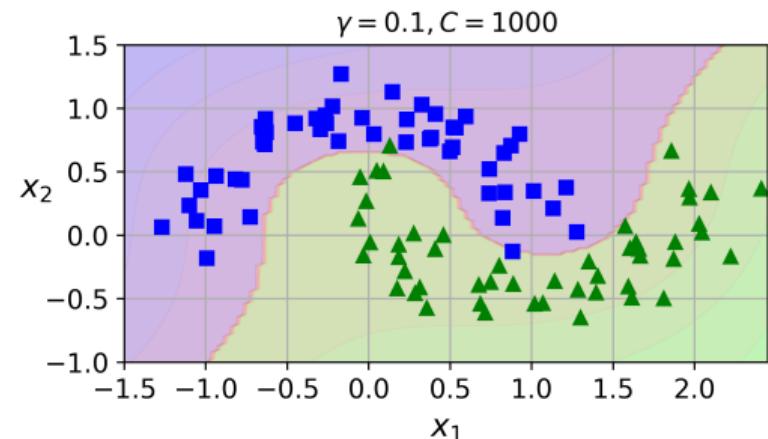


Radiális bázis függvények kernelként

Ahogy a polinomiális jellemzők esetében, úgy a **hasonlósági jellemzők** esetén is működik a kernel trükk.

A diagramon különböző γ és C értékekkel tanított modellek láthatóak.

A γ növelése szűkebb haranggörbékkel végzi az illesztést, ezért minden egyed hasonlósági tartománya kisebb. A γ csökkentése nagyobb befolyási teret ad a pontoknak.

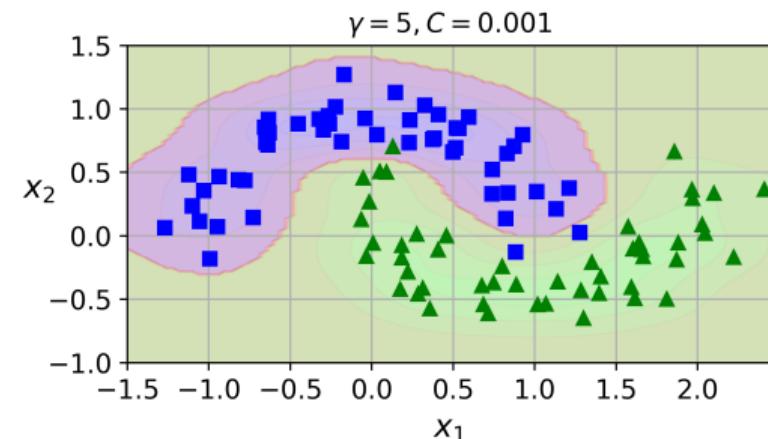


Radiális bázis függvények kernelként

Ahogy a polinomiális jellemzők esetében, úgy a hasonlósági jellemzők esetén is működik a kernel trükk.

A diagramon különböző γ és C értékekkel tanított modellek láthatóak.

A γ növelése szűkebb haranggörbékkel végzi az illesztést, ezért minden egyed hasonlósági tartománya kisebb. A γ csökkentése nagyobb befolyási teret ad a pontoknak.

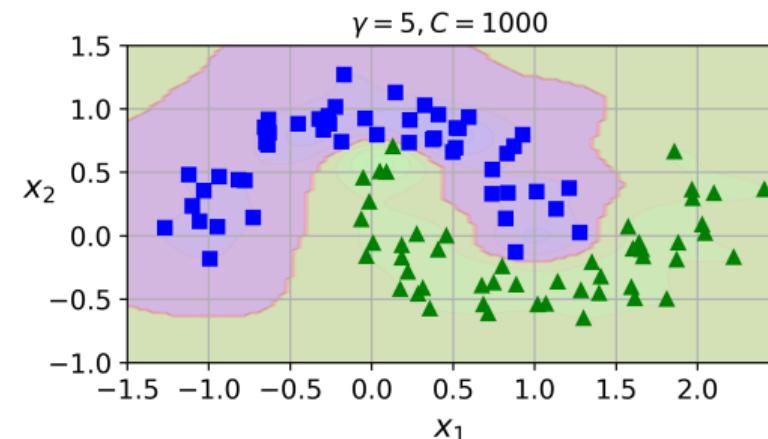


Radiális bázis függvények kernelként

Ahogy a polinomiális jellemzők esetében, úgy a hasonlósági jellemzők esetén is működik a kernel trükk.

A diagramon különböző γ és C értékekkel tanított modellek láthatóak.

A γ növelése szűkebb haranggörbékkel végzi az illesztést, ezért minden egyed hasonlósági tartománya kisebb. A γ csökkentése nagyobb befolyási teret ad a pontoknak.



1 SVM modellek

2 Nemlineáris SVM

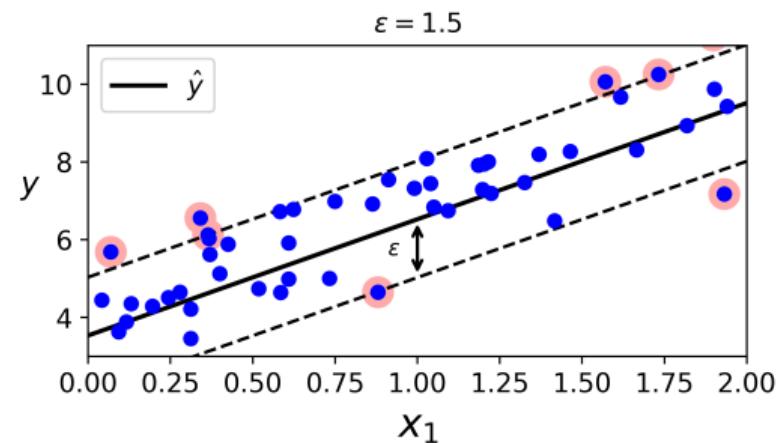
3 SVM regresszió

4 Az SVM matematikai alapjai

SVM regresszió

Az SVM regresszió esetében a cél ellentétes az osztályozáséval. A legnagyobb út helyett az algoritmus megpróbálja a lehető legtöbb egyedet az útra illeszteni a margósértések minimumon tartása mellett.

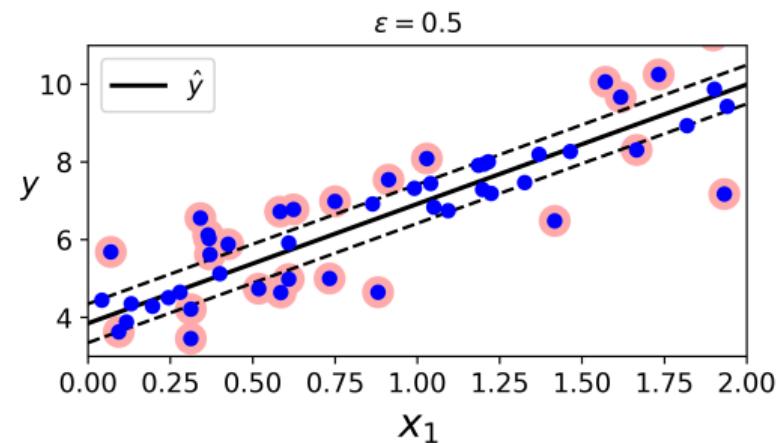
Az út szélességét a ε paraméter szabályozza. Nagyobb ε paraméter nagyobb margót jelent. A margón belül hozzáadott egyedek nem befolyásolják a modell definíciót. Ez az ε -érzéketlenség.



SVM regresszió

Az SVM regresszió esetében a cél ellentétes az osztályozáséval. A legnagyobb út helyett az algoritmus megpróbálja a lehető legtöbb egyedet az útra illeszteni a margósértések minimumon tartása mellett.

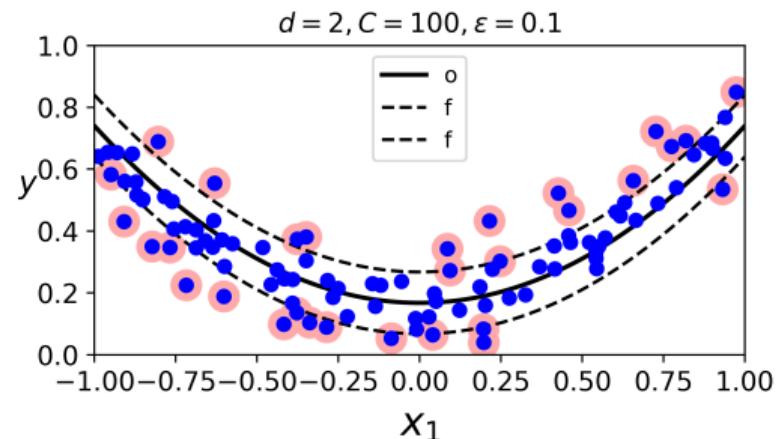
Az út szélességét a ε paraméter szabályozza. Nagyobb ε paraméter nagyobb margót jelent. A margón belül hozzáadott egyedek nem befolyásolják a modell definíciót. Ez az ε -érzéketlenség.



Nemlineáris SVM regresszió

Kernelizált SVM regresszorok használata
lehetséges regressziós problémák esetén is.

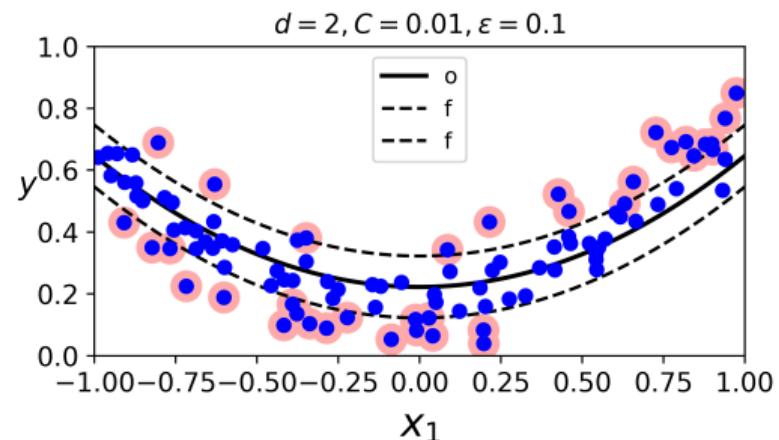
A következő diagramokon másodfokú
polinomiális kernellel tanított SVM
regresszorok láthatóak.



Nemlineáris SVM regresszió

Kernelizált SVM regresszorok használata
lehetséges regressziós problémák esetén is.

A következő diagramokon másodfokú
polinomiális kernellel tanított SVM
regresszorok láthatóak.



1 SVM modellek

2 Nemlineáris SVM

3 SVM regresszió

4 Az SVM matematikai alapjai

Az SVM működésének részletei

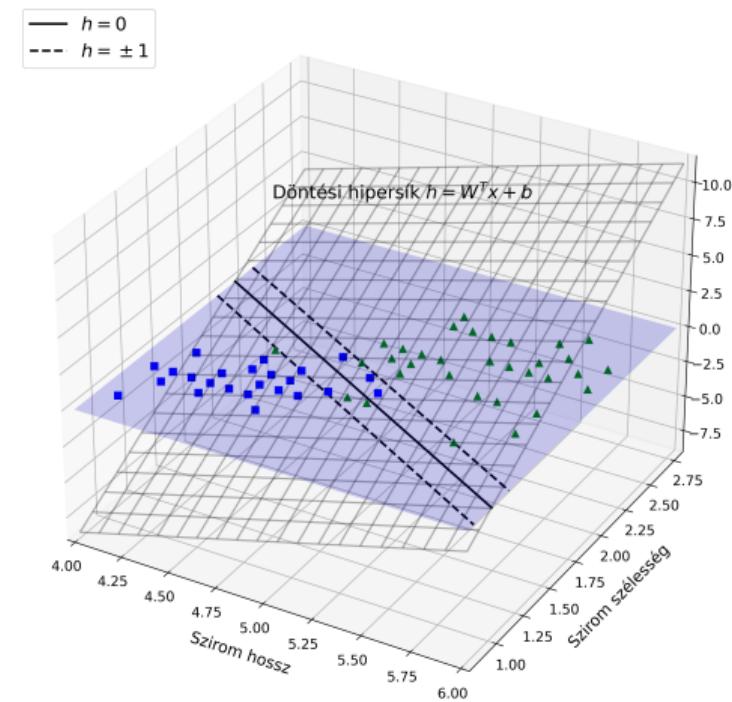
Az SVM modellek egy h döntési síkot keresnek a térben amelyre:

$$h = w^T x + b = w_1 x_1 + w_2 x_2 + \dots + w_n x_n$$

Attól függően, hogy a mintaegyed a hipersík melyik oldalára esik, lesz beosztályozva a 0, 1 osztályok valamelyikébe:

$$\hat{y} = \begin{cases} 0, & \text{ha } w^T x + b < 0 \\ 1, & \text{ha } w^T x + b \geq 0 \end{cases}$$

A döntési határ azon pontok halmaza, amelyre $h = 0$.



A keménymargós SVM célfüggvénye

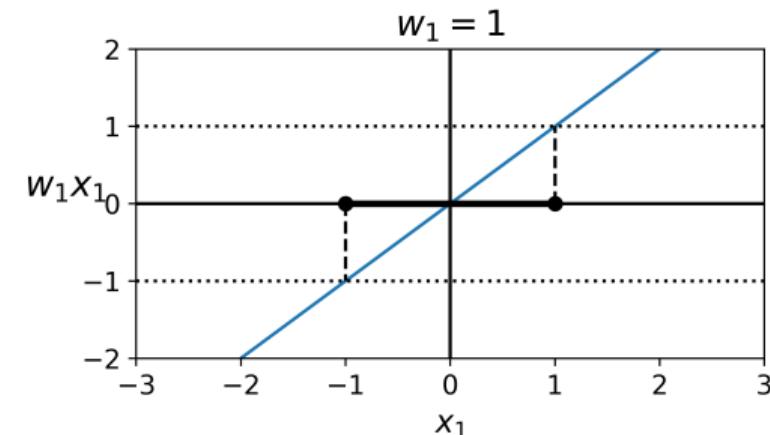
A döntési hipersík meredeksége a w vektor normáltja, $\|w\|$. Ez a meredekség fordítottan arányos a margó méretével: nagyobb meredekség szűkebb margót eredményez.

A célfüggvény a $\|w\|$ minimalizálására irányul annak érdekében, hogy a margó minél nagyobb legyen:

$$\min_{w,b} \frac{1}{2} \|w\|^2$$

A következő megkötésekkel:

$$y_i (w_i \cdot x_i + b) \geq 1 \quad \forall i$$



A keménymargós SVM célfüggvénye

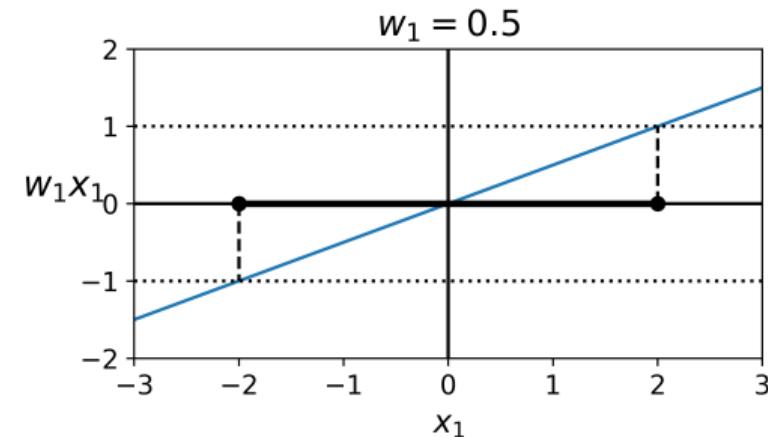
A döntési hipersík meredeksége a w vektor normáltja, $\|w\|$. Ez a meredekség fordítottan arányos a margó méretével: nagyobb meredekség szűkebb margót eredményez.

A célfüggvény a $\|w\|$ minimalizálására irányul annak érdekében, hogy a margó minél nagyobb legyen:

$$\min_{w,b} \frac{1}{2} \|w\|^2$$

A következő megkötésekkel:

$$y_i (w_i \cdot x_i + b) \geq 1 \quad \forall i$$



A lágymargós SVM tanításának célfüggvénye

A lágymargós SVM bevezet egy $\xi > 0$ kiegyenlítő változót amely azt adja meg, hogy egy adott mintaegyed mekkora mértékben sértheti a margót.

Ezáltal két egymással ellentétes cél jön létre a problémában. Az egyik, hogy a kiegyenlítő változót minimalizálni kell, a másik pedig, hogy az $\frac{1}{2}w^T w$ kifejezést minimalizálni, hogy a margó a lehető legnagyobb legyen.

A C hiperparaméter definiálhatóvá teszi az egyensúlyt a két egymással ellentétes cél között.

Ennek megfelelően a következő optimalizálási probléma áll elő:

$$\min_{w,b,\xi} \frac{1}{2}w^T w + C \sum_{i=1}^n \xi_i$$

A következő megkötésekkel:

$$y_i (w_i \cdot x_i + b) \geq 1 - \xi \quad \forall i$$

és

$$\xi_i > 0 \quad \forall i$$

Mercer tézise

A tézis szerint ha egy kernel $K(a, b)$ folytonos, és argumentumok szerint szimmetrikus, tehát $K(a, b) = K(b, a)$, akkor létezik egy olyan φ leképezés, amely a -t és b -t olyan magasabb dimenziós térbe képezi le, amelyre igaz, hogy $K(a, b) = \varphi(a)^T \varphi(b)$.

Ennek megfelelően egy K leképezést lehet használni kernelként, ha φ nem is ismert. Elég, ha annyi tudott, hogy φ létezik.

