Support vektor gépek (support vector machines)

A support vektor gépek az előrecsatolt osztályozó/regressziós hálózatok strukturális kockázat minimalizáláson (structural risk minimisation) alapuló változatai. Alapváltozatuk lineáris szeparálásra képes, amely azonban kiterjeszthető nemlineáris szeparálásra és nemlineáris regressziós feladatokra is. Ez az összefoglaló a lineáris szeparálásra alkalmas változat bemutatásával indul, majd röviden foglalkozik a további feladatokra alkalmas általánosításokkal is, de nem foglalkozik részletesen az elméleti háttérrel (method of structual risk minimisation, Vapnik-Chervonenkis (VC) dimension).

Kétosztályos lineáris osztályozási feladat:

Adottak az alábbi tanítópontok: $\{(\mathbf{x}_i, y_i)\}_{i=1}^{P}$, mely $\mathbf{x}_i$ bemeneti pontok két osztály elemei:

$\mathbf{x}_i \in X^1 \quad y_i = +1, \ \ \mathbf{x}_i \in X^2 \quad y_i = -1$

Keressük azt a szeparáló hipersíkot, melynek egyenlete: $\mathbf{w}^T\mathbf{x} + b = 0$

és amely a tanítópontokat hiba nélkül osztályozza, továbbá ahol a hipersíkhoz legközelebb álló tanítópontok távolsága maximális. A hipersík paramétereinek megfelelő skálázásával biztosítható, hogy:

$$\begin{aligned} &\mathbf{w}^T\mathbf{x}_i + b \geq +1, \text{ ha } \mathbf{x}_i \in X^1 \text{ és} \\ &\mathbf{w}^T\mathbf{x}_i + b \leq -1, \text{ ha } \mathbf{x}_i \in X^2 \end{aligned} \tag{1}$$

ami az alábbi formában is írható: $(\mathbf{w}^T\mathbf{x}_i + b)y_i \geq 1, \ \forall i - re.$

Ez azt jelenti, hogy a hipersíkhoz legközelebbi tanítópontokra : $\min\limits_{\mathbf{x}_i} \left|\mathbf{w}^T\mathbf{x}_i + b\right| = 1$

Egy **x** pont távolsága a **w** és *b* paraméterekkel megadott hiprersíktól: $d(\mathbf{w}, b, \mathbf{x}) = \dfrac{\left|\mathbf{w}^T\mathbf{x} + b\right|}{\|\mathbf{w}\|}$

Az optimális hipersík az alábbi tartalék (a hipersíkhoz legközelebbi tanítópontok távolsága) maximumát biztosítja :

$$\begin{aligned} \rho(\mathbf{w}, b) &= \min_{\{\mathbf{x}_i; y_i = 1\}} d(\mathbf{w}, b, \mathbf{x}_i) + \min_{\{\mathbf{x}_i; y_i = -1\}} d(\mathbf{w}, b, \mathbf{x}_i) = \\ &= \min_{\{\mathbf{x}_i; y_i = 1\}} \frac{\left|\mathbf{w}^T\mathbf{x}_i + b\right|}{\|\mathbf{w}\|} + \min_{\{\mathbf{x}_i; y_i = -1\}} \frac{\left|\mathbf{w}^T\mathbf{x}_i + b\right|}{\|\mathbf{w}\|} \\ &= \frac{1}{\|\mathbf{w}\|}\Big[\min_{\{\mathbf{x}_i; y_i = 1\}} \left|\mathbf{w}^T\mathbf{x}_i + b\right| + \min_{\{\mathbf{x}_i; y_i = -1\}} \left|\mathbf{w}^T\mathbf{x}_i + b\right|\Big] = \frac{2}{\|\mathbf{w}\|} \end{aligned} \tag{2}$$

Az osztályozási tartalék maximális akkor lesz ha $\|\mathbf{w}\|$ minimális értékű, vagyis minimalizálni kell az alábbi kifejezést:

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

ami az (1) egyenlet figyelembevételével az alábbi feltételes szélsőérték keresési problémára vezet

$$J(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{P} \alpha_i \{[\mathbf{w}^T \mathbf{x}_i + b]y_i - 1\} \tag{3}$$

ahol az $\alpha_i -$ k a Lagrange multiplikátorok.

A fenti Lagrange egyenletet kell minimalizálni **w** és *b* szerint és maximalizálni α_i szerint. (nyeregpont meghatározása)
A megoldást két lépesben érjük el. először **w** és *b* szerinti szélsőértéket keresünk, majd ezek eredményét behelyettesítve a Lagrange egyenletbe kapjuk az ún. másodlagos feladatot, melynek megoldásai a Lagrange multiplikátorok.

A **w** szerinti szélsőérték, $\frac{\partial J}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{P} \alpha_i \mathbf{x}_i y_i = 0$ illetve ebből adódóan $\mathbf{w} = \sum_{i=1}^{P} \alpha_i \mathbf{x}_i y_i$

A *b* szerinti szélsőértékből adódik: $\frac{\partial J}{\partial b} = \sum_{i=1}^{P} \alpha_i y_i = 0$

Behelyettesítve **w** összefüggését a Lagrange egyenletbe a duális probléma:

$$\max_{\alpha} W(\alpha) = \max_{\alpha} -\frac{1}{2}\sum_{i=1}^{P}\sum_{j=1}^{P} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \mathbf{x}_j) + \sum_{i=1}^{P} \alpha_i$$

azzal a feltétellel, hogy $\alpha_i \geq 0$ $i=1,\ldots,P$, és $\sum_{i=1}^{P} \alpha_i y_i = 0$.

A megoldás tehát a duális feladat: kvadratikus optimalizálási feladat két mellékfeltétellel. Az vagy pozitív értéket vagy nullát vesznek fel. Azon tanítópontok, melyekhöz pozitív α_i tartozik, az ún. support vektorok, ezek a vektorok lesznek a legközelebb az elválasztó hipersíkhoz, így a hipersík egyenletét is csak ezek a tanítópontok befolyásolják:

$\mathbf{w}^0 = \sum_{i=1}^{P} \alpha_i y_i \mathbf{x}_i$, míg b^0 értékét a support vektorokra alkalmazott (1) egyenletből határozhatjuk meg. Az optimális hipersík a két osztályba tartozó support vektorok között helyezkedik el olyan módon, hogy a support vektoroktól való távolsága a lehető legnagyobb legyen.

Lineáris szeparálás a biztonsági sávba eső mintapontokkal

A tartalékkal lineárisan szeparálható feladaton túl a ξ_i support vektor gépek általánosíthatók olyan lineárisan szeparálható esetekre, ahol a biztonsági sávon belül is lehet tanítópont, illetve olyan esetekre is, ahol a tanítópontok részben átlapolódnak, tehát az optimális hipersík "rossz" oldalán is lehetnek tanítópontok. Ezen esetek kezelésére ún. nemnegatív gyengítő változókat vezetnek be. Ennek megfelelően az alapegyenlet:

$$y_i[\mathbf{w}^T\mathbf{x}_i + b] \geq 1 - \xi_i \quad i = 1,...,P \tag{4}$$

Azon esetek, amikor $\xi_i = 0$, visszakapjuk az alapfeladatot, ha 0 < ξ_i < 1, a megfelelő tanítópontok a hipersík megfelelő oldalán, de a biztonsági sávban helyezkednek el, ha pedig ξ_i >1 az adott tanítópont a sík ellenkező oldalán (a hibás oldalon) van. A hipersíkot úgy kell meghatározni, hogy a hibás osztályozások száma minimális legyen, vagyis a minimalizálandó kifejezés:

$$\Phi(w,\xi) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{P}\xi_i \tag{5}$$

ahol *C* egy adott érték. A megfelelő Lagrange egyenlet

$$J(\mathbf{w},b,\xi,\alpha,\beta) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{P}\xi_i - \sum\alpha_i\{y_i[\mathbf{w}^T\mathbf{x}_i + b] - 1 + \xi_i\} - \sum_{i=1}^{P}\beta_i\xi_i \tag{6}$$

melyet **w**, *b* és ξ_i szerint kell minimalizálni és α_i és β_i szerint maximalizálni. Hasonlóan az eredeti feladathoz kapjuk, hogy :

$$\frac{\partial J}{\partial b} = \sum_{i=1}^{P}\alpha_i y_i = 0$$

illetve $\frac{\partial J}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{P}\alpha_i\mathbf{x}_i y_i = 0$ és ebből adódóan, $\mathbf{w} = \sum_{i=1}^{P}\alpha_i\mathbf{x}_i y_i$

továbbá a $\frac{\partial J}{\partial \xi_i} = 0$ feltételből : $\alpha_i + \beta_i = C$

Ezek figyelembevételével a duális feladat megegyezik a szeparálható eset duális feladatával az alábbi peremfeltételekkel:

$\sum_{i=1}^{P}\alpha_i y_i = 0$ és $0 \leq \alpha_i \leq C$, minden i=1,. . . , P-re.

Nemlineáris szeparálás SV gépekkel.

Ha egy feladat nemlineárisan szeparálható a bemenet megfelelő nemlineáris (és általában dimenziónövelő) transzformációjával lineárisan szeparálhatóvá transzformálható (ld. a Perceptron kapacitására vonatkozó Cover tételt) . A nemlineárisan szeparálható feladat SVM-nel történő megoldásában tehát két lépés szerepel: (*i*) nemlineáris transzformáció, (*ii*) a transzformált térben az optimális lineáris szeparáló hipersík meghatározása.
A hipersík egyenlete:

$$\mathbf{w}^T \varphi(\mathbf{x}) + b = 0 \qquad (7)$$

ahol a $\{\varphi_j(\mathbf{x})\}_{j=1}^{M}$ függvények a nemlineáris transzformáció függvényei, melyek a bemeneti *N*-dimenziós térből az *M*-dimenziós transzformált térbe képeznek le. Bevezetve a $\varphi_0(\mathbf{x}) = 1$ és a $w_o = b$ jelölést az elválasztó felület egyenlete: $\sum_{j=0}^{M} w_j \varphi_j(\mathbf{x}) = 0$

A megoldás menete innen megegyezik a lineárisan szeparálható eset menetével azzal a különbséggel, hogy $\mathbf{x}$ helyére mindenhol $\varphi(\mathbf{x})$ és $\mathbf{x}_i^T \mathbf{x}_j$ helyére $\varphi^T(\mathbf{x}_i)\varphi(\mathbf{x}_j)$írandó.

A bemeneti tanítópontok a $K(\mathbf{x}_i, \mathbf{x}_j) = \varphi^T(\mathbf{x}_i)\varphi(\mathbf{x}_j)$ belső szorzat magfüggvényen keresztül határozzák meg az elválasztó felületet, mivel:

$$\sum_{i=1}^{P} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) = \sum_{i=1}^{P} \left(\alpha_i y_i \sum_{j=0}^{M} \varphi_j(\mathbf{x}_i) \varphi_j(\mathbf{x}) \right) = 0 \qquad (8)$$

A magfüggvénynek fontos szerepe van a nemlineárisan szeparálható SV gépes megoldásban. A magfüggvény fontos tulajdonságai:

- szimmetrikus $K(\mathbf{x}_i, \mathbf{x}) = K(\mathbf{x}, \mathbf{x}_i)$
- a magfüggvény az alábbi sorba fejthető: $K(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{M} \lambda_k \varphi_k(\mathbf{x}_i) \varphi_k(\mathbf{x}_j)$,

  ahol a λ_k értékek a magfüggvény sajátértékei, a $\varphi_k(\mathbf{x})$ függvények pedig a saját-függvényei.

A fenti sorfejtés érvényességéhez és abszolút egyenletes konvergenciájához szükséges és elégséges, hogy:

$$\int\int K(\mathbf{x}_i, \mathbf{x}_j) g(\mathbf{x}_i) g(\mathbf{x}_j) d\mathbf{x}_i \, d\mathbf{x}_j \geq 0$$

minden olyan $g \neq 0$ esetre, ahol $\int g^2(\mathbf{x}) d\mathbf{x} \langle \infty$. (Mercer tétel)

A Mercer tétel annak eldöntésében segít, hogy egy magfüggvény belső szorzat magfüggvénye, de nem ad segítséget a sajátfüggvények megtalálásában. A SV gépeknél belső szorzat magfüggvény szerepel az elválasztó felület egyenletében.

A SVM tervezése

Adott $\{(\mathbf{x}_i, y_i)\}_{i=1}^{P}$ tanítópontok mellett keressük azokat a Lagrange multiplikátor értékeket, melyek maximalizálják a következő függvényt (duális feladat):

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

az alábbi feltételek mellett:

$\sum_{i=1}^{P} \alpha_i d_i = 0$ és $0 \le \alpha_i \le C$, i=1,...,P, ahol C egy a felhasználó által specifikált pozitív paraméter.

A Lagrange multiplikátorok meghatározása után meghatározható az optimális súlyvektor, $\mathbf{w}^0$ és b^0.

Gyakrabban alkalmazott magfüggvények SVM konstrukciójánál:

- polinomiális $K(\mathbf{x}, \mathbf{x}_i) = (\mathbf{x}^T \mathbf{x}_i + 1)^d$, $d = 1,...$
- radiális bázis függvény $K(\mathbf{x}, \mathbf{x}_i) = \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x} - \mathbf{x}_i\|^2\right)$
- Kétrétegű perceptron $K(\mathbf{x}, \mathbf{x}_i) = \tanh\left(\beta_0 \mathbf{x}^T \mathbf{x}_i + \beta_1\right)$

Az utóbbi a Mercer tételt csak bizonyos β_0 és β_1 értékekre elégíti ki.

Polinomiális esetben a magfüggvény és a bázisfüggvények (a magfüggvény sajátfüggvényei) kétdimenziós bemeneti tér $\mathbf{x}=[x_1\ x_2]^T$ mellett.

$$K(\mathbf{x}, \mathbf{x}_i) = 1 + x_1^2 x_{i1}^2 + 2x_1 x_2 x_{i1} x_{i2} + x_2^2 x_{i2}^2 + 2x_1 x_{i1} + 2x_2 x_{i2}$$

$$\varphi(\mathbf{x}_i) = [1,\ x_{i1}^2,\ \sqrt{2}x_{i1}x_{i2},\ x_{i2}^2,\ \sqrt{2}x_{i1},\ \sqrt{2}x_{i2}]^T$$

Kiterjesztés regressziós feladatra

Függvényapproximációs feladatnál az átlagos négyzetes hiba helyett az ε érzéketlenségi sávval rendelkező abszolútérték függvényt alkalmazzák az eltérés mérésére.

$$L_\varepsilon(y, f(\mathbf{x},\alpha)) = |y - f(\mathbf{x},\alpha)|_\varepsilon = \begin{cases} \varepsilon & \text{ha } |y - f(\mathbf{x},\alpha)| \le \varepsilon \\ |y - f(\mathbf{x},\alpha)| & \text{egyebkent} \end{cases}$$

A kimenet előállítása nemlineáris bázisfüggvények lineáris kombinációjaként történik:

$$y = \sum_{j=0}^{M} w_j \varphi_j(\mathbf{x}).$$

A kielégítendő feltételek (gyengítő változók bevezetésével):

$$\begin{aligned} y_i - \mathbf{w}^T \varphi(\mathbf{x}_i) &\le \varepsilon + \xi_i \qquad & i = 1,\dots,P \\ \mathbf{w}^T \varphi(\mathbf{x}_i) - y_i &\le \varepsilon + \xi_i' \qquad & i = 1,\dots,P \\ \xi_i &\ge 0 \qquad & i = 1,\dots,P \\ \xi_i' &\ge 0 \qquad & i = 1,\dots,P \end{aligned} \tag{9}$$

A minimalizálandó költségfüggvény:

$$\Phi(\mathbf{w},\xi,\xi') = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{P}(\xi_i + \xi_i') \tag{10}$$

A megfelelő Lagrange függvény:

$$\begin{aligned} J(\mathbf{w},\xi,\xi',\alpha,\alpha',\gamma,\gamma') = C\sum_{i=1}^{P}(\xi_i + \xi_i') + \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{P}\alpha_i\left[\mathbf{w}^T\varphi(\mathbf{x}_i) - y_i + \varepsilon + \xi_i\right] - \\ - \sum_{i=1}^{P}\alpha_i'\left[y_i - \mathbf{w}^T\varphi(\mathbf{x}_i) + \varepsilon + \xi_i'\right] - \sum_{i=1}^{P}(\gamma_i\xi_i + \gamma_i'\xi') \end{aligned} \tag{11}$$

A Lagrange függvényt minimalizálni kell $\mathbf{w}$ és a ξ és ξ' gyengítő változók szerint és maximalizálni kell α, α', γ és γ' szerint. Az optimalizáció eredménye:

$$\begin{aligned} \mathbf{w} &= \sum_{i=1}^{P}(\alpha_i - \alpha_i')\varphi(\mathbf{x}_i) \\ \gamma_i &= C - \alpha_i \qquad \text{és} \\ \gamma_i' &= C - \alpha_i' \end{aligned} \tag{12}$$

Ezek figyelembevételével kapjuk a duális feladatot:

$$W(\alpha_i\alpha_i') = \sum_{i=1}^{P} y_i(\alpha_i - \alpha_i') - \varepsilon\sum_{i=1}^{P}(\alpha_i + \alpha_i') - \frac{1}{2}\sum_{i=1}^{P}\sum_{j=1}^{P}(\alpha_i - \alpha_i')(\alpha_j - \alpha_j')K(\mathbf{x}_i,\mathbf{x}_j) \tag{13}$$

Ennek megoldása a megfelelő korlátozó feltételek

$$\sum_{i=1}^{P}(\alpha_i - \alpha_i') = 0 \quad , \quad 0 \le \alpha_i \le C, \quad i=1,...,P \quad \text{és} \quad 0 \le \alpha_i' \le C, \; i=1,...,P$$

mellett adja a Lagrange multiplikátorokat, melyek és a magfüggvényértékek segítségével kapjuk az optimális approximáló függvény súlyvektorát, **w**-t. Azon tanítópontok melyekre $\alpha_i \ne \alpha_i'$ lesznek a support vektorok, hiszen a súlyvektor (**w**) meghatározásában (12) szerint csak ezek a pontok vesznek részt.

Az eljárásban ε és C értéke a felhasználó által megválasztandó.

Néhány további irodalom:

A. J. Smola, B. Schölkopf: "A tutorial on support vector regression" NeuroCOLT2 Technical Report Series, NC2-TR-1998-030. http://www.neurocolt.com

V. Vapnik: "The Nature of Statistical Learning Theory" Springer, N.Y. 1995.

C. J. C. Burges: "A Tutorial on Support Vector Machines for Pattern Recognition" Knowledge Discovery and Data Mining, 1998. pp. 121-167.