

0.1 Tukivektorikone

(‘A Tutorial on Support Vector Machines for Pattern Recognition’,

<http://www.kernel-machines.org/papers/Burges98.ps.gz>)

Tukivektorikoneen (‘Support Vector Machines’, SVM) yleisessä tapauksessa piirrevektoreille suoritetaan aluksi epälineaarinen kuvaus ja pyritään sitten löytämään yhdensuuntaiset hypertasot, jotka maksimoivat luokkien väliin jäävän marginaalin ja minimoivat luokitteluvirheet

Tarkastellaan aluksi lineaarista SVM-menetelmää lineaarisesti separoituvalla ongelmalla ja laajennetaan sitten menetelmää ongelmaan, joka ei ole lineaarisesti separoituva.

Menetelmän epälineaariseen yleistykseen palataan myöhemmin epälineaaristen luokittimien yhteydessä.

Lineaarisesti separoituvat luokat

Tarkastellaan kahden luokan tapausta

Ol., että luokat ovat lineaarisesti separoituvat

Etsitään kaksi yhdensuuntaista hypertasoa, H_1 ja H_2 , siten, että seuraavat ehdot toteutuvat kaikille opetusnäytteille $\mathbf{x}(i)$, $1 \leq i \leq N$:

$$\begin{aligned} \mathbf{w}^T \mathbf{x}(i) + w_0 &\geq +1, & \text{kun } \mathbf{x}(i) \in \omega_1 \\ \mathbf{w}^T \mathbf{x}(i) + w_0 &\leq -1, & \text{kun } \mathbf{x}(i) \in \omega_2, \end{aligned} \tag{1}$$

missä \mathbf{w} on hypertasojen normaali

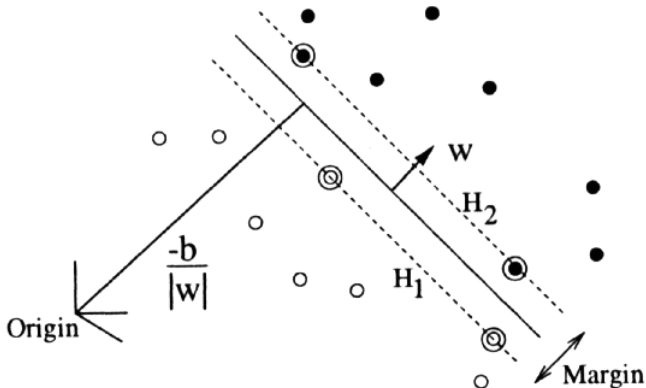
Hypertasot on määritelty seuraavasti:

$$\begin{aligned} H_1 : \mathbf{w}^T \mathbf{x} + w_0 &= +1 \\ H_2 : \mathbf{w}^T \mathbf{x} + w_0 &= -1 \end{aligned} \tag{2}$$

Ehdot (1) voidaan yhdistää muuttujien $y(i)$ avulla, jotka saavat arvon $+1$ tai -1 riippuen siitä kuuluuko $\mathbf{x}(i)$ luokkaan ω_1 vai ω_2

$$y(i)(\mathbf{w}^T \mathbf{x}(i) + w_0) - 1 \geq 0, \forall 1 \leq i \leq N \quad (3)$$

Koska luokat oletettiin lineaarisesti separoituviksi, hypertasojen H_1 ja H_2 väliin jää marginaali, jossa ei ole lainkaan opetusnäytteitä



Olkoot d_+ (d_-) etäisyys hypertason $\mathbf{w}^T \mathbf{x} + w_0 = 0$ ja sitä lähinnä olevan luokan ω_1 (ω_2) pisteen välinen Euklidinen etäisyys. Silloin marginaalin leveys on $d_+ + d_- = 2/\|\mathbf{w}\|$

Yritetään etsiä sellaiset hypertasot H_1 ja H_2 , että marginaali maksimoituu (tai $\|\mathbf{w}\|^2$ minimoituu) ja ehdot (3) toteutuvat

Muodostetaan minimoitava kustannusfunktio $J(\mathbf{w}, w_0, \lambda)$ käyttäen Lagrangen kertoimia $\lambda = [\lambda_1, \dots, \lambda_N]^T \geq \mathbf{0}$:

$$J(\mathbf{w}, w_0, \lambda) = \frac{1}{2} \mathbf{w}^T \mathbf{w} - \sum_{i=1}^N \lambda_i (y(i)(\mathbf{w}^T \mathbf{x}(i) + w_0) - 1) \quad (4)$$

Minimiratkaisu löytyy J :n gradientin nollakohdasta:

$$\frac{\partial J}{\partial \mathbf{w}} = \mathbf{0} \Rightarrow \mathbf{w} = \sum_{i=1}^N \lambda_i y(i) \mathbf{x}(i) \quad (5)$$

$$\frac{\partial J}{\partial w_0} = 0 \Rightarrow \sum_{i=1}^N \lambda_i y(i) = 0 \quad (6)$$

Sijoitetaan nämä ehdot kaavaan (4) ja muodostetaan ns duaaliongelman λ :n löytämiseksi:

$$\max_{\lambda} Q(\lambda) = \sum_{i=1}^N \lambda_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \lambda_i \lambda_j y(i) y(j) \mathbf{x}^T(i) \mathbf{x}(j) \quad (7)$$

siten, että

$$\sum_{i=1}^N \lambda_i y(i) = 0 \quad (8)$$

$$\lambda_i \geq 0, \forall 1 \leq i \leq N \quad (9)$$

Kun on löydetty duaaliuongelman ratkaisu λ^* , saadaan optimaaliset \mathbf{w}^* ja w_0^* seuraavasti:

$$\mathbf{w}^* = \sum_{i=1}^N \lambda_i^* y(i) \mathbf{x}(i) \quad (10)$$

$$w_0^* = 1 - \mathbf{x}^T(i) \mathbf{w}^*, \text{ missä } \mathbf{x}(i) \in \omega_1, \lambda_i^* > 0 \quad (11)$$

Saadaan siis ratkaisuksi marginaalin maksimoiva diskriminanttifunktio $g(\mathbf{x})$, joka kulkee hypertasojen H_1 ja H_2 puolessavälissä:

$$g(\mathbf{x}) = \mathbf{x}^T \mathbf{w}^* + w_0^* = \sum_{i=1}^N \lambda_i^* y(i) \mathbf{x}^T(i) \mathbf{x} + w_0^* = 0 \quad (12)$$

Edellinen kaava voidaan tulkita siten, että diskriminanttifunktiota approksimoidaan lineaarisilla kantafunktiolla $\mathbf{x}^T(i) \mathbf{x}$

Niitä opetusnäytteitä $\mathbf{x}(i)$, joita vastaava Lagrangen kerroin $\lambda_i > 0$, kutsutaan *tukivektoreiksi*. Ainoastaan tukivektorit huomioidaan diskriminanttifunktiossa

Lineaarisesti separoitumattomat luokat

Tarkastellaan edelleen kahden luokan tapausta, mutta oletetaan, että luokat eivät ole lineaarisesti separoituvat

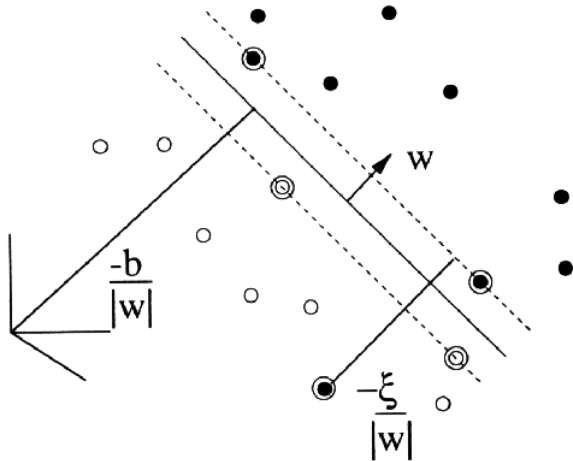
Nyt pyritään maksimoimaan luokkien väliin jäävää marginaalia ja minimoimaan luokitteluvirheiden lkm:ää

Muodostetaan opetusnäytteille seuraavat ehdot:

$$y(i)(\mathbf{w}^T \mathbf{x}(i) + w_0) \geq 1 - \xi_i \quad (13)$$

missä $\xi_i \geq 0$ on ns slack-muuttuja, joka kertoo onko opetusnäyte $\mathbf{x}(i)$ marginaalin sisällä ($0 \leq \xi_i \leq 1$) tai väärällä puolella ($\xi_i > 1$)

$\sum_{i=1}^N \xi_i$ on yläraja luokitteluvirheiden lkm:lle



Minimoitava kustannusfunktio, jossa edelliset ehdot on huomioitu käyttäen Lagrangen kertoimia $\lambda \geq \mathbf{0}$ ja $\mu \geq \mathbf{0}$:

$$\begin{aligned}
 J(\mathbf{w}, w_0, \xi, \lambda, \mu) = & \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^N \xi_i \\
 & - \sum_{i=1}^N \lambda_i (y(i) (\mathbf{w}^T \mathbf{x}(i) + w_0) - 1 + \xi_i) - \sum_{i=1}^N \mu_i \xi_i
 \end{aligned} \tag{14}$$

missä $\xi = [\xi_1, \dots, \xi_N]^T$ ovat slack-muuttajat ja $\mu = [\mu_1, \dots, \mu_N]^T$ niihin liittyvät Lagrangen kertoimet. C on käyttäjän määräämä parametri, joka kertoo kuinka tärkeää luokitteluvirheiden minimointi on suhteessa marginaalin leveyden maksimointiin

Kun toimitaan kuten lineaarisesti separoituvassa tapauksessa, saadaan seuraava duaaliongelman:

$$\max_{\lambda} Q(\lambda) = \sum_{i=1}^N \lambda_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \lambda_i \lambda_j y(i) y(j) \mathbf{x}^T(i) \mathbf{x}(j) \tag{15}$$

siten, että

$$\sum_{i=1}^N \lambda_i y(i) = 0 \quad (16)$$

$$0 \leq \lambda_i \leq C, \quad \forall 1 \leq i \leq N \quad (17)$$

Kun on löydetty duaaliongelman ratkaisu λ^* , optimaalinen \mathbf{w}^* saadaan seuraavasti:

$$\mathbf{w}^* = \sum_{i=1}^{N_s} \lambda_i^* y(i) \mathbf{x}(i), \quad (18)$$

missä N_s on tukivektoreiden lkm

Loput tuntemattomat eli w_0^* , ξ^* ja μ^* voidaan ratkaista seuraavista yhtälöistä:

$$\lambda_i^* (y(i) (\mathbf{x}^T(i) \mathbf{w}^* + w_0) - 1 + \xi_i) = 0, \quad \forall 1 \leq i \leq N \quad (19)$$

$$\mu_i \xi_i = 0, \quad \forall 1 \leq i \leq N \quad (20)$$

$$\lambda_i^* + \mu_i = C, \quad \forall 1 \leq i \leq N \quad (21)$$

Tukivektoreita ovat ne opetusnäytteet $\mathbf{x}(i)$, joille pätee:

$$y(i)(\mathbf{x}^T(i)\mathbf{w}^* + w_0^*) = 1 - \xi_i^* \quad (22)$$

eli $0 < \lambda_i \leq C$

Mikäli luokkia on enemmän kuin kaksi, muodostetaan SVM jokaisen luokkaparin välille

SVM-menetelmässä kantafunktioiden lkm:n ja sijoittelun valinta tapahtuu automaattisesti

Toisaalta, jos luokat ovat kovin päällekkäiset tai datassa on paljon ns outlier-pisteitä, menetelmä löytää runsaasti tukivektoreita

1. EPÄLINEAARISET LUOKITTIMET

Yleensä luokitteluongelma ei ole lineaarisesti separoituva eivätkä lineaarisilla luokittimilla saadut ratkaisut ole tyydyttäviä

Tällä luennolla tarkastellaan menetelmiä, jotka jakavat piirreavaruuden *epälineaarisesti* eri luokkia vastaaviksi päätösalueiksi

Aluksi annetaan teoreettinen motivointi epälineaarille luokittimille, seuraavaksi käsitellään yleisimmin käytetyt menetelmät

Eräs perinteinen esimerkki epälineaarisesta luokitteluongelmasta on XOR-funktio

Esimerkki: XOR-ongelma

XOR ('Exclusive OR') on Boolean-funktio, jonka arvo on tosi (1) vain jos täsmälleen yksi sen argumenteista on tosi. Muulloin sen arvo on epätosi (0)

Tarkastellaan 2-ulotteista tapausta. Silloin $\mathbf{x} = [x_1, x_2]^T$, $x_i \in \{0, 1\}$, ja luokkia ω_1 ja ω_2 vastaavat seuraavat havainnot:

$$\omega_1 : [1, 0]^T, [0, 1]^T \text{ (tosi)}$$

$$\omega_2 : [0, 0]^T, [1, 1]^T \text{ (epätosi)}$$

Kun piirretään luokitteluongelmasta kuva, nähdään helposti, että luokat eivät ole lineaarisesti separoituvia

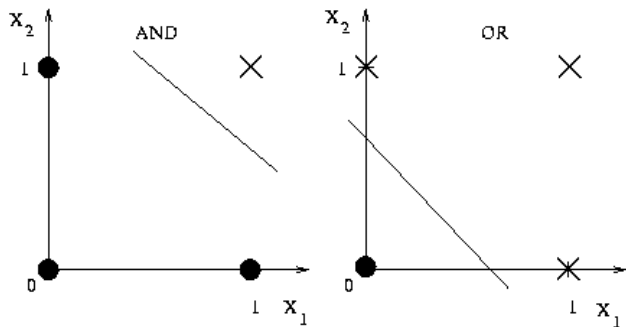
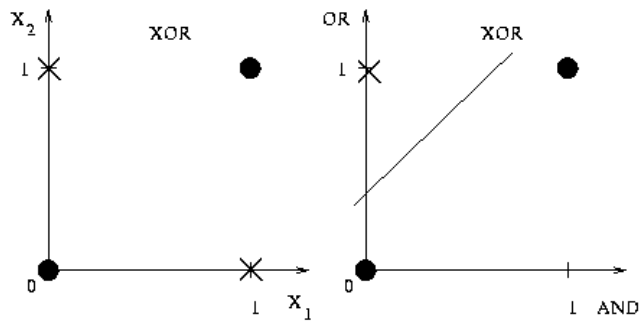
Toisaalta AND- ja OR-funktiota vastaavat luokitteluongelmat ovat lineaarisesti separoituvia

Jos tehdään piirteille epälineaarinen muunnos

$$[x_1, x_2]^T \rightarrow [\text{AND}(x_1, x_2), \text{OR}(x_1, x_2)]^T,$$

saadaan XOR-ongelmasta lineaarisesti separoituva

x_1	x_2	AND	class	OR	class	XOR	class
0	0	0	ω_2	0	ω_2	0	ω_2
0	1	0	ω_2	1	ω_1	1	ω_1
1	0	0	ω_2	1	ω_1	1	ω_1
1	1	1	ω_1	1	ω_1	0	ω_2



✕ tosi ● epätosi

1.1 Yleistetty lineaarinen luokitin

Tarkastellaan edellisessä esimerkissä esiteltyä tapaa tehdä luokitteluongelmasta lineaarisesti separoituva piirteiden epälineaarisen kuvauksen avulla yleisemmällä tasolla:

- Tarkastellaan kahden luokan tapausta: ω_1 ja ω_2
- Ol., että luokat ω_1 ja ω_2 eivät ole lineaarisesti separoituvia
- Ol., että piirrevektorit ovat l -ulotteisia ja niiden alkiot ovat reaalityyppisiä eli $\mathbf{x} \in \mathbb{R}^l$
- Suoritetaan piirrevektoreille seuraava epälineaarinen kuvaus:

$$\begin{aligned}\mathbf{x} \in \mathbb{R}^l &\rightarrow \mathbf{y} \in \mathbb{R}^k \\ \mathbf{y} &= [f_1(\mathbf{x}), \dots, f_k(\mathbf{x})]^T,\end{aligned}\tag{23}$$

missä $f_i : \mathbb{R}^l \rightarrow \mathbb{R}$ on jokin epälineaarinen funktio

- Pyritään valitsemaan uusien piirteiden lkm k ja funktiot $f_i(\cdot)$ siten, että luokitteluongelmasta tulee lineaarisesti separoituva eli voidaan löytää seuraavanlainen separoiva hypertaso \mathbf{w} :

$$\begin{aligned}\omega_1 : \mathbf{w}^T \mathbf{y} &> 0 \\ \omega_2 : \mathbf{w}^T \mathbf{y} &< 0\end{aligned}\tag{24}$$

Edellä kuvatun piirteiden epälineaaristen kuvausten seurauksena eri luokkiin liittyvät diskriminanttifunktiot ovat muotoa

$$g(\mathbf{x}) = w_0 + \sum_{i=1}^k w_i f_i(\mathbf{x}) = 0\tag{25}$$

Uudessa piirreavaruudessa määriteltyjen hypertasojen painokertoimet voidaan etsiä käyttämällä edellisellä luennolla esiteltyjä menetelmiä

Kaava (25) voidaan tulkita diskriminanttifunktion approksimoinniksi interpolaatio- tai kantafunktioden $f_i(\cdot)$ avulla. Interpolaatiofunktiot voivat olla tyyppiltään esim. eksponentiaalisia tai polynomeja

Coverin teoreema antaa teoreettisen perustelun, miksi tällainen piirteiden epälineaarinen kuvaus kannattaa tehdä

1.2 Coverin teoreema

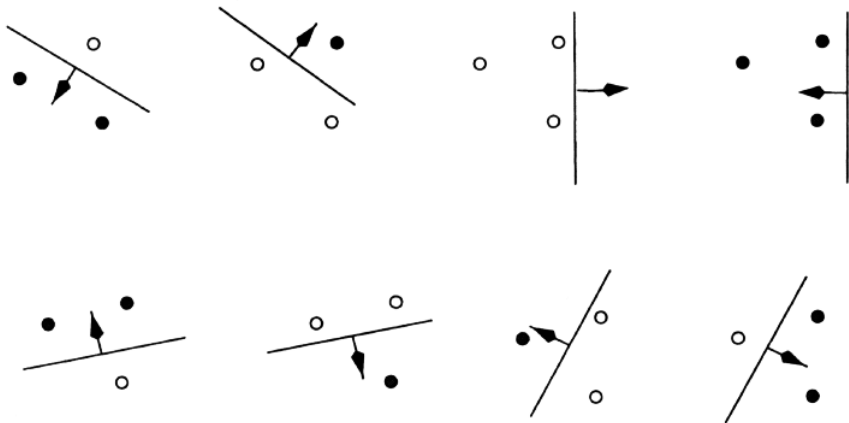
Todennäköisyys sille, että luokat ovat lineaarisesti separoituvia kasvaa, kun suoritetaan epälineaarinen kuvaus alkuperäisestä piirreavaruudesta uuteen, korkeampiulotteisempaan piirreavaruuteen (Cover, 1965)

Todistus:

- Voidaan osoittaa, että $(l - 1)$ -ulotteinen hypertaso voi jakaa N mielivaltaisesti valittua pistettä kahteen luokkaan $O(N, l)$:llä tavalla:

$$O(N, l) = 2 \sum_{i=0}^l \binom{N-1}{i} \tag{26}$$
$$\binom{N-1}{i} = \frac{(N-1)!}{(N-1-i)!i!}$$

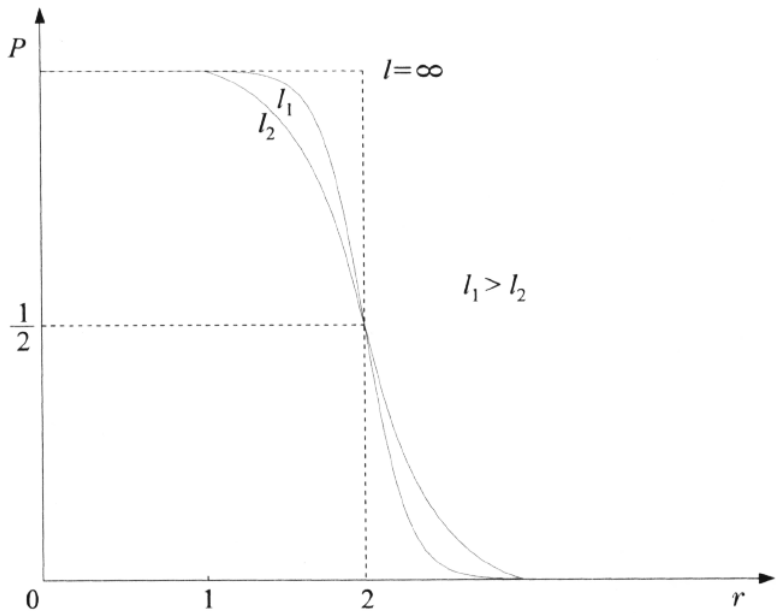
Näitä erilaisia kahtiajakoja kutsutaan dikotomioiksi ('dichotomies').
 Kun $N \leq l + 1$, $O(N, l) = 2^N$. Esim. $O(3, 2) = 8$



- Tn, että N :n pisteen dikotomia l -ulotteisessa piirreavaruudessa vastaa kahta lineaarisesti separoituvaa luokkaa on P_N^l :

$$P_N^l = \frac{O(N, l)}{2^N} = \begin{cases} \frac{1}{2^{N-1}} \sum_{i=0}^l \binom{N-1}{i}, & \text{kun } N > l + 1 \\ 1, & \text{muulloin} \end{cases} \quad (27)$$

- Tarkastellaan P_N^l :n riippuvuutta N :stä ja l :stä muuttujan r avulla olettamalla, että $N = r(l + 1)$. Kohdassa $r = 2$ $P_N^l = 1/2$. Lisäksi, kun $l \rightarrow \infty$, $P_N^l \rightarrow 1$, jos $N < 2(l + 1)$
- Toisin sanoen, tn että luokat ovat lineaarisesti separoituvia lähestyy ykköistä, kun piirteiden lkm l lähestyy ääretöntä ja havaintojen lkm N on sopivasti rajoitettu



Edellisen tarkastelun perusteella piirrevektorit kannattaa kuvata korkeampiulotteiseen avaruuteen ja yrittää vasta sitten luokkien lineaarista separointia. Tn onnistumiselle on tällöin parempi

1.3 Polynomiluokitin

Tarkastellaan diskriminanttifunktioita, jotka on muodostettu kaavan (25) avulla ja käyttämällä kantafunktioina polynomeja:

- Ol., että polynomien asteluku r on 2. Silloin kaava (25) voidaan kirjoittaa seuraavasti:

$$g(\mathbf{x}) = w_0 + \sum_{i=1}^l w_i x_i + \sum_{i=1}^{l-1} \sum_{m=i+1}^l w_{im} x_i x_m + \sum_{i=1}^l w_{ii} x_i^2 = 0 \quad (28)$$

- Esim. jos $\mathbf{x} = [x_1, x_2]^T$, silloin $\mathbf{y} = [1, x_1, x_2, x_1x_2, x_1^2, x_2^2]^T$ ja

$$g(\mathbf{x}) = \mathbf{w}^T \mathbf{y} = 0, \quad (29)$$

missä $\mathbf{w} = [w_0, w_1, w_2, w_{12}, w_{11}, w_{22}]^T$

- Vapaiden parametrien eli erilaisten muotoa $x_1^{p_1} \cdots x_l^{p_l}$, $0 \leq p_1 + \cdots + p_l \leq r$, olevien termien lkm k :

$$k = \frac{(l+r)!}{r!l!} \quad (30)$$

Huom! Kun $l = 10$ ja $r = 10$, $k = 184756$

Vapaat parametrit voidaan taas valita esim. edellisellä luennolla esitetyillä menetelmillä

Esimerkki: XOR ja polynomiluokitin

Palataan takaisin XOR-ongelmaan ja muodostetaan sille polynomiluokitin

Valitaan kantafunktiot seuraavasti:

$$\mathbf{y} = [1, x_1, x_2, x_1x_2]^T$$

Erilaisten havaintojen sijainnit uudessa piirreavaruudessa:

$$\omega_1 : [1, 0]^T \rightarrow [1, 1, 0, 0]^T, [0, 1]^T \rightarrow [1, 0, 1, 0]^T$$

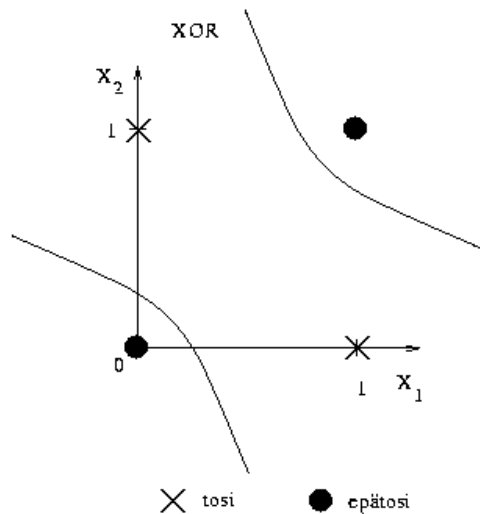
$$\omega_2 : [0, 0]^T \rightarrow [1, 0, 0, 0]^T, [1, 1]^T \rightarrow [1, 1, 1, 1]^T$$

Nämä uudet pisteet separoituvat lineaarisesti esim. hypertason $\mathbf{w}^T \mathbf{y} = 0$ avulla, missä $\mathbf{w} = [-1/4, 1, 1, -2]^T$

Alkuperäisessä piirreavaruudessa tätä hypertasoa vastaa diskriminanttifunktio $g(\mathbf{x})$:

$$g(\mathbf{x}) = -\frac{1}{4} + x_1 + x_2 - 2x_1x_2 = 0, \quad (31)$$

$g(\mathbf{x})$ näyttää suurinpiirtein tältä:



1.4 RBF-luokitin

(RBF-luokitin voidaan ymmärtää myös neuroverkkona...)

Seuraavaksi muodostetaan diskriminanttifunktiot käyttäen kaavaa (25) ja kantafunktiota, jotka ovat muotoa $f_i(\|\mathbf{x} - \mathbf{c}_i\|)$

Kantafunktioiden arvo siis riippuu vain piirvektorin \mathbf{x} ja keskusvektorin \mathbf{c}_i välisestä Euklidisestä etäisyydestä. Siitä kantafunktioiden englanninkielinen nimitys, 'Radial Basis Functions' (RBF)

Kantafunktiot voivat olla tyypiltään esim. seuraavanlaisia:

$$f_i(\mathbf{x}) = \exp\left(-\frac{1}{2\sigma_i^2}\|\mathbf{x} - \mathbf{c}_i\|^2\right) \quad (32)$$

$$f_i(\mathbf{x}) = \frac{\sigma_i^2}{\sigma_i^2 + \|\mathbf{x} - \mathbf{c}_i\|^2}, \quad (33)$$

joista ensimmäinen, gaussinen, on hyvin yleisesti käytetty

Kun käytetään gaussisia kantafunktiota, kaava (25) voidaan kirjoittaa näin:

$$g(\mathbf{x}) = w_0 + \sum_{i=1}^k w_i \exp\left(-\frac{(\mathbf{x} - \mathbf{c}_i)^T(\mathbf{x} - \mathbf{c}_i)}{2\sigma_i^2}\right) = 0 \quad (34)$$

Edellisen kaavan voi tulkita siten, että diskriminanttifunktiota approksimoidaan summaamalla yhteen keskusvektoreiden ympärille asetettuja kumpareita. Kumpareiden leveyttä voidaan säätää muuttamalla σ_i :ja

RBF-luokittimella ja tnjakauman estimoinnilla Partzen-ikkunoilla on paljon yhteistä. Jälkimmäisessä menetelmässä jokaisen opetusnäytteen päälle asetetaan kumpare, ensimmäisessä vain keskusvektoreiden (k kpl:tta) päälle

Kuinka valitaan sopivat parametrit?

- RBF-luokittimille pitää valita seuraavat parametrit: k ,
 $\mathbf{w} = [w_0, \dots, w_k]^T$, $\mathbf{C} = [\mathbf{c}_1, \dots, \mathbf{c}_k]$ ja $\sigma = [\sigma_1, \dots, \sigma_k]^T$
- Usein valitaan ensin \mathbf{C} ja σ , ja sitten painokertoimet \mathbf{w} käyttäen esim. edellisellä luennolla esiteltyjä menetelmiä

- k :n, \mathbf{C} :n ja σ :n valinta voi perustua esim. ongelmaan liittyvään *a priori* tietoon tai klusterointimenetelmillä (näistä myöhemmin) saatuihin tuloksiin
- Mikäli opetusjoukko on edustava otos kaikista mahdollisista piirrevektoreista ja hyvin jakautunut, voidaan keskusfunktiot myös arpoa niiden joukosta
- Parametrit (paitsi k) voidaan myös valita minimoimalla seuraavan tyypistä kustannusfunktiota $J(\mathbf{w}, \mathbf{C}, \sigma)$:

$$J(\mathbf{w}, \mathbf{C}, \sigma) = \sum_{j=1}^N \Phi(y(j) - g(\mathbf{x}(j))), \quad (35)$$

missä $y(j)$ on opetusnäytettä $\mathbf{x}(j)$ vastaava diskriminanttifunktion taivoitearvo ja $\Phi(\cdot)$ jokin differentioituva funktio, esim. $(\cdot)^2$

- Edellä esitetty ongelma on tyypillinen epälineaarinen optimointiongelma, jonka ratkaisu löytyy parametrien suhteen lasketun J :n gradientin

nollakohdasta. Käytännössä tällaisen ongelman ratkaisu ei ole täysin mutkatonta

Esimerkki: XOR ja RBF-luokitin

Lähestytään jo tutuksi käynyttä XOR-ongelmaa käyttäen RBF-luokitinta

Tehdään seuraavanlainen epälineaarinen kuvaus piirrevektoreille:

$$\mathbf{y} = \mathbf{y}(\mathbf{x}) = \begin{bmatrix} \exp(-\|\mathbf{x} - \mathbf{c}_1\|^2) \\ \exp(-\|\mathbf{x} - \mathbf{c}_2\|^2) \end{bmatrix},$$

missä $\mathbf{c}_1 = [1, 1]^T$ ja $\mathbf{c}_2 = [0, 0]^T$

Tällöin piirrevektorit kuvautuvat seuraaviin paikkoihin uudessa piirreavaruudessa:

$$\omega_1 : [1, 0]^T \rightarrow [0.368, 0.368]^T, [0, 1]^T \rightarrow [0.368, 0.368]^T$$

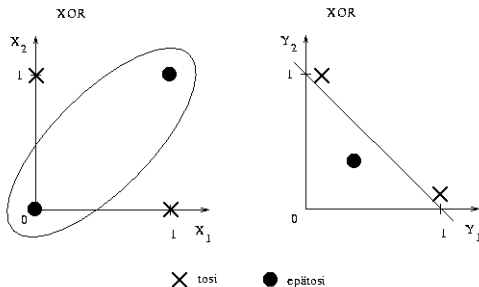
$$\omega_2 : [0, 0]^T \rightarrow [0.135, 1]^T, [1, 1]^T \rightarrow [1, 0.135]^T$$

ja luokista tulee lineaarisesti separoituvat

Valitaan separoiva hypertaso uudessa piirreavaruudessa seuraavasti:

$$g(\mathbf{y}) = y_1 + y_2 - 1 = \exp(-\|\mathbf{x} - \mathbf{c}_1\|^2) + \exp(-\|\mathbf{x} - \mathbf{c}_2\|^2) - 1 = 0$$

Luokitteluongelman ratkaisu uudessa ja alkuperäisessä piirreavaruudessa:



Alkuperäisessä piirreavaruudessa päätösraja ei ole täsmälleen ellipsi kuten kuvassa, vaan vain sitä muistuttava käyrä. Kirjassa on parempi kuva

1.5 Tukivektorikoneen epälineaarinen yleistys

Tukivektorikoneen epälineaarisen yleistyksen idea on hyvin samankaltainen kuin RBF-luokittimen

Epälineaarinen yleistys korvaamalla kaavassa

(12) esiintyvät sisätulot $\mathbf{x}^T(i)\mathbf{x}$ epälineaarisilla kantafunktioilla

$$K(\mathbf{x}, \mathbf{x}(i)) = \Phi(\mathbf{x})^T \Phi(\mathbf{x}(i)), \quad (36)$$

missä $\Phi(\cdot) : \mathbb{R}^l \rightarrow \mathbb{R}^k$ on jokin epälineaarinen kuvaus

Tällainen ratkaisu saadaan, jos ensiksi kuvataan piirrevektorit funktion $\Phi(\cdot)$ avulla uuteen piirreavaruuteen ja muodostetaan sitten lineaarinen SVM

Epälineaarisen SVM:n ratkaisussa esiintyviä sisätuloja ei välttämättä tarvitse laskea uudessa piirreavaruudessa

Kantafunktioiden tulee toteuttaa Mercerin teoreeman ehdot. Mm seuraavat

funktiot ovat käypiä kantafunktioita:

$$(\mathbf{x}^T \mathbf{x}(i) + 1)^p \quad (\text{polynomiaalinen}) \quad (37)$$

$$\exp\left(-\frac{1}{2\sigma^2} \|\mathbf{x} - \mathbf{x}(i)\|^2\right) \quad (\text{RBF}) \quad (38)$$

$$\tanh(\beta \mathbf{x}^T \mathbf{x}(i) + \gamma) \quad (\text{Hyperbolinen tangentti}) \quad (39)$$

Näistä viimeinen sopivilla β :n ja γ :n arvoilla, jotka toteuttavat Mercerin teoreeman ehdot.

Tapauksen, jossa luokat eivät ole lineaarisesti separoituvia eli on käytettävä ns. slack muuttujia, epälineaarinen yleistys muodostetaan samalla tavalla kuin lineaarisesti separoituvassa tapauksessa korvaamalla ratkaisussa esiintyvät sisätulot $\mathbf{x}^T(i)\mathbf{x}$ epälineaarisilla kantafunktioilla

SVM-menetelmän etuna on se, että joudutaan ratkaisemaan vain kvadraattinen optimointiongelma, jossa rajoitusehdot ovat lineaarisia

Esimerkki: XOR ja SVM-menetelmä

Ratkaistaan jälleen kerran XOR-ongelma, mutta nyt epälineaarisella SVM-

menetelmällä

Merkitään $+1$ =tosi, -1 =epätosi eli:

i	$\mathbf{x}(i)$	class	$y(i)$
1	$[-1, -1]^T$	ω_2	-1
2	$[-1, +1]^T$	ω_1	+1
3	$[+1, -1]^T$	ω_1	+1
4	$[+1, +1]^T$	ω_2	-1

Käytetään polynomiaalisia kantafunktioita, joiden asteluku on 2:

$$\begin{aligned}K(\mathbf{x}, \mathbf{x}(i)) &= (\mathbf{x}^T \mathbf{x}(i) + 1)^2 \\ &= 1 + x_1^2 x_1^2(i) + 2x_1 x_2 x_1(i) x_2(i) \\ &\quad + x_2^2 x_2^2(i) + 2x_1 x_1(i) + 2x_2 x_2(i)\end{aligned}$$

Tämä kantafunktio vastaa seuraavaa piirteiden epälineaarista kuvausta $\mathbf{f} : \mathbf{x} \in \mathbb{R}^2 \rightarrow \mathbf{y} \in \mathbb{R}^6$

$$\mathbf{y} = \mathbf{f}(\mathbf{x}) = [1, x_1^2, \sqrt{2}x_1x_2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2]^T$$

Koska uusissa piirvektoreissa oli mukana vakiotermi 1, kynnyisarvo w_0 on turha ja se voidaan asettaa suoraan nolllaksi

Tehdään ratkaisulle (15) epälineaarinen yleistys korvaamalla alkuperäiset piirvektorit uusilla piirvektoreilla

Uusien piirvektoreiden sisätulot voidaan korvata kantafunktioilla

Silloin saadaan seuraava duaaliingelman kustannusfunktio:

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

$Q(\lambda)$:n maksimi löytyy gradientin nolllakohdasta:

$$\frac{\partial Q(\lambda)}{\partial \lambda} = \mathbf{0} \Rightarrow \lambda^* = \frac{1}{8}[1, 1, 1, 1]^T$$

Edellisen tuloksen perusteella tiedetään, että kaikki opetusnäytteet ovat tuki-

vektoreita. Silloin kaavan (18) epälineaarisen yleistyksen perusteella

$$\begin{aligned}\mathbf{w}^* &= \sum_{i=1}^{N_s} \lambda_i^* y(i) \mathbf{f}(\mathbf{x}(i)) \\ &= [0, 0, -1/\sqrt{2}, 0, 0, 0]^T\end{aligned}$$

Optimaalinen diskriminanttifunktio on siis

$$g(\mathbf{x}; \mathbf{w}^*) = \mathbf{f}^T(\mathbf{x})\mathbf{w}^* = 0 \Rightarrow -x_1x_2 = 0$$

ja sen antama ratkaisu näyttää alkuperäisessä piirreavaruudessa tältä:

