Stödvektormaskiner
Linjära hyperplan i Hilbertrum

Oscar Granlund

Kandidatavhandling i matematik
Fakulteten för naturvetenskaper och teknik
Åbo Akademi

9 november 2018

Stödvektormaskinen utvecklades under den senare halvan av 1900-talet i huvudsak av den ryske statistikern/datavetaren Vladimir Vapnik.

- Tog sin början år 1963 med en *linjär klassificerare* som endast gick att tillämpa på några problem.
- År 1992 presenterades en version som gick att tillämpa på alla problem.

Stödvektormaskinen går ut på att man skjuter in ett hyperplan mellan två klasser och använder hyperplanet för att klassificera nya observationer.

Parallellt med forskningen om stödvektormaskiner fann statistiker att en speciell typ av funktion, kärnor, kunde användas för att generalisera linjära algoritmer.

- Kärnorna föreslogs redan år 1964 för att generalisera en annan typ av linjär klassificerare.
- De användes även för att studera till exempel spline-modeller.
- År 1992 tillämpades kärnor på den ursprungliga stödvektormaskinmetoden.

Snart därefter (1995) tillämpades kärnor på den mera generaliserade algoritmen som presenterades 1992. Resultatet är en (tidsmässigt och resultatmässigt) effektiv *olinjär klassificerare* som ännu idag används.

De flesta av algoritmerna inom statistik och maskininlärning går att skriva om som konvexa optimeringsproblem. Ett optimeringsproblem är kvadratiskt och konvext om det går att skriva om på formen

$$\begin{aligned} \min_{\mathbf{x}} \quad & \frac{1}{2}\mathbf{x}^\intercal \mathbf{P}\mathbf{x} + \mathbf{q}^\intercal \mathbf{x} + r \\ \text{så att} \quad & g_i(\mathbf{x}) \leq 0, \quad i = 1, \ \ldots, n \\ & h_i(\mathbf{x}) = 0, \quad i = 1, \ \ldots, m, \end{aligned}$$

där $\mathbf{P}$ är en positivt semidefinit matris, $g_i(\mathbf{x})$ är högst en kvadratisk funktion och alla krav g_i och h_i är satisfierbara samtidigt.

Ett konvext optimeringsproblem med olikhetskrav och likhetskrav kan lösas med hjälp av Lagrangemultiplikatorer. För ett kvadratiskt optimeringsproblem blir den primala Lagrangefunktionen

$$L_P = f(\mathbf{x}) + \sum_{i=1}^{n} \lambda_i g_i(\mathbf{x}) + \sum_{i=1}^{m} v_i h_i(\mathbf{x})$$

där $f(\mathbf{x})$ är objektfunktionen och g_i, h_i är kraven. Den duala Lagrangefunktionen fås om man tar infimum över alla giltiga $\mathbf{x}$:

$$L_D = \inf_{\mathbf{x} \text{ är giltig}} \left(f(\mathbf{x}) + \sum_{i=1}^{n} \lambda_i g_i(\mathbf{x}) + \sum_{i=1}^{m} v_i h_i(\mathbf{x}) \right).$$

Den duala Lagrangefunktionen ger en undre gräns för värdet av objektivfunktionen i den optimala punkten. Genom att maximera den duala Lagrangefunktionen borde man få en så bra undre gräns som möjligt. Karush-Kuhn-Tucker villkoren ger krav på hurudana de optimala värdena $\mathbf{x}^*$, λ_i^* och v_i^* måste vara för att lösningen ska vara optimal.

$$
\begin{aligned}
g_i(\mathbf{x}^*) &\leq 0, & i &= 1, \ldots, n, \\
h_i(\mathbf{x}^*) &= 0, & i &= 1, \ldots, m, \\
\lambda_i^* &\geq 0, & i &= 1, \ldots, n, \\
\lambda_i^* g_i(\mathbf{x}^*) &\geq 0, & i &= 1, \ldots, n,
\end{aligned}
$$

$$
\nabla f(\mathbf{x}^*) + \sum_{i=1}^{n} \lambda_i \nabla g_i(\mathbf{x}^*) + \sum_{i=1}^{m} v_i \nabla h_i(\mathbf{x}^*) = 0.
$$

Definition

Låt X vara ett vektorrum. En *inreprodukt* är en funktion $\langle \cdot, \cdot \rangle : X \times X \longmapsto \mathbb{R}$ sådan att, för alla $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z} \in X$ och alla $\lambda \in \mathbb{R}$, gäller:

IP1 $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$,

IP2 $\langle \lambda \mathbf{x}, \mathbf{y} \rangle = \lambda \langle \mathbf{x}, \mathbf{y} \rangle$,

IP3 $\langle \mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle$,

IP4 $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$ där likhet gäller om och endast om $\mathbf{x} = \mathbf{0}$.

Definition

Den *inducerade normen* $\| \cdot \|_{\mathcal{H}}$ i ett Hilbertrum $\mathcal{H}$ med en inreprodukt $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ definieras genom

$$\|\mathbf{x}\|_{\mathcal{H}} := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{H}}} \qquad \text{där } \mathbf{x} \in \mathcal{H}.$$

Definition

Två vektorer $\mathbf{x},\ \mathbf{y} \in \mathcal{H}$ är *ortogonala* om $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{H}} = 0$. Dessutom är vektorerna *ortonormala* ifall de är både ortogonala och normaliserade det vill säga $\|\mathbf{x}\|_{\mathcal{H}} = \|\mathbf{y}\|_{\mathcal{H}} = 1$.

Definition

För två vektorer $\mathbf{x}, \mathbf{y} \in \mathcal{H}$ olika $\mathbf{0}$. Definiera *komponenten* av $\mathbf{x}$ i $\mathbf{y}$:s riktning, $\mathrm{comp}_{\mathbf{y}}(\mathbf{x})$, som talet

$$\mathrm{comp}_{\mathbf{y}}(\mathbf{x}) = \frac{\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{H}}}{\langle \mathbf{y}, \mathbf{y} \rangle_{\mathcal{H}}}$$

och *projektionen* av $\mathbf{x}$ på $\mathbf{y}$, $\mathrm{proj}_{\mathbf{y}}(\mathbf{x})$, som

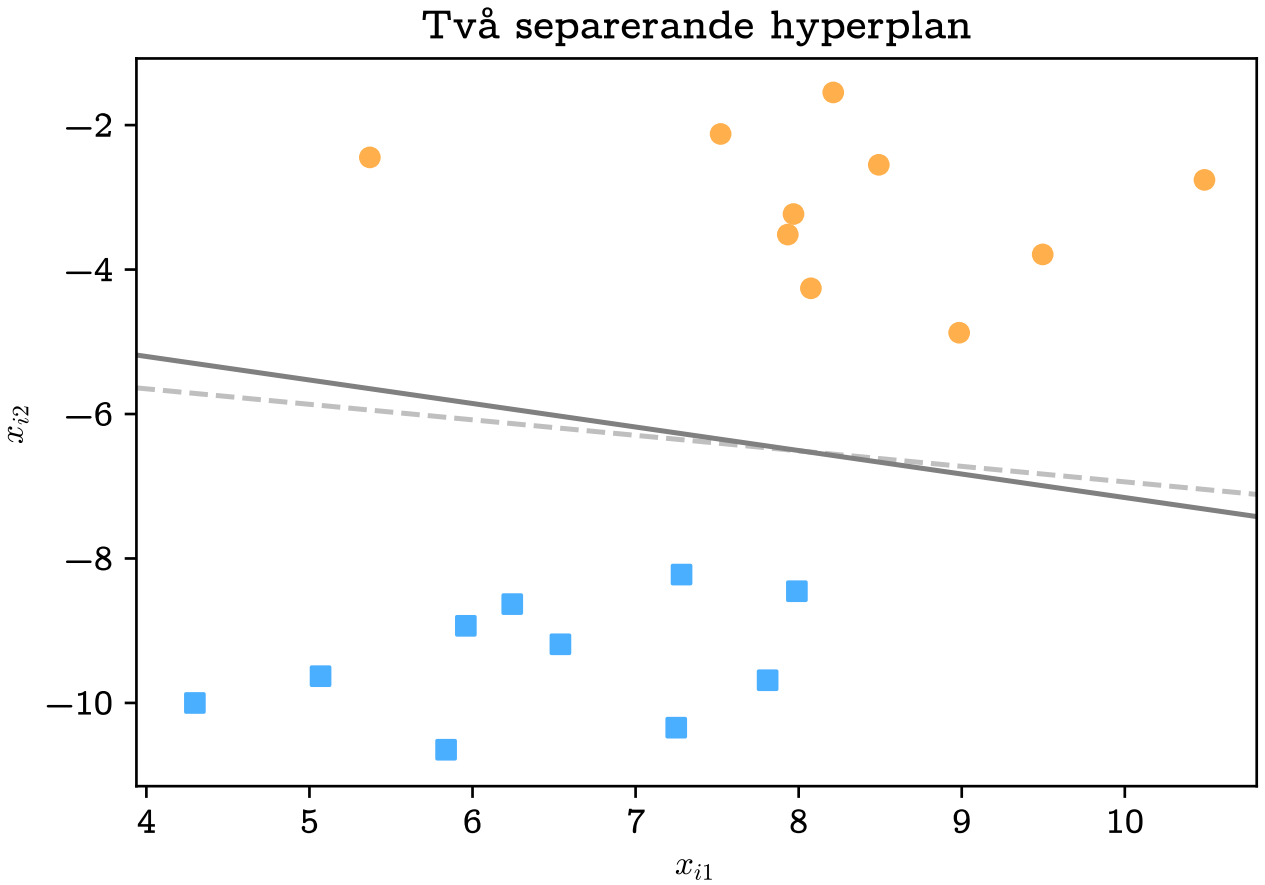
$$\mathrm{proj}_{\mathbf{y}}(\mathbf{x}) = \mathrm{comp}_{\mathbf{y}}(\mathbf{x})\,\mathbf{y} = \frac{\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{H}}}{\langle \mathbf{y}, \mathbf{y} \rangle_{\mathcal{H}}} \mathbf{y}.$$

Definition

Ett *hyperplan* i ett inreproduktrum $\mathcal{H}$ är ett affint underrum av $\mathcal{H}$ definierat som mängden $\{\mathbf{x} : \langle \mathbf{x}, \boldsymbol{\beta} \rangle_{\mathcal{H}} + \beta_0 = 0\}$, med $\mathbf{x}$, $\boldsymbol{\beta} \in X$ och $\beta_0 \in \mathbb{R}$.

Definition

Ett klassificeringsproblem eller en mängd observationspar $(\mathbf{x}_i, y_i)$ är *linjärt separabelt* om det existerar ett hyperplan $L = \{\mathbf{x} : f(\mathbf{x}) = \langle \mathbf{x}, \boldsymbol{\beta} \rangle_{\mathcal{H}} + \beta_0 = 0\}$ sådant att punkten $\mathbf{x}_i$ ligger över hyperplanet om $y_i = 1$ och under om $y_i = -1$. Ett sådant hyperplan kallas ett *separerande hyperplan*.



Figur: 20 datapunkter med två separerande hyperplan (linjer) där klassen $y_i = 1$ framställs som blå fyrkanter och klassen $y_i = -1$ som orangea cirklar.

Tanken är att man vill hitta ett hyperplan sådant att:

- alla observationer klassificeras rätt och,
- hyperplanet samtidigt maximerar det kortaste avståndet från hyperplanet till det närmsta observationsparet.

Matematiskt kan man uttrycka problemet som följande optimeringsproblem

$$\max_{\widehat{\boldsymbol{\beta}},\widehat{\beta}_0,\left\|\widehat{\boldsymbol{\beta}}\right\|_p=1} C$$

$$\text{så att} \quad y_i\left(\left\langle \mathbf{x}_i,\widehat{\boldsymbol{\beta}}\right\rangle_p+\widehat{\beta}_0\right)\geq C, \quad i=1,\ \ldots,\ N$$

där C kallas *marginalen* och betecknar avståndet från hyperplanet till de närmaste observationerna.

Transformera optimeringsproblemet

$$\max_{\widehat{\boldsymbol{\beta}},\widehat{\beta}_0,\left\|\widehat{\boldsymbol{\beta}}\right\|_p=1} \quad C$$
$$\text{så att} \quad y_i\left(\left\langle \mathbf{x}_i,\widehat{\boldsymbol{\beta}}\right\rangle_p+\widehat{\beta}_0\right)\geq C, \quad i=1,\ \ldots,\ N,$$

genom att välja $C=\frac{1}{\|\boldsymbol{\beta}\|_p}$ och kvadrera objektfunktionen, då fås optimeringsproblemet

$$\min_{\boldsymbol{\beta},\beta_0} \quad \frac{1}{2}\left\|\boldsymbol{\beta}\right\|_p^2$$
$$\text{så att} \quad y_i\left(\left\langle \mathbf{x}_i,\boldsymbol{\beta}\right\rangle_p+\beta_0\right)\geq 1, \quad i=1,\ \ldots,\ N.$$

Optimeringsproblemet löses genom Lagrangefunktionen som ges av

$$L_P = \frac{1}{2} \langle \boldsymbol{\beta}, \boldsymbol{\beta} \rangle_p - \sum_{i=1}^{N} \lambda_i \left(y_i \left(\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle_p + \beta_0 \right) - 1 \right)$$

som ska minimeras med avseende på $\boldsymbol{\beta}$ och β_0. Efter differentiering och insättning av nollpunkterna i L_P fås den duala Lagrangefunktionen

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

som ska maximeras med avseende på λ_i, $i = 1, \ldots, N$, och kravet

$$\lambda_i \geq 0, \quad i = 1, \ldots, N.$$

Då måste den optimala lösningen $\boldsymbol{\beta}^*, \beta^*, \lambda_i^*$ uppfylla följande villkor:

$\frac{\partial L_P}{\partial \boldsymbol{\beta}}$ ger

$$\boldsymbol{\beta}^* = \sum_{i=1}^{N} \lambda_i^* y_i \mathbf{x}_i, \tag{1}$$

$\frac{\partial L_P}{\partial \beta}$ ger

$$0 = \sum_{i=1}^{N} \lambda_i^* y_i, \tag{2}$$

från differentieringen av L_P och

$$\lambda_i^* \geq 0, \quad i = 1, \ldots, N, \tag{3}$$

$$\lambda_i^* \left(y_i \left(\langle \mathbf{x}, \boldsymbol{\beta}^* \rangle_p + \beta_0^* \right) - 1 \right) = 0, \quad i = 1, \ldots, N. \tag{4}$$

- Krav (1) säger att vektorn $\boldsymbol{\beta}^*$ är en linjär kombination av vektorerna $\mathbf{x}_i$, $i = 1, \ldots, N$.
- Ifall $\lambda_i^* > 0$ så ger krav (4) att $y_i \left(\langle \mathbf{x}_i, \boldsymbol{\beta}^* \rangle_p + \beta_0^* \right) = 1$. Punkten $\mathbf{x}_i$ är med andra ord en av punkterna som ligger närmast det separerande hyperplanet.
- Ifall $y_i \left(\langle \mathbf{x}_i, \boldsymbol{\beta}^* \rangle_p + \beta_0^* \right) > 1$ så är $\lambda_i^* = 0$ och punkten $\mathbf{x}_i$ är inte en av punkterna som ligger närmast det separerande hyperplanet.
- Parametern β_0^* kan bestämmas genom att man utnyttjar relationen $y_i \left(\langle \mathbf{x}_i, \boldsymbol{\beta}^* \rangle_p + \beta_0^* \right) = 1$ för någon av punkterna där $\lambda_i^* > 0$.

Baserat på de tre tidigare slutsatserna kan man dra slutsatsen att $\boldsymbol{\beta}^*$ är en linjär kombination av endast de punkter $\mathbf{x}_i$ som ligger på randen av marginalen. En punkt som ligger på randen av marginalen kallas *stödvektor*.

Ifall ett optimeringsproblems krav gör det olösbart kan man tillåta lösningar som strider mot kraven och samtidigt försöka reglera hur långt från de ursprungliga kraven man tillåter lösningar. I praktiken åstadkoms detta med hjälp av *slackvariabler* och lösningarna blir (separerande) *hyperplan med mjuka marginaler*.
För optimalt separerande hyperplan finns bara kraven

$$y_i \left(\left\langle \mathbf{x}_i, \widehat{\boldsymbol{\beta}} \right\rangle_p + \widehat{\beta}_0 \right) \geq C, \qquad i = 1, \ \ldots, \ N$$

det vill säga kravet att observationsparen är linjärt separabla.

För det ursprungliga optimeringsproblemet finns två naturliga sätt att ändra på kraven, endera låter man

$$y_i \left(\left\langle \mathbf{x}_i, \widehat{\boldsymbol{\beta}} \right\rangle_p + \widehat{\beta}_0 \right) \geq C - s_i, \tag{5}$$

eller

$$y_i \left(\left\langle \mathbf{x}_i, \widehat{\boldsymbol{\beta}} \right\rangle_p + \widehat{\beta}_0 \right) \geq C \left(1 - s_i\right), \tag{6}$$

där slackvariablerna $s_i \in \mathbb{R}$ är nedåt begränsade av noll samt uppåt begränsade så att summan av alla slackvariabler blir mindre än någon konstant K, det vill säga

$$s_i \geq 0, \qquad i = 1, \ \ldots, N,$$

$$\sum_{i=1}^{N} s_i \leq K.$$

För hyperplan med mjuka marginaler blir det ursprungliga optimeringsproblemet:

$$\begin{aligned} \max_{\widehat{\boldsymbol{\beta}},\widehat{\beta}_0,\left\|\widehat{\boldsymbol{\beta}}\right\|_p=1} \quad & C \\ \text{så att} \quad & y_i\left(\left\langle \mathbf{x}_i,\widehat{\boldsymbol{\beta}}\right\rangle_p+\widehat{\beta}_0\right)\geq C\left(1-s_i\right), \quad i=1,\ \ldots,\ N, \\ & s_i\geq 0, \quad i=1,\ \ldots,\ N, \\ & \sum_{i=1}^{N} s_i \leq K. \end{aligned}$$

Optimeringsproblemet på föregående sida kan arbetas om till följande optimeringsproblem:

$$
\begin{aligned}
\min_{\boldsymbol{\beta},\beta_0} \quad & \frac{1}{2}\left\|\boldsymbol{\beta}\right\|_p^2 \\
\text{så att} \quad & y_i\left(\left\langle \mathbf{x}_i, \boldsymbol{\beta}\right\rangle_p + \beta_0\right) \geq 1 - s_i, \quad i = 1, \ \ldots, \ N, \\
& s_i \geq 0, \quad i = 1, \ \ldots, \ N, \\
& \sum_{i=1}^{N} s_i \leq K.
\end{aligned}
$$

Efter approximering av det sista kravet i objektfunktionen fås optimeringsproblemet:

$$\begin{aligned} \min_{\boldsymbol{\beta},\beta_0} \quad & \frac{1}{2}\left\|\boldsymbol{\beta}\right\|_p^2 + \gamma \sum_{i=1}^{N} s_i \\ \text{så att} \quad & y_i \left(\left\langle \mathbf{x}_i, \boldsymbol{\beta} \right\rangle_p + \beta_0 \right) \geq 1 - s_i, \quad i = 1, \ \ldots, \ N, \\ & s_i \geq 0, \quad i = 1, \ \ldots, \ N, \end{aligned}$$

som alltid är lösbart och kan lösas med Lagrangemultiplikatorer.

Den primala Lagrangefunktionen ges av

$$L_P = \frac{1}{2}\langle \boldsymbol{\beta}, \boldsymbol{\beta} \rangle_p + \gamma \sum_{i=1}^{N} s_i - \sum_{i=1}^{N} \lambda_i \left(y_i \left(\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle_p + \beta_0 \right) - (1 - s_i) \right) - \sum_{i=1}^{N} \mu_i s_i.$$

Efter samma manipulationer som tidigare fås den duala Lagrangefunktionen:

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

som ska maximeras med avseende på λ_i, med kraven $0 \leq \lambda_i \leq \gamma$ och $\sum_{i=1}^{N} \lambda_i y_i = 0$.

Dessutom fås följande krav:

$$\boldsymbol{\beta}^* = \sum_{i=1}^{N} \lambda_i^* y_i \mathbf{x}_i, \tag{7}$$

$$0 = \sum_{i=1}^{N} \lambda_i^* y_i, \tag{8}$$

$$\lambda_i^* = \gamma - \mu_i^* \qquad i = 1, \ldots, N, \tag{9}$$

$$\lambda_i^*, \, \mu_i^*, \, s_i \geq 0, \qquad i = 1, \ldots, N. \tag{10}$$

Och följande krav

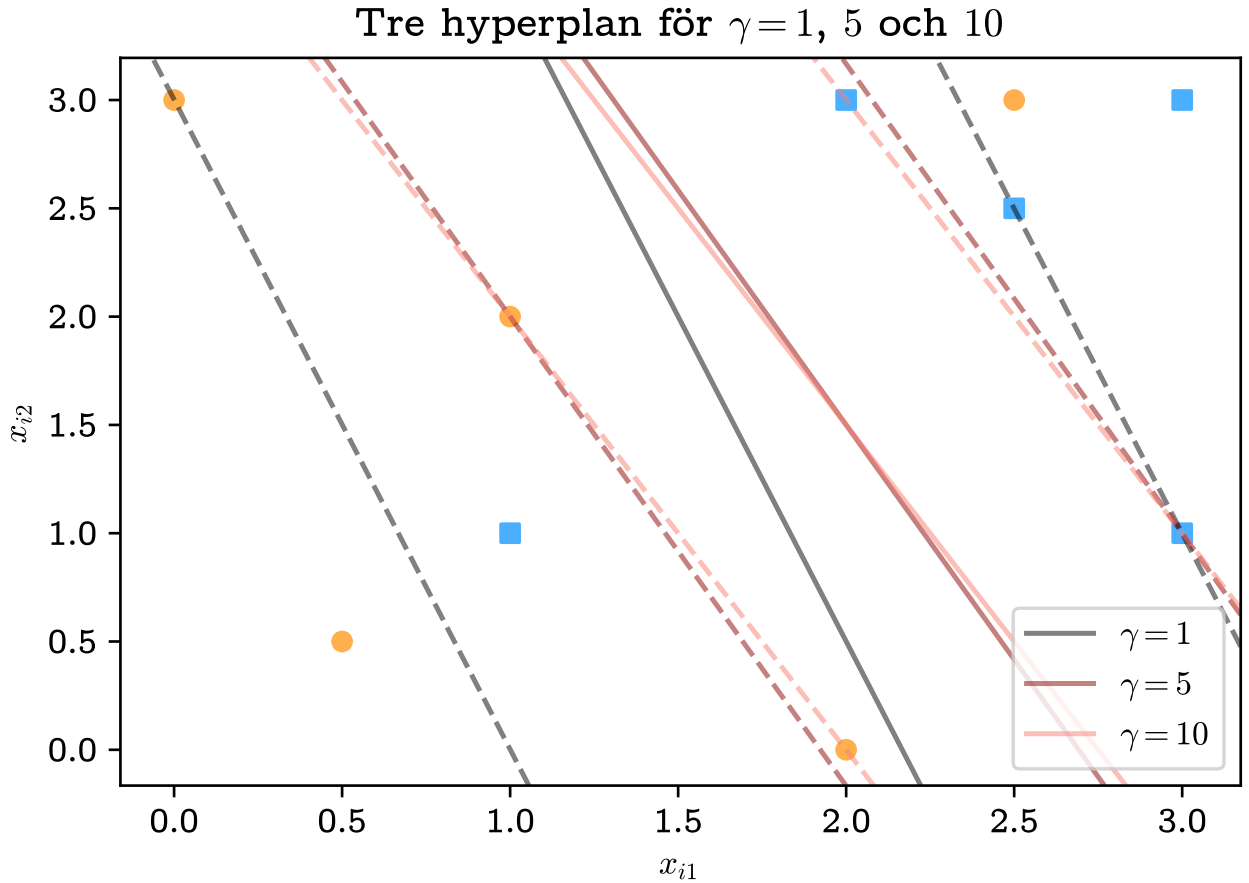
$$\lambda_i^* \left(y_i \left(\langle \mathbf{x}_i, \boldsymbol{\beta}^* \rangle_p + \beta_0^* \right) - (1 - s_i^*) \right) = 0, \qquad i = 1, \ldots, N, \quad (11)$$

$$\mu_i^* s_i^* = 0, \qquad i = 1, \ldots, N, \quad (12)$$

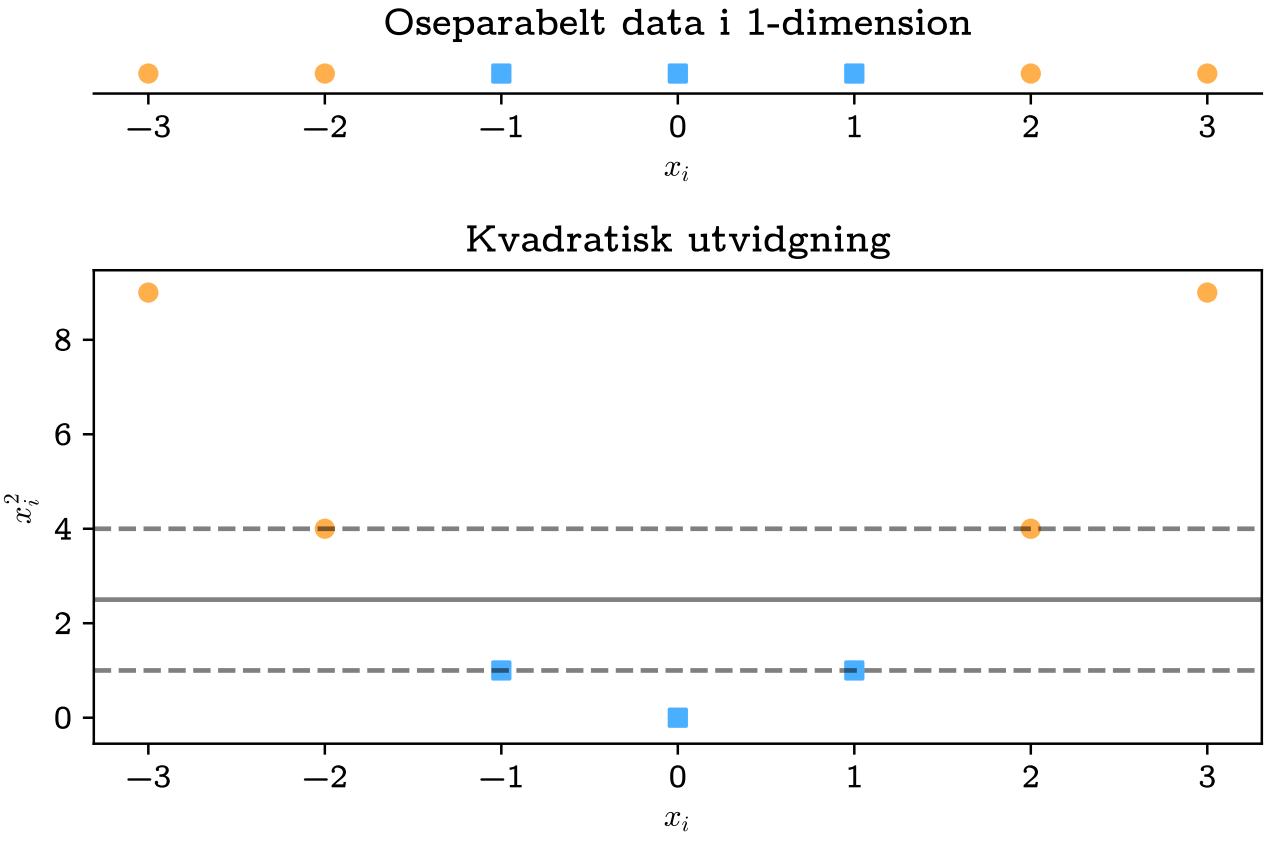
$$y_i \left(\langle \mathbf{x}_i, \boldsymbol{\beta}^* \rangle_p + \beta_0^* \right) - (1 - s_i^*) \geq 0, \qquad i = 1, \ldots, N. \quad (13)$$

Precis som för algoritmen med optimala separerande hyperplan kan man karaktärisera lösningen för hyperplan med mjuka marginaler med hjälp av kraven (7) till (13).

- Krav (7) och (11) ger att den optimala lösningen $\boldsymbol{\beta}^*$ ges som den linjära kombinationen $\boldsymbol{\beta}^* = \sum_{i=1}^{N} \lambda_i^* y_i \mathbf{x}_i$, av punkter $\mathbf{x}_i$ på eller i marginalen. För punkterna på eller i marginalen gäller att $\lambda_i^* > 0$, de kallas *stödvektorer* eftersom de är de enda punkterna som behövs för att representera $\boldsymbol{\beta}^*$.
- För stödvektorer ($\lambda_i^* > 0$) som ligger på marginalen ($s_i^* = 0$) ger kraven (9) och (12) att $0 < \lambda_i^* < \gamma$.
- För de resterande stödvektorerna ($\lambda_i^* > 0$) gäller $\lambda_i^* = \gamma$.
- Vilken som helst av punkterna på marginalen ($\lambda_i^* > 0$, $s_i^* = 0$) kan användas för att lösa för β_0^*.



Figur: Löst exempel för linjärt oseparabelt data för 3 olika värden på γ. De streckade linjerna är marginalernas ränder.



Figur: En lösning med optimala separerande hyperplan och kvadratisk utvidgning där endast hyperplan med mjuka marginaler inte hade fungerat.

Klart är att observationsparen är linjärt oseparabla men nu kan inte heller separerande hyperplan med mjuka marginaler ge vettiga lösningar. Istället kan man lägga till en dimension och definiera att $\mathbf{x}_i \in \mathbb{R}^2$ och $\mathbf{x}_{i2} = \mathbf{x}_{i1}^2$. Då får man situationen som illustreras nederst i föregående figur och observationsparen är nu linjärt separabla. Det optimala separerande hyperplanet bestämdes med hjälp av `sklearn`:s metod `SVC` med `kernel='linear'` och `C=1000`.
Moralen är att hyperplan med mjuka marginaler inte alltid räcker till utan flera verktyg behövs. Ett sådant verktyg är olinjära utvidgningar av det ursprungliga rummet $\mathbf{x}_i \in \mathbb{R}^p$ till ett större rum där det kan vara enklare att hitta vettiga klassificeringsregler.

Betrakta funktionen $\phi : [\mathbf{x}] \longmapsto \left[\mathbf{x}^2, \sqrt{2}\mathbf{x}, 1\right]^\intercal$ där $\mathbf{x} \in \mathbb{R}$, denna funktion motsvarar den olinjära utvidgningen i föregående. Följande inreprodukt mellan två observationer $\mathbf{x}_1$ och $\mathbf{x}_2$ ska beräknas:

$$\begin{aligned}\langle \phi(\mathbf{x}_1), \phi(\mathbf{x}_2) \rangle_3 &= \mathbf{x}_1^2\mathbf{x}_2^2 + 2\mathbf{x}_1\mathbf{x}_2 + 1 \\ &= \langle \mathbf{x}_1, \mathbf{x}_2 \rangle_1^2 + 2\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_1 + 1 \\ &= (\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_1 + 1)^2 = k(\mathbf{x}_1, \mathbf{x}_2)\end{aligned}$$

där $\langle \cdot, \cdot \rangle_p$ är den vanliga inreprodukten i $\mathbb{R}^p$.

Man byter alltså ut inreprodukten $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_3$ mot funktionen $k(\mathbf{x}_i, \mathbf{x}_j) = (\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_1 + 1)^2$ när man löser optimeringsproblemet. I det här fallet gör man inga större besparingar när man räknar ut matrisen $\mathbf{K}_{ij} = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$, eftersom inreprodukten $\left\langle [\mathbf{x}_i^2, \mathbf{x}_i]^\intercal, [\mathbf{x}_j^2, \mathbf{x}_j]^\intercal \right\rangle_2$ endast kräver 3 operationer att beräkna, lika många som funktionen k. Ifall man hade använt femte gradens polynom hade funktionen $k(\mathbf{x}_i, \mathbf{x}_j) = (\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_1 + 1)^5$ krävt mellan 3 och 5 operationer att beräkna (beroende på hur exponentiering är implementerat[1]) medan inreprodukten $\langle \cdot, \cdot \rangle_4$ kräver 7 operationer att räkna. Man behöver dessutom inte räkna ut utvidgningen på förhand.

[1] Antalet operationer blir 3 ifall datorn kan räkna ut a^5 med en operation, annars är $a^5 = a^2 a^2 a$ där man inte behöver räkna ut a^2 två gånger och antalet operationer blir 5.

Definition

Givet en funktion $\phi : \mathbb{R}^p \longmapsto \mathbb{R}^P$ definieras *kärnan* k som funktionen $k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle_P$ där $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$ och $\langle \cdot, \cdot \rangle_P$ är den vanliga inreprodukten i $\mathbb{R}^P$. Vidare om man fixerar ett $\mathbf{y} \in \mathbb{R}^p$ så betecknar vi $\Phi_{\mathbf{y}}(\mathbf{x}) = k(\mathbf{x}, \mathbf{y})$ där $\mathbf{x} \in \mathbb{R}^p$.

Givet en mängd observationer $\mathbf{x}_i \in \mathbb{R}$, $i = 1, \ldots, N$, samt den polynomiella kärnan $k(\mathbf{x}, \mathbf{y}) := (\langle \mathbf{x}, \mathbf{y} \rangle_1 + 1)^2 = (\mathbf{xy})^2 + 2\mathbf{xy} + 1$, kan man definiera ett vektorrum av funktioner genom

$$f(\mathbf{x}) := \sum_{i=1}^{N} \alpha_i k(\mathbf{x}, \mathbf{x}_i) \quad , \alpha_i \in \mathbb{R}.$$

Varje funktion $f(\mathbf{x})$ är alltså en linjär kombination av funktionerna $\Phi_{\mathbf{x}_i}(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}_i)$ där $\mathbf{x}_i$, $i = 1, \ldots, N$ är fixerade.

För en annan funktion $g(\mathbf{x}) := \sum_{j=1}^{m} \beta_j k(\mathbf{x}, \mathbf{x}_j)$ i samma vektorrum kan man definiera inreprodukten

$$\langle f, g \rangle_k := \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \beta_j k(\mathbf{x}_i, \mathbf{x}_j).$$

I beviset för att detta är en inreprodukt stöder man sig på att funktionen k är symmetrisk och positivt semidefinit (eftersom att kärnan är en inreprodukt).
Om man istället definierar en kärna på följande sätt borde beviset fortfarande fungera:

Definition

En kärna är en symmetrisk positivt semidefinit funktion $k : X \times X \longmapsto \mathbb{R}$.

Betrakta funktionen $\Phi_{\mathbf{x}_l}(\mathbf{x}) := k(\mathbf{x}, \mathbf{x}_l)$, där $k(\mathbf{x}, \mathbf{y})$ är en symmetrisk positivt semidefinit funktion. Funktionen $\Phi_{\mathbf{x}_l}(\mathbf{x})$ kan skrivas som $\sum_{j=1}^{N} \beta_j k(\mathbf{x}, \mathbf{x}_l)$ med $\beta_j = 1$ om $j = l$, 0 annars, för att passa in i definitionen för inreprodukten $\langle \cdot, \cdot \rangle_k$. För $\Phi_{\mathbf{x}_l}(\mathbf{x})$ och en annan funktion $f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i k(\mathbf{x}, \mathbf{x}_i)$ gäller då att

$$\langle \Phi_{\mathbf{x}_l}, f \rangle_k = \langle k(\cdot, \mathbf{x}_i), f \rangle_k = \sum_{j=1}^{N} \sum_{i=1}^{N} \beta_j \alpha_i k(\mathbf{x}_j, \mathbf{x}_i)$$

där $\beta_j = 0$ om $j \neq l$ och 1 om $j = l$, det vill säga

$$\langle \Phi_{\mathbf{x}_l}, f \rangle_k = \sum_{i=1}^{N} \alpha_i k(\mathbf{x}_l, \mathbf{x}_i) = f(\mathbf{x}_l).$$

Tolkningen är att inreprodukten mellan en funktion f och $k(\cdot, \mathbf{x}_i)$ är samma sak som evaluering av funktionen f i punkten $\mathbf{x}_i$ men $\mathbf{x}_i$ måste vara en av de punkterna som man byggde upp inreprodukten av.

Speciellt för en annan funktion $\Phi_{\mathbf{x}_h}(\mathbf{x}) := k(\mathbf{x}, \mathbf{x}_h)$ fås att

$$\langle \Phi_{\mathbf{x}_l}, \Phi_{\mathbf{x}_h} \rangle_k = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \beta_j k(\mathbf{x}_i, \mathbf{x}_j)$$

där $\alpha_i = \beta_j = 1$ endast om $i = l$ och $j = h$, $\alpha_i = \beta_j = 0$ annars. Då fås

$$\begin{aligned} \langle \Phi_{\mathbf{x}_l}, \Phi_{\mathbf{x}_h} \rangle_k &= \langle k(\cdot, \mathbf{x}_l), k(\cdot, \mathbf{x}_h) \rangle_k \\ &= \alpha_l \beta_h k(\mathbf{x}_l, \mathbf{x}_h) \\ &= k(\mathbf{x}_l, \mathbf{x}_h). \end{aligned}$$

Här är tolkningen att även om $k(\mathbf{x}_l, \mathbf{x}_h)$ inte är definierad genom en inreprodukt så kan k tolkas som en inreprodukt i något (olinjärt) rum. De här två egenskaperna är orsaken till att man pratar om *reproducerande* kärnor, den första ekvationen brukar också ibland användas som definitionen på en reproducerande kärna.

Med andra ord borde man givet en symmetrisk positivt semidefinit funktion k kunna skriva den som en inreprodukt i något rum. Låt $\mathbf{x}_i$, $\mathbf{x}_j \in \mathbb{R}$ och betrakta funktionen

$$
\begin{aligned}
k\left(\mathbf{x}_i, \mathbf{x}_j\right) &= e^{-\frac{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_1^2}{2\sigma^2}} \\
&= e^{-\frac{\left\langle \mathbf{x}_i - \mathbf{x}_j, \mathbf{x}_i - \mathbf{x}_j \right\rangle_1}{2\sigma^2}} \\
&= e^{-\frac{\mathbf{x}_i^2 + \mathbf{x}_j^2 - 2\mathbf{x}_i\mathbf{x}_j}{2\sigma^2}} \\
&= e^{-\frac{\mathbf{x}_i^2 + \mathbf{x}_j^2}{2\sigma^2}} \cdot e^{\frac{2\mathbf{x}_i\mathbf{x}_j}{2\sigma^2}} \\
&= e^{-\left(\frac{\mathbf{x}_i}{\sqrt{2}\sigma}\right)^2} \cdot e^{-\left(\frac{\mathbf{x}_j}{\sqrt{2}\sigma}\right)^2} \cdot e^{\frac{\mathbf{x}_i\mathbf{x}_j}{\sigma^2}}.
\end{aligned}
$$

Genom Taylorutvecklingen för e^z fås

$$\begin{aligned} k(\mathbf{x}_i, \mathbf{x}_j) &= e^{-\left(\frac{\mathbf{x}_i}{\sqrt{2}\sigma}\right)^2} \cdot e^{-\left(\frac{\mathbf{x}_j}{\sqrt{2}\sigma}\right)^2} \cdot \sum_{n=0}^{\infty} \frac{\mathbf{x}_i^n \mathbf{x}_j^n}{n!} \\ &= \sum_{n=0}^{\infty} \left(\frac{e^{-\left(\frac{\mathbf{x}_i}{\sqrt{2}\sigma}\right)^2} \mathbf{x}_i^n}{\sqrt{n!}} \right) \left(\frac{e^{-\left(\frac{\mathbf{x}_j}{\sqrt{2}\sigma}\right)^2} \mathbf{x}_j^n}{\sqrt{n!}} \right) \\ &= \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\ell^2} \end{aligned}$$

där $\langle \mathbf{x}, \mathbf{y} \rangle_{\ell^2}$ är inreprodukten given av summan $\sum_{h=1}^{\infty} \mathbf{x}_h \mathbf{y}_h$ om den är ändlig. Här betecknar $\mathbf{x}_h$ och $\mathbf{y}_h$ de h:te komponenterna av de oändligtdimensionella vektorerna $\mathbf{x}$ och $\mathbf{y}$. Summan konvergerar om $\mathbf{x}$ och $\mathbf{y}$ tillhör Hilbertrummet ℓ^2 det vill säga rummet av alla följder $\mathbf{x}$ för vilka summan $\sum_{h=1}^{\infty} \mathbf{x}_h^2 = \langle \mathbf{x}, \mathbf{x} \rangle_{\ell^2} = \|\mathbf{x}\|_{\ell^2}^2$ är ändlig.

Kärnan $k(\mathbf{x}, \mathbf{y}) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}}$ kan med andra ord skrivas som inreprodukten $\langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle_{\ell^2}$ i rummet som ges av utvidgningen

$$\phi(\mathbf{x}) = \left[\frac{e^{-\left(\frac{\mathbf{x}}{\sqrt{2}\sigma}\right)^2} \mathbf{x}^0}{\sqrt{0!}}, \frac{e^{-\left(\frac{\mathbf{x}}{\sqrt{2}\sigma}\right)^2} \mathbf{x}^1}{\sqrt{1!}}, \frac{e^{-\left(\frac{\mathbf{x}}{\sqrt{2}\sigma}\right)^2} \mathbf{x}^2}{\sqrt{2!}}, \dots \right]^\mathsf{T}.$$

Implikationen är att motsvarande olinjära transformation skulle ge ett oändligtdimensionellt rum att jobba med om man gör transformationen direkt medan man genom kärnan k implicit kan jobba i ett oändligtdimensionellt rum, något som inte hade varit möjligt om man försökte operera med vanliga inreprodukter på det utvidgade rummet.