

# Hilbertrum med Reproducerande Kärnor

Oscar Granlund

22 augusti 2018

## Sammanfattning

Testtesttesttesttest

# Kapitel 1

## Stödvektormaskiner (SVM)

### 1.1 Klassificering med hjälp av separerande hyperplan

INTRODUKTION OM VARFÖR KLASSIFICERING, EXEMPEL MED SPAM-FILTER

**Definition 1.1.1.** Ett *klassificeringsproblem* är ett problem var man utgående från en mängd observationspar (*träningsdata*)  $(\mathbf{x}_i, y_i)$ ,  $\mathbf{x}_i \in \mathbb{R}^p$ ,  $y_i \in \{-1, 1\}$ ,  $i = 1, \dots, N$ , försöker hitta en regel  $g : \mathbb{R}^p \mapsto \{-1, 1\}$  sådan att  $g(\mathbf{x}_i) = y_i$  för så många träningspar  $(\mathbf{x}_i, y_i)$  som möjligt.

Inom statistiken och maskininläringen finns många olika metoder för att försöka lösa klassificeringsproblem, till exempel med hjälp av regressionsmodeller eller klusteranalys. I detta kapitel behandlas en metod där en affin mängd med dimensionen  $p - 1$  används för att definiera en regel som klassificerar *observationerna*  $\mathbf{x}_i$  i *klasserna*  $y_i \in \{-1, 1\}$  genom separering.

**Definition 1.1.2.** Ett *hyperplan* i ett vektorrum med dimensionen  $p$  är ett underrum med dimensionen  $p - 1$ , i figur 1.1 illustreras ett separerande hyperplan för fallet  $p = 2$ . Klassificeringsregeln  $g$  för separerande hyperplan blir  $g(\mathbf{x}_i) = \text{sign}(\mathbf{x}_i^\top \boldsymbol{\beta} + \beta_0)$  där mängden  $\{\mathbf{x} : \mathbf{x}^\top \boldsymbol{\beta} + \beta_0 = 0\}$ , med  $\mathbf{x}, \boldsymbol{\beta} \in \mathbb{R}^p$ , definierar ett hyperplan parametriserad av  $\boldsymbol{\beta}$  och  $\beta_0$ .

**Sats 1.1.1.** Ett hyperplan definierat som den affina mängden  $L = \{\mathbf{x} : f(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta} + \beta_0 = 0\}$  har följande egenskaper [2]:

1.  $\boldsymbol{\beta}$  är en normalvektor till  $L$  och kan normaliseras genom

$$\hat{\boldsymbol{\beta}} = \frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|}.$$

2.  $\mathbf{x}_0^\top \boldsymbol{\beta} = -\beta_0$  för alla  $\mathbf{x}_0$  i  $L$ .

3. Det signerade avståndet från en punkt  $\mathbf{x}$  till hyperplanet  $L$ ,  $d^\pm(\mathbf{x}, L)$ , ges av

$$\begin{aligned} (\mathbf{x} - \mathbf{x}_0)^\top \hat{\boldsymbol{\beta}} &= \frac{1}{\|\boldsymbol{\beta}\|} (\mathbf{x}^\top \boldsymbol{\beta} + \beta_0) \\ &= \frac{1}{\|f'(\mathbf{x})\|} f(\mathbf{x}). \end{aligned}$$

*Bevis.*

1. Låt  $\mathbf{x}_1$  och  $\mathbf{x}_2$  vara två punkter i  $L$ . Då gäller att  $f(\mathbf{x}_1) = f(\mathbf{x}_2) = 0$  och

$$\begin{aligned} 0 &= f(\mathbf{x}_1) - f(\mathbf{x}_2) \\ &= \mathbf{x}_1^\top \boldsymbol{\beta} + \beta_0 - \mathbf{x}_2^\top \boldsymbol{\beta} - \beta_0 \\ &= (\mathbf{x}_1 - \mathbf{x}_2)^\top \boldsymbol{\beta} \end{aligned}$$

alltså uppfyller  $\boldsymbol{\beta}$  kravet för normalvektorer och  $\hat{\boldsymbol{\beta}} := \frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|}$  är den normaliserade normalvektorn till hyperplanet  $L$ . ■

2. Låt  $\mathbf{x}_0$  vara en punkt i  $L$ . Då gäller att  $f(\mathbf{x}_0) = \mathbf{x}_0^\top \boldsymbol{\beta} + \beta_0 = 0$  alltså är  $\mathbf{x}_0^\top \boldsymbol{\beta} = -\beta_0$ . ■

3. Låt  $\mathbf{x}_0$  vara en punkt i hyperplanet  $L$ . Då är avståndet från hyperplanet till punkten  $\mathbf{x}$  lika med längden av projektionen av vektorn  $(\mathbf{x} - \mathbf{x}_0)$  på hyperplanet normaliserade normal  $\hat{\boldsymbol{\beta}}$ ,  $\text{comp}_{\hat{\boldsymbol{\beta}}}(\mathbf{x} - \mathbf{x}_0)$ . Vi får alltså att

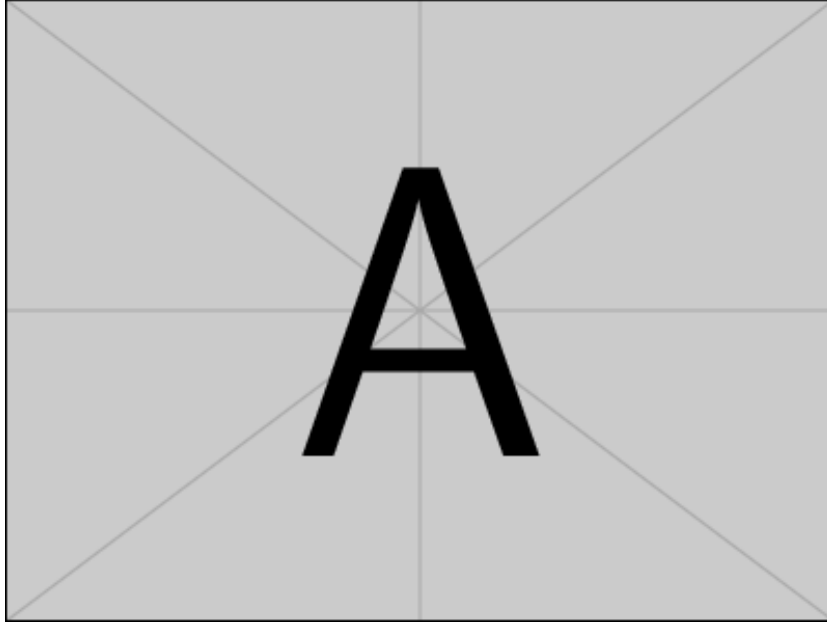
$$\begin{aligned} d^\pm(\mathbf{x}, L) &= \text{comp}_{\hat{\boldsymbol{\beta}}}(\mathbf{x} - \mathbf{x}_0) = \frac{(\mathbf{x} - \mathbf{x}_0)^\top \hat{\boldsymbol{\beta}}}{\|\hat{\boldsymbol{\beta}}\|} \\ &= \frac{1}{\|\boldsymbol{\beta}\|} (\mathbf{x}^\top \boldsymbol{\beta} - \mathbf{x}_0^\top \boldsymbol{\beta}) = \frac{1}{\|\boldsymbol{\beta}\|} (\mathbf{x}^\top \boldsymbol{\beta} + \beta_0) \end{aligned}$$

och om man noterar att  $f(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta} + \beta_0$  och  $f'(\mathbf{x}) = \boldsymbol{\beta}$  så fås även att

$$\frac{1}{\|\boldsymbol{\beta}\|} (\mathbf{x}^\top \boldsymbol{\beta} + \beta_0) = \frac{1}{\|f'(\mathbf{x})\|} f(\mathbf{x}).$$

■

*Observation.* Definitionen  $L = \{\mathbf{x} : f(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta} + \beta_0 = 0\}$  för hyperplanet  $L$  är inte entydig.



Figur 1.1: 20 datapunkter med ett separerande hyperplan (linje) där klassen  $y = 1$  har färgats blå och klassen  $y = -1$  har färgats orange.

*Orsak.* Betrakta hyperplanen  $L_1 = \{\mathbf{x} : f(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta} + \beta_0 = 0\}$  och  $L_2 = \{\mathbf{x} : g(\mathbf{x}) = \mathbf{x}^\top (-\boldsymbol{\beta}) + (-\beta_0) = 0\}$ . Eftersom att  $g(\mathbf{x}) = -f(\mathbf{x})$  så gäller att om  $\mathbf{x}_0$  tillhör  $L_1$  så tillhör  $\mathbf{x}_0$  även  $L_2$ . Betrakta vidare hyperplanet  $L_3 = \{\mathbf{x} : h(\mathbf{x}) = \frac{\mathbf{x}^\top \boldsymbol{\beta}}{\|\boldsymbol{\beta}\|} + \frac{\beta_0}{\|\boldsymbol{\beta}\|} = 0\}$ . Om  $\mathbf{x}_0$  då tillhör  $L_1$  så tillhör  $\mathbf{x}_0$  även  $L_3$  eftersom att  $h(\mathbf{x}) = \frac{f(\mathbf{x})}{\|\boldsymbol{\beta}\|} = 0$ . Notera även att  $\frac{1}{\|\boldsymbol{\beta}\|}$  kunde ha varit vilket reellt tal som helst.

*Observation.* För att få entydiga hyperplan för klassificering kan man lägga till villkor. Om man kräver att  $\|\boldsymbol{\beta}\| = 1$  och  $y_i(\mathbf{x}_i^\top \boldsymbol{\beta} + \beta_0) \geq 0$  för alla  $i = 1, \dots, N$ , där  $y_i$  är klasserna i klassificeringsproblemet, så får man en entydig definition av hyperplanet där vektorn  $\boldsymbol{\beta}$  ”pekar mot” klassen där  $y_i = 1$  och  $\beta_0$  anger det signerade avståndet (i relation till riktningen på  $\boldsymbol{\beta}$ ) från origo till hyperplanet.

*Orsak.* De extra villkoren gör att man inte längre kan göra manipulationerna som påvisade icke-entydigheten. Om man sätter  $\mathbf{x} = \mathbf{0}$  så får man med hjälp av sats 1.1.1 att avståndet från origo till planet är lika med  $\frac{1}{\|\boldsymbol{\beta}\|}(\mathbf{x}^\top \boldsymbol{\beta} + \beta_0) = \frac{1}{\|\boldsymbol{\beta}\|}(\mathbf{0}^\top \boldsymbol{\beta} + \beta_0) = \beta_0$ .

**Definition 1.1.3.** Ett klassificeringsproblem eller en mängd observationspar  $(\mathbf{x}_i, y_i)$  är *linjärt separabelt* om det existerar ett hyperplan  $L = \{\mathbf{x} : f(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta} + \beta_0 = 0\}$  som separerar klasserna.

**Sats 1.1.2.** [1] För ett hyperplan  $L = \{\mathbf{x} : f(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta} + \beta_0 = 0\}$  som separerar två klasser gäller att

$$y_i(\mathbf{x}_i^\top \boldsymbol{\beta} + \beta_0) > 0$$

eller

$$y_i(\mathbf{x}_i^\top \boldsymbol{\beta} + \beta_0) < 0$$

för alla  $i = 1, \dots, N$ .

*Bevis.* Ifall ett klassificeringsproblem är linjärt separabelt så ligger alla observationer  $y_i$  på rätt sida av hyperplanet definierat genom  $\mathbf{x}^\top \boldsymbol{\beta} + \beta_0$ ; eller så ligger alla observationer på fel sida av hyperplanet. Vilket betyder att ifall  $y_i = 1$  så är  $\mathbf{x}_i^\top \boldsymbol{\beta} + \beta_0 > 0$  och om  $y_i = -1$  så är  $\mathbf{x}_i^\top \boldsymbol{\beta} + \beta_0 < 0$ . Detta betyder att  $y_i(\mathbf{x}_i^\top \boldsymbol{\beta} + \beta_0) > 0$ . Ifall  $\mathbf{x}_i^\top \boldsymbol{\beta} + \beta_0 = 0$  är problemet inte linjärt separabelt. ■

**Exempel 1.1.1.** Låt observationsparen vara  $([2, 2]^\top, 1)$ ,  $([1, 2]^\top, -1)$ . Då är

$$L_1 = \{\mathbf{x} \in \mathbb{R}^2 : \mathbf{x}^\top \begin{bmatrix} 1 \\ 0 \end{bmatrix} - 1.5 = 0\}$$

och

$$L_2 = \{\mathbf{x} \in \mathbb{R}^2 : \mathbf{x}^\top \begin{bmatrix} \sqrt{2} \\ \sqrt{2} \end{bmatrix} - 3.5\sqrt{2} = 0\}$$

två separerande hyperplan (linjer i detta fall).

*Bevis.* För  $L_1$ :

$$y_1(\mathbf{x}_1^\top \begin{bmatrix} 1 \\ 0 \end{bmatrix} - 1.5) = [2, 2]^\top \begin{bmatrix} 1 \\ 0 \end{bmatrix} - 1.5 = 0.5 > 0$$

och

$$y_2(\mathbf{x}_2^\top \begin{bmatrix} 1 \\ 0 \end{bmatrix} - 1.5) = -1([1, 2]^\top \begin{bmatrix} 1 \\ 0 \end{bmatrix} - 1.5) = (-1)(-0.5) = 0.5 > 0.$$

Och för  $L_2$ :

$$y_1(\mathbf{x}_1^\top \begin{bmatrix} \sqrt{2} \\ \sqrt{2} \end{bmatrix} - 3.5\sqrt{2}) = [2, 2]^\top \begin{bmatrix} \sqrt{2} \\ \sqrt{2} \end{bmatrix} - 3.5\sqrt{2} = 0.5\sqrt{2} > 0$$

och

$$y_2(\mathbf{x}_2^\top \begin{bmatrix} \sqrt{2} \\ \sqrt{2} \end{bmatrix} - 3.5\sqrt{2}) = -1([1, 2]^\top \begin{bmatrix} \sqrt{2} \\ \sqrt{2} \end{bmatrix} - 3.5\sqrt{2}) = (-1)(-0.5\sqrt{2}) = 0.5\sqrt{2} > 0$$

■

*Observation.* Hyperplan kan konstrueras enkelt genom att man i  $\mathbb{R}^p$  väljer  $p$  stycken punkter  $\mathbf{x}_i$  som man vill att planet ska gå igenom, sedan löser man ekvationssystemet  $X\boldsymbol{\beta} = -\beta_0\mathbf{1}$ , i vilket  $X$  är en matris där raderna består av punkterna  $\mathbf{x}_i$ ,  $i = 1, \dots, p$ , och  $\beta_0\mathbf{1}$  är en vektor med värdet  $\beta_0$  i alla rader.

Som syns i exempel 1.1.1 finns det många separerande hyperplan om ett klassificeringsproblem är linjärt separabelt och frågan är då vilket separerande hyperplan man borde välja.

## 1.2 Optimala separerande hyperplan

Inom statistiken finns många olika metoder som en modell till data och metoderna kan ofta visas vara ekvivalenta med något optimeringsproblem, till exempel maximum likelihood-metoden (ML-metoden) för linjär regression, som är ekvivalent med minstakvadratmetoden. Optimeringsproblemen kan oftast ändras genom att man lägger till eller tar bort termer i objektivfunktionen eller ändrar på kraven.

För separerande hyperplan kommer jag att behandla ett optimeringsproblem som är utformat så att det kortaste avståndet från hyperplanet till de närmaste observationsparen från vardera klass maximeras [?]. Med andra ord fås följande optimeringsproblem

$$\begin{aligned} \max_{\hat{\boldsymbol{\beta}}, \hat{\beta}_0, \|\hat{\boldsymbol{\beta}}\|=1} \quad & C \\ \text{så att} \quad & y_i(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}} + \hat{\beta}_0) \geq C, \quad i = 1, \dots, N \end{aligned} \tag{1.1}$$

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

*Observation.* Ifall alla punkter är rätt klassificerade så anger  $y_i(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}} + \hat{\beta}_0)$  det absoluta avståndet mellan hyperplanet och punkten  $\mathbf{x}_i$ .

Förhoppningen är att om man väljer det separerande hyperplan som befinner sig så långt som möjligt från båda klasserna får ett hyperplan som även generaliserar väl till ny data. Dessutom är detta även ett unikt sätt att välja ett separerande hyperplan det vill säga optimeringsproblemet är konvext.

För att visa att optimeringsproblemet (1.1) är *konvext* måste det skrivas om. Idén är här att man låter inversen av längden på vektorn  $\boldsymbol{\beta}$  beskriva avståndet till närmast punkt. På så sätt skapas en direktare länk mellan kraven och objektfunktionen i optimeringsproblemet.

Först måste alltså kravet  $\|\hat{\boldsymbol{\beta}}\| = 1$  bytas ut. Detta görs genom att man byter ut kraven

$$y_i(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}} + \hat{\beta}_0) \geq C, \quad i = 1, \dots, N$$

mot kraven

$$y_i \left( \mathbf{x}_i^\top \frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|} + \frac{\beta_0}{\|\boldsymbol{\beta}\|} \right) = \frac{1}{\|\boldsymbol{\beta}\|} y_i (\mathbf{x}_i^\top \boldsymbol{\beta} + \beta_0) \geq C, \quad i = 1, \dots, N$$

eller ekvivalent

$$y_i (\mathbf{x}_i^\top \boldsymbol{\beta} + \beta_0) \geq C \|\boldsymbol{\beta}\|, \quad i = 1, \dots, N,$$

där man valt en av de andra representationerna för samma hyperplan genom att skala om  $\hat{\boldsymbol{\beta}}$  och  $\hat{\beta}_0$ . Vidare kan  $C$  elimineras genom att man väljer  $C = \frac{1}{\|\boldsymbol{\beta}\|}$ , då fås

$$y_i (\mathbf{x}_i^\top \boldsymbol{\beta} + \beta_0) \geq 1, \quad i = 1, \dots, N$$

och eftersom  $C = \frac{1}{\|\boldsymbol{\beta}\|}$  är en avtagande funktion med avseende på  $\|\boldsymbol{\beta}\|$  är maximering av  $C$  ekvivalent med minimering av  $\|\boldsymbol{\beta}\|$  och motsvarande optimeringsproblemet blir

$$\begin{aligned} & \min_{\boldsymbol{\beta}, \beta_0} \|\boldsymbol{\beta}\| \\ & \text{så att} \quad y_i (\mathbf{x}_i^\top \boldsymbol{\beta} + \beta_0) \geq 1, \quad i = 1, \dots, N. \end{aligned}$$

Därefter görs ännu en konvexitetsbevarande kvadratisk transformering av *kostfunktionen*  $\|\boldsymbol{\beta}\|$ , det vill säga man noterar att  $\operatorname{argmin}_{\boldsymbol{\beta}, \beta_0} \|\boldsymbol{\beta}\| = \operatorname{argmin}_{\boldsymbol{\beta}, \beta_0} \frac{1}{2} \|\boldsymbol{\beta}\|^2$ . Med andra ord är detta ett optimeringsproblem med kvadratisk objektfunktion och linjära krav det vill säga ett konvext optimeringsproblem med entydig lösning.

*Observation.* För två vektorer  $\mathbf{a}$  och  $\mathbf{b}$  i  $\mathbb{R}^p$  kan produkten  $\mathbf{a}^\top \mathbf{b}$  uttryckas som den normala inre produkten  $\langle \mathbf{a}, \mathbf{b} \rangle$  i  $\mathbb{R}^p$ . Detta kommer till nytta i kapitel 2 där konvexiteten för en generalisering av det linjära problemet utforskas.

Ovanstående resonemang är ett bevis för sats 1.2.1.

**Sats 1.2.1.** Låt  $\hat{\boldsymbol{\beta}}, \boldsymbol{\beta} \in \mathbb{R}^p$  och  $\hat{\beta}_0, \beta_0 \in \mathbb{R}$ . Låt dessutom observationsparen  $(\mathbf{x}_i, y_i)$  vara linjärt separabla. Då är optimeringsproblemet

$$\begin{aligned} & \max_{\hat{\boldsymbol{\beta}}, \hat{\beta}_0, \|\hat{\boldsymbol{\beta}}\|=1} C \\ & \text{så att} \quad y_i (\mathbf{x}_i^\top \hat{\boldsymbol{\beta}} + \hat{\beta}_0) \geq C, \quad i = 1, \dots, N \end{aligned}$$

konvext och ekvivalent med optimeringsproblemet

$$\begin{aligned} & \min_{\boldsymbol{\beta}, \beta_0} \frac{1}{2} \|\boldsymbol{\beta}\|^2 \\ & \text{så att} \quad y_i (\mathbf{x}_i^\top \boldsymbol{\beta} + \beta_0) \geq 1, \quad i = 1, \dots, N. \end{aligned}$$



Vidare ger observationen ovan ytterligare ett ekvivalent optimeringsproblem:

**Korollarium 1.2.2.** Optimeringsproblemen i sats 1.2.1 är ekvivalenta med optimeringsproblemet

$$\begin{aligned} \min_{\beta, \beta_0} \quad & \frac{1}{2} \langle \beta, \beta \rangle \\ \text{så att} \quad & y_i (\langle \mathbf{x}_i, \beta \rangle + \beta_0) \geq 1, \quad i = 1, \dots, N \end{aligned}$$

med samma antaganden.

*Bevis.* Normen  $\|\beta\| = (\beta^\top \beta)^{\frac{1}{2}}$  kan uttryckas som  $\langle \beta, \beta \rangle^{\frac{1}{2}}$  alltså är  $\frac{1}{2} \|\beta\|^2 = \frac{1}{2} \langle \beta, \beta \rangle$ . Resten följer från observationen. ■

Framställningen i korollarium 1.2.2 används i många källor, bland annat i den ursprungliga framställningen för stödvektormaskinen, och är en av de mer generella framställningarna för optimeringsproblemet som stödvektormaskinen bygger på.

## 1.2.1 Primala och duala problem

För att hitta alla extrempunkter till ett optimeringsproblem, det vill säga lösa ett konvext optimeringsproblem, används Lagrangemultiplikatorer. Den primala Lagrangefunktionen  $L_P$  för optimeringsproblemet

$$\begin{aligned} \min_{\beta, \beta_0} \quad & \frac{1}{2} \|\beta\|^2 \\ \text{så att} \quad & y_i (\mathbf{x}_i^\top \beta + \beta_0) \geq 1, \quad i = 1, \dots, N \end{aligned}$$

ges då av

$$L_P = \frac{1}{2} \|\beta\|^2 - \sum_{i=1}^N \lambda_i (y_i (\mathbf{x}_i^\top \beta + \beta_0) - 1) \quad (1.2)$$

som ska minimeras med avseende på  $\beta$  och  $\beta_0$ .

För att minimera  $L_P$  sätts derivatorna med avseende på elementen  $[\beta]_j$

av  $\beta$  och  $\beta_0$  till 0, och följande relationer erhålls:

$$\begin{aligned}
D_{[\beta]_j}(L_P) &= D_{[\beta]_j} \left( \frac{1}{2} \beta^\top \beta \right) - D_{[\beta]_j} \left( \sum_{i=1}^N (\lambda_i y_i (\mathbf{x}_i^\top \beta) + \lambda_i y_i \beta_0 - \lambda_i) \right) \\
&= D_{[\beta]_j} \left( \frac{1}{2} \sum_{k=1}^p [\beta]_k^2 \right) - \sum_{i=1}^N D_{[\beta]_j} \left( \lambda_i y_i \left( \sum_{k=1}^p [\mathbf{x}_i]_k [\beta]_k \right) + \lambda_i y_i \beta_0 - \lambda_i \right) \\
&= [\beta]_j - \sum_{i=1}^N D_{[\beta]_j} \left( \sum_{k=1}^p \lambda_i y_i [\mathbf{x}_i]_k [\beta]_k \right) + 0 \\
&= [\beta]_j - \sum_{i=1}^N \lambda_i y_i [\mathbf{x}_i]_j
\end{aligned} \tag{1.3}$$

där  $j = 1, \dots, p$  och

$$D_{\beta_0}(L_P) = D_{\beta_0} \left( - \sum_{i=1}^N \lambda_i y_i \beta_0 \right) = - \sum_{i=1}^N \lambda_i y_i.$$

Vidare kan 1.3 skrivas om som derivatan med avseende på hela  $\beta$  eftersom att  $[D_\beta(L_P)]_j = D_{[\beta]_j}(L_P)$ . Efter att man tar i beaktande kraven att  $D_\beta(L_P) = \mathbf{0}$  och  $D_{\beta_0}(L_P) = 0$  fås följande krav:

$$\beta = \sum_{i=1}^N \lambda_i y_i \mathbf{x}_i \tag{1.4}$$

$$0 = \sum_{i=1}^N \lambda_i y_i. \tag{1.5}$$

Efter omskrivning av  $\|\beta\|^2$  som  $\beta^\top \beta$  ger insättning av kraven 1.4 och 1.5 i  $L_P$  följande duala problem

$$\begin{aligned}
L_D &= \frac{1}{2} \left( \sum_{i=1}^N \lambda_i y_i \mathbf{x}_i \right)^\top \left( \sum_{j=1}^N \lambda_j y_j \mathbf{x}_j \right) \\
&\quad - \sum_{i=1}^N \lambda_i \left( y_i \left( \mathbf{x}_i^\top \left( \sum_{j=1}^N \lambda_j y_j \mathbf{x}_j \right) + \beta_0 \right) - 1 \right) \\
&= \frac{1}{2} \sum_{i=1}^N \lambda_i y_i \mathbf{x}_i^\top \left( \sum_{j=1}^N \lambda_j y_j \mathbf{x}_j \right) \\
&\quad - \sum_{i=1}^N \lambda_i y_i \mathbf{x}_i^\top \left( \sum_{j=1}^N \lambda_j y_j \mathbf{x}_j \right) - \beta_0 \sum_{i=1}^N \lambda_i y_i + \sum_{i=1}^N \lambda_i \\
&= - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \lambda_i \lambda_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j + \sum_{i=1}^N \lambda_i \quad \left( \sum_{i=1}^N \lambda_i y_i = 0 \right)
\end{aligned}$$

som ska maximeras med avseende på  $\lambda_i$ ,  $i = 1, \dots, N$ , och kravet

$$\lambda_i \geq 0, \quad i = 1, \dots, N. \quad (1.6)$$

Uträkningarna och kravet  $\lambda_i \geq 0$ ,  $i = 1, \dots, N$ , kan motiveras genom Karush-Kuhn-Tucker kraven för konvexa problem, det vill säga kraven 1.4, 1.5 och 1.6 samt kravet

$$\lambda_i (y_i (\mathbf{x}_i^\top \boldsymbol{\beta} + \beta_0) - 1) = 0, \quad i = 1, \dots, N. \quad (1.7)$$

*Observation.* Kraven 1.4 till 1.7 säger något om hurudan den optimala lösningen  $(\boldsymbol{\beta}^*, \beta_0^*, \lambda_1^*, \dots, \lambda_N^*)$  måste vara:

- Krav 1.4 säger att vektorn  $\boldsymbol{\beta}^*$  är en linjär kombination av vektorerna  $\mathbf{x}_i$ ,  $i = 1, \dots, N$ .
- Ifall  $\lambda_i^* > 0$  så ger krav 1.7 att  $y_i (\mathbf{x}_i^\top \boldsymbol{\beta}^* + \beta_0^*) = 1$  vilket enligt det ursprungliga optimeringsproblemet 1.1 ska tolkas som att punkten  $\mathbf{x}_i$  ligger på avståndet  $C$  från det separerande hyperplanet, det vill säga punkten  $\mathbf{x}_i$  är en av punkterna som ligger närmast det separerande hyperplanet.
- Ifall  $y_i (\mathbf{x}_i^\top \boldsymbol{\beta}^* + \beta_0^*) > 1$  så är  $\lambda_i^* = 0$  och punkten  $\mathbf{x}_i$  är inte en av de punkter som ligger närmast det separerande hyperplanet.
- Parametern  $\beta_0^*$  kan bestämmas genom att man utnyttjar relationen  $y_i (\mathbf{x}_i^\top \boldsymbol{\beta}^* + \beta_0^*) = 1$  för någon av punkterna där  $\lambda_i^* > 0$ .

Baserat på de tre tidigare slutsatserna kan vidare slutsatsen att  $\boldsymbol{\beta}^*$  inte bara är en linjär kombination av observationerna  $\mathbf{x}_i$ , utan mer specifikt en linjär kombination av endast de punkter  $\mathbf{x}_i$  som ligger på randen marginalen. Dessa punkter kallas *stödvektorer*.

Kvar finns också möjligheten att  $\lambda_i^* = 0$  och  $y_i (\mathbf{x}_i^\top \boldsymbol{\beta}^* + \beta_0^*) = 1$ . Detta är endast möjligt om åtminstone  $p + 1$  stycken punkter med  $y_i (\mathbf{x}_i^\top \boldsymbol{\beta}^* + \beta_0^*) = 1$  existerar och dessutom måste punkterna ligga på samma  $p$ -dimensionella hyperplan. Då kan vilken som helst av punkterna skrivas som en linjär kombination av de andra punkterna. Existensen av en entydig lösning till optimeringsproblemet kan då inte garanteras men sannolikheten att punkterna som ligger närmast det optimala separerande hyperplanet ligger på samma hyperplan är mycket liten, speciellt om man tar datorernas begränsade värderymd i beaktande.

## 1.3 Det oseparabla fallet

Antag att observationsparen  $(\mathbf{x}_i, y_i)$  inte är linjärt separabla, det vill säga inget hyperplan  $\{\mathbf{x} : f(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta} + \beta_0\}$  med  $y_i(\mathbf{x}_i^\top \boldsymbol{\beta} + \beta_0) > 0$  för alla träningspar  $(\mathbf{x}_i, y_i)$ ,  $i = 1, \dots, N$  existerar. Oseparabla observationspar leder till att de tidigare optimeringsproblemen 1.1 samt de i sats 1.2.1 inte längre är lösbara. I sådana fall tillåter man lösningar som strider mot kraven men försöker samtidigt reglera hur långt från de ursprungliga kraven man tillåter lösningar. I praktiken åstadkoms detta med hjälp av *slackvariabler*.

## 1.4 En enkel utvidgning med olinjära faktorer?

Visa att problemet lösbart

# Kapitel 2

## Hilbertrumteori, reproducerande kärnor

Varför utvidga faktoterna?

### 2.1 Grundläggande teori

Bevis av Mercers villkor för positivsemidefinita ekvationer/operatorer.

### 2.2 SVM som exempel

Något exempel. Introducera SVM för regression?

# Litteraturförteckning

- [1] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, seventh printing edition, 2009.
- [2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning : data mining, inference, and prediction*. Springer series in statistics. Springer New York Inc., 2001.