Stödvektormaskiner

Linjära hyperplan i Hilbertrum

Oscar Granlund

20 september 2018

Sammanfattning

Skrivs till sist

Kapitel 1

Inledning

Kort kapitel med avstamp i Vapniks ursprungliga algoritm för optimala separerande hyperplan (1963) och utvecklingen fram till 1995 då Vapnik och Cortes presenterade den första algoritmen med både mjuk marginal (Smith 1968, Bennet och Mangasarian 1992) och olinjära kärnor (Aizerman 1964, Poggio och Girosi 1990?) baserade på Aronszajns Theory of Reproducing Kernels", 1950.

Hur uppsatsen relaterar till kapitlen om SVMs i boken The Elements och Statistical Learning Theory".

Kapitel 2

Stödvektormaskiner (SVM)

2.1 Klassificering med hjälp av separerande hyperplan

Definition 2.1.1. Ett *klassificeringsproblem* är ett problem var man utgående från en mängd observationspar (*träningsdata*) $(\mathbf{x}_i, y_i)$, $\mathbf{x}_i \in \mathbb{R}^p$, $y_i \in \{-1, 1\}$, $i = 1, \ldots, N$, försöker hitta en regel $g : \mathbb{R}^p \longmapsto \{-1, 1\}$ sådan att $g(\mathbf{x}_i) = y_i$ för så många träningspar $(\mathbf{x}_i, y_i)$ som möjligt.

Inom statistiken och maskininlärningen 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 2.1.2. Ett *hyperplan* i ett vektorrum med dimensionen p är ett underrum med dimensionen $p - 1$, i figur 2.1 illustreras ett separerande hyperplan för fallet $p = 2$. Klassificeringsregeln g för separerande hyperplan blir $g(\mathbf{x}_i) = \text{sign}(\mathbf{x}_i^\intercal \boldsymbol{\beta} + \beta_0)$ där mängden $\{\mathbf{x} : \mathbf{x}^\intercal \boldsymbol{\beta} + \beta_0 = 0\}$, med $\mathbf{x},\ \boldsymbol{\beta} \in \mathbb{R}^p$, definierar ett hyperplan parametriserad av $\boldsymbol{\beta}$ och β_0.

Sats 2.1.1. Ett hyperplan definierat som den affina mängden $L = \{\mathbf{x} : f(\mathbf{x}) = \mathbf{x}^\intercal \boldsymbol{\beta} + \beta_0 = 0\}$ har följande egenskaper [**?**]:

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

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

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

3. Det signerade avståndet från en punkt $\mathbf{x}$ till hyperplanet L, $\mathrm{d}^{\pm}(\mathbf{x}, L)$, ges av
$$\begin{aligned}(\mathbf{x}-\mathbf{x}_0)^{\intercal}\widehat{\boldsymbol{\beta}} &= \frac{1}{\|\boldsymbol{\beta}\|}(\mathbf{x}^{\intercal}\boldsymbol{\beta}+\beta_0)\\ &= \frac{1}{\|f'(\mathbf{x})\|}f(\mathbf{x}).\end{aligned}$$

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
$$\begin{aligned}0 &= f(\mathbf{x}_1) - f(\mathbf{x}_2)\\ &= \mathbf{x}_1^{\intercal}\boldsymbol{\beta}+\beta_0-\mathbf{x}_2^{\intercal}\boldsymbol{\beta}-\beta_0\\ &= (\mathbf{x}_1-\mathbf{x}_2)^{\intercal}\boldsymbol{\beta}\end{aligned}$$
alltså uppfyller $\boldsymbol{\beta}$ kravet för normalvektorer och $\widehat{\boldsymbol{\beta}} := \frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|}$ är den normaliserade normalvektorn till hyperplanet L. ■

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

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 hyperplanets normaliserade normal $\widehat{\beta}$, $\mathrm{comp}_{\widehat{\boldsymbol{\beta}}}(\mathbf{x}-\mathbf{x}_0)$. Vi får alltså att
$$\begin{aligned}\mathrm{d}^{\pm}(\mathbf{x}, L) = \mathrm{comp}_{\widehat{\boldsymbol{\beta}}}(\mathbf{x}-\mathbf{x}_0) &= \underline{\underline{(\mathbf{x}-\mathbf{x}_0)^{\intercal}\widehat{\boldsymbol{\beta}}}}\\ &= \frac{1}{\|\boldsymbol{\beta}\|}(\mathbf{x}^{\intercal}\boldsymbol{\beta}-\mathbf{x}_0^{\intercal}\boldsymbol{\beta}) = \underline{\underline{\frac{1}{\|\boldsymbol{\beta}\|}(\mathbf{x}^{\intercal}\boldsymbol{\beta}+\boldsymbol{\beta}_0)}}\end{aligned}$$
och om man noterar att $f(\mathbf{x}) = \mathbf{x}^{\intercal}\boldsymbol{\beta}+\beta_0$ och $f'(\mathbf{x}) = \boldsymbol{\beta}$ så fås även att
$$\frac{1}{\|\boldsymbol{\beta}\|}(\mathbf{x}^{\intercal}\boldsymbol{\beta}+\boldsymbol{\beta}_0) = \frac{1}{\|f'(\mathbf{x})\|}f(\mathbf{x}).$$

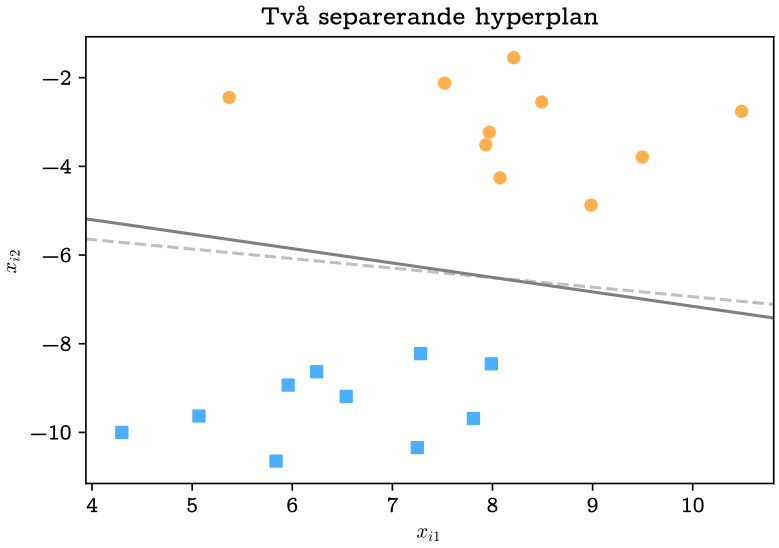
■

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

Orsak. Betrakta hyperplanen $L_1 = \{\mathbf{x} : f(\mathbf{x}) = \mathbf{x}^{\intercal}\boldsymbol{\beta}+\beta_0 = 0\}$ och $L_2 = \{\mathbf{x} : g(\mathbf{x}) = \mathbf{x}^{\intercal}(-\boldsymbol{\beta})+(-\beta_0) = 0\}$. 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 $L_3 = \left\{\mathbf{x} : h(\mathbf{x}) = \frac{\mathbf{x}^{\intercal}\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|}+\frac{\beta_0}{\|\boldsymbol{\beta}\|} = 0\right\}$. 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 $y_i(\mathbf{x}_i^\intercal \boldsymbol{\beta} + \beta_0) \geq 0$ för alla $i = 1, \ldots, N$, 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 β_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 2.1.1 att avståndet från origo till planet är lika med $\frac{1}{\|\boldsymbol{\beta}\|}(\mathbf{x}^\intercal \boldsymbol{\beta} + \beta_0) = \frac{1}{\|\boldsymbol{\beta}\|}(\mathbf{0}^\intercal \boldsymbol{\beta} + \beta_0) = \beta_0$.



Figur 2.1: 20 datapunkter med två separerande hyperplan (linje) där klassen $y_i = 1$ framställs som blå fyrkanter och klassen $y_i = -1$ som orangea cirklar.

Definition 2.1.3. Ett klassificeringsproblem eller en mängd observationspar $(\mathbf{x}_i, y_i)$ är *linjärt separabelt* om det existerar ett hyperplan $L = \{\mathbf{x} : f(\mathbf{x}) = \mathbf{x}^\intercal \boldsymbol{\beta} + \beta_0 = 0\}$ som separerar klasserna.

Sats 2.1.2. [?] För ett hyperplan $L = \{\mathbf{x} : f(\mathbf{x}) = \mathbf{x}^\intercal \boldsymbol{\beta} + \beta_0 = 0\}$ som separerar två klasser gäller att

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

eller

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

för alla $i = 1, \ldots, N$.

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

Exempel 2.1.1. Låt observationsparen vara $([2,\ 2]^\intercal,\ 1),\ ([1,\ 2]^\intercal, -1)$. Då är

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

och

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

två separerande hyperplan (linjer i detta fall).

Bevis. För L_1:

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

och

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

Och för L_2:

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

och

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

■

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 $\mathbf{x}_i$, $i = 1, \ldots, p$, och $\beta_0\mathbf{1}$ är en vektor med värdet β_0 i alla rader.

Som syns i exempel 2.1.1 finns det många separerande hyperplan om ett klassificeringsproblem är linjärt separabelt och frågan är då vilket separerande hyperplan man borde välja.

2.2 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

$$\begin{aligned} \max_{\widehat{\boldsymbol{\beta}},\widehat{\beta}_0,\|\widehat{\boldsymbol{\beta}}\|=1} \quad & C \\ \text{så att} \quad & y_i(\mathbf{x}_i^\intercal \widehat{\boldsymbol{\beta}} + \widehat{\beta}_0) \geq C, \quad i = 1, \ldots, N \end{aligned} \tag{2.1}$$

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

Observation. Ifall alla punkter är rätt klassificerade så anger $y_i(\mathbf{x}_i^\intercal \widehat{\boldsymbol{\beta}} + \widehat{\beta}_0)$ 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 (2.1) 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 $\|\widehat{\boldsymbol{\beta}}\| = 1$ bytas ut. Detta görs genom att man byter ut kraven

$$y_i(\mathbf{x}_i^\intercal \widehat{\boldsymbol{\beta}} + \widehat{\beta}_0) \geq C, \quad i = 1, \ldots, N$$

mot kraven

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

eller ekvivalent

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

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

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

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

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

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 $\mathbf{a}^\intercal\mathbf{b}$ uttryckas som den normala inre produkten $\langle \mathbf{a}, \mathbf{b} \rangle$ i $\mathbb{R}^p$. Detta kommer till nytta i kapitel 3 där konvexiteten för en generalisering av det linjära problemet utforskas.

Ovanstående resonemang är ett bevis för sats 2.2.1.

Sats 2.2.1. Låt $\widehat{\boldsymbol{\beta}},\ \boldsymbol{\beta} \in \mathbb{R}^p$ och $\widehat{\beta}_0,\ \beta_0 \in \mathbb{R}$. Låt dessutom observationsparen $(\mathbf{x}_i, y_i)$ vara linjärt separabla. Då är optimeringsproblemet

$$\begin{aligned} \max_{\widehat{\boldsymbol{\beta}},\widehat{\beta}_0,\|\widehat{\boldsymbol{\beta}}\|=1} \quad & C \\ \text{så att} \quad & y_i(\mathbf{x}_i^\intercal \widehat{\boldsymbol{\beta}} + \beta_0) \geq C, \quad i = 1, \ldots, N \end{aligned}$$

konvext och ekvivalent med optimeringsproblemet

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

Vidare ger observationen ovan ytterligare ett ekvivalent optimeringsproblem:

Korollarium 2.2.2. Optimeringsproblemen i sats 2.2.1 är ekvivalenta med optimeringsproblemet

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

med samma antaganden.

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

Framställningen i korollarium 2.2.2 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.

2.2.1 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

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

ges då av

$$L_P = \frac{1}{2}\|\boldsymbol{\beta}\|^2 - \sum_{i=1}^{N} \lambda_i \left(y_i \left(\mathbf{x}_i^\intercal \boldsymbol{\beta} + \beta_0 \right) - 1 \right) \tag{2.2}$$

som ska minimeras med avseende på $\boldsymbol{\beta}$ och β_0.

För att minimera L_P sätts derivatorna med avseende på elementen $[\boldsymbol{\beta}]_j$ av $\boldsymbol{\beta}$ och β_0 till 0, och följande relationer erhålls:

$$\begin{aligned} D_{[\boldsymbol{\beta}]_j}(L_P) &= D_{[\boldsymbol{\beta}]_j}\left(\frac{1}{2}\boldsymbol{\beta}^\intercal\boldsymbol{\beta}\right) && - D_{[\boldsymbol{\beta}]_j}\left(\sum_{i=1}^{N}\left(\lambda_i y_i \left(\mathbf{x}_i^\intercal \boldsymbol{\beta}\right) + \lambda_i y_i \beta_0 - \lambda_i\right)\right) \\ &= D_{[\boldsymbol{\beta}]_j}\left(\frac{1}{2}\sum_{k=1}^{p}[\boldsymbol{\beta}]_k^2\right) && - \sum_{i=1}^{N} D_{[\boldsymbol{\beta}]_j}\left(\lambda_i y_i \left(\sum_{k=1}^{p}[\mathbf{x}_i]_k [\boldsymbol{\beta}]_k\right) + \lambda_i y_i \beta_0 - \lambda_i\right) \\ &= [\boldsymbol{\beta}]_j && - \sum_{i=1}^{N} D_{[\boldsymbol{\beta}]_j}\left(\sum_{k=1}^{p}\lambda_i y_i [\mathbf{x}_i]_k [\boldsymbol{\beta}]_k\right) + 0 \\ &= [\boldsymbol{\beta}]_j && - \sum_{i=1}^{N}\lambda_i y_i [\mathbf{x}_i]_j \end{aligned} \tag{2.3}$$

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

$$\boldsymbol{\beta} = \sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_i \tag{2.4}$$

$$0 = \sum_{i=1}^{N} \lambda_i y_i. \tag{2.5}$$

Efter omskrivning av $\|\boldsymbol{\beta}\|^2$ som $\boldsymbol{\beta}^\intercal \boldsymbol{\beta}$ ger insättning av kraven (2.4) och (2.5) i L_P följande duala problem

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

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

$$\lambda_i \geq 0,\ i = 1, \dots, N. \tag{2.6}$$

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 (2.4), (2.5) och (2.6) samt kravet

$$\lambda_i \left(y_i \left(\mathbf{x}_i^\intercal \boldsymbol{\beta} + \beta_0\right) - 1\right) = 0,\ i = 1, \dots, N. \tag{2.7}$$

Observation. Kraven (2.4) till (2.7) säger något om hurudan den optimala lösningen $(\boldsymbol{\beta}^*, \beta_0^*, \lambda_1^*, \ldots, \lambda_N^*)$ måste vara:

- Krav (2.4) 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 (2.7) att $y_i\left(\mathbf{x}_i^\intercal \boldsymbol{\beta}^* + \beta_0^*\right) = 1$ vilket enligt det ursprungliga optimeringsproblemet (2.1) 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 $y_i\left(\mathbf{x}_i^\intercal \boldsymbol{\beta}^* + \beta_0^*\right) > 1$ så är $\lambda_i^* = 0$ och punkten $\mathbf{x}_i$ är inte en av de punkter som ligger närmast det separerande hyperplanet.
- Parametern β_0^* kan bestämmas genom att man utnyttjar relationen $y_i\left(\mathbf{x}_i^\intercal \boldsymbol{\beta}^* + \beta_0^*\right) = 1$ 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 $y_i\left(\mathbf{x}_i^\intercal \boldsymbol{\beta}^* + \beta_0^*\right) = 1$. Detta är endast möjligt om åtminstone $p+1$ stycken punkter med $y_i\left(\mathbf{x}_i^\intercal \boldsymbol{\beta}^* + \beta_0^*\right) = 1$ 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.

2.3 Det oseparabla fallet

Antag att observationsparen $(\mathbf{x}_i, y_i)$ inte är linjärt separabla, det vill säga inget hyperplan $\{\mathbf{x} : f(\mathbf{x}) = \mathbf{x}^\intercal \boldsymbol{\beta} + \beta_0\}$ med $y_i\left(\mathbf{x}_i^\intercal \boldsymbol{\beta} + \beta_0\right) > 0$ för alla träningspar $(\mathbf{x}_i, y_i),\ i = 1, \ldots, N$ existerar. Oseparabla observationspar leder till att optimeringsproblemet (2.1) samt optimeringsproblemen i sats 2.2.1 inte längre är lösbara.

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

För optimeringsproblemet (2.1) finns två omedelbara sätt att ändra på kraven, endera låter man

$$y_i \left(\mathbf{x}_i^\intercal \widehat{\boldsymbol{\beta}} + \widehat{\beta}_0 \right) \geq C - s_i \tag{2.8}$$

eller så

$$y_i \left(\mathbf{x}_i^\intercal \widehat{\boldsymbol{\beta}} + \widehat{\beta}_0 \right) \geq C \left(1 - s_i \right) \tag{2.9}$$

där slackvariablerna är nedåt begränsade av noll samt uppåt begränsade så att summan av alla slackvariabler blir mindre än någon konstant K, det vill säga

$$\begin{aligned} s_i &\geq 0, \ i = 1, \ldots, N, \\ \sum_{i=1}^{N} s_i &\leq K. \end{aligned}$$

Alternativ (2.8) kan tolkas som att man låter observationen $\mathbf{x}_i$ vara på avståndet s_i från marginalens rand, på ”fel“ sida om randen. Observationen $\mathbf{x}_i$ blir då felklassificerad om $s_i > C$. För alternativ (2.9) gäller istället att observationen $\mathbf{x}_i$ tillåts vara Cs_i enheter innanför marginalens rand. Då gäller att felklassificering händer om $s_i \geq 1$. Kravet $\sum_{i=1}^{N} s_i \leq K$ kan i det andra fallet tolkas som att K är det största antalet felklassificeringar man tillåter, medan det för det första fallet inte finns någon motsvarande tolkning om man inte låter K variera i proportion till C. Detta är en bidragande orsak till att alternativ (2.9) är det mest allmänt använda.

För hyperplan med mjuka marginaler blir då det ursprungliga optimeringsproblemet:

$$\begin{aligned} \max_{\widehat{\boldsymbol{\beta}}, \widehat{\beta}_0, \|\widehat{\boldsymbol{\beta}}\|=1} \quad & C \\ \text{så att} \quad & y_i(\mathbf{x}_i^\intercal \widehat{\boldsymbol{\beta}} + \widehat{\beta}_0) \geq C \left(1 - s_i \right), \quad i = 1, \ldots, N, \\ & s_i \geq 0, \quad i = 1, \ldots, N, \\ & \sum_{i=1}^{N} s_i \leq K. \end{aligned} \tag{2.10}$$

En annan orsak till att det andra alternativet föredras är att om man försöker skriva om motsvarande optimeringsproblem på samma sätt som i beviset för sats 2.2.1 så stöter man på problem; efter att man sätter $\widehat{\boldsymbol{\beta}} = \frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|}$

och $C = \frac{1}{\|\boldsymbol{\beta}\|}$ får man nämligen optimeringsproblemet

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

vilket inte genast går att skriva om i någon standardform för optimeringsproblem.

För optimeringsproblemet (2.10) ger stegen i beviset för sats 2.2.1 istället ett bevis för sats 2.3.1:

Sats 2.3.1. Låt $\widehat{\boldsymbol{\beta}},\ \boldsymbol{\beta} \in \mathbb{R}^p$ och $\widehat{\beta}_0,\ \beta_0 \in \mathbb{R}$. Låt dessutom konstanten K vara vald så att optimeringsproblemen är lösbara för givna observationspar $(\mathbf{x}_i, y_i)$. Då är optimeringsproblemet

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

konvext och ekvivalent med optimeringsproblemet

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

Även korrolarium 2.2.2 har en mjuk motsvarighet:

Korollarium 2.3.2. Optimeringsproblemen i sats 2.3.1 är ekvivalenta med

optimeringsproblemet

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

med samma antaganden.

Bevis. Se beviset för korrolarium 2.2.2. ■

Observation. Hyperplan med mjuka marginaler kan användas även ifall observationsparen $(\mathbf{x}_i, y_i)$ är linjärt separabla; man kan till och med få det optimala separerande hyperplanet som lösning genom att välja $K = 0$. Att använda hyperplan med mjuka marginaler kan vara en bra idé till exempel om man har outliers i mätdatan som väljs till stödvektorer. I sådana fall kan man få ett hyperplan som generaliserar bättre om man inte kräver att alla observationer klassificeras rätt.

För optimeringsproblem med krav av typen $\sum_{i=1}^{N} s_i \leq K$ kan man använda barriärmetoden vid optimering för att approximera kravet med en term i objektfunktionen [**?**]. I [**?**] används en liknande approximation där man istället för att följa uppdateringsstrategin för vägningen av straffunktionen använder andra metoder för att bestämma vägningen. Optimeringsproblemet som oftast löses för hyperplan med mjuka marginaler blir då

$$\begin{aligned} \min_{\boldsymbol{\beta},\beta_0} \quad & \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \gamma \sum_{i=1}^{N} s_i \\ \text{så att} \quad & y_i(\mathbf{x}_i^\intercal \boldsymbol{\beta} + \beta_0) \geq 1 - s_i, \quad i = 1, \ldots, N, \\ & s_i \geq 0, \quad i = 1, \ldots N, \end{aligned} \tag{2.11}$$

eller

$$\begin{aligned} \min_{\boldsymbol{\beta},\beta_0} \quad & \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \gamma \left(\sum_{i=1}^{N} s_i \right)^2 \\ \text{så att} \quad & y_i(\mathbf{x}_i^\intercal \boldsymbol{\beta} + \beta_0) \geq 1 - s_i, \quad i = 1, \ldots, N, \\ & s_i \geq 0, \quad i = 1, \ldots N, \end{aligned} \tag{2.12}$$

i huvudsak eftersom att de båda är kvadratiska optimeringsproblem och således kan lösas relativt enkelt.

Tolkningen av optimeringsproblemen (2.11) och (2.12) är att man istället för kravet $\sum_{i=1}^{N} s_i \leq K$ ger ett straff baserat på storleken av slackvariablerna $s_i,\ i = 1, \ldots, N$. Märk här att om marginalen ökar så ökar även straffet medan om marginalen minskar så minskar straffet. Avvägningen mellan minskning av straff eller ökning av marginal bestäms med hjälp av parametern γ som kan jämföras med parametern K i sats 2.3.1. Skillnaden är att om γ är litet så tillåts slackvariablerna vara större och ifall γ är stort så är det viktigare att slackvariablerna hålls små. Det separabla fallet fås när γ går mot ∞.

En viktig skillnad mellan formuleringen i sats 2.3.1 och (2.11) är att optimeringsproblemet (2.11) alltid är lösbart medan optimeringsproblemen i sats 2.3.1 är lösbara endast om K väljs tillräckligt stort.

Av de två alternativen (2.11) och (2.12) är (2.11) vanligare och behandlas således i resten av uppsatsen.

2.3.1 Primala och Duala Lagrangeproblem för mjuka marginaler

Precis som med separabelt data kan lösningen för optimeringsproblemet (2.11) karaktäriseras med hjälp av Lagrangemultiplikatorer. Den primala Lagrangefunktionen ges av

$$L_P = \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \gamma \sum_{i=1}^{N} s_i - \sum_{i=1}^{N} \lambda_i \left(y_i \left(\mathbf{x}_i^\intercal \boldsymbol{\beta} + \beta_0\right) - (1 - s_i)\right) - \sum_{i=1}^{N} \mu_i s_i \quad (2.13)$$

som ska minimeras med avseende på $\boldsymbol{\beta},\ \beta_0$ och s_i. För att hitta extrempunkterna räknas först derivatorna med avseende på $\boldsymbol{\beta}$, β_0 och s_i ut, då fås:

$$\begin{aligned}
D_{\boldsymbol{\beta}}\left(L_P\right) &= D_{\boldsymbol{\beta}}\left(\frac{1}{2}\|\boldsymbol{\beta}\|^2 - \sum_{i=1}^{N} \lambda_i \left(y_i \left(\mathbf{x}_i^\intercal \boldsymbol{\beta} + \beta_0\right) - (1 - s_i)\right)\right) \\
&= \boldsymbol{\beta} - \sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_i, \\
D_{\beta_0}\left(L_P\right) &= D_{\beta_0}\left(-\sum_{i=1}^{N} \lambda_i \left(y_i \left(\mathbf{x}_i^\intercal \boldsymbol{\beta} + \beta_0\right) - (1 - s_i)\right)\right) \\
&= -\sum_{i=1}^{N} \lambda_i y_i
\end{aligned}$$

och

$$\begin{aligned} D_{s_j}(L_P) &= D_{s_j}\left(\gamma\sum_{i=1}^{N} s_i - \sum_{i=1}^{N}\lambda_i\left(y_i\left(\mathbf{x}_i^\intercal\boldsymbol{\beta}+\beta_0\right)-(1-s_i)\right) - \sum_{i=1}^{N}\mu_i s_i\right) \\ &= D_{s_j}\left(\gamma s_j - \lambda_j\left(y_j\left(\mathbf{x}_j^\intercal\boldsymbol{\beta}+\beta_0\right)-(1-s_j)\right) - \mu_j s_j\right) \\ &= \gamma - \lambda_j - \mu_j. \end{aligned}$$

Efter att man sedan kräver att derivatorna skalla vara lika med 0 får man följande krav:

$$\boldsymbol{\beta} = \sum_{i=1}^{N}\lambda_i y_i \mathbf{x}_i, \tag{2.14}$$

$$0 = \sum_{i=1}^{N}\lambda_i y_i, \tag{2.15}$$

$$\lambda_i = \gamma - \mu_i \qquad i = 1,\ldots,N \tag{2.16}$$

samt kraven

$$\lambda_i \geq 0, \qquad i = 1,\ldots,N, \tag{2.17}$$

$$\mu_i \geq 0, \qquad i = 1,\ldots,N, \tag{2.18}$$

$$s_i \geq 0 \qquad i = 1,\ldots,N. \tag{2.19}$$

Insättning av kraven (2.14) till (2.15) och (2.16) i den primala Lagrangefunktionen (2.13) ger den duala Lagrangefunktionen

$$\begin{aligned} L_D =& \frac{1}{2}\boldsymbol{\beta}^\intercal\boldsymbol{\beta} + \gamma\sum_{i=1}^{N} s_i - \sum_{i=1}^{N}\lambda_i\left(y_i\left(\mathbf{x}_i^\intercal\boldsymbol{\beta}+\beta_0\right)-(1-s_i)\right) - \sum_{i=1}^{N}\mu_i s_i \\ =& \frac{1}{2}\left(\sum_{i=1}^{N}\lambda_i y_i\mathbf{x}_i\right)^\intercal\left(\sum_{j=1}^{N}\lambda_j y_j\mathbf{x}_j\right) - \left(\sum_{i=1}^{N}\lambda_i y_i\mathbf{x}_i\right)^\intercal\left(\sum_{j=1}^{N}\lambda_j y_j\mathbf{x}_j\right) \\ & - \sum_{i=1}^{N}\lambda_i y_i\beta_0 + \sum_{i=1}^{N}\lambda_i + \sum_{i=1}^{N}\left(\gamma-\lambda_i-\mu_i\right)s_i \\ =& \sum_{i=1}^{N}\lambda_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j \mathbf{x}_i^\intercal\mathbf{x}_j \end{aligned}$$

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

$$\lambda_i\left(y_i\left(\mathbf{x}_i^\intercal\boldsymbol{\beta}+\beta_0\right)-(1-s_i)\right) = 0, \qquad i = 1,\ldots,N, \tag{2.20}$$

$$\mu_i s_i = 0, \qquad i = 1,\ldots,N, \tag{2.21}$$

$$y_i\left(\mathbf{x}_i^\intercal\boldsymbol{\beta}+\beta_0\right)-(1-s_i) \geq 0 \qquad i = 1,\ldots,N, \tag{2.22}$$

från Karush-Kuhn-Tucker kraven för konvexa problem.

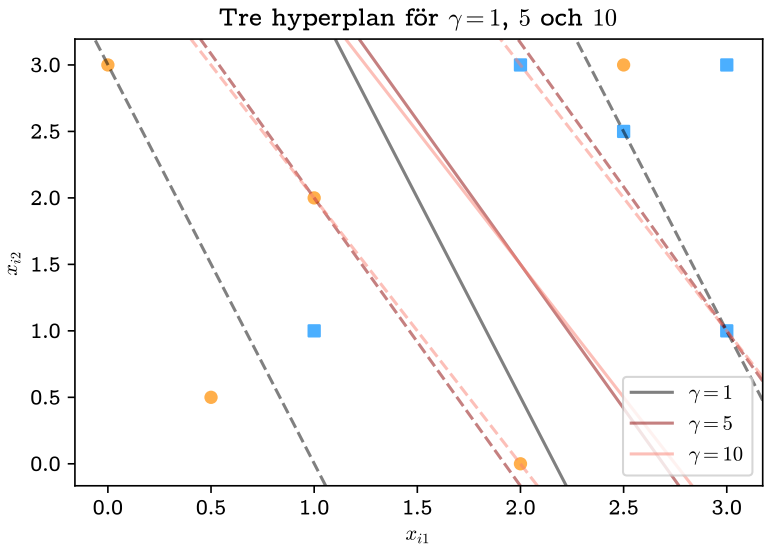
Observation. Precis som för algoritmen för optimala separerande hyperplan kan man karaktärisera lösningen för hyperplan med mjuka marginaler med hjälp av kraven (2.14) till (2.22).

- Krav (2.14) och (2.20) ger att den optimala lösningen $\boldsymbol{\beta}^*$ ges som den linjära kombinationen
$$\boldsymbol{\beta}^* = \sum_{i=1}^{N} \lambda_i^* y_i x_i,$$
av punkterna x_i på eller i marginalen för vilka $\lambda_i^* > 0$, Punkterna med $\lambda_i^* > 0$ kallas *stödvektorer* eftersom att 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 (2.16) och (2.21) 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^*.

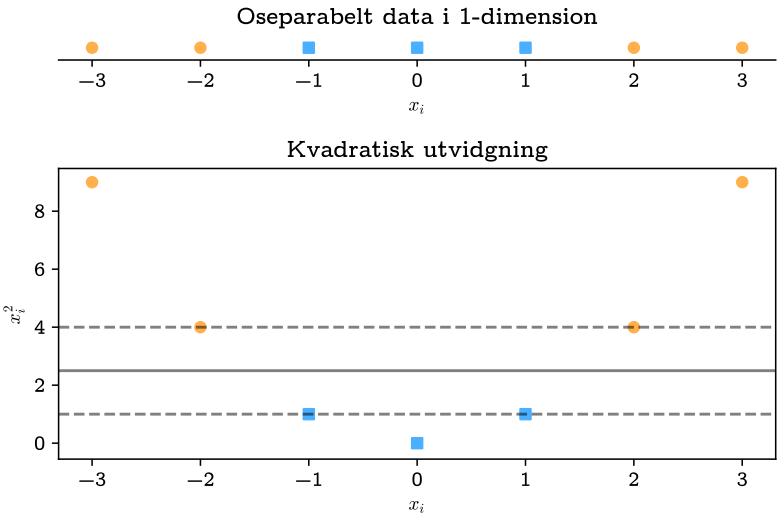


Figur 2.2: Löst exempel för linjärt oseparabelt data för 3 olika värden på γ.

Exempel 2.3.1. Låt observationsparen $(\mathbf{x}_i,\ y_i)$ vara sådana som i figur 2.2 där blåa rutor är klassen $y_i = 1$ och orangea cirklar är klassen $y_i = -1$. Axlarna motsvarar här $\mathbf{x}_i$:s första respektive andra komponenter. Klart är

här att observationsparen inte är linjärt separabla men det verkar som att en punkt från vardera klassen kanske mätts fel. För att bestämma en klassificeringsregel används metoden med hyperplan med mjuka marginaler för 3 olika värden på γ. Funktionen `SVC` med `kernel='linear'` från paketet `sklearn` [?] användes för att beräkna hyperplanen.

Observera hur parametern γ påverkar lösningen. Ju mindre γ är desto större marginaler vilket betyder att flera punkter används som stödvektorer.



Figur 2.3: En lösning med optimala separerande hyperplan och kvadratisk utvidgning där endast hyperplan med mjuka marginaler inte hade fungerat.

Exempel 2.3.2. Låt $\mathbf{x}_i \in \mathbb{R}$ och observationsparen $(\mathbf{x}_i,\ y_i)$ vara sådana att klassen $y_i = 1$ befinner sig mitt i klassen $y_i = -1$, situationen finns illustrerad överst i figur 2.3. Klart är även här att observationsparen är linjärt separabla men nu kan inte heller metoden med 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 figur 2.3 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 här är då 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.

Kapitel 3

Hilbertrumteori, reproducerande kärnor

Exempel 2.3.2 antyder att det kunde vara en bra idé att utvidga observationerna $\mathbf{x}_i$ med olinjära faktorer, frågan är bara hur detta görs bäst. Klart är att man alltid kan bilda n:te gradens polynom men ifall den ursprungliga dimensionen p, mängden observationspar N eller graden n är stor så kan detta bli övermäktigt för även den snabbaste datorn. Det kommer att visa sig att eftersom observationerna $\mathbf{x}_i$ endast förekommer i inreprodukter (se till exempel korrolarium 2.2.2 och 2.3.2) så finns det ett behagligare alternativ men först några definitioner och resultat.

3.1 Grundläggande teori

Definition 3.1.1. En *inreprodukt* är en funktion $\langle \cdot, \cdot \rangle : XX \longmapsto \mathbb{R}$ sådan att:

3.2 Hilbertrum med Reproducerande Kärnor

Bygg vidare på korrolarium 2.3.2:s mjuka motsvarighet genom att byta ut inreprodukten mot en annan, ge krav på inreprodukten. Visa att just precis de inreprodukter som samtidigt är reproducerande kärnor uppfyller de kraven.

(Bevis av Mercers villkor för positivsemidefinita ekvationer/operatorer, behöver åtminstone Riesz representationssats och någon sats om eigenvärden/funktioner, material från kursen fördjupad analys I/II borde räcka". Måste kanske hoppa över vissa tekniska detaljer.)

Kapitel 4

Avslutning

Har hittills inte gått in på statistiska detaljer, till exempel tränings/valideringsdata eller gränser på felklassificering.

Kan också nämna andra varianter som till exempel användning inom regression, glesa"(sparse) varianter.