Stödvektormaskiner

Linjära hyperplan i Hilbertrum

Kandidatavhandling i matematik

Fakulteten för naturvetenskaper och teknik
Åbo Akademi

Oscar Granlund
37920

Handledare:
Anne-Maria Ernvall-Hytönen

16 november 2018

Sammanfattning

Stödvektormaskinen (SVM) är en metod för att klassificera observationer som var mycket populär under den senare delen av 90-talet och början av 00-talet. Metoden går ut på att man passar in ett hyperplan mellan observationernas klasser och använder hyperplanet som gräns för övergången mellan klasserna. I denna avhandling undersöks och härleds metoden med start i ett enkelt optimeringsproblem som i tre steg görs mera allmänt. Först undersöks ett optimeringsproblem som kräver att alla observationer går att klassificera rätt, speciellt läggs vikt på att med Lagrangemultiplikatorer karaktärisera den optimala lösningen. Till näst undersöks några varianter av det första optimeringsproblemet där man tillåter att observationerna klassificeras fel, igen karaktäriseras den optimala lösningen. Till sist presenteras en introduktion till teorin om Hilbertrum med reproducerande kärnor som ger ett sätt att generalisera de hittills linjära klassificeringsmetoderna (samt andra linjära metoder) till en olinjär klassificeringsmetod, det är denna metod som man vanligtvis avser med begreppet stödvektormaskin.

Innehåll

Kapitel 1

Inledning

I takt med att datorerna blev kraftfullare och användarvänligare under den senare delen av 1900-talet forskades det flitigt i hur man på bästa sätt utnyttjar datorer för beräkning och lösning av statistiska problem. I samband med den flitiga forskningen uppstod maskininlärning som ett forskningsområde gemensamt för statistiken och datavetenskapen.

En tidig klassificeringsalgoritm var Frank Rosenblatts *perceptron*-algoritm (år 1957) [13] där man med inspiration från hjärnans neuroner försökt klassificera *observationer* genom att dra ett *hyperplan* mellan klasserna. År 1963 presenterade Aleksandr Lerner och Vladimir Vapnik en variant av Rosenblatts perceptron där *optimala separerande hyperplan* används för att klassificera observationerna [18]. Lerner och Vapniks algoritm är matematiskt mera tilltalande än Rosenblatts eftersom den optimala lösningen kan visas vara unik, men det finns fortfarande några problem. Bland annat går algoritmen bara att använda om observationsparen är *linjärt separabla*. År 1968 föreslog Fred Smith [16] en generaliserad algoritm som använde *slackvariabler* för att även fungera för linjärt oseparabla observationspar. Smiths arbete med slackvariabler verkar inte varit speciellt välkänt men undersöktes vidare av Kristin Bennet och Olvi Mangasarian år 1992 [3].

Parallellt med forskningen i de linjära optimala separerande hyperplanen forskades det om tillämpningar av speciella funktioner som kallades *kärnor*, med avstamp i James Mercers forskning (år 1909) i *positiva* funktioner [11] och Nachman Aronszajns fortsatta forskning (år 1950) om *reproducerande* kärnor [2]. Kärnor föreslogs av Mark Aizerman, Emmanuil Braverman och Lev Rozonoer [1] för att generalisera perceptron-algoritmen till en algoritm för olinjär klassificering. Efter att användningen av olinjära kärnor visades endera ge bra resultat vid tillämpningar eller ge nya sätt att analysera algoritmer, se till exempel Grace Wahbas bok om spline-modeller (år 1990) [19], tillämpades kärnor även på den ursprungliga algoritmen med optimala sepa-

rerande hyperplan av Bernhard Boser, Isabelle Gyuon och Vladimir Vapnik år 1992 [4]. Snart därefter generaliserades även Bennets och Mangasarians algoritm av Corinna Cortes och Vladimir Vapnik år 1995 [6], detta är den algoritm som vanligtvis associeras med begreppet *stödvektormaskin* (Support Vector Machine, SVM) och med andra ord den algoritm som i denna avhandling behandlas.

I avhandlingen kommer till först den ursprungliga algoritmen att undersökas för att sedan modifieras med slackvariabler, upplägget följer i stora drag Trevor Hasties, Robert Tibshiranis och Jerome Friedmans bok [8], speciellt kapitlet om optimala separerande hyperplan samt kapitlet som börjar med utvidgningen med hjälp av slackvariabler. Därefter kommer de reproducerande kärnorna att undersökas, för det följs i stora drag Bernhard Schölkopfs och Alexander Smolas bok [14].

Precis som många andra metoder inom statistiken och maskininlärningen bygger stödvektormaskinen på ett konvext optimeringsproblem och därför borde läsaren vara bekant med några optimeringsrelaterade koncept. En bra introduktion är Stephen Boyd och Lieven Vandenberghes bok [5]. Främst kommer teorin om kvadratiska optimeringsproblem och analys av duala problem med hjälp av Lagrangemultiplikatorer att användas.

Kapitel 2

Hilbertrumteori

2.1 Geometriska begrepp

I många statistiska metoder används enkla geometriska koncept, till exempel plan eller linjer, för att dra slutsatser angående insamlat data. Ofta vill man även hitta den bästa modellen, till exempel den modell som minimerar avståndet mellan observationerna och modellens predikterade värden (tänk som i linjär regression) eller den modell som maximerar det minsta avståndet mellan två klasser. För att effektivt kunna resonera om hur rummen man arbetar i ser ut visar det sig att teorin om *inreproduktrum*, eller närmare bestämt *Hilbertrum*, ger många bra verktyg. Ett *Hilbertrum*, $\mathcal{H}$, är ett vektorrum X försett med en *inreprodukt*, $\langle \cdot, \cdot \rangle$, som dessutom är *fullständigt*.

Definition 2.1.1 (Enligt [20])**.** 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 2.1.2. Ett inreproduktrum X är *fullständigt* om varje Cauchyföljd $\mathbf{x}_n$ konvergerar (med avseende på inreproduktens inducerade norm) till en punkt $\mathbf{x}$ i X.

Definition 2.1.3. Den *inducerade normen* $\| \cdot \|_{\mathcal{H}}$ i ett Hilbertrum $\mathcal{H}$ med en inreprodukt $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ definieras genom

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

För att bevisa att normen definierad ovan är en norm krävs ett välkänt resultat:

Sats 2.1.1 (Cauchy-Scwharz olikhet enligt [20])**.** För $\mathbf{x}$, $\mathbf{y}$ i ett inreproduktrum X gäller

$$\langle \mathbf{x}, \mathbf{y} \rangle^2 \leq \langle \mathbf{x}, \mathbf{x} \rangle \langle \mathbf{y}, \mathbf{y} \rangle$$

med likhet om och endast om $\mathbf{y} = \lambda \mathbf{x}$ för något $\lambda \in \mathbb{R}$, det vill säga om $\mathbf{x}$ och $\mathbf{y}$ är linjärt beroende.

Bevis. Om $\mathbf{y} = \lambda \mathbf{x}$ så gäller

$$\langle \mathbf{x}, \mathbf{y} \rangle^2 = \langle \mathbf{x}, \lambda \mathbf{x} \rangle^2 = \lambda^2 \langle \mathbf{x}, \mathbf{x} \rangle^2 = \langle \lambda \mathbf{x}, \lambda \mathbf{x} \rangle \langle \mathbf{x}, \mathbf{x} \rangle = \langle \mathbf{y}, \mathbf{y} \rangle \langle \mathbf{x}, \mathbf{x} \rangle$$

vilket skulle visas.

Ifall $\mathbf{y} \neq \lambda \mathbf{x}$ så måste följande gälla enligt **IP4**:

$$\begin{aligned} 0 &\leq \langle \mathbf{x} - \lambda \mathbf{y}, \mathbf{x} - \lambda \mathbf{y} \rangle \\ &= \langle \mathbf{x}, \mathbf{x} \rangle + \langle \mathbf{x}, -\lambda \mathbf{y} \rangle + \langle -\lambda \mathbf{y}, \mathbf{x} \rangle + \langle -\lambda \mathbf{y}, -\lambda \mathbf{y} \rangle \\ &= \langle \mathbf{y}, \mathbf{y} \rangle \lambda^2 - 2 \langle \mathbf{x}, \mathbf{y} \rangle \lambda + \langle \mathbf{x}, \mathbf{x} \rangle \end{aligned}$$

där sista raden är en kvadratisk ekvation av λ med högst en unik reell rot. Då följer att diskriminanten $\Delta = (-2 \langle \mathbf{x}, \mathbf{y} \rangle)^2 - 4 \langle \mathbf{y}, \mathbf{y} \rangle \langle \mathbf{x}, \mathbf{x} \rangle \leq 0$. Efter omarrangerande och division med 4 fås då

$$\langle \mathbf{x}, \mathbf{y} \rangle^2 \leq \langle \mathbf{x}, \mathbf{x} \rangle \langle \mathbf{y}, \mathbf{y} \rangle . \qquad \blacksquare$$

Nedan följer ett bevis för att $\|\mathbf{x}\|_{\mathcal{H}} := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{H}}}$, $\mathbf{x} \in \mathcal{H}$ är en norm.

Bevis. Låt $\lambda \in \mathbb{R}$, $\mathbf{x}$, $\mathbf{y} \in \mathcal{M}$:

N1 $\|\mathbf{x} + \mathbf{y}\|_{\mathcal{H}} \leq \|\mathbf{x}\|_{\mathcal{H}} + \|\mathbf{y}\|_{\mathcal{H}}$ (subadditiv alternativt uppfyller triangelolikheten):

$$\begin{aligned} \|\mathbf{x} + \mathbf{y}\|_{\mathcal{H}}^2 &= \langle \mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle_{\mathcal{H}} \\ &= \langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{H}} + 2 \langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{H}} + \langle \mathbf{y}, \mathbf{y} \rangle_{\mathcal{H}} \\ &\leq \langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{H}} + 2\sqrt{\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{H}}^2} + \langle \mathbf{y}, \mathbf{y} \rangle_{\mathcal{H}} \\ \text{(Enligt sats 2.1.1)} \quad &\leq \langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{H}} + 2\sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{H}} \langle \mathbf{y}, \mathbf{y} \rangle_{\mathcal{H}}} + \langle \mathbf{y}, \mathbf{y} \rangle_{\mathcal{H}} \\ &= \|\mathbf{x}\|_{\mathcal{H}}^2 + 2 \|\mathbf{x}\|_{\mathcal{H}} \|\mathbf{y}\|_{\mathcal{H}} + \|\mathbf{y}\|_{\mathcal{H}}^2 \\ &= (\|\mathbf{x}\|_{\mathcal{H}} + \|\mathbf{y}\|_{\mathcal{H}})^2 \end{aligned}$$

där olikheten fås efter att man tagit kvadratrötter av båda sidorna.

N2 $\|\lambda\mathbf{x}\|_{\mathcal{H}} = |\lambda| \|\mathbf{x}\|_{\mathcal{H}}$ (absolut homogen):

$$\|\lambda\mathbf{x}\|_{\mathcal{H}}^2 = \langle\lambda\mathbf{x}, \lambda\mathbf{x}\rangle_{\mathcal{H}} = \lambda^2 \langle\mathbf{x}, \mathbf{x}\rangle_{\mathcal{H}},$$

efter att man tar kvadratrötter på båda sidorna fås då

$$\|\lambda\mathbf{x}\|_{\mathcal{H}} = \sqrt{\lambda^2} \|\mathbf{x}\|_{\mathcal{H}} = |\lambda| \|\mathbf{x}\|_{\mathcal{H}}.$$

N3 $\|\mathbf{x}\|_{\mathcal{H}} = 0$ om och endast om $\mathbf{x} = \mathbf{0}$ (positivt definit):
Detta följer genast ur $\|\mathbf{x}\|_{\mathcal{H}} = \sqrt{\langle\mathbf{x}, \mathbf{x}\rangle_{\mathcal{H}}}$ och **IP4**. ■

Hilbertrum kan definieras för många olika vektorrum, till exempel det p-dimensionella vektorrummet med den vanliga inreprodukten $\langle\mathbf{x}, \mathbf{y}\rangle_d = \mathbf{x}^\intercal\mathbf{y}$. Man kan även definiera en inreprodukt för vektorrum bestående av funktioner på intervallet $[a, b]$, då brukar inreprodukten definieras som

$$\langle f, g\rangle_{\mathcal{L}^2} = \int_a^b f(x)\, g(x)\, dx,$$

men man måste dessutom kräva att normen $\|f\|_{\mathcal{L}^2} = \left(\int_a^b f(x)^2\, dx\right)^{\frac{1}{2}}$ är ändlig för alla funktioner i vektorrummet.

Ett hjälpmedel när man jobbar med Hilbertrum kan vara att kunna visualisera och föreställa sig relationer mellan vektorer, speciellt räta vinklar är viktiga och spelar en betydande roll. I Hilbertrum säger man att vinkeln mellan två vektorer är 90° om de är ortogonala och olika $\mathbf{0}$:

Definition 2.1.4 (Enligt [10])**.** 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 $\|\mathbf{x}\|_{\mathcal{H}} = \|\mathbf{y}\|_{\mathcal{H}} = 1$.

Exempel 2.1.1. Låt $\mathbf{x} = [1, 0]^\intercal$ och $\mathbf{y} = [0, 1]^\intercal$. Då är $\langle\mathbf{x}, \mathbf{y}\rangle_2 = \mathbf{x}^\intercal\mathbf{y} = [1, 0]\,[0, 1]^\intercal = 1 \cdot 0 + 0 \cdot 1 = 0$ och vektorerna är ortogonala. Normen av vektorerna är också 1 så de är dessutom ortonormala.

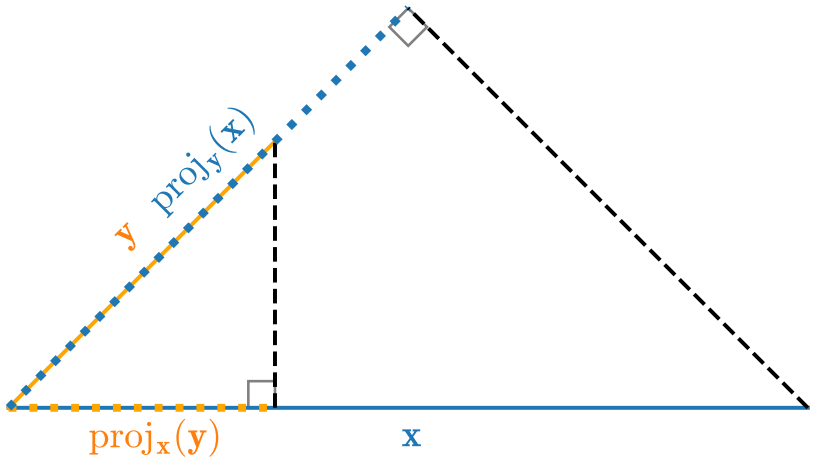
Med hjälp av ortogonaliteten kan man härleda konceptet för komponenten av en vektor längs en annan vektor och projektionen av en vektor på en annan vektor. Figur 2.1 illustrerar konceptet.

Låt $\mathbf{x}, \mathbf{y} \in \mathcal{H}$ vara två vektorer olika $\mathbf{0}$. Välj $\lambda \in \mathbb{R}$ så att vektorn $(\mathbf{x} - \lambda\mathbf{y})$ är ortogonal till $\mathbf{y}$ det vill säga

$$\langle\mathbf{x} - \lambda\mathbf{y}, \mathbf{y}\rangle_{\mathcal{H}} = \langle\mathbf{x}, \mathbf{y}\rangle_{\mathcal{H}} - \lambda \langle\mathbf{y}, \mathbf{y}\rangle_{\mathcal{H}} = 0.$$

När man löser för λ får man då

$$\lambda = \frac{\langle\mathbf{x}, \mathbf{y}\rangle_{\mathcal{H}}}{\langle\mathbf{y}, \mathbf{y}\rangle_{\mathcal{H}}}.$$



Figur 2.1: Två vektorer, $\mathbf{x}$ i blått och $\mathbf{y}$ i orange, samt projektionen av $\mathbf{x}$ på $\mathbf{y}$ streckat i blått och projektionen av $\mathbf{y}$ på $\mathbf{x}$ streckat i orange. Streckat i svart finns de ortogonala vektorerna.

Skalären λ anger alltså hur långt längs med $\mathbf{y}$ man ska ta sig för att vektorn $(\mathbf{x} - \lambda\mathbf{y})$ ska vara ortogonal till $\mathbf{y}$. Man kallar skalären λ för vektorn $\mathbf{x}$:s komponent i $\mathbf{y}$:s riktning. Vektorn $\lambda\mathbf{y}$ kallar man projektionen av $\mathbf{x}$ på $\mathbf{y}$. Man får då definitionen:

Definition 2.1.5 (Enligt [10])**.** För två vektorer $\mathbf{x}, \mathbf{y} \in \mathcal{H}$ olika $\mathbf{0}$. Definiera *komponenten* av $\mathbf{x}$ i $\mathbf{y}$:s riktning, $\text{comp}_{\mathbf{y}}(\mathbf{x})$, som skalären

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

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

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

Observation. Ifall vektorn $\mathbf{y}$ är normaliserad, det vill säga $\|\mathbf{y}\|_{\mathcal{H}} = 1$, så är $\text{comp}_{\mathbf{y}}(\mathbf{x}) = \left\|\text{proj}_{\mathbf{y}}(\mathbf{x})\right\|_{\mathcal{H}}$ ifall $\text{comp}_{\mathbf{y}}(\mathbf{x}) \geq 0$ och $\text{comp}_{\mathbf{y}}(\mathbf{x}) = -\left\|\text{proj}_{\mathbf{y}}(\mathbf{x})\right\|_{\mathcal{H}}$ ifall $\text{comp}_{\mathbf{y}}(\mathbf{x}) \leq 0$. Märk även att för $\mathbf{x} = \mathbf{0}$ går det att definiera komponenter och projektioner på samma sätt men de är inte speciellt intressanta, $\text{comp}_{\mathbf{y}}(\mathbf{x}) = 0$ och $\text{proj}_{\mathbf{y}}(\mathbf{x}) = \text{comp}_{\mathbf{y}}(\mathbf{x})\,\mathbf{y} = \mathbf{0}$. Vektorn $\mathbf{y}$ måste däremot vara olika $\mathbf{0}$ för att undvika division med 0. Ifall vektorerna $\mathbf{x}$, $\mathbf{y}$ är parallella, det vill säga $\mathbf{y} = \lambda\mathbf{x}$, gäller $\text{comp}_{\mathbf{x}}(\mathbf{y}) = \text{comp}_{\mathbf{x}}(\lambda\mathbf{x}) = \frac{\langle \lambda\mathbf{x}, \mathbf{x} \rangle_{\mathcal{H}}}{\langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{H}}} = \frac{\lambda\langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{H}}}{\langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{H}}} = \lambda$ och $\text{proj}_{\mathbf{x}}(\mathbf{y}) = \lambda\mathbf{x} = \mathbf{y}$. För $\text{proj}_{\mathbf{y}}(\mathbf{x})$ gäller $\text{proj}_{\mathbf{y}}(\mathbf{x}) = \text{proj}_{\mathbf{y}}\left(\frac{1}{\lambda}\mathbf{y}\right) = \text{comp}_{\mathbf{y}}(\mathbf{x})\,\mathbf{y} = \frac{\left\langle \frac{1}{\lambda}\mathbf{y}, \mathbf{y} \right\rangle_{\mathcal{H}}}{\langle \mathbf{y}, \mathbf{y} \rangle_{\mathcal{H}}}\mathbf{y} = \frac{1}{\lambda}\mathbf{y} = \mathbf{x}$.

Märk hur dimensionen på vektorrummet X inte nämns i definitionen för inreprodukten och därför kan Hilbertrum även vara oändligtdimensionella. Många av de bekanta egenskaperna för ändligtdimensionella inreproduktrum gäller även för oändligtdimensionella inreproduktrum, ett exempel är Hilbertrummet $\mathcal{L}^2$ med inreprodukten $\langle f, g \rangle_{\mathcal{L}^2}$ som behandlades tidigare. För $\mathcal{L}^2$ gäller fortfarande att två funktioner f och g är ortogonala om

$$\langle f, g \rangle_{\mathcal{L}^2} = \int_a^b f(x)g(x)\,dx = 0,$$

även om det kan vara svårt att visualisera för oändligtdimensionella rum.

För att hjälpa till med visualiseringen av oändligtdimensionella rum finns ett till verktyg som ger en motsvarighet till de ändligtdimensionella vektorrummens koordinatsystem och basvektorer. För att undersöka saken är det naturligt att leta efter något som kunde agera som motsvarighet till basvektorerna. Ett sådant koncept finns om det existerar en *ortonormal följd* vektorer, $\mathbf{e}_i$, $i = 1, 2, 3, \ldots$, det vill säga en följd vektorer sådana att $\|\mathbf{e}_i\|_{\mathcal{H}} = 1$, $i = 1, 2, 3, \ldots$, och $\langle \mathbf{e}_i, \mathbf{e}_j \rangle_{\mathcal{H}} = 0$ om $i \neq j$.

Exempel 2.1.2. I $\mathcal{L}^2$, $x \in [-\pi, \pi]$ är följande en ortonormal följd [20]:

$$\mathbf{e}_1 = \frac{1}{\sqrt{2\pi}}, \quad \mathbf{e}_2 = \frac{1}{\sqrt{\pi}}\cos(x), \quad \mathbf{e}_3 = \frac{1}{\sqrt{\pi}}\sin(x), \quad \mathbf{e}_2 = \frac{1}{\sqrt{\pi}}\cos(2x), \ \ldots$$

Definition 2.1.6 (Enligt [20])**.** En ortonormal följd basvektorer $\mathbf{e}_i$ i ett Hilbertrum $\mathcal{H}$ är *fullständig* (inte samma som definition 2.1.2) ifall nollvektorn $\mathbf{0}$ är den enda vektorn i $\mathcal{H}$ ortogonal till varje basvektor $\mathbf{e}_i$. Vidare är ett Hilbertrum *separabelt* ifall det existerar en fullständig ortonormal följd basvektorer $\mathbf{e}_i \in \mathcal{H}$.

Det visar sig att inte alla Hilbertrum har en motsvarighet till basvektorer men om Hilbertrummet är separabelt så existerar det en ortonormal följd vektorer sådan att för varje vektor $\mathbf{x} \in \mathcal{H}$ gäller [20]:

$$\mathbf{x} = \sum_{i=1}^{\infty} \langle \mathbf{x}, \mathbf{e}_i \rangle_{\mathcal{H}} \, \mathbf{e}_i,$$

där vektorerna $\mathbf{e}_i$ agerar bas och koefficienten $\langle \mathbf{x}, \mathbf{e}_i \rangle_{\mathcal{H}}$ kallas den i:te Fourierkoefficienten med avseende på basen $\mathbf{e}_i$, $i = 1, 2, 3, \ldots$. Märk att man projicerar $\mathbf{x}$ på varje basvektor $\mathbf{e}_i$ och summerar de resulterande projektionerna. Antalet basvektorer $\mathbf{e}_i$ bestämmer i princip dimensionen på Hilbertrummet.

Kapitel 3

Stödvektormaskiner (SVM)

Inom statistiken och maskininlärningen finns många olika metoder för att försöka lösa *klassificeringsproblem* till exempel med hjälp av regressionsmodeller eller klusteranalys. De flesta av metoderna går ofta att uttrycka som något optimeringsproblem. I detta kapitel kommer ett antal optimeringsproblem att undersökas men speciellt kommer två av optimeringsproblemen att undersökas noggrant. Det visar sig att det senare optimeringsproblemet är en generalisering av det första eller omvänt att det första optimeringsproblemet är ett specialfall av det andra.

Gemensamt för båda optimeringsproblemen är den speciella egenskapen att lösningen kommer bero på endast ett fåtal av de observationer man matar in i algoritmen. Med andra ord ger algoritmen samma lösning även om man tar bort alla andra observationer. Observationerna som lösningen beror på kallas stödvektorer och därifrån kommer även termen stödvektormaskin.

Egenskapen att lösningen endast beror på ett fåtal av observationerna medför några fördelar, bland annat ger egenskapen ofta bra poäng i de vanligaste testen för hur väl en algoritm generaliserar. Betrakta N stycken observationer, ifall n stycken av dem är stödvektorer, så betyder det att algoritmen ger precis samma lösning ifall man tar bort någon av de andra $N-n$ stycken observationerna. Ifall man slumpmässigt väljer en observation att ta bort så får man samma lösning med sannolikheten $\frac{N-n}{N}$. Om man då låter antalet observationer N gå mot oändligheten så går sannolikheten att man får samma lösning mot 1.

3.1 Klassificering med hjälp av separerande hyperplan

Definition 3.1.1. Ett *klassificeringsproblem* är ett problem var man utgående från en mängd observationspar (*träningsdata*) $(\mathbf{x}_i, y_i)$, med $\mathbf{x}_i \in \mathbb{R}^p$ och $y_i \in \{-1, 1\}$ för $i = 1, \ldots, N$, försöker hitta en regel $g : \mathbb{R}^p \longmapsto \{-1, 1\}$ sådan att $g(\mathbf{x}_i) = y_i$ för så många observationspar $(\mathbf{x}_i, y_i)$ som möjligt.

För att lösa klassificeringsproblem använder stödvektormaskinerna *hyperplan*, det vill säga affina underrum med dimensionen $p-1$, för att klassificera *observationerna* $\mathbf{x}_i$ i *klasserna* $y_i \in \{-1, 1\}$.

Definition 3.1.2. Ett *hyperplan* i ett inreproduktrum $\mathcal{H}$ är ett affint underrum i $\mathcal{H}$ definierat som mängden $\{\mathbf{x} : \langle \mathbf{x}, \boldsymbol{\beta} \rangle_{\mathcal{H}} + \beta_0 = 0\}$, med $\mathbf{x}$, $\boldsymbol{\beta} \in X$ och $\beta_0 \in \mathbb{R}$.

Exempel 3.1.1. I figur 3.1 illustreras två separerande hyperplan i $\mathbb{R}^2$. I $\mathbb{R}^2$ blir hyperplanet en linje, med andra ord ett affint underrum med dimensionen $2 - 1 = 1$. Allmänt gäller att i ett rum med dimensionen p blir hyperplanet ett affint underrum med dimensionen $p - 1$. Märk att om punkten[1] $\mathbf{x} = \mathbf{0}$ tillhör hyperplanet så är hyperplanet ett underrum.

För nästa sats behövs några konventioner för ett hyperplan $L = \{\mathbf{x} : f(\mathbf{x}) = \langle \mathbf{x}, \boldsymbol{\beta} \rangle_{\mathcal{H}} + \beta_0 = 0\}$ i relation till en punkt $\mathbf{y}$ i $\mathcal{H}$:

- Man säger att punkten $\mathbf{y}$ ligger över hyperplanet L om $f(\mathbf{y}) > 0$ och under om $f(\mathbf{y}) < 0$.
- Det signerade avståndet från punkten $\mathbf{y}$ till hyperplanet L, $\mathrm{d}^{\pm}(\mathbf{y}, L)$, definieras som $\inf_{\mathbf{x} \in L} \|\mathbf{y} - \mathbf{x}\|_{\mathcal{H}}$ om $f(\mathbf{y}) \geq 0$ och $-(\inf_{\mathbf{x} \in L} \|\mathbf{y} - \mathbf{x}\|_{\mathcal{H}})$ om $f(\mathbf{y}) \leq 0$. Med andra ord är $\mathrm{d}^{\pm}(\mathbf{y}, L)$ det kortaste avståndet (med avseende på den inducerade normen i $\mathcal{H}$) från punkten $\mathbf{y}$ till alla punkter $\mathbf{x} \in L$ om $\mathbf{y}$ ligger över L och minus det kortaste avståndet från $\mathbf{y}$ till L om $\mathbf{y}$ ligger under L.

[1] I resten av avhandlingen kommer begreppen *punkt* och *vektor* att användas om vartannat. Egentligen är alla punkter också vektorer i samma vektorrum som resten av vektorerna men i litteraturen används ofta begreppet punkter eftersom det är hur man intuitivt brukar tänka på uppmätta observationer. Begreppet vektor brukar användas om man vill poängtera att riktningen är viktigt.

Sats 3.1.1 (Enligt [8])**.** Ett hyperplan i ett inreproduktrum $\mathcal{H}$ definierat som den affina mängden $L = \{\mathbf{x} : f(\mathbf{x}) = \langle \mathbf{x}, \boldsymbol{\beta} \rangle_{\mathcal{H}} + \beta_0 = 0\}$ har följande egenskaper:

1. Vektorn $\boldsymbol{\beta}$ är ortogonal till alla vektorer i L (det vill säga alla vektorer sådana att ändpunkterna ligger i L) och kan *ortonormeras* (göras ortonormal) genom
$$\widehat{\boldsymbol{\beta}} = \frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|_{\mathcal{H}}}.$$

2. $$\langle \mathbf{x}_0, \boldsymbol{\beta} \rangle_{\mathcal{H}} = -\beta_0 \qquad \text{för alla } \mathbf{x}_0 \text{ i } L.$$

3. För det signerade avståndet från en punkt $\mathbf{y}$ till hyperplanet L gäller att
$$\begin{aligned} \mathrm{d}^{\pm}(\mathbf{y}, L) &= \left\langle \mathbf{y} - \mathbf{x}_0, \widehat{\boldsymbol{\beta}} \right\rangle_{\mathcal{H}} \\ &= \frac{1}{\|\boldsymbol{\beta}\|_{\mathcal{H}}} \left(\langle \mathbf{y}, \boldsymbol{\beta} \rangle_{\mathcal{H}} + \beta_0 \right) \end{aligned}$$
där $\mathbf{x}_0$ är en godtycklig punkt i hyperplanet L. Om $\mathcal{H}$ är lika med $\mathbb{R}^p$ med den vanliga inreprodukten fås dessutom
$$\mathrm{d}^{\pm}(\mathbf{y}, L) = \frac{1}{\|f'(\mathbf{y})\|_p} f(\mathbf{y}).$$

Bevis.

1. Låt $\mathbf{x}_1$ och $\mathbf{x}_2$ vara två godtyckliga punkter i L. Då gäller att $f(\mathbf{x}_1) = f(\mathbf{x}_2) = 0$ och
$$\begin{aligned} 0 &= f(\mathbf{x}_1) - f(\mathbf{x}_2) \\ &= \langle \mathbf{x}_1, \boldsymbol{\beta} \rangle_{\mathcal{H}} + \beta_0 - \langle \mathbf{x}_2, \boldsymbol{\beta} \rangle_{\mathcal{H}} - \beta_0 \\ &= \langle \mathbf{x}_1 - \mathbf{x}_2, \boldsymbol{\beta} \rangle_{\mathcal{H}} \end{aligned}$$
med andra ord är $\boldsymbol{\beta}$ ortogonal till vektorn $(\mathbf{x}_1 - \mathbf{x}_2)$, det vill säga vektorn från en punkt i L till en annan punkt i L. Dessutom gäller för $\widehat{\boldsymbol{\beta}} := \frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|_p}$ att $\left\|\widehat{\boldsymbol{\beta}}\right\|_{\mathcal{H}} = 1$ och vektorn $\widehat{\boldsymbol{\beta}}$ är ortonormal till alla vektorer i L. ■

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

3. Låt $\mathbf{x}_0$ vara en punkt i hyperplanet L, då är avståndet från punkt $\mathbf{x}_0$ till punkt $\mathbf{y}$ minimerat om vektorn $(\mathbf{y} - \mathbf{x}_0)$ är ortogonal till hyperplanet L. I $\mathbb{R}^2$ är det här principen att det kortaste avståndet från en linje till en punkt är avståndet mätt längs med en linje vinkelrät mot den ursprungliga linjen. Eftersom $\widehat{\boldsymbol{\beta}}$ är ortonormal till varje punkt i L blir det kortaste avståndet från $\mathbf{y}$ till L längden av projektionen av vektorn från $\mathbf{x}_0$ till $\mathbf{y}$ på $\widehat{\boldsymbol{\beta}}$ det vill säga $\mathrm{d}^{\pm}(\mathbf{y}, L) = \mathrm{comp}_{\widehat{\boldsymbol{\beta}}}(\mathbf{y} - \mathbf{x}_0)$. Vidare fås då

$$\begin{aligned}\mathrm{d}^{\pm}(\mathbf{y}, L) = \mathrm{comp}_{\widehat{\boldsymbol{\beta}}}(\mathbf{y} - \mathbf{x}_0) &= \underline{\underline{\left\langle \mathbf{y} - \mathbf{x}_0, \widehat{\boldsymbol{\beta}} \right\rangle_{\mathcal{H}}}} \\ &= \frac{1}{\|\boldsymbol{\beta}\|_{\mathcal{H}}} \left(\langle \mathbf{y}, \boldsymbol{\beta} \rangle_{\mathcal{H}} - \langle \mathbf{x}_0, \boldsymbol{\beta} \rangle_{\mathcal{H}} \right) = \underline{\underline{\frac{1}{\|\boldsymbol{\beta}\|_{\mathcal{H}}} \left(\langle \mathbf{y}, \boldsymbol{\beta} \rangle_{\mathcal{H}} + \beta_0 \right)}}\end{aligned}$$

där det sista steget följer från egenskap 2 då $\mathbf{x}_0$ är en punkt i L. Om $\mathcal{H}$ är lika med $\mathbb{R}^p$ och inreprodukten ges av $\langle \mathbf{x}, \mathbf{y} \rangle_p = \mathbf{x}^\intercal \mathbf{y}$ kan man notera att $f(\mathbf{y}) = \langle \mathbf{y}, \boldsymbol{\beta} \rangle_p + \beta_0$ och $f'(\mathbf{y}) = \boldsymbol{\beta}$, då fås även att

$$\mathrm{d}^{\pm}(\mathbf{y}, L) = \frac{1}{\|\boldsymbol{\beta}\|_p} \left(\langle \mathbf{y}, \boldsymbol{\beta} \rangle_p + \beta_0 \right) = \frac{1}{\|f'(\mathbf{y})\|_p} f(\mathbf{y}).$$

■

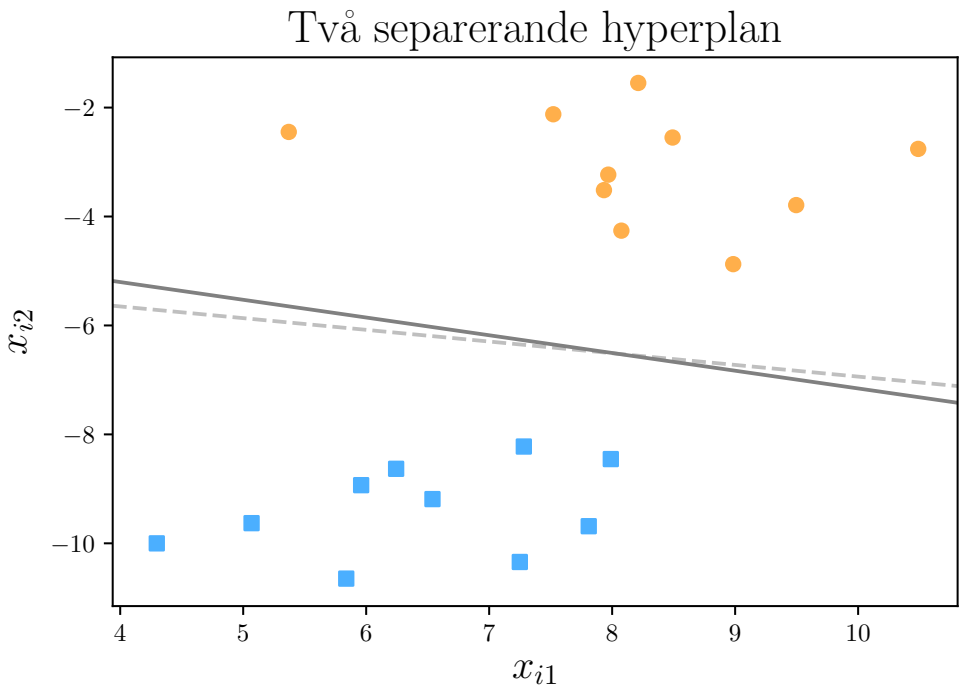
Från beviset av egenskap 1 i sats 3.1.1 och efter man arbetat med definitionen för ett hyperplan L lite kan man ana att parametrarna $\boldsymbol{\beta}$ och β_0 inte är entydigt bestämda, med andra ord kanske det finns andra parametrar som ger samma hyperplan. Detta utreds i följande exempel.

Exempel 3.1.2. Betrakta hyperplanen $L_1 = \{\mathbf{x} : f(\mathbf{x}) = \langle \mathbf{x}, \boldsymbol{\beta} \rangle_{\mathcal{H}} + \beta_0 = 0\}$ och $L_2 = \{\mathbf{x} : g(\mathbf{x}) = \langle \mathbf{x}, -\boldsymbol{\beta} \rangle_{\mathcal{H}} + (-\beta_0) = 0\}$. Eftersom $g(\mathbf{x}) = -f(\mathbf{x})$ gäller att om $\mathbf{x}_0$ tillhör L_1 så tillhör $\mathbf{x}_0$ även L_2. Betrakta vidare hyperplanet $L_3 = \left\{\mathbf{x} : h(\mathbf{x}) = \frac{\langle \mathbf{x}, \boldsymbol{\beta} \rangle_{\mathcal{H}}}{\|\boldsymbol{\beta}\|_{\mathcal{H}}} + \frac{\beta_0}{\|\boldsymbol{\beta}\|_{\mathcal{H}}} = 0\right\}$. Om punkten $\mathbf{x}_0$ tillhör L_1 så tillhör punkten $\mathbf{x}_0$ även L_3 eftersom $h(\mathbf{x}) = \frac{f(\mathbf{x})}{\|\boldsymbol{\beta}\|_{\mathcal{H}}} = 0$. Notera även att $\frac{1}{\|\boldsymbol{\beta}\|_{\mathcal{H}}}$ kunde ha varit vilket reellt tal som helst.

För att få entydiga hyperplan för klassificering kan man lägga till villkor. Om man kräver att $\|\boldsymbol{\beta}\|_{\mathcal{H}} = 1$ och $\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle_{\mathcal{H}} + \beta_0 \geq 0$ för alla i:n sådana att $y_i = 1$, där $(\mathbf{x}_i, y_i)$ är observationsparen i klassificeringsproblemet, så får man en entydig definition av hyperplanet där observationerna $\mathbf{x}_i$ i klassen $y_i = 1$ ligger över hyperplanet. Det första villkoret gör att hyperplanen L_1 och L_3 har samma parametrar eftersom $\|\boldsymbol{\beta}\|_{\mathcal{H}} = 1$. Vidare är endera L_1 eller L_2 sådant att det andra tilläggsvillkoret inte uppfylls. De extra villkoren

gör alltså att man inte längre kan göra manipulationerna som påvisade icke-entydighet. Om man dessutom sätter $\mathbf{x} = \mathbf{0}$ så får man med hjälp av sats 3.1.1 att det signerade avståndet från origo till hyperplanet (i relation till riktningen på $\boldsymbol{\beta}$) är lika med

$$\frac{1}{\|\boldsymbol{\beta}\|_p}\left(\langle \mathbf{x}, \boldsymbol{\beta}\rangle_{\mathcal{H}} + \beta_0\right) = \frac{1}{\|\boldsymbol{\beta}\|_p}\left(\langle \mathbf{0}, \boldsymbol{\beta}\rangle_{\mathcal{H}} + \beta_0\right) = \beta_0.$$



Figur 3.1: 20 datapunkter med två separerande hyperplan (linjer) där klassen $y_i = 1$ framställs som blå fyrkanter och klassen $y_i = -1$ som orangea cirklar.

Definition 3.1.3. Ett klassificeringsproblem eller en mängd observationspar $(\mathbf{x}_i, y_i)$ är *linjärt separabelt* om det existerar ett hyperplan $L = \{\mathbf{x} : f(\mathbf{x}) = \langle \mathbf{x}, \boldsymbol{\beta}\rangle_{\mathcal{H}} + \beta_0 = 0\}$ 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*.

Genom att byta tecken på $\boldsymbol{\beta}$ och β_0 kan man se att ifall det existerar ett hyperplan $\{\mathbf{x} : f(\mathbf{x}) = \langle \mathbf{x}, \boldsymbol{\beta}\rangle_{\mathcal{H}} + \beta_0 = 0\}$ sådant att punkten $\mathbf{x}_i$ ligger under hyperplanet om $y_i = 1$ och över om $y_i = -1$ så finns även ett sådant att punkten $\mathbf{x}_i$ ligger över hyperplanet om $y_i = 1$ och under om $y_i = -1$. Konventionen är att man väljer det hyperplan som passar ovanstående definition.

Sats 3.1.2 (Enligt [5])**.** För ett separerande hyperplan $L = \{\mathbf{x} : f(\mathbf{x}) = \langle \mathbf{x}, \boldsymbol{\beta} \rangle_{\mathcal{H}} + \beta_0 = 0\}$ gäller att

$$y_i \left(\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle_{\mathcal{H}} + \beta_0 \right) > 0, \qquad i = 1, \ \ldots, \ N.$$

Bevis. Ifall ett klassificeringsproblem är linjärt separabelt så ligger alla observationer y_i på rätt sida om det separerande hyperplanet. Det betyder att om $y_i = 1$ så är $\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle_{\mathcal{H}} + \beta_0 > 0$ och om $y_i = -1$ så är $\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle_{\mathcal{H}} + \beta_0 < 0$. Då fås $y_i \left(\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle_{\mathcal{H}} + \beta_0 \right) > 0$, $i = 1, \ \ldots, \ N$. Ifall $\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle_{\mathcal{H}} + \beta_0 = 0$ är problemet inte linjärt separabelt. ■

Klassificeringsregeln g för separerande hyperplan blir

$$g(\mathbf{x}_i) = \begin{cases} 1 & \text{om } \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle_{\mathcal{H}} + \beta_0 > 0, \\ -1 & \text{om } \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle_{\mathcal{H}} + \beta_0 < 0, \end{cases} \qquad \text{där } \mathbf{x}_i \text{ är en observation.}$$

Exempel 3.1.3. Låt observationsparen vara $([2, \, 2]^\intercal, \, 1)$, $([1, \, 2]^\intercal, -1)$, inreproduktrummet i fråga är då $\mathbb{R}^2$ och inreprodukten $\langle \mathbf{x}, \mathbf{y} \rangle_2 := \mathbf{x}^\intercal \mathbf{y}$. 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:

$$y_1 \left(\mathbf{x}_1^\intercal \begin{bmatrix} 1 \\ 0 \end{bmatrix} - 1{,}5 \right) = [2, \, 2] \begin{bmatrix} 1 \\ 0 \end{bmatrix} - 1{,}5 = 0{,}5 > 0$$

och

$$y_2 \left(\mathbf{x}_2^\intercal \begin{bmatrix} 1 \\ 0 \end{bmatrix} - 1{,}5 \right) = -1 \left([1, \, 2] \begin{bmatrix} 1 \\ 0 \end{bmatrix} - 1{,}5 \right) = (-1)(-0{,}5) = 0{,}5 > 0.$$

Och för L_2:

$$y_1 \left(\mathbf{x}_1^\intercal \begin{bmatrix} \sqrt{2} \\ \sqrt{2} \end{bmatrix} - 3{,}5\sqrt{2} \right) = [2, \, 2] \begin{bmatrix} \sqrt{2} \\ \sqrt{2} \end{bmatrix} - 3{,}5\sqrt{2} = 0{,}5\sqrt{2} > 0$$

och

$$\begin{aligned} y_2 \left(\mathbf{x}_2^\intercal \begin{bmatrix} \sqrt{2} \\ \sqrt{2} \end{bmatrix} - 3{,}5\sqrt{2} \right) &= -1 \left([1, \, 2] \begin{bmatrix} \sqrt{2} \\ \sqrt{2} \end{bmatrix} - 3{,}5\sqrt{2} \right) \\ &= (-1) \left(-0{,}5\sqrt{2} \right) = 0{,}5\sqrt{2} > 0 \end{aligned}$$

■

Observation. Hyperplan i $\mathbb{R}^p$ kan konstrueras genom att man väljer p stycken punkter $\mathbf{x}_i$ som man vill att hyperplanet ska gå igenom och sedan löser ekvationssystemet $X\boldsymbol{\beta} = -\beta_0\mathbf{1}$, i vilket X är en matris där raderna består av punkterna $\mathbf{x}_i^\intercal$, $i = 1, \ldots, p$, och $\beta_0\mathbf{1}$ är en vektor med värdet β_0 i alla rader. Med andra ord löser man ekvationssystemet

$$\begin{bmatrix} \mathbf{x}_{1,1} & \mathbf{x}_{1,2} & \cdots & \mathbf{x}_{1,p-1} & \mathbf{x}_{1,p} \\ \mathbf{x}_{2,1} & \mathbf{x}_{2,2} & & & \mathbf{x}_{2,p} \\ \vdots & & \ddots & & \vdots \\ \mathbf{x}_{p-1,1} & & & \mathbf{x}_{p-1,p-1} & \mathbf{x}_{p-1,p} \\ \mathbf{x}_{p,1} & \mathbf{x}_{p,2} & \cdots & \mathbf{x}_{p,p-1} & \mathbf{x}_{p,p} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_{p-1} \\ \boldsymbol{\beta}_p \end{bmatrix} = \begin{bmatrix} -\beta_0 \\ -\beta_0 \\ \vdots \\ -\beta_0 \\ -\beta_0 \end{bmatrix}.$$

Ekvationssystemet kan aldrig vara överbestämt men nog underbestämt ifall punkterna inte är linjärt oberoende.

Som syns i exempel 3.1.3 finns det ofta många separerande hyperplan om ett klassificeringsproblem är linjärt separabelt och frågan är då vilket separerande hyperplan man borde välja.

3.2 Optimala separerande hyperplan

Metoder för att modellera data inom statistik och maskininlärning kan ofta visas vara ekvivalenta med något optimeringsproblem, till exempel maximum likelihood-metoden för linjär regression som är ekvivalent med minstakvadratmetoden [7]. Optimeringsproblemen kan ofta modifieras genom att man lägger till eller tar bort termer i objektfunktionen eller ändrar på kraven. På så sätt får man en ny metod (för att modellera data) med andra egenskaper.

För metoden med optimalt separerande hyperplan är tanken att om man hittar ett hyperplan sådant att:

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

så borde hyperplanet även fungera bra för att separera och klassificera nya observationer [18]. Matematiskt kan man uttrycka problemet som följande optimeringsproblem

$$\begin{aligned} \max_{\widehat{\boldsymbol{\beta}},\widehat{\beta}_0,\left\|\widehat{\boldsymbol{\beta}}\right\|_p=1} \quad & C \\ \text{så att} \quad & y_i\left(\left\langle \mathbf{x}_i, \widehat{\boldsymbol{\beta}} \right\rangle_p + \widehat{\beta}_0\right) \geq C, \quad i = 1, \ldots, N, \end{aligned} \tag{3.1}$$

där C kallas *marginalen* och betecknar avståndet från hyperplanet till de närmaste observationerna. Här betecknar $(\mathbf{x}_i, y_i)$ observationsparen man tränar algoritmen på, där $\mathbf{x}_i \in \mathbb{R}^p$ och $y_i \in \{-1, 1\}$ för $i = 1, \ldots, N$, detta gäller även för resten av kapitlet om inte annat anges.

Kraven $y_i \left(\left\langle \mathbf{x}_i, \widehat{\boldsymbol{\beta}} \right\rangle_p + \widehat{\beta}_0 \right) \geq C, \ i = 1, \ \ldots, \ N$, ska tolkas med hjälp av satserna 3.1.1 och 3.1.2. Ifall alla punkter är rätt klassificerade ger satserna att $y_i \left(\left\langle \mathbf{x}_i, \widehat{\boldsymbol{\beta}} \right\rangle_p + \widehat{\beta}_0 \right)$ är det absoluta avståndet mellan hyperplanet och punkten $\mathbf{x}_i$, som man kräver att är större än C, det man försöker maximera. Förhoppningen är att om man väljer det separerande hyperplan som befinner sig så långt som möjligt från båda klasserna får man ett hyperplan som även generaliserar väl till ny data. Dessutom är detta ett unikt sätt att välja ett separerande hyperplan det vill säga optimeringsproblemet är konvext [8] vilket ska visas.

För att visa att optimeringsproblemet (3.1) är konvext måste det skrivas om. Idén är att man låter inversen av längden på vektorn $\boldsymbol{\beta}$ beskriva avståndet till närmast punkt, då skapas en direktare länk mellan kraven och objektfunktionen i optimeringsproblemet.

Först måste alltså kravet $\left\| \widehat{\boldsymbol{\beta}} \right\|_p = 1$ bytas ut. Det görs genom att man byter ut kraven

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

mot kraven

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

eller ekvivalent

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

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 $C = \frac{1}{\|\boldsymbol{\beta}\|_p}$, då fås

$$y_i \left(\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle_p + \beta_0 \right) \geq 1, \quad i = 1, \ \ldots, \ N,$$

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

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

Därefter görs ännu en konvexitetsbevarande kvadratisk transformering av objektfunktionen[2] $\|\boldsymbol{\beta}\|_p$, det vill säga man noterar att om $\boldsymbol{\beta}^*$ är sådan att $\min_{\boldsymbol{\beta},\beta_0} \|\boldsymbol{\beta}\|_p = \|\boldsymbol{\beta}^*\|_p$, så gäller även $\min_{\boldsymbol{\beta},\beta_0} \frac{1}{2}\|\boldsymbol{\beta}\|_p^2 = \frac{1}{2}\|\boldsymbol{\beta}^*\|_p^2$ för samma $\boldsymbol{\beta}^*$. Detta brukar betecknas med funktionen $\operatorname{argmin}_{\mathbf{y}}(f(\mathbf{y}))$ som ger som resultat det $\mathbf{y}^*$ som minimerar funktionen $f(\mathbf{y})$.

En orsak till att göra den kvadratiska transformeringen är att man då kan garantera att objektfunktionen är deriverbar:

Exempel 3.2.1. Låt $x \in \mathbb{R}$, $f(x) := |x - x_0|$ och $g(x) := (f(x))^2$. Då är $f(x)$ inte deriverbar i punkten x_0 medan $D_x(g(x)) = D_x\left((x-x_0)^2\right) = D_x(x^2 - 2xx_0 + x_0^2) = 2x - 2x_0 = 2(x - x_0)$ och $D_x(g(x_0)) = 0$ det vill säga $\operatorname{argmin}_x f(x) = x_0 = \operatorname{argmin}_x g(x)$.

Optimeringsproblemet (3.1) kan alltså skrivas på formen

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

där $\left(\frac{1}{2}I\right)$ är en *positivt semi-definit* matris, det vill säga den uppfyller kraven för att ett kvadratiskt optimeringsproblem med linjära lösbara krav ska vara konvext [5].

Ovanstående resonemang ger ett bevis för sats 3.2.1:

Sats 3.2.1 (Enligt [8])**.** 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

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

konvext och man får en optimal lösning $\left(\widehat{\boldsymbol{\beta}^*},\widehat{\beta_0^*}\right) = \left(\frac{\boldsymbol{\beta}^*}{\|\boldsymbol{\beta}^*\|_p}, \frac{\beta_0^*}{\|\boldsymbol{\beta}^*\|_p}\right)$ genom att lösa optimeringsproblemet

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

[2]Inom statistik- och maskininlärningslitteraturen kallas objektfunktionen ibland även för *kostfunktionen*.

3.2.1 Primala och duala Lagrangeproblem

För att hitta alla extrempunkter till ett optimeringsproblem, det vill säga lösa ett konvext optimeringsproblem, används Lagrangemultiplikatorer [5]. Den primala Lagrangefunktionen L_P för optimeringsproblemet

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

ges av

$$L_P = \frac{1}{2}\left\langle \boldsymbol{\beta}, \boldsymbol{\beta} \right\rangle_p - \sum_{i=1}^{N} \lambda_i \left(y_i\left(\left\langle \mathbf{x}_i, \boldsymbol{\beta} \right\rangle_p + \beta_0\right) - 1\right) \tag{3.2}$$

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

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

$$\begin{aligned} D_{\left[\boldsymbol{\beta}\right]_j}\left(L_P\right) &= D_{\left[\boldsymbol{\beta}\right]_j}\left(\frac{1}{2}\boldsymbol{\beta}^\intercal\boldsymbol{\beta}\right) && -D_{\left[\boldsymbol{\beta}\right]_j}\left(\sum_{i=1}^{N}\left(\lambda_i y_i\left(\mathbf{x}_i^\intercal\boldsymbol{\beta}\right) + \lambda_i y_i\beta_0 - \lambda_i\right)\right) \\ &= D_{\left[\boldsymbol{\beta}\right]_j}\left(\frac{1}{2}\sum_{k=1}^{p}\left[\boldsymbol{\beta}\right]_k^2\right) && -\sum_{i=1}^{N} D_{\left[\boldsymbol{\beta}\right]_j}\Bigg(\lambda_i y_i\left(\sum_{k=1}^{p}\left[\mathbf{x}_i\right]_k\left[\boldsymbol{\beta}\right]_k\right) \\ & && \qquad + \lambda_i y_i\beta_0 - \lambda_i\Bigg) \\ &= \left[\boldsymbol{\beta}\right]_j && -\sum_{i=1}^{N} D_{\left[\boldsymbol{\beta}\right]_j}\left(\sum_{k=1}^{p}\lambda_i y_i\left[\mathbf{x}_i\right]_k\left[\boldsymbol{\beta}\right]_k\right) + 0 \\ &= \left[\boldsymbol{\beta}\right]_j && -\sum_{i=1}^{N}\lambda_i y_i\left[\mathbf{x}_i\right]_j \end{aligned} \tag{3.3}$$

där $j = 1,\ \ldots,\ p$ och

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

Vidare kan härledningen (3.3) skrivas om som derivatan med avseende på hela $\boldsymbol{\beta}$ eftersom $\left[D_{\boldsymbol{\beta}}\left(L_p\right)\right]_j = D_{\left[\boldsymbol{\beta}\right]_j}\left(L_p\right)$. Efter att man löser derivatorna för nollställen får man följande krav:

$$\boldsymbol{\beta} = \sum_{i=1}^{N}\lambda_i y_i\mathbf{x}_i, \tag{3.4}$$

$$0 = \sum_{i=1}^{N}\lambda_i y_i. \tag{3.5}$$

Insättning av kraven (3.4) och (3.5) i L_P ger det duala Lagrangeproblemet med följande duala Lagrangefunktion

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

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

$$\lambda_i \geq 0, \quad i = 1, \ldots, N. \tag{3.6}$$

Uträkningarna och kravet $\lambda_i \geq 0$, $i = 1, \ldots, N$, kan motiveras genom Karush-Kuhn-Tucker kraven [5] för konvexa problem, det vill säga kraven (3.4), (3.5) och (3.6) samt kravet

$$\lambda_i \left(y_i \left(\left\langle \mathbf{x}_i, \boldsymbol{\beta} \right\rangle_p + \beta_0 \right) - 1 \right) = 0, \quad i = 1, \ldots, N. \tag{3.7}$$

Observation. Kraven (3.4) till (3.7) säger något om hurudan den optimala lösningen $(\boldsymbol{\beta}^*, \beta_0^*, \lambda_1^*, \ldots, \lambda_N^*)$ måste vara:

- Krav (3.4) 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 (3.7) att $y_i \left(\left\langle \mathbf{x}_i, \boldsymbol{\beta}^* \right\rangle_p + \beta_0^* \right) = 1$ vilket enligt det ursprungliga optimeringsproblemet (3.1) ska tolkas som att punkten $\mathbf{x}_i$ ligger på avståndet C från det separerande hyperplanet. Punkten $\mathbf{x}_i$ är med andra ord en av punkterna som ligger närmast det separerande hyperplanet.
- Ifall $y_i \left(\left\langle \mathbf{x}_i, \boldsymbol{\beta}^* \right\rangle_p + \beta_0^* \right) > 1$ 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 $\boldsymbol{\beta}^*$ inte bara är en linjär kombination av observationerna $\mathbf{x}_i$, utan mer exakt 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*.

Kvar finns också möjligheten att $\lambda_i^* = 0$ och $y_i\left(\langle \mathbf{x}_i, \boldsymbol{\beta}^* \rangle_p + \beta_0^*\right) = 1$ för något i. Detta kan tolkas som att punkten $\mathbf{x}_i$ ligger på randen av marginalen men inte behövs för att beskriva vektorn $\boldsymbol{\beta}^*$, punkten $\mathbf{x}_i$ är alltså redan en linjär kombination av andra punkter på marginalen och ligger på samma hyperplan som resten av punkterna på marginalen. Existensen av en entydig lösning till optimeringsproblemet kan då inte garanteras men sannolikheten att punkterna som ligger närmast det optimala separerande hyperplanet ligger på exakt samma hyperplan är mycket liten (sannolikheten är lika med 0 om punkternas koordinater dras från en kontinuerlig fördelning).

3.3 Det oseparabla fallet

Antag att observationsparen $(\mathbf{x}_i, y_i)$ inte är linjärt separabla, det vill säga inget hyperplan $\left\{\mathbf{x} : f(\mathbf{x}) = \langle \mathbf{x}, \boldsymbol{\beta} \rangle_p + \beta_0 = 0\right\}$ med $y_i\left(\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle_p + \beta_0\right) > 0$ för alla träningspar $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, N$ existerar. Oseparabla observationspar leder till att optimeringsproblemet (3.1) samt optimeringsproblemen i sats 3.2.1 inte längre är lösbara eftersom det inte existerar något hyperplan som satisfierar alla krav.

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 optimeringsproblemet (3.1) finns två naturliga sätt att ändra på kraven [8], endera låter man

$$y_i\left(\left\langle \mathbf{x}_i, \widehat{\boldsymbol{\beta}} \right\rangle_p + \widehat{\beta}_0\right) \geq C - s_i, \tag{3.8}$$

eller

$$y_i\left(\left\langle \mathbf{x}_i, \widehat{\boldsymbol{\beta}} \right\rangle_p + \widehat{\beta}_0\right) \geq C\left(1 - s_i\right), \tag{3.9}$$

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

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

Alternativ (3.8) 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 (3.9) gäller istället att observationen $\mathbf{x}_i$ tillåts vara $C \cdot s_i$ 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 (3.9) är det mest allmänt använda [8].

För hyperplan med mjuka marginaler blir det ursprungliga optimeringsproblemet:

$$\begin{aligned} \max_{\widehat{\boldsymbol{\beta}}, \widehat{\beta}_0, \left\|\widehat{\boldsymbol{\beta}}\right\|_p = 1} \quad & C \\ \text{så att} \quad & y_i \left(\left\langle \mathbf{x}_i, \widehat{\boldsymbol{\beta}} \right\rangle_p + \widehat{\beta}_0 \right) \geq C \left(1 - s_i\right), \quad i = 1, \ \ldots, \ N, \\ & s_i \geq 0, \quad i = 1, \ \ldots, \ N, \\ & \sum_{i=1}^{N} s_i \leq K. \end{aligned} \tag{3.10}$$

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 3.2.1 så stöter man på problem – efter att man sätter $\widehat{\boldsymbol{\beta}} = \frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|_p}$ och $C = \frac{1}{\|\boldsymbol{\beta}\|_p}$ får man optimeringsproblemet

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

där slackvariablerna direkt beror på $\|\boldsymbol{\beta}\|_p$ som man indirekt försöker minimera. Beräkningar blir då mera komplicerade.

För optimeringsproblemet (3.10) ger stegen i beviset för sats 3.2.1 istället ett bevis för sats 3.3.1:

Sats 3.3.1 (Enligt [8])**.** 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 optimeringsproblemens krav är lösbara för givna observationspar $(\mathbf{x}_i, y_i)$. Då är optimeringsproblemet

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

konvext och man får en optimal lösning $\left(\widehat{\boldsymbol{\beta}^*}, \widehat{\beta_0^*}\right) = \left(\frac{\boldsymbol{\beta}^*}{\|\boldsymbol{\beta}^*\|_p}, \frac{\beta_0^*}{\|\boldsymbol{\beta}^*\|_p}\right)$ genom att lösa optimeringsproblemet

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

Eftersom kraven för optimeringsproblemen i sats 3.3.1 är mer tillåtande än motsvarande krav i sats 3.2.1 kan hyperplan med mjuka marginaler 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 $K = 0$, då måste alla s_i vara lika med 0 och då är kraven i sats 3.3.1 de samma som de i sats 3.2.1 (man måste givetvis också kräva att observationsparen $(\mathbf{x}_i, y_i)$ är linjärt separabla).

Att använda hyperplan med mjuka marginaler kan vara en bra idé till exempel om man har outliers eller felklassificerade observationer 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 för att approximera kravet med en term i objektfunktionen, en så kallad straffunktion [5]. I [6] 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

$$\begin{aligned} \min_{\boldsymbol{\beta},\beta_0} \quad & \frac{1}{2}\left\|\boldsymbol{\beta}\right\|_p^2 + \gamma \sum_{i=1}^{N} s_i \\ \text{så att} \quad & y_i\left(\left\langle \mathbf{x}_i, \boldsymbol{\beta}\right\rangle_p + \beta_0\right) \geq 1 - s_i, \quad i = 1,\ \ldots,\ N, \\ & s_i \geq 0, \quad i = 1,\ \ldots,\ N, \end{aligned} \tag{3.11}$$

eller

$$\begin{aligned} \min_{\boldsymbol{\beta},\beta_0} \quad & \frac{1}{2}\left\|\boldsymbol{\beta}\right\|_p^2 + \gamma \left(\sum_{i=1}^{N} s_i\right)^2 \\ \text{så att} \quad & y_i\left(\left\langle \mathbf{x}_i, \boldsymbol{\beta}\right\rangle_p + \beta_0\right) \geq 1 - s_i, \quad i = 1,\ \ldots,\ N, \\ & s_i \geq 0, \quad i = 1,\ \ldots,\ N, \end{aligned} \tag{3.12}$$

där γ är en på förhand bestämd parameter. Märk att båda problemen är kvadratiska optimeringsproblem och således kan lösas relativt enkelt.

Tolkningen av optimeringsproblemen (3.11) och (3.12) 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 att om marginalen ökar så ökar även straffet medan om marginalen minskar så minskar straffet. Avvägningen mellan minskning av straff och ökning av marginal bestäms med hjälp av parametern γ som kan jämföras med parametern K i sats 3.3.1. Skillnaden är att om γ är litet så tillåts slackvariablerna vara större och om 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 3.3.1 och optimeringsproblemet (3.11) är att optimeringsproblemet (3.11) alltid är lösbart medan optimeringsproblemen i sats 3.3.1 är lösbara endast om K väljs tillräckligt stort.

Av de två optimeringsproblemen (3.11) och (3.12) är (3.11) det vanligare och behandlas således i resten av avhandlingen.

3.3.1 Primala och duala Lagrangeproblem för mjuka marginaler

Precis som för den ursprungliga algoritmen med hårda marginaler kan den optimala lösningen för optimeringsproblemet (3.11) karaktäriseras med hjälp av Lagrangemultiplikatorer. Den primala Lagrangefunktionen ges av

$$L_P = \frac{1}{2}\left\langle \boldsymbol{\beta}, \boldsymbol{\beta}\right\rangle_p + \gamma \sum_{i=1}^{N} s_i - \sum_{i=1}^{N} \lambda_i \left(y_i\left(\left\langle \mathbf{x}_i, \boldsymbol{\beta}\right\rangle_p + \beta_0\right) - (1 - s_i)\right) - \sum_{i=1}^{N} \mu_i s_i \tag{3.13}$$

som ska minimeras med avseende på $\boldsymbol{\beta}$, β_0 och s_i. För att hitta extrempunkterna räknas först derivatorna med avseende på $\boldsymbol{\beta}$, β_0 och s_i ut, om man följer liknande steg som i härledning (3.3) får man:

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

och

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

Efter att man löser för nollställen får man följande krav:

$$\boldsymbol{\beta} = \sum_{i=1}^{N}\lambda_i y_i \mathbf{x}_i, \tag{3.14}$$

$$0 = \sum_{i=1}^{N}\lambda_i y_i, \tag{3.15}$$

$$\lambda_i = \gamma - \mu_i, \qquad i = 1,\ \ldots,\ N, \tag{3.16}$$

samt kraven

$$\lambda_i \geq 0, \qquad i = 1,\ \ldots,\ N, \tag{3.17}$$

$$\mu_i \geq 0, \qquad i = 1,\ \ldots,\ N, \tag{3.18}$$

$$s_i \geq 0, \qquad i = 1,\ \ldots,\ N. \tag{3.19}$$

Insättning av kraven (3.14), (3.15) och (3.16) i den primala Lagrange-

funktionen (3.13) ger den duala Lagrangefunktionen

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

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

$$\lambda_i \left(y_i \left(\left\langle \mathbf{x}_i, \boldsymbol{\beta} \right\rangle_p + \beta_0 \right) - (1 - s_i) \right) = 0, \qquad i = 1, \ \ldots, \ N, \tag{3.20}$$

$$\mu_i s_i = 0, \qquad i = 1, \ \ldots, \ N, \tag{3.21}$$

$$y_i \left(\left\langle \mathbf{x}_i, \boldsymbol{\beta} \right\rangle_p + \beta_0 \right) - (1 - s_i) \geq 0, \qquad i = 1, \ \ldots, \ N, \tag{3.22}$$

från Karush-Kuhn-Tucker kraven för konvexa problem.

Observation. Precis som för algoritmen med optimala separerande hyperplan kan man karaktärisera lösningen för hyperplan med mjuka marginaler med hjälp av kraven (3.14) till (3.22).

- Krav (3.14) och (3.20) ger att den optimala lösningen $\boldsymbol{\beta}^*$ ges som den linjära kombinationen

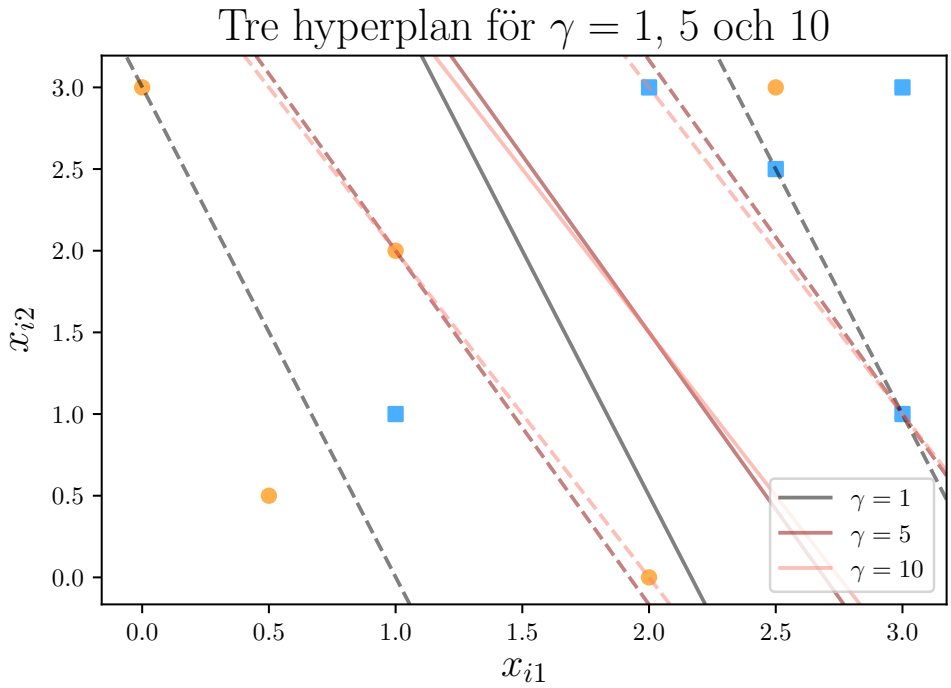
  $$\boldsymbol{\beta}^* = \sum_{i=1}^{N} \lambda_i^* y_i \mathbf{x}_i,$$

  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 (3.16) och (3.21) att $0 < \lambda_i^* < \gamma$.

- För de resterande stödvektorerna ($\lambda_i^* > 0$) gäller $\lambda_i^* = \gamma$.

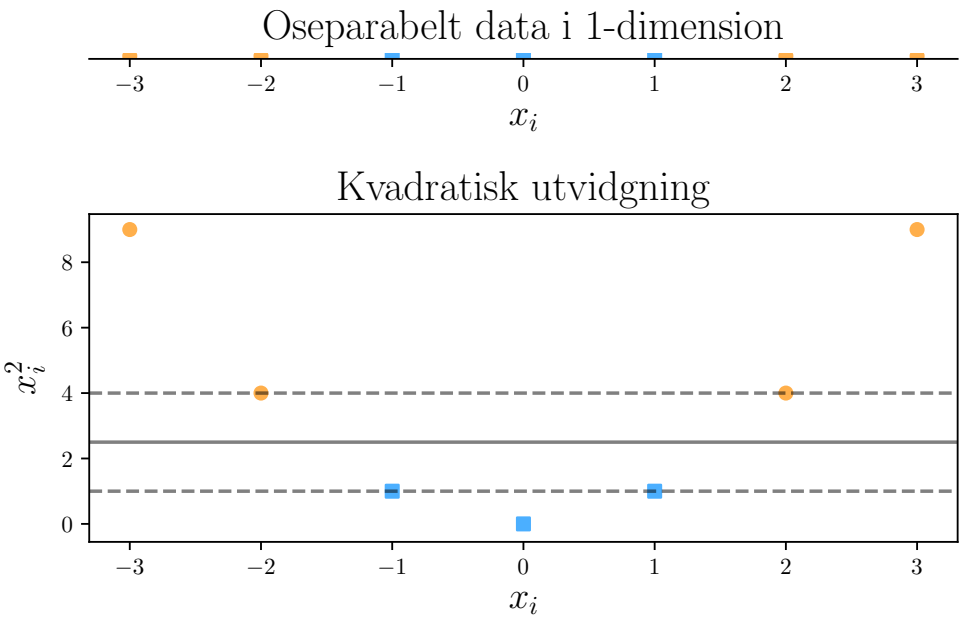
- Vilken som helst av punkterna på marginalen ($\lambda_i^* > 0$, $s_i^* = 0$) kan användas för att lösa för β_0^*.



Figur 3.2: Löst exempel för linjärt oseparabelt data för 3 olika värden på γ. De streckade linjerna är marginalernas ränder.

Exempel 3.3.1. Låt observationsparen $(\mathbf{x}_i, y_i)$ vara sådana som i figur 3.2, 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 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 separerande hyperplan med mjuka marginaler för tre olika värden på γ. Funktionen `SVC` med `kernel='linear'` från paketet `sklearn` [12] användes för att beräkna hyperplanen.

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



Figur 3.3: En lösning med optimala separerande hyperplan och kvadratisk utvidgning där endast hyperplan med mjuka marginaler inte hade fungerat.

Exempel 3.3.2. 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 3.3. 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 figur 3.3 och observationsparen är nu linjärt separabla. Det optimala separerande hyperplanet bestämdes med hjälp av `sklearn`:s metod `SVC` med `kernel='linear'` och `C=1000` [12].

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.

Kapitel 4

Reproducerande kärnor

I detta kapitel kommer grunden till ännu ett sätt att generalisera optimeringsproblemen i kapitel 3 att presenteras. En olinjär generalisering av optimeringsproblemet 3.11 är vad man vanligtvis avser med begreppet stödvektormaskin.

I exempel 3.3.2 bildades andra gradens polynom för varje observation $\mathbf{x}_i$ så att de nya observationerna $\mathbf{x}'_i$ blev $[\mathbf{x}_i, \mathbf{x}_i^2]^\intercal$. Att räkna ut de nya $\mathbf{x}'_i$:na kommer då att kräva N multiplikationer (kvadrering av $\mathbf{x}_i$). Hade man istället bildat tredje gradens polynom hade det krävt $2N$ multiplikationer och att bilda n:te gradens polynom hade krävt nN multiplikationer. Efter utvidgningen löstes optimeringsproblemet

$$\begin{aligned} \min_{\boldsymbol{\beta},\beta_0} \quad & \frac{1}{2}\left\|\boldsymbol{\beta}\right\|_2^2 \\ \text{så att} \quad & y_i\left(\left\langle \mathbf{x}'_i, \boldsymbol{\beta}\right\rangle_2 + \beta_0\right) \geq 1, \quad i = 1,\ \ldots,\ N, \end{aligned}$$

där $\boldsymbol{\beta}$ nu är en 2-dimensionell vektor. För att lösa ovanstående optimeringsproblem undersöks det duala Lagrangeproblemet

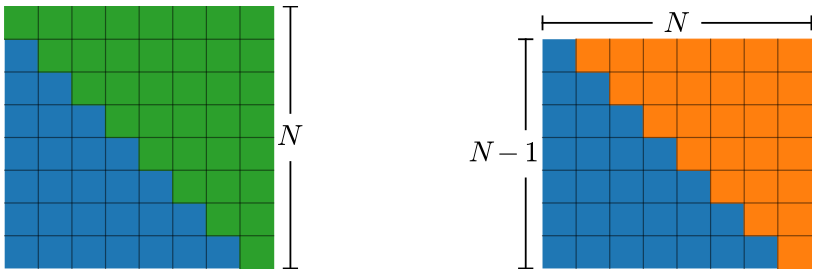
$$L_D = \sum_{i=1}^{N} \lambda_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \left\langle \mathbf{x}'_i, \mathbf{x}'_j \right\rangle_2$$

där den sista dubbelsumman kan representeras som matrismultiplikationen

$$\boldsymbol{\lambda}^\intercal \mathbf{K} \boldsymbol{\lambda}$$

där $\boldsymbol{\lambda}$ är en vektor bestående av värdena på $y_i\lambda_i$ och matrisen $\mathbf{K}$ har värdena $\mathbf{K}_{ij} = \left\langle \mathbf{x}'_i, \mathbf{x}'_j \right\rangle_2$. För att räkna ut $\mathbf{K}$ räknas inreprodukten $\left\langle \mathbf{x}'_i, \mathbf{x}'_j \right\rangle_2$ ut för alla kombinationer men på grund av symmetri måste inte elementen under diagonalen beräknas. Inreprodukten tar 3 operationer i anspråk, 2 för att

multiplicera $\mathbf{x}'_i$ och $\mathbf{x}'_j$:s komponenter komponentvis och 1 för att summera. Inreprodukten måste räknas ut $\frac{N(N+1)}{2}$ gånger (se figur 4.1) så totalt måste en dator genomföra åtminstone $N + 3\frac{N(N+1)}{2}$ operationer före optimeringsproblemet kan börja lösas.



Figur 4.1: Visuellt bevis för hur många gånger man måste räkna ut inreprodukten $\langle \mathbf{x}_i, \mathbf{x}_j \rangle_{\mathcal{H}}$ där $i, j = 1, \ldots, N$. Svaret blir $N^2 - \frac{N(N-1)}{2}$ det vill säga antalet rutor i en $N \times N$ kvadrat minus antalet rutor under diagonalen som räknas till höger. Svaret ger antalet gröna rutor.

Hade man bildat n:te gradens polynom hade antalet operationer varit nN för att räkna ut de nya observationerna och $(n + n - 1)\frac{N(N+1)}{2}$ operationer för matrisen. Dessutom måste inreprodukten $\langle \mathbf{x}'_i, \boldsymbol{\beta} \rangle_n$ räknas ut flera gånger när problemet löses.

Dimensionen på det ursprungliga rummet har inte ännu tagits i beaktande, i exempel 3.3.2 var dimensionen 1 men vad händer om dimensionen är 2 eller man har 16×16 pixlar stora bilder? Det visar sig att om den ursprungliga dimensionen p är stor så får till och med enkla utvidgningar en mycket stor dimension P vilket gör det arbetsamt att räkna ut inreprodukten $\left\langle \mathbf{x}'_i, \mathbf{x}'_j \right\rangle_P$. Om man till exempel vill samla alla monom[1] av graden 5 för en 16×16 pixlar stor bild blir dimensionen nästan 10^{10} [14].

Det visar sig att speciellt inreprodukten går att räkna ut effektivare. Istället för att först göra utvidgningarna kan man byta ut inreprodukten mot en annan funktion med vissa egenskaper. För exempel 3.3.2 skulle den funktionen se ut på följande sätt:

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 exempel 3.3.2. Följande inreprodukt

[1] Till exempel $\mathbf{x}_{i1}\mathbf{x}_{i2}\mathbf{x}_{i3}^2$ är ett monom av graden 5 medan $\mathbf{x}_{i1}\mathbf{x}_{i2}\mathbf{x}_{i3}$ inte är det och inte $\mathbf{x}_{i1}\mathbf{x}_{i2}\mathbf{x}_{i3}^4$ heller.

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 $\mathbb{R}^p$. Med andra ord kan man istället för att använda inreprodukten $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_3$ räkna med 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 $\mathbf{K}$, 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 (här används inreprodukten i $\mathbb{R}^2$ då den sista komponenten i $\phi(\mathbf{x})$ är konstant det vill säga onödig), 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[2]), medan inreprodukten $\langle \cdot, \cdot \rangle_4$ kräver 7 operationer att räkna.

Man behöver dessutom inte på förhand räkna ut de nya komponenterna för alla observationer och sparar således in $2N$ operationer. För mera realistiska exempel med större dimensioner sparar man ofta mycket mera tid.

Funktioner $k(\mathbf{x}_1, \mathbf{x}_2)$ som kan uttryckas som en inreprodukt av en funktion ϕ evaluerad i två olika punkter $\mathbf{x}_1$ och $\mathbf{x}_2$ brukar kallas *kärnor* och studerades först av David Hilbert [9] i samband med studien av integraloperatorn $T_k f(x) = \int_X k(x_1, x_2) f(x_2)\, dx_2$ där funktionen k är operatorn T_k:s kärna [14].

4.1 Kärnor som inreprodukter

Ovanstående exempel motiverar studien samt definitionen av *kärnor*:

Definition 4.1.1 (Enligt [14])**.** 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 man $\Phi_{\mathbf{y}}(\mathbf{x}) = k(\mathbf{x}, \mathbf{y})$ 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 $k(\mathbf{x}, \mathbf{y}) := (\langle \mathbf{x}, \mathbf{y} \rangle_1 + 1)^2 = (\mathbf{x}\mathbf{y})^2 + 2\mathbf{x}\mathbf{y} + 1$, kan ett vektorrum

[2]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.

av funktioner definiera 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

$$\langle f, g \rangle_k := \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \beta_j k(\mathbf{x}_i, \mathbf{x}_j). \tag{4.1}$$

Definitionen för $\langle f, g \rangle_k$ innehåller samma koefficienter som de linjära kombinationer som definierar f och g men eftersom

$$\langle f, g \rangle_k = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \beta_j k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i=1}^{N} \alpha_i \sum_{j=1}^{N} \beta_j k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i=1}^{N} \alpha_i g(\mathbf{x}_i) \tag{4.2}$$

och

$$\begin{aligned} \langle f, g \rangle_k &= \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \beta_j \left((\mathbf{x}_i \mathbf{x}_j)^2 + 2\mathbf{x}_i \mathbf{x}_j + 1 \right) \\ &= \sum_{j=1}^{N} \sum_{i=1}^{N} \alpha_i \beta_j \left((\mathbf{x}_j \mathbf{x}_i)^2 + 2\mathbf{x}_j \mathbf{x}_i + 1 \right) = \sum_{j=1}^{N} \beta_j \sum_{i=1}^{N} \alpha_i k(\mathbf{x}_j, \mathbf{x}_i) \\ &= \sum_{j=1}^{N} \beta_j f(\mathbf{x}_j) \end{aligned} \tag{4.3}$$

beror inte summan på vilka linjära kombinationer man väljer för funktionerna f och g ifall de inte går att välja unikt. Som mellansteg visades även att $k(\mathbf{x}_i, \mathbf{x}_j) = k(\mathbf{x}_j, \mathbf{x}_i)$.

För att visa att $\langle f, g \rangle_k$ är en inreprodukt måste man kolla att alla villkor i definition 2.1.1 är uppfyllda:

IP1 $\langle f, g \rangle_k = \langle g, f \rangle_k$:

$$\langle f, g \rangle_k = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \beta_j k(\mathbf{x}_i, \mathbf{x}_j) \overset{(4.3)}{=} \sum_{j=1}^{N} \sum_{i=1}^{N} \beta_j \alpha_i k(\mathbf{x}_j, \mathbf{x}_i) = \langle g, f \rangle_k$$

där man kan kasta om summorna eftersom de är ändliga. ■

IP2 $\langle \lambda f, g \rangle_k = \lambda \langle f, g \rangle_k$, $\lambda \in \mathbb{R}$:

$$\langle \lambda f, g \rangle_k \overset{(4.3)}{=} \sum_{j=1}^{N} \beta_j \lambda f(\mathbf{x}_j) = \lambda \sum_{j=1}^{N} \beta_j f(\mathbf{x}_j) = \lambda \langle f, g \rangle_k. \qquad \blacksquare$$

IP3 $\langle f + h, g \rangle_k = \langle f, g \rangle_k + \langle h, g \rangle_k$:

$$\begin{aligned} \langle f + h, g \rangle_k &\overset{(4.3)}{=} \sum_{j=1}^{N} \beta_j \left(f(\mathbf{x}_j) + h(\mathbf{x}_j) \right) \\ &= \sum_{j=1}^{N} \beta_j f(\mathbf{x}_j) + \sum_{j=1}^{N} \beta_j h(\mathbf{x}_j) \\ &= \langle f, g \rangle_k + \langle h, g \rangle_k \end{aligned}$$

där även h kan skrivas som en linjär kombination men beviset beror inte på vilken kombination man väljer. $\blacksquare$

IP4 $\langle f, f \rangle_k \geq 0$:

$$\begin{aligned} \langle f, f \rangle_k &= \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_3 \\ &= \sum_{i=1}^{N} \alpha_i \sum_{j=1}^{N} \alpha_j \langle \phi(\mathbf{x}_j), \phi(\mathbf{x}_i) \rangle_3 = \sum_{i=1}^{N} \alpha_i \left\langle \sum_{j=1}^{N} \alpha_j \phi(\mathbf{x}_j), \phi(\mathbf{x}_i) \right\rangle_3 \\ &= \sum_{i=1}^{N} \alpha_i \left\langle \phi(\mathbf{x}_i), \sum_{j=1}^{N} \alpha_j \phi(\mathbf{x}_j) \right\rangle_3 = \left\langle \sum_{i=1}^{N} \alpha_i \phi(\mathbf{x}_i), \sum_{j=1}^{N} \alpha_j \phi(\mathbf{x}_j) \right\rangle_3 \\ &= \left\| \sum_{i=1}^{N} \alpha_i \phi(\mathbf{x}_i) \right\|_3^2 \geq 0 \end{aligned}$$

där likhet gäller om och endast om $\sum_{i=1}^{N} \alpha_i \phi(\mathbf{x}_i) = \mathbf{0}$.

För

$$f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i k(\mathbf{x}, \mathbf{x}_i) = \sum_{i=1}^{N} \alpha_i \langle \phi(\mathbf{x}), \phi(\mathbf{x}_i) \rangle_3 = \left\langle \phi(\mathbf{x}), \sum_{i=1}^{N} \alpha_i \phi(\mathbf{x}_i) \right\rangle_3$$

gäller då att om $\langle f, f \rangle_k = 0$ så är $\sum_{i=1}^{N} \alpha_i \phi(\mathbf{x}_i) = \mathbf{0}$ och då även

$$f(\mathbf{x}) = \left\langle \phi(\mathbf{x}), \sum_{i=1}^{N} \alpha_i \phi(\mathbf{x}_i) \right\rangle_3 = \langle \phi(\mathbf{x}), \mathbf{0} \rangle_3 = 0.$$

Med andra ord är $\langle f, f \rangle_k \geq 0$ där likhet gäller om och endast om $f(\mathbf{x}) = 0$ för alla $\mathbf{x}$. $\blacksquare$

Eftersom alla villkor i definition 2.1.1 är uppfyllda definierar

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

en inreprodukt. Här är $k\left(\mathbf{x}, \mathbf{y}\right) = \left(\langle \mathbf{x}, \mathbf{y}\rangle_1 + 1\right)^2 = \langle \phi\left(\mathbf{x}\right), \phi\left(\mathbf{y}\right)\rangle_3$ och $\phi\left(\mathbf{x}\right) = \left[\mathbf{x}^2, \sqrt{2}\mathbf{x}, 1\right]^\intercal$. Om man dessutom kräver att alla Cauchyföljder konvergerar till en funktion i samma rum så är rummet ett Hilbertrum vilket betyder att allting i kapitel 2 gäller även i det nya Hilbertrummet.

I beviset användes egenskap (4.3) flitigt. Det visar sig att det räcker med att kärnan k är symmetrisk det vill säga $k\left(\mathbf{x}_i, \mathbf{x}_j\right) = k\left(\mathbf{x}_j, \mathbf{x}_i\right)$ som garanteras om kärnan kan definieras som inreprodukten mellan någon funktion $\phi\left(\mathbf{x}\right)$ evaluerad i punkterna $\mathbf{x}_i$ och $\mathbf{x}_j$. Med andra ord kan man dra slutsatsen att följande sats gäller:

Sats 4.1.1 (Enligt [14])**.** För en kärna k definierad som i definition 4.1.1, det vill säga $k\left(\mathbf{x}, \mathbf{y}\right) = \langle \phi\left(\mathbf{x}\right), \phi\left(\mathbf{y}\right)\rangle_P$ där $\phi\left(\cdot\right)$ är någon utvidgning, är rummet av funktioner definierade genom $f\left(\mathbf{x}\right) = \sum_{i=1}^{N} \alpha_i k\left(\mathbf{x}, \mathbf{x}_i\right)$ med inreprodukten $\langle f, g\rangle_k = \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\beta_j k\left(\mathbf{x}_i, \mathbf{x}_j\right)$ ett inreproduktrum.

Korollarium 4.1.2 (Enligt [14])**.** Om dessutom alla Cauchyföljder konvergerar till en funktion i samma rum, med avseende på normen inducerad av inreprodukten $\langle f, g\rangle_k$, så är rummet ett Hilbertrum.

Nu kan man fundera på om kärnan k måste vara definierad genom en inreprodukt, det kan ju hända att även andra funktioner $X \times X \longmapsto \mathbb{R}$ ger upphov till ett liknande Hilbertrum. Som sagt så garanterar inreprodukten att k är symmetrisk men det finns säkert symmetriska funktioner som inte kan definieras genom inreprodukter. På samma sätt garanterar inreprodukten att egenskap **IP4** uppfylls men man kan istället genast kräva att

$$\langle f, f\rangle_k = \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\alpha_j k\left(\mathbf{x}_i, \mathbf{x}_j\right) \geq 0$$

med likhet om och endast om $f = 0$. Detta krav kan skrivas om som en matrismultiplikation:

$$\langle f, f\rangle_k = \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\alpha_j k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \boldsymbol{\alpha}^\intercal \mathbf{K}\boldsymbol{\alpha} \geq 0 \tag{4.4}$$

där $\boldsymbol{\alpha}$ är en vektor med komponenterna $\alpha_1,\ \alpha_2,\ \ldots,\ \alpha_N$ och $\mathbf{K}$ är matrisen med elementen $\mathbf{K}_{ij} = k\left(\mathbf{x}_i, \mathbf{x}_j\right)$. Matrisen $\mathbf{K}$ kallas Gram-matrisen och olikheten till sist i ekvation (4.4) brukar användas som definition på *positiv semidefinithet* för en matris $\mathbf{K}$. En funktion $k\left(\mathbf{x}_i, \mathbf{x}_j\right)$ sägs vara positivt semidefinit om Gram-matrisen $\mathbf{K}_{ij} = k\left(\mathbf{x}_i, \mathbf{x}_j\right)$ är positivt semidefinit för alla val av $\mathbf{x}_i$, $\mathbf{x}_j$.

4.2 Kärnor som positivt semidefinita funktioner

Ovan gavs en motivering för varför man borde utvidga definitionen av kärnor.

Definition 4.2.1 (Enligt [14])**.** En kärna är en symmetrisk positivt semidefinit funktion $k : X \times X \longmapsto \mathbb{R}$.

Med små modifikationer gäller då beviset för sats 4.1.1 och korollarium 4.1.2 även för följande sats och korollarium:

Sats 4.2.1 (Enligt [14])**.** För en kärna det vill säga en symmetrisk positivt semidefinit funktion $k : X \times X \longmapsto \mathbb{R}$ är rummet med funktioner definierade genom $f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i k(\mathbf{x}, \mathbf{x}_i)$ och inreprodukten $\langle f, g \rangle_k = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \beta_j k(\mathbf{x}_i, \mathbf{x}_j)$ ett inreproduktrum.

Korollarium 4.2.2 (Enligt [14])**.** Om dessutom alla Cauchyföljder konvergerar till en funktion i samma rum, med avseende på normen inducerad av inreprodukten $\langle f, g \rangle_k$, så är rummet ett Hilbertrum.

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 $j = l$, $\beta_j = 0$ 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

$$\langle \Phi_{\mathbf{x}_l}, f \rangle_k = \sum_{i=1}^{N} \alpha_i k(\mathbf{x}_l, \mathbf{x}_i) = f(\mathbf{x}_l). \tag{4.5}$$

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

$$\begin{aligned} \langle \Phi_{\mathbf{x}_l}, \Phi_{\mathbf{x}_h} \rangle_k &= \langle k(\cdot, \mathbf{x}_l), k(\cdot, \mathbf{x}_h) \rangle_k \\ &= \alpha_l \beta_h k(\mathbf{x}_l, \mathbf{x}_h) \\ &= k(\mathbf{x}_l, \mathbf{x}_h). \end{aligned} \tag{4.6}$$

Tolkningen av ekvation (4.5) 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.

För ekvation (4.6) 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 (möjligtvis olinjärt) rum. De här två egenskaperna är orsaken till benämningen *reproducerande* kärnor, den första ekvationen brukar också ibland användas som definitionen på en reproducerande kärna. En annan tolkning är att givet en symmetrisk positivt semidefinit funktion k borde man kunna skriva den som en inreprodukt i något rum, nedan följer ett exempel som samtidigt visar en av de största fördelarna med att använda kärnor istället för att räkna ut en olinjär utvidgning på förhand.

Exempel 4.2.1. Låt $\mathbf{x}_i$, $\mathbf{x}_j \in \mathbb{R}$ och betrakta funktionen

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

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 $\mathbf{x}_h$, $\mathbf{y}_h$ 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, detta eftersom $\langle \mathbf{x}, \mathbf{y} \rangle_{\ell^2}^2 \leq \langle \mathbf{x}, \mathbf{x} \rangle_{\ell^2} \langle \mathbf{y}, \mathbf{y} \rangle_{\ell^2} < \infty$ enligt sats 2.1.1.

Kärnan $k(\mathbf{x}, \mathbf{y}) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|_1^2}{2\sigma^2}}$ kan med andra ord skrivas som inreprodukten

$\langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle_{\ell^2}$ i rummet som ges av utvidgningen

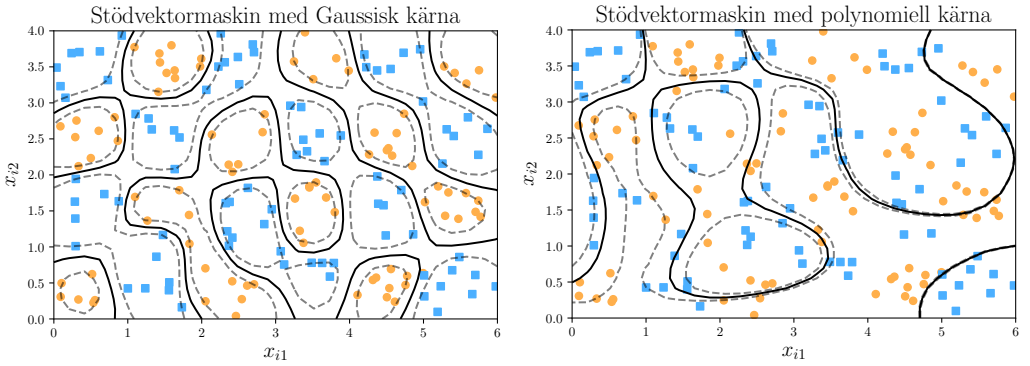
$$\phi(\mathbf{x}) = \left[\frac{e^{-\left(\frac{\mathbf{x}}{\sqrt{2}\sigma}\right)^2} \mathbf{x}^0}{\sqrt{0!}}, \frac{e^{-\left(\frac{\mathbf{x}}{\sqrt{2}\sigma}\right)^2} \mathbf{x}^1}{\sqrt{1!}}, \frac{e^{-\left(\frac{\mathbf{x}}{\sqrt{2}\sigma}\right)^2} \mathbf{x}^2}{\sqrt{2!}}, \dots \right]^\mathsf{T}.$$

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.

Kärnan

$$k(\mathbf{x}, \mathbf{y}) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|_p^2}{2\sigma^2}}$$

brukar kallas den Gaussiska kärnan och används ofta med goda resultat. Den Gaussiska kärnans nytta syns i figur 4.2 där den Gaussiska kärnan jämförs med den polynomiella kärnan med graden 7.



(a) Gaussisk kärna med $\gamma = 100$ och $\sigma^2 = 0{,}25$. Beräkningstid 2 sekunder.

(b) Polynomiell kärna $k(\langle \mathbf{x}_i, \mathbf{x}_j \rangle + 1)^7$ med $\gamma = 1$. Beräkningstid 3,5 minuter.

Figur 4.2: Jämförelse mellan ett olinjärt klassificeringsproblem löst med hjälp av två olika instansieringar av en stödvektormaskin, en med Gaussisk kärna och en annan med en polynomiell kärna. Lösningarna beräknades med funktionen `SVC` från paketet `sklearn` [12]. Parametrarna var `kernel='rbf'` och `kernel='poly'` med kärn-specifika parametrar nämnda i figurtexten. Den heldragna kurvan i svart är det olinjära hyperplanet och de streckade kurvorna är marginalens ränder. Observationerna är ritade som blåa fyrkanter eller orangea cirklar beroende på klasstillhörighet.

Kapitel 5

Avslutning

Det finns också andra sätt att härleda stödvektormaskinernas optimeringsproblem, till exempel genom Tikhonov-regularisation. Andra härledningar kan ge nya insikter, till exempel kan man se stödvektormaskiner som en medlem i en större grupp av modeller innefattande bland annat spline-modeller [8]. Med några små förändringar kan man också använda stödvektormaskiner för regression. Man kan också ändra på objektfunktionen för att uppmuntra lösningarna att ha vissa egenskaper, till exempel bara vara olika 0 i ett fåtal dimensioner (gles) [14].

Stödvektormaskinerna utvecklades under 1990-talet ungefär på samma gång som Vapnik-Chervonenkis teorin där man försöker karaktärisera när och varför metoder inom maskininlärning och statistik generaliserar väl till nya observationer [17]. Vapnik-Chervonenkis teorin kan ses som ett försök att möjliggöra till exempel analys av konfidensintervall för mera komplicerade algoritmer än linjär regression. För stödvektormaskiner med mjuka marginaler och kärnor forskas det ännu i vilka teoretiska gränser man kan ge för generalisering till ny data [15].

För att algoritmen ska fungera så väl som möjligt i praktiken måste man ofta bestämma olika parametrar som inte kan lösas genom konvex optimering. Till exempel måste parametern γ och kärnan bestämmas på något sätt. Oftast görs detta genom att man löser algoritmen för cirka 80 % av observationerna med olika val av kärna och γ för att sedan välja den kombination som fungerar bäst på de resterande 20 %:en av observationerna [8].

Reproducerande kärnor kan också användas för att göra andra algoritmer mera flexibla, det huvudsakliga kravet är att observationerna ska förekomma endast i inreprodukter [8, 14]. Till exempel finns det versioner av principalkomponentsanalys (principal component analysis, PCA) och linjär diskriminantanalys (linear discriminant analysis, LDA) som gjorts olinjära med hjälp av reproducerande kärnor [8].

Litteraturförteckning

[1] Mark. A. Aizerman, Emmanuil A. Braverman och Lev Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control,*, 25(6):821–837, 1964.

[2] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 63(3):337–404, 1950.

[3] Kristin P. Bennet och Olvi L. Mangasarian. Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, 1(1):23–34, 1992.

[4] Bernhard E. Boser, Isabelle M. Gyuon och Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. I: *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, ss 144–152. ACM Press, 1992.

[5] Stephen Boyd och Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2009.

[6] Corinna Cortes och Vladimir N. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[7] Anders Hald. On the History of Maximum Likelihood in Relation to Inverse Probability and Least Squares. *Statistical Science*, 14(2):214–222, 1999.

[8] Trevor Hastie, Robert Tibshirani och Jerome Friedman. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. Springer series in statistics. Springer New York Inc., 2001.

[9] David Hilbert. Grundzüge einer allgemeinen Theorie der linearen Integralgleichungen. *Nachrichten der Göttinger Akademie der Wissenschaften, Matematisch-Physikalische Klasse*, ss 49–91, 1904.

[10] Serge Lang. *Introduction to Linear Algebra*. Undergraduate Texts in Mathematics. Springer New York, 2 utgåvan, 1986.

[11] James Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society A*, 209:441–458, 1909.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot och E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[13] Frank Rosenblatt. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*, 65(6):386–408, 1958.

[14] Bernhard Schölkopf och Alexander J. Smola. *Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive Computation and Machine Learning series. The MIT Press, 2002.

[15] John Shawe-Taylor och Nello Christianini. Margin distribution and soft margin. I: Alexander J. Smola, Peter L. Bartlett, Bernhard Schölkopf och Dale Schuurmans, redaktörer, *Advances in Large Margin Classifiers*, ss 349–358. The MIT Press, 200.

[16] Fred W. Smith. Pattern Classifier Design by Linear Programming. *IEEE Transactions on Computers*, C-17(4):367–372, 1968.

[17] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, Berlin, Heidelberg, 1995.

[18] Vladimir N. Vapnik och Aleksandr Ya. Lerner. Pattern Recognition Using Generalized Portraits. *Avtomatika i Telemekhanika*, 24(6):774–780, 1963.

[19] Grace Wahba. *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in applied mathematics, 59. Society for Industrial and Applied Mathematics, 1990.

[20] Nicholas Young. *An introduction to Hilbert space*. Cambridge University Press, 1988.