

Stödvektormaskiner

Linjära hyperplan i Hilbertrum

Oscar Granlund

25 oktober 2018

Sammanfattning

Skrivs till sist

Kapitel 1

Inledning

Tack vare en ökande beräkningsförmåga samt användbarhet hos datorer under senare delar av 1900-talet forskades det flitigt om hur man bästa använder datorer för beräkning och lösning av statistiska metoder och problem. I samband med detta uppstod även forskningsområdet maskininläring där tyngdpunkten låg mera på datorerna. En tidig algoritm var Frank Rosenblatts *Perceptron* (år 1957) [13] där man med inspiration från hjärnans neuroner försökt klassificera *observationer* \mathbf{x} genom att dra ett *hyperplan* mellan klasserna. År 1963 gav Aleksandr Lerner och Vladimir Vapnik en variant av Rosenblatts perceptron där *optimala separerande hyperplan* används för att klassificera observationerna [16]. Lerner och Vapniks algoritm är matematiskt mera tilltalande än Rosenblatts eftersom att 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 [15] en generaliserad algoritm som använde *slackvariabler* för att även fungera för icke linjärt separabla observationspar. Smiths arbete med slackvariabler undersöktes vidare av Kristin Bennet och Olvi Mangasarian år 1992 [3].

Parallellt med forskningen i de linjära optimalt separerande hyperplanen forskades om tillämpningar av funktioner kallade *kärnor*, med avstamp i James Mercers forskning (1909) i *positiva* funktioner [11] och Nachman Aronszajns fortsatta forskning (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 kärnorna visades vara nyttiga för andra algoritmer, se till exempel Grace Wahbas bok om splin-modeller (1990) [17] tillämpades kärnor även på den ursprungliga algoritmen med optimalt separerande av Bernhard Boser, Isabelle Gyuon och Vladimir Vapnik i 1992 [4]. Snart därefter generaliserades även Bennet och Mangasarians algoritm av Corinna Cortes och

Vladimir Vapnik i 1995 [6], detta är den algoritm som vanligtvis associeras med begreppet *stödvektormaskin* (Support Vector Machine, SVM) och därmed är det den jag vill härleda.

I uppsatsen kommer till först den ursprungliga algoritmen att undersökas för att sedan modifieras med slackvariabler, upplägget följer långt Trevor Hastie, Robert Tibshirani och Jerome Friedmans bok [8], speciellt kapitlet om optimalt separerande hyperplan samt den mjuka utvidgningen med slackvariabler. Därefter kommer de reproducerande kärnorna att undersökas, för det följs i stora drag Bernhard Schölkopf och Alexander Smolas bok [14].

Precis som många andra metoder inom statistiken och maskininlärningen bygger stödvektormaskinen på ett konvext optimeringsproblem och på grund av detta borde läsaren vara bekant med koncept ur konvex optimering. En bra introduktion är Stephen Boyd och Lieven Vandenberghe's bok [5]. Främst kommer teorin om kvadratiska optimeringsproblem och analys av duala problem med hjälp av Lagrangemultiplikatorer att användas.

Kapitel 2

Lite Hilbertrumteori

2.1 Geometrisk 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 rummet man arbetar i ser ut visar det sig att teorin om *inreprodukttrum*, 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 [18]). Låt X vara ett vektorrum. En *inreprodukt* är en funktion $\langle \cdot, \cdot \rangle : X \times X \longrightarrow \mathbb{R}$ 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 inreprodukttrum X är *fullständigt* om varje Cauchy-följd \mathbf{x}_n konvergerar (med avseende på normen inducerad av inreprodukten) till en punkt \mathbf{x} i X .

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

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

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

Sats 2.1.1 (Cauchy-Schwarz olikhet enligt [18]). För \mathbf{x}, \mathbf{y} i ett inreprodukt-rum 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**:

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

det vill säga 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

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

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 triangel-olikheten):

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

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

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

$$\|\lambda \mathbf{x}\|_{\mathcal{H}}^2 = \langle \lambda \mathbf{x}, \lambda \mathbf{x} \rangle_{\mathcal{H}} = \lambda^2 \langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{H}}, \quad (2.1)$$

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 $\mathbf{x} = \mathbf{0}$ (positiv 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 $\langle \mathbf{x}, \mathbf{y} \rangle_d = \mathbf{x}^T \mathbf{y}$. 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. För den vanliga inreprodukten i \mathbb{R}^2 och \mathbb{R}^3 skulle det vara bra om intuitionen för till exempel när vinkeln mellan två vektorer är 90° det vill säga när två vektorer är ortogonala:

Definition 2.1.3 (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 $\mathbf{x} = [1, 0]^T$ och $\mathbf{y} = [0, 1]^T$. Då är $\langle \mathbf{x}, \mathbf{y} \rangle_2 = \mathbf{x}^T \mathbf{y} = [1, 0] [0, 1]^T = 1 \cdot 0 + 0 \cdot 1 = 0$ och vektorerna är vinkelräta. Normen av vektorerna är också 1 så de är dessutom ortogonala.

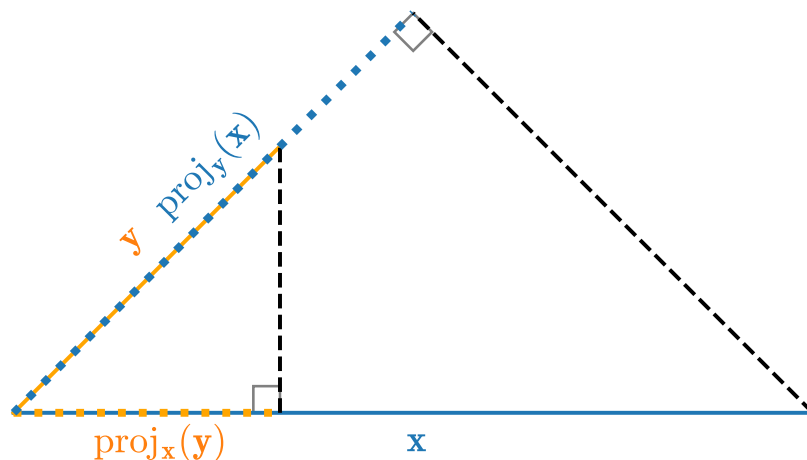
Med hjälp av ortogonaliteten kan man härleda konceptet för en komponenten av en vektor längs en annan vektor och projektionen av en vektor på en annan vektor. Figur 2.1 illustrerar konceptet.

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 oranget, samt projektionen av \mathbf{x} på \mathbf{y} streckat i blått och projektionen av \mathbf{y} på \mathbf{x} streckat i oranget. Streckat i svart finns de ortogonala vektorerna.

Talet λ är 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 talet 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 alltså definitionen:

Definition 2.1.4 (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 talet

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

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 $\text{comp}_{\mathbf{y}}(\mathbf{x}) = \|\text{proj}_{\mathbf{y}}(\mathbf{x})\|_{\mathcal{H}}$ ifall $\text{comp}_{\mathbf{y}}(\mathbf{x}) \geq 0$ och $\text{comp}_{\mathbf{y}}(\mathbf{x}) = -\|\text{proj}_{\mathbf{y}}(\mathbf{x})\|_{\mathcal{H}}$ ifall $\text{comp}_{\mathbf{y}}(\mathbf{x}) \leq 0$. Märk även att för $\mathbf{x} = \mathbf{0}$ så 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.

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 parallell till de ändligtdimensionella vektorrummens koordinatsystem och basvektorer. Det visar sig att inte alla Hilbertrum har en motsvarighet till basvektorer men ifall Hilbertrummet är *separabelt* så existerar det en ortonormal följd vektorer \mathbf{e}_i , $\langle \mathbf{e}_i, \mathbf{e}_j \rangle_{\mathcal{H}} = 0$ om $i \neq j$, $\|\mathbf{e}_i\|_{\mathcal{H}} = 1$, sådan att för varje vektor $\mathbf{x} \in \mathcal{H}$ gäller [18]:

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

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 , $i = 1, 2, 3, \dots$. Märk att här har man projicerat \mathbf{x} på varje basvektor \mathbf{e}_i och summerat de resulterande projektionerna. Antalet basvektorer \mathbf{e}_i bestämmer i princip dimensionen på Hilbertrummet.

Exempel 2.1.2. I \mathcal{L}^2 , $x \in [-\pi, \pi]$ är följande en ortonormal följd basvektorer [18]:

$$\mathbf{e}_1 = \frac{1}{\sqrt{2\pi}}, \quad \mathbf{e}_2 = \frac{1}{\sqrt{\pi}} \cos(x), \quad \mathbf{e}_3 = \frac{1}{\sqrt{\pi}} \sin(x), \quad \mathbf{e}_4 = \frac{1}{\sqrt{\pi}} \cos(2x), \dots$$

Definition 2.1.5 (Enligt [18]). 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 $\mathbf{e}_i \in \mathcal{H}$.

I fortsättningen behandlas bara separabla Hilbertrum, så att varje vektor \mathbf{x} kan skrivas som en linjär kombination av basvektorerna.

Kapitel 3

Stödvektormaskiner (SVM)

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) , $\mathbf{x}_i \in \mathbb{R}^p$, $y_i \in \{-1, 1\}$, $i = 1, \dots, N$, försöker hitta en regel $g : \mathbb{R}^p \mapsto \{-1, 1\}$ sådan att $g(\mathbf{x}_i) = y_i$ för så många träningspar (\mathbf{x}_i, y_i) som möjligt.

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

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

Exempel 3.1.1. I figur 3.1 illustreras två separerande hyperplan i \mathbb{R}^2 . I \mathbb{R}^2 blir hyperplanet en linje det vill säga ett delrum med dimensionen $2 - 1 = 1$. Allmänt gäller att i ett rum med dimensionen p blir hyperplanet ett delrum med dimensionen $p - 1$. Märk att ifall $\mathbf{x} = 0$ tillhör hyperplanet så är hyperplanet ett underrum.

För nästa sats behövs några konventioner för hyperplan $L = \{\mathbf{x} : f(\mathbf{x}) = \langle \mathbf{x}, \boldsymbol{\beta} \rangle_{\mathcal{H}} + \beta_0 = 0\}$ i relation till en punkt ¹ \mathbf{y} i \mathcal{H} :

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

- Man säger att \mathbf{y} ligger över hyperplanet L om $f(\mathbf{y}) > 0$ och under om $f(\mathbf{y}) < 0$.
- Med det signerade avståndet från punkten \mathbf{y} till hyperplanet L , $d^\pm(\mathbf{y}, L)$, menar man $\inf_{\mathbf{x} \in L} \|\mathbf{y} - \mathbf{x}\|_{\mathcal{H}}$ 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 $d^\pm(\mathbf{y}, L)$ 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 .

Sats 3.1.1 (Enligt [8]). Ett hyperplan i ett inreproduktum \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

$$\hat{\boldsymbol{\beta}} = \frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|_{\mathcal{H}}}.$$

2. $\langle \mathbf{x}_0, \boldsymbol{\beta} \rangle_{\mathcal{H}} = -\beta_0$ för alla \mathbf{x}_0 i L .
3. Det signerade avståndet från en punkt \mathbf{y} till hyperplanet L , $d^\pm(\mathbf{y}, L)$, ges av

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

där \mathbf{x}_0 är en godtycklig punkt i hyperplanet L . Om \mathcal{H} är lika med \mathbb{R}^p med den vanliga inreprodukten så fås dessutom

$$d^\pm(\mathbf{y}, L) = \frac{1}{\|f'(\mathbf{y})\|_p} f(\mathbf{y}).$$

Bevis. (Inte från [8].)

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 β ortogonal till vektorn $(\mathbf{x}_1 - \mathbf{x}_2)$ som beskriver riktningen från en punkt i L till en annan punkt i L . Dessutom gäller för $\hat{\beta} := \frac{\beta}{\|\beta\|_p}$ att $\|\hat{\beta}\|_{\mathcal{H}} = 1$ så $\hat{\beta}$ är ortonormal till alla vektorer i L . ■

2. Låt \mathbf{x}_0 vara en punkt i L . Då gäller att $f(\mathbf{x}_0) = \langle \mathbf{x}_0, \beta \rangle_{\mathcal{H}} + \beta_0 = 0$ alltså är $\langle \mathbf{x}_0, \beta \rangle_{\mathcal{H}} = -\beta_0$. ■
3. Låt \mathbf{x}_0 vara en punkt i hyperplanet L . Då är avståndet från punkten \mathbf{x}_0 till punkten \mathbf{y} minimerat om vektorn $(\mathbf{y} - \mathbf{x}_0)$ är ortogonal till hyperplanet, i \mathbb{R}^2 är detta principen att det kortaste avståndet från en linje till en punkt är avståndet mätt längs med en linje vinkelrät mot den ursprungliga linjen. Eftersom att $\hat{\beta}$ är ortonormal till varje punkt i L så blir det kortaste avståndet från \mathbf{y} till L längden av projektionen av vektorn från $\mathbf{x}_0 \in L$ till \mathbf{y} på $\hat{\beta}$ det vill säga $d^{\pm}(\mathbf{y}, L) = \text{comp}_{\hat{\beta}}(\mathbf{y} - \mathbf{x}_0)$. Vidare fås då

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

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 $\langle \mathbf{x}, \mathbf{y} \rangle_p = \mathbf{x}^T \mathbf{y}$ kan man noterar att $f(\mathbf{y}) = \langle \mathbf{y}, \beta \rangle_p + \beta_0$ och $f'(\mathbf{y}) = \beta$, då fås även att

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

■

Observation. Definitionen $L = \{\mathbf{x} : f(\mathbf{x}) = \langle \mathbf{x}, \beta \rangle_{\mathcal{H}} + \beta_0 = 0\}$ för hyperplanet L är inte entydig.

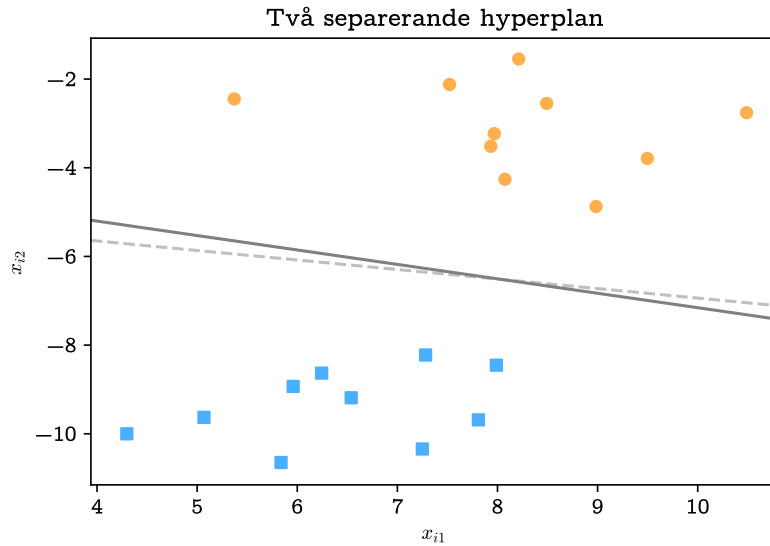
Orsak. Betrakta hyperplanen $L_1 = \{\mathbf{x} : f(\mathbf{x}) = \langle \mathbf{x}, \beta \rangle_{\mathcal{H}} + \beta_0 = 0\}$ och $L_2 = \{\mathbf{x} : g(\mathbf{x}) = \langle \mathbf{x}, -\beta \rangle_{\mathcal{H}} + (-\beta_0) = 0\}$. Eftersom att $g(\mathbf{x}) = -f(\mathbf{x})$ så gäller att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

Observation. För att få entydiga hyperplan för klassificering kan man lägga till villkor. Om man kräver att $\|\beta\|_{\mathcal{H}} = 1$ och $y_i (\langle \mathbf{x}_i, \beta \rangle_{\mathcal{H}} + \beta_0) \geq 0$ för alla $i = 1, \dots, N$, där y_i är klasserna i klassificeringsproblemet, så får man en entydig definition av hyperplanet där observationerna \mathbf{x}_i i klassen $y_i = 1$

ligger över hyperplanet medan observationerna \mathbf{x}_i i klassen $y_i = -1$ ligger under. Dessutom anger β_0 det signerade avståndet från origo till hyperplanet (i relation till riktningen på β).

Orsak. De extra villkoren gör att man inte längre kan göra manipulationerna som påvisade icke-entydigheten. Om man sätter $\mathbf{x} = \mathbf{0}$ så får man med hjälp av sats 3.1.1 att det signerade avståndet från origo till hyperplanet är lika med

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



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

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

Observation. Genom att byta tecken på β och β_0 kan man se att ifall det existerar ett hyperplan $\{\mathbf{x} : f(\mathbf{x}) = \langle \mathbf{x}, \beta \rangle_{\mathcal{H}} + \beta_0 = 0\}$ 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 $y_i = -1$. Som sagt tidigare så är konventionen 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

$$y_i (\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle_{\mathcal{H}} + \beta_0) > 0$$

för alla $i = 1, \dots, N$.

Bevis. Ifall ett klassificeringsproblem är linjärt separabelt så ligger alla observationer y_i på rätt sida av det separerande hyperplanet. Detta betyder att ifall $y_i = 1$ 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 $y_i (\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle_{\mathcal{H}} + \beta_0) > 0$. Ifall $\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle_{\mathcal{H}} + \beta_0 = 0$ är problemet inte linjärt separabelt. ■

Klassificeringsregeln g för separerande hyperplan blir

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

Exempel 3.1.2. Låt observationsparen vara $([2, 2]^T, 1)$, $([1, 2]^T, -1)$, inreprodukttrummet i fråga är då \mathbb{R}^2 och inreprodukten $\langle \mathbf{x}, \mathbf{y} \rangle_2 := \mathbf{x}^T \mathbf{y}$. Då är

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

och

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

två separerande hyperplan (linjer i detta fall).

Bevis. För L_1 :

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

och

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

Och för L_2 :

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

och

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

■

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

$$\begin{bmatrix} \mathbf{x}_{1,1} & \mathbf{x}_{1,2} & \cdots & \mathbf{x}_{1,p-1} & \mathbf{x}_{1,p} \\ \mathbf{x}_{2,1} & \mathbf{x}_{2,2} & & & \mathbf{x}_{2,p} \\ \vdots & & \ddots & & \vdots \\ \mathbf{x}_{p-1,1} & & & \mathbf{x}_{p-1,p-1} & \mathbf{x}_{p-1,p} \\ \mathbf{x}_{p,1} & \mathbf{x}_{p,2} & \cdots & \mathbf{x}_{p,p-1} & \mathbf{x}_{p,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \\ \beta_p \end{bmatrix} = \begin{bmatrix} -\beta_0 \\ -\beta_0 \\ \vdots \\ -\beta_0 \\ -\beta_0 \end{bmatrix},$$

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

Som syns i exempel 3.1.2 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

Inom statistiken finns många olika metoder för att modellera data och metoderna 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 ändras genom att man lägger till eller tar bort termer i objektfunktionen eller ändrar på kraven och på så sätt får 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 [16]. Matematiskt kan man uttrycka problemet som följande optimeringsproblem

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

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

Observation. Ifall alla punkter är rätt klassificerade ger sats 3.1.1 och 3.1.1 att $y_i \left(\langle \mathbf{x}_i, \hat{\beta} \rangle_p + \hat{\beta}_0 \right)$ ger det absoluta avståndet mellan hyperplanet och punkten \mathbf{x}_i .

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

För att visa att optimeringsproblemet (3.1) är *konvext* måste det skrivas om. Idén är här att man låter inversen av längden på vektorn β beskriva avståndet till närmast punkt. På så sätt skapas en direktare länk mellan kraven och objektfunktionen i optimeringsproblemet.

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

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

mot kraven

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

eller ekvivalent

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

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

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

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

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

Därefter görs ännu en konvexitetsbevarande kvadratisk transformering av objektfunktionen² $\|\beta\|_p$, det vill säga man noterar att om β^* är sådan att $\min_{\beta, \beta_0} \|\beta\|_p = \|\beta^*\|_p$ gäller så gäller även $\min_{\beta, \beta_0} \frac{1}{2} \|\beta\|_p^2 = \frac{1}{2} \|\beta^*\|_p^2$ för samma β^* . Detta brukar betecknas med funktionen $\operatorname{argmin}_{\beta, \beta_0} (f(\mathbf{y}))$ som ger som resultat det \mathbf{y}^* som minimerar funktionen $f(\mathbf{y})$.

En orsak till att göra den kvadratiske transformeringen är att man på så sätt 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 $D_x(g(x)) = D_x((x - x_0)^2) = D_x(x^2 - 2xx_0 + x_0^2) = 2x - 2x_0 = 2(x - x_0)$ 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

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

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

Ovanstående resonemang är ett bevis för sats 3.2.1.

Sats 3.2.1 (Enligt [8]). Låt $\hat{\beta}$, $\beta \in \mathbb{R}^p$ och $\hat{\beta}_0$, $\beta_0 \in \mathbb{R}$. Låt dessutom observationsparen (\mathbf{x}_i, y_i) vara linjärt separabla. Då är optimeringsproblemet

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

²Inom statistik- och maskinlärningslitteraturen kallas objektfunktionen ibland även för *kostfunktionen*.

konvext och man får en lösning $(\widehat{\boldsymbol{\beta}}^*, \widehat{\beta}_0^*) = \left(\frac{\boldsymbol{\beta}^*}{\|\boldsymbol{\beta}^*\|_p}, \frac{\beta_0^*}{\|\boldsymbol{\beta}^*\|_p} \right)$ genom att lösa optimeringsproblemet

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

3.2.1 Primala och duala problem

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

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

ges av

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

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

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

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

där $j = 1, \dots, p$ och

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

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

$$\beta = \sum_{i=1}^N \lambda_i y_i \mathbf{x}_i \quad (3.4)$$

$$0 = \sum_{i=1}^N \lambda_i y_i. \quad (3.5)$$

Insättning av kraven (3.4) och (3.5) i L_P ger följande duala problem

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

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

$$\lambda_i \geq 0, \quad i = 1, \dots, N. \quad (3.6)$$

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

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

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

- Krav (3.4) säger att vektorn β^* är en linjär kombination av vektorerna \mathbf{x}_i , $i = 1, \dots, N$.

- Ifall $\lambda_i^* > 0$ så ger krav (3.7) att $y_i (\langle \mathbf{x}_i, \boldsymbol{\beta}^* \rangle_p + \beta_0^*) = 1$ vilket enligt det ursprungliga optimeringsproblemet (3.1) ska tolkas som att punkten \mathbf{x}_i ligger på avståndet C från det separerande hyperplanet, det vill säga punkten \mathbf{x}_i är en av punkterna som ligger närmast det separerande hyperplanet.
- Ifall $y_i (\langle \mathbf{x}_i, \boldsymbol{\beta}^* \rangle_p + \beta_0^*) > 1$ så är $\lambda_i^* = 0$ och punkten \mathbf{x}_i är inte en av de punkter som ligger närmast det separerande hyperplanet.
- Parametern β_0^* kan bestämmas genom att man utnyttjar relationen $y_i (\langle \mathbf{x}_i, \boldsymbol{\beta}^* \rangle_p + \beta_0^*) = 1$ 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. Dessa punkter kallas *stödvektorer*.

Kvar finns också möjligheten att $\lambda_i^* = 0$ och $y_i (\langle \mathbf{x}_i, \boldsymbol{\beta}^* \rangle_p + \beta_0^*) = 1$ för något i . Detta kan tolkas som att punkten \mathbf{x}_i ligger på randen av marginalen men behövs inte för att beskriva $\boldsymbol{\beta}^*$, 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 $\{\mathbf{x} : f(\mathbf{x}) = \langle \mathbf{x}, \boldsymbol{\beta} \rangle_p + \beta_0 = 0\}$ med $y_i (\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle_p + \beta_0) > 0$ för alla träningspar (\mathbf{x}_i, y_i) , $i = 1, \dots, N$ existerar. Oseparabla observationspar leder till att optimeringsproblemet (3.1) samt optimeringsproblemen i sats 3.2.1 inte längre är lösbara eftersom att 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 *hyperplan med mjuka marginaler*.

För optimeringsproblemet (3.1) finns två omedelbara sätt att ändra på

kraven [8], endera låter man

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

eller så

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

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

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

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:

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

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

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

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

där slackvariablerna direkt beror på $\|\beta\|_p$ och beräkningar blir 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. Låt $\hat{\beta}$, $\beta \in \mathbb{R}^p$ och $\hat{\beta}_0$, $\beta_0 \in \mathbb{R}$. Låt dessutom konstanten K vara vald så att optimeringsproblemets krav är lösbara för givna observationspar (\mathbf{x}_i, y_i) . Då är optimeringsproblemet

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

konvext och man får en lösning $(\hat{\beta}^*, \hat{\beta}_0^*) = \left(\frac{\beta^*}{\|\beta^*\|_p}, \frac{\beta_0^*}{\|\beta^*\|_p} \right)$ genom att lösa optimeringsproblemet

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

Observation. Hyperplan med mjuka marginaler kan användas även ifall observationsparen (\mathbf{x}_i, y_i) är linjärt separabla; man kan till och med få det

optimala separerande hyperplanet som lösning genom att välja $K = 0$. Att använda hyperplan med mjuka marginaler kan vara en bra idé till exempel om man har outliers 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 vid optimering för att approximera kravet med en term i objektfunktionen [5]. I [6] används en liknande approximation där man istället för att följa uppdateringsstrategin för vägningen av strafffunktionen använder andra metoder för att bestämma vägningen. Optimeringsproblemet som oftast löses för hyperplan med mjuka marginaler blir då

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

eller

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

där γ är en på förhand bestämd variabel. Märk att båda problemen är kvadratiske 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 , $i = 1, \dots, N$. Märk att om marginalen ökar så ökar även straffet medan om marginalen minskar så minskar straffet. Avvägningen mellan minskning av straff eller ökning av marginal bestäms med hjälp av parametern γ som kan jämföras med parametern K i sats 3.3.1. Skillnaden är att om γ är litet så tillåts slackvariablerna vara större och ifall γ är stort så är det viktigare att slackvariablerna hålls små. Det separabla fallet fås när γ går mot ∞ .

En viktig skillnad mellan formuleringen i sats 3.3.1 och (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 alternativen (3.11) och (3.12) är (3.11) vanligare och behandlas således i resten av uppsatsen.

3.3.1 Primala och Duala Lagrangeproblem för mjuka marginaler

Precis som med separabelt data kan lösningen för optimeringsproblemet (3.11) karaktäriseras med hjälp av Lagrangemultiplikatorer. Den primala Lagrangefunktionen ges av

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

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 (3.3) får man:

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

och

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

Efter att man löser för nollställen får man följande krav:

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

$$0 = \sum_{i=1}^N \lambda_i y_i, \quad (3.15)$$

$$\lambda_i = \gamma - \mu_i \quad i = 1, \dots, N \quad (3.16)$$

samt kraven

$$\lambda_i \geq 0, \quad i = 1, \dots, N, \quad (3.17)$$

$$\mu_i \geq 0, \quad i = 1, \dots, N, \quad (3.18)$$

$$s_i \geq 0 \quad i = 1, \dots, N. \quad (3.19)$$

Insättning av kraven (3.14) till (3.15) och (3.16) i den primala Lagrangefunktionen (3.13) ger den duala Lagrangefunktionen

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

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

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

$$\mu_i s_i = 0, \quad i = 1, \dots, N, \quad (3.21)$$

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

från Karush-Kuhn-Tucker kraven för konvexa problem.

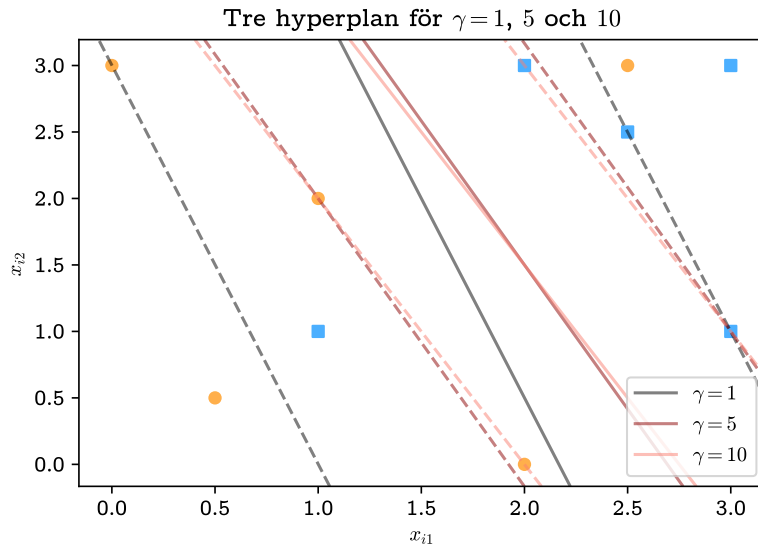
Observation. Precis som för algoritmen med optimala separerande hyperplan kan man karaktärisera lösningen för hyperplan med mjuka marginaler med hjälp av kraven (3.14) till (3.22).

- 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 $\lambda_i^* > 0$ och de kallas *stödvektorer* eftersom att de är de enda punkterna som behövs för att representera $\boldsymbol{\beta}^*$.

- För stödvektorer ($\lambda_i^* > 0$) som ligger på marginalen ($s_i^* = 0$) ger kraven (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^* .



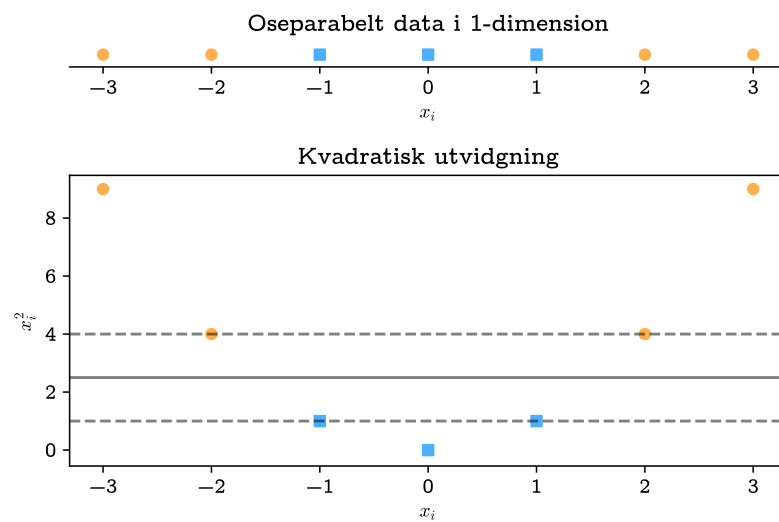
Figur 3.2: Löst exempel för linjärt oseparabelt data för 3 olika värden på γ . De streckade linjerna är marginalens rand.

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

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 även här att observationsparen är linjärt oseparabla men nu kan inte heller metoden med hyperplan med mjuka marginaler ge vettiga lösningar. Istället kan man lägga till en dimension och definiera att $\mathbf{x}_i \in \mathbb{R}^2$ och $\mathbf{x}_{i2} = \mathbf{x}_{i1}^2$. Då får man situationen som illustreras nederst i figur 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



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

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

Exempel 3.3.2 antyder att det kunde vara en bra idé att utvidga observationerna \mathbf{x}_i med olinjära faktorer, frågan är bara hur detta görs bäst. Klart är att man alltid kan bilda n :te gradens polynom men ifall den ursprungliga dimensionen p , mängden observationspar N eller graden n är stor så kan detta bli övermäktigt för även den snabbaste datorn. Det kommer att visa sig att eftersom observationerna \mathbf{x}_i endast förekommer i inreprodukter (se till exempel korrolarium ?? och ??) så finns det ett behagligare alternativ.

Exempel 4.0.1. Betrakta funktionen $\phi : (\mathbf{x}) \mapsto (\mathbf{x}^2, \sqrt{2}\mathbf{x}, 1)$ där $\mathbf{x} \in \mathbb{R}$, denna funktion motsvarar den olinjära utvidgningen i exempel 3.3.2. För att lösa exempel 3.3.2 borde följande inreprodukt mellan två observationer \mathbf{x}_1 och \mathbf{x}_2 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 .

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].

Föregående exempel motiverar alltså definitionen av en *kärna*:

Definition 4.0.1. Givet en funktion $\phi : \mathbb{R}^p \mapsto \mathbb{R}^P$ definieras *kärnan* k som funktionen $k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle_P$ där $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$ och $\langle \cdot, \cdot \rangle_P$ är den vanliga inreprodukten i \mathbb{R}^P . Vidare om man fixerar ett $\mathbf{y} \in \mathbb{R}^p$ så betecknar vi $\Phi_{\mathbf{y}}(\mathbf{x}) = k(\mathbf{x}, \mathbf{y})$ där $\mathbf{x} \in \mathbb{R}^p$.

Den sista delen i definition 4.0.1 ger med andra ord att varje val av ett fixt \mathbf{y} ger upphov till en reellvärd funktion $\Phi_{\mathbf{y}}$ av \mathbf{x} .

Givet en mängd observationer \mathbf{x}_i , $i = 1, \dots, N$ samt kärnan i exempel 4.0.1, det vill säga 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 man definiera ett vektorrum av funktioner genom

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

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

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

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

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

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

och

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

så 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 definitionen 2.1.1 är uppfyllda:

$$1. \quad \langle f, g \rangle_k = \langle g, f \rangle_k:$$

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

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

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

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

3. $\langle f + h, g \rangle_k = \langle f, g \rangle_k + \langle h, g \rangle_k$:

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

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

4. $\langle f, f \rangle_k \geq 0$:

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

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 $\langle f, f \rangle \geq 0$ där likhet gäller om och endast om $f = 0$. \blacksquare

Eftersom att alla villkor i 2.1.1 är uppfyllda så är $\langle f, g \rangle_k = \sum_{i=1}^N \sum_{j=1}^N \alpha_i \beta_j k(\mathbf{x}_i, \mathbf{x}_j)$ där $k(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle_1 + 1)^2 = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle_3$ och $\phi(\mathbf{x}) = [\mathbf{x}^2, \sqrt{2}\mathbf{x}, 1]^T$ en inreprodukt. Om man dessutom kräver att alla Cauchyföljder konvergerar till en funktion i rummet så har man ett Hilbertrum vilket betyder att man enkelt kan definiera till exempel projektioner.

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

Sats 4.0.1. För en kärna k definierad genom $k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle_P$ 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 inreprodukttrum.

Dessutom fås även följande korollarium:

Korollarium 4.0.2. 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 $X \times X \mapsto \mathbb{R}$ 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 så garanterade inreprodukten att egenskap 4 uppfylldes men man kan istället genast kräva att

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

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

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

där $\boldsymbol{\alpha}$ är en vektor med elementen $\alpha_1, \alpha_2, \dots, \alpha_N$ och \mathbf{K} är matrisen med elementen $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. Matrisen \mathbf{K} kallas Gram-matrisen och kravet i (4.4) brukar användas som definition på *positiv semidefinitet*. En funktion $k(\mathbf{x}_i, \mathbf{x}_j)$ sägs vara positiv semidefinit om Gram matrisen $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ är positiv semidefinit för alla val av $\mathbf{x}_i, \mathbf{x}_j$. På så sätt kan man motivera följande definition av kärnor:

Definition 4.0.2. En kärna är en symmetrisk positiv semidefinit funktion $k : X \times X \mapsto \mathbb{R}$.

Med modifikationer gäller då beviset för sats 4.0.1 och korollarium 4.0.2 även för följande sats och korollarium:

Sats 4.0.3. För en kärna det vill säga en symmetrisk positiv semidefinit funktion $k : X \times X \mapsto \mathbb{R}$ är rummet med funktioner definierade genom $f(x) = \sum_{i=1}^N \alpha_i k(\mathbf{x}, \mathbf{x}_i)$ 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 inreprodukttrum.

Korollarium 4.0.4. 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.

Kapitel 5

Avslutning

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

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

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. Guyon 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, Mathematisch-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] Fred W. Smith. Pattern Classifier Design by Linear Programming. *IEEE Transactions on Computers*, C-17(4):367–372, 1968.
- [16] Vladimir N. Vapnik och Aleksandr Ya. Lerner. Pattern Recognition Using Generalized Portraits. *Avtomatika i Telemekhanika*, 24(6):774–780, 1963.
- [17] Grace Wahba. *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in applied mathematics, 59. Society for Industrial and Applied Mathematics, 1990.
- [18] Nicholas Young. *An introduction to Hilbert space*. Cambridge University Press, 1988.