TEK5020/9020 Mønstergjenkjenning
Høsten 2023

Forelesning 10 – Lineære og generaliserte diskriminantfunksjoner (2)

Idar Dyrdal (idar.dyrdal@its.uio.no)

UiO : Institutt for teknologisystemer

14. oktober 2023

Innhold i kurset

- Introduksjon til mønstergjenkjenning
- Beslutningsteori (desisjonsteori)
- Parametriske metoder
- Ikke-parametriske metoder
- Lineære og generaliserte diskriminantfunksjoner (forts.)
- Evaluering av klassifikatorer
- Ikke-ledet læring
- Klyngeanalyse.

Relaksasjonsmetoden

Vi har så langt sett på perceptronkriteriet

$$J_p(\boldsymbol{a}) = -\sum_{\boldsymbol{y} \in \mathscr{Y}} \boldsymbol{a}^t \boldsymbol{y},$$

som er en lineær funksjon av $\boldsymbol{a}$ og følgelig har en diskontinuerlig gradient, men J_p er ikke den eneste mulige kriteriefunksjonen.

Et alternativ er å kvadrere leddene i summen over $\mathscr{Y}$ (de feilklassifiserte samplene) slik at kriteriefunksjonen blir

$$J_q(\boldsymbol{a}) = \sum_{\boldsymbol{y} \in \mathscr{Y}} (\boldsymbol{a}^t \boldsymbol{y})^2.$$

Dette gir kontinuerlig gradient og en glattere flate å søke over. J_q er imidlertid så glatt nær randen randen av løsningsregionen at gradientsøk vil kunne gi langsom konvergens mot et randpunkt, og derved en dårlig vektvektor med tanke på klassifisering av nye sampler. J_q er også dominert av de lengste egenskapsvektorene i datasettet.

Relaksasjonskriteriet

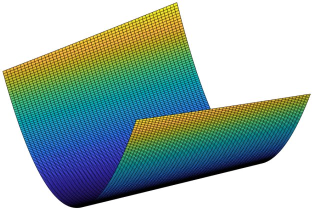
Relaksasjonskriteriet gitt ved

$$J_r(\mathbf{a}) = \frac{1}{2} \sum_{\mathbf{y} \in \mathscr{Y}} \frac{(\mathbf{a}^t \mathbf{y} - b)^2}{||\mathbf{y}||^2} \quad \text{der} \quad \mathscr{Y} = \{\mathbf{y} : \mathbf{a}^t \mathbf{y} \leq b\},$$

motvirker disse problemene ved normalisering (divisjon med normen til $\mathbf{y}$) og introduksjon av marginen $b > 0$, som forhindrer konvergens mot randen av løsningsregionen.

Egenskaper for $J_r(\mathbf{a})$:

- $J_r(\mathbf{a}) \geq 0$ dvs. det skal søkes etter et minimum,
- $J_r(\mathbf{a}) \overset{def}{=} 0$ hvis $\mathscr{Y} = \emptyset$,
- $J_r(\mathbf{a}) = 0$ hvis og bare hvis $\mathbf{a}^t \mathbf{y} > b \forall \mathbf{y}$.



Relaksasjonsalgoritmene

Gradienten til J_r med hensyn på $\boldsymbol{a}$ blir

$$\nabla_{\boldsymbol{a}} J_r(\boldsymbol{a}) = \sum_{\boldsymbol{y} \in \mathscr{Y}} \frac{\boldsymbol{a}^t \boldsymbol{y} - b}{||\boldsymbol{y}||^2} \boldsymbol{y}.$$

Gradientsøkalgoritmen (*Relaksasjonsalgoritmen*) blir da

$$\left.\begin{aligned} \boldsymbol{a}_1 &= \text{ vilkårlig startvektor} \\ \boldsymbol{a}_{k+1} &= \boldsymbol{a}_k + \rho_k \sum_{\boldsymbol{y} \in \mathscr{Y}_k} \frac{b - \boldsymbol{a}_k^t \boldsymbol{y}}{||\boldsymbol{y}||^2} \boldsymbol{y}, \quad \mathscr{Y}_k = \{\boldsymbol{y} : \boldsymbol{a}_k^t \boldsymbol{y} \leq b\} \end{aligned}\right\} \quad \textit{Relaksasjonsalgoritmen},$$

og den tilsvarende enkeltsamplealgoritmen (*Relaksasjonsregelen*) blir

$$\left.\begin{aligned} \boldsymbol{a}_1 &= \text{ vilkårlig startvektor} \\ \boldsymbol{a}_{k+1} &= \boldsymbol{a}_k + \rho \frac{b - \boldsymbol{a}_k^t \boldsymbol{y}^k}{||\boldsymbol{y}^k||^2} \boldsymbol{y}^k, \quad \text{der } \rho_k = \rho = \text{konst. og } \boldsymbol{a}_k^t \boldsymbol{y}^k \leq b \end{aligned}\right\} \textit{Relaksasjonsregelen.}$$

Oppdateringen i relaksasjonsregelen

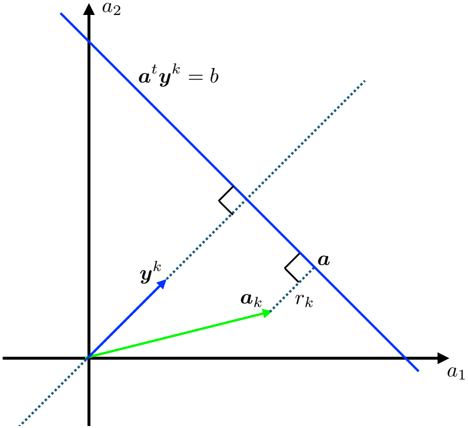
Samplet $\mathbf{y}^k$ i figuren er feilklassifisert av $\mathbf{a}_k$ fordi $\mathbf{a}_k^t\mathbf{y}^k < b$.

Skal finne punktet $\mathbf{a}$ på hyperplanet $\mathbf{a}^t\mathbf{y}^k = b$ som er nærmest $\mathbf{a}_k$.

Dette punktet er gitt ved

$$\mathbf{a} = \mathbf{a}_k + r_k \frac{\mathbf{y}^k}{||\mathbf{y}^k||}$$

der r_k er avstanden fra $\mathbf{a}_k$ til hyperplanet.



Oppdateringen i relaksasjonsregelen (forts.)

Dette gir

$$\boldsymbol{a}^t\boldsymbol{y}^k = \boldsymbol{a}_k^t\boldsymbol{y}^k + r_k\frac{(\boldsymbol{y}^k)^t}{\|\boldsymbol{y}^k\|}\boldsymbol{y}^k$$

$$\Downarrow$$

$$b = \boldsymbol{a}_k^t\boldsymbol{y}^k + r_k\|\boldsymbol{y}^k\| \quad (\text{siden } \boldsymbol{a}^t\boldsymbol{y}^k = b)$$

$$\Downarrow$$

$$r_k = \frac{b - \boldsymbol{a}_k^t\boldsymbol{y}^k}{\|\boldsymbol{y}^k\|},$$

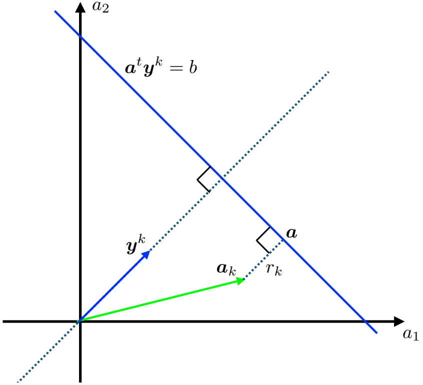
slik at oppdateringen i relaksasjonsregelen derved kan skrives som

$$\boldsymbol{a}_{k+1} = \boldsymbol{a}_k + \rho\, r_k\frac{\boldsymbol{y}^k}{\|\boldsymbol{y}^k\|}.$$

Oppdateringen i relaksasjonsregelen (forts.)

Vektvektoren $\boldsymbol{a}_k$ oppdateres til $\boldsymbol{a}_{k+1}$ ved å gi et tillegg $\rho \cdot r_k$ i retning mot hyperplanet.

- $\rho = 1$: vektvektoren flyttes direkte til hyperplanet slik at *spenningen* forbundet med ulikheten $\boldsymbol{a}_k^t \boldsymbol{y}^k < b$ fjernes (relaksasjon),
- $\rho > 1$: vektvektoren flyttes til andre siden av hyperplanet (overrelaksasjon),
- $\rho < 1$: vektvektoren flyttes nærmere, men ikke helt frem til hyperplanet (underrelaksasjon).



Det kan vises at relaksasjonsregelen (og derved også relaksasjonsalgoritmen) konvergerer til løsningsvektor for lineært separable sett dersom $0 < \rho < 2$.

Feilrettingsmetoder

Vi har sett på trening av lineære diskriminantfunksjoner ved hjelp av:

- Perceptron-metodene
  - Perceptron-algoritmen
  - Variabelt inkrement regelen
  - Fast-inkrement regelen

- Relaksasjonsmetodene
  - Relaksasjonsalgoritmen
  - Relaksasjonsregelen

Her ønskes å tilfredsstille et sett av ulikheter der $\boldsymbol{a}^t \boldsymbol{y} > 0$ for alle sampler i treningssettet ved å rette feilklassifiseringer som påtreffes; derav betegnelsen *feilrettingsmetoder*.

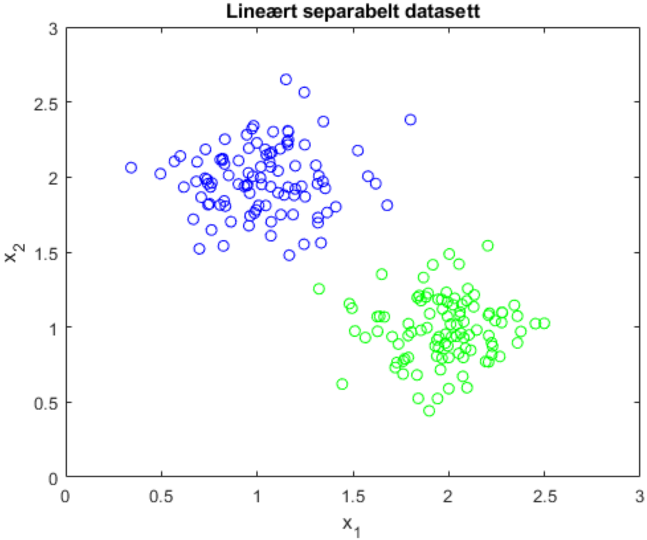
Ikke-separable problemer

Feilrettingsmetodene konvergerer under gitte betingelser til løsningsvektorer for lineært separable problemer, men kan også gi gode resultater på ikke-separable problemer.

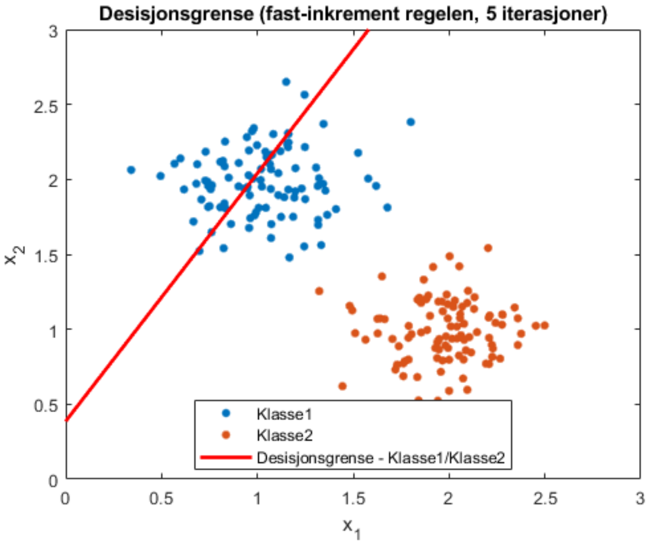
Muligheter å prøve ut:

- Stopp etter et maksimalt antall iterasjoner,
- Stopp etter et gitt antall iterasjoner uten noen forbedring av resultatet,
- Bruke middelet av de siste vektvektorene før algoritmen stopper som endelig vektvektor (med håp om mer robust løsning),
- *Pocket-algoritmen* (ta vare på beste vektvektor så langt i iterasjonsprosessen),
- Forskjellige valg av inkrement ρ_k og startvektor $\boldsymbol{a}_1$ (kjøre algoritmen flere ganger med forskjellig utgangspunkt i håp om å finne et globalt minimum av kriteriefunksjonen).

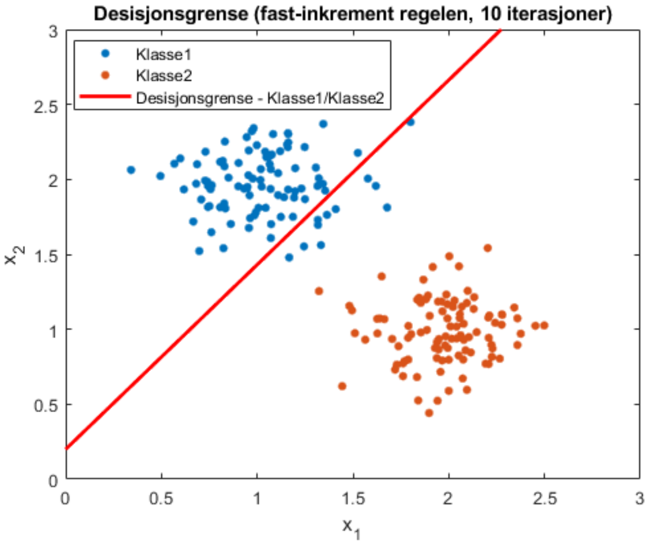
Eksempel – lineært separabelt datasett med to klasser



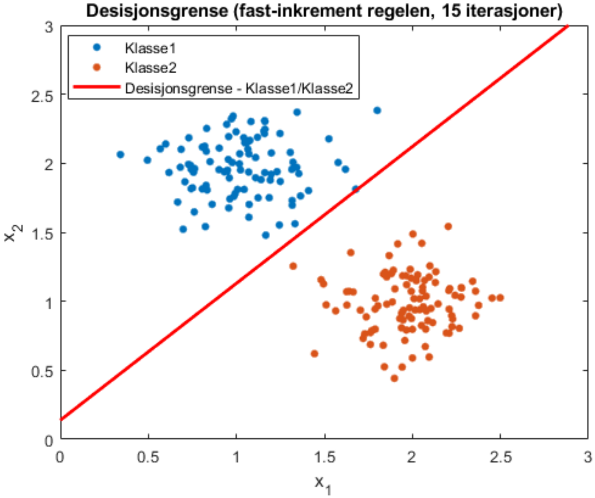
Eksempel – lineært separabelt datasett med to klasser (forts.)



Eksempel – lineært separabelt datasett med to klasser (forts.)

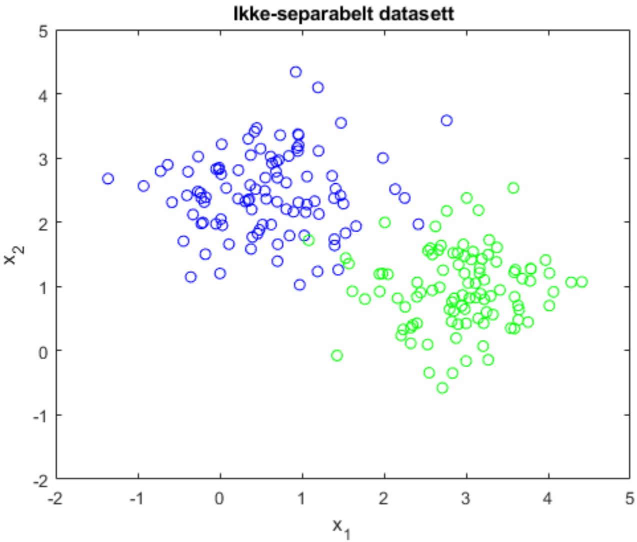


Eksempel – lineært separabelt datasett med to klasser (forts.)

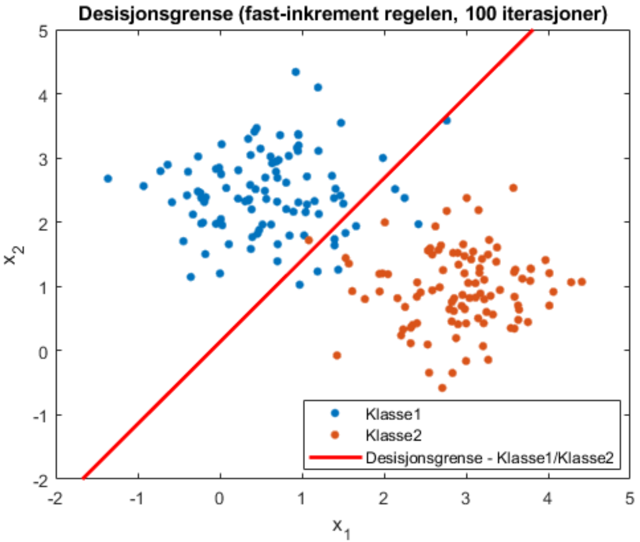


Konvergens til løsningsvektor etter 15 iterasjoner.

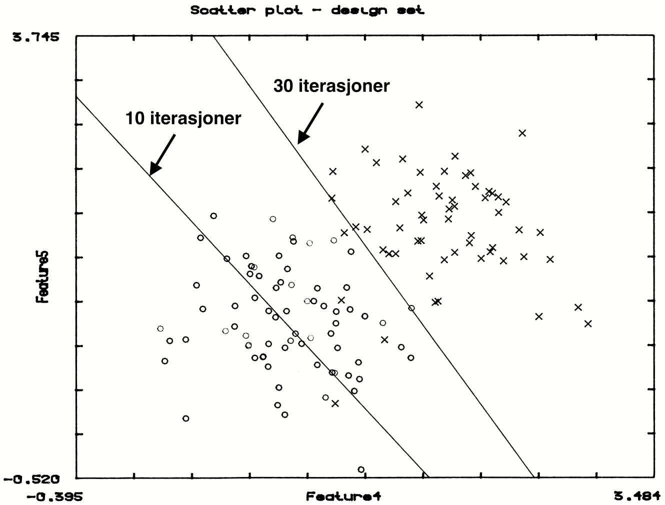
Eksempel – ikke-separabelt datasett med to klasser (forts.)



Eksempel – ikke-separabelt datasett med to klasser (forts.)

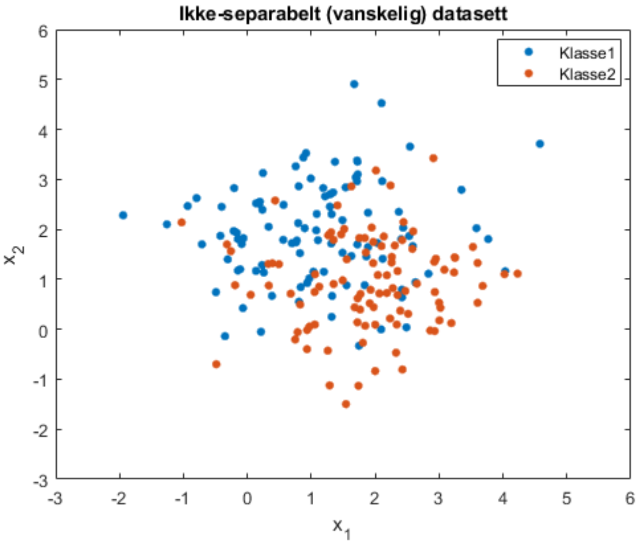


Eksempel – Perceptron-algoritmen på ikke-separabelt datasett

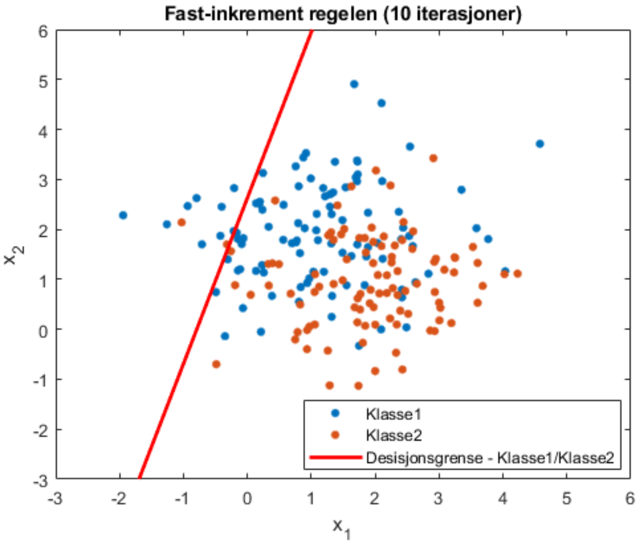


Desisjonsgrenser etter 10 og 30 iterasjoner (sammensatt oppdatering).

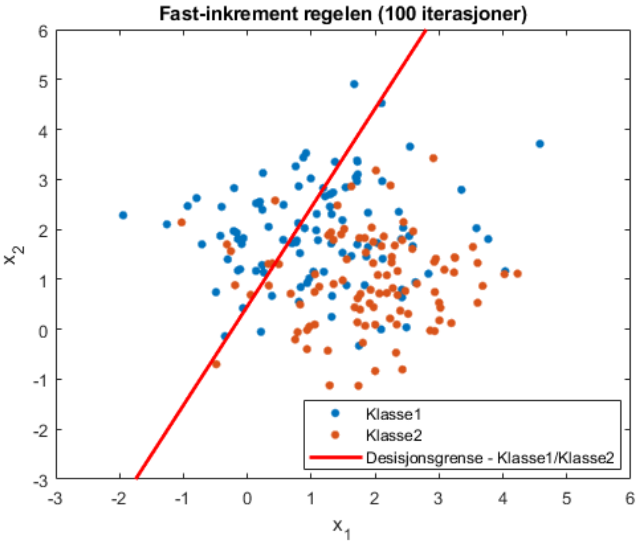
Eksempel – stor overlapp mellom klassene



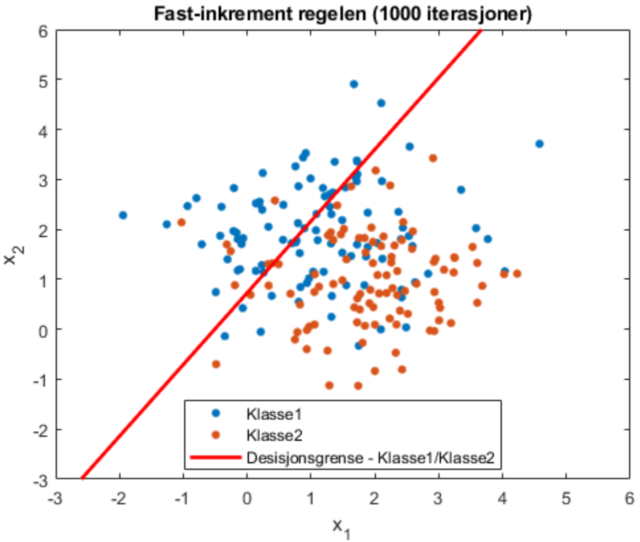
Eksempel – stor overlapp mellom klassene (forts.)



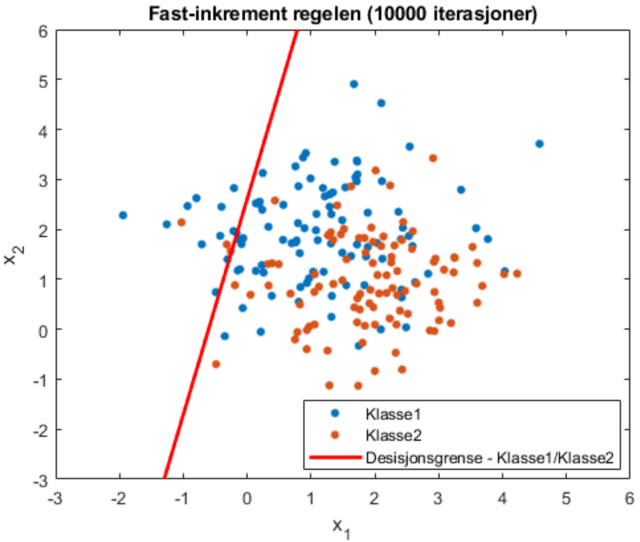
Eksempel – stor overlapp mellom klassene (forts.)



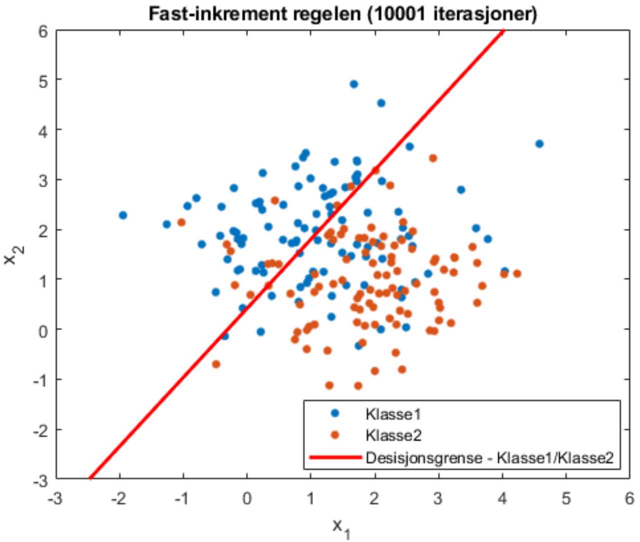
Eksempel – stor overlapp mellom klassene (forts.)



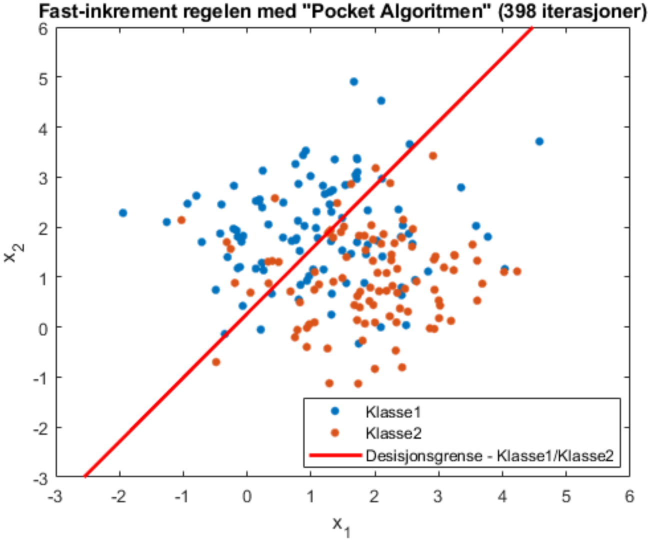
Eksempel – stor overlapp mellom klassene (forts.)



Eksempel – stor overlapp mellom klassene (forts.)



Eksempel – stor overlapp mellom klassene (forts.)



Minste kvadraters metode

Ønsker vektvektor $\boldsymbol{a}$ som tilfredsstiller likningssystemet

$$\boldsymbol{a}^t \boldsymbol{y}_i = b_i \quad \text{der } b_i > 0, \quad i =, 1, \ldots, n \quad \text{(positive marginer)}$$

slik at samplene $\boldsymbol{y}_i$ er riktig klassifisert av $\boldsymbol{a}$. Definerer en datamatrise Y og marginvektor $\boldsymbol{b}$:

$$Y = \begin{bmatrix} \boldsymbol{y}_1^t \\ \vdots \\ \boldsymbol{y}_n^t \end{bmatrix} \quad (n \times \hat{d}) \quad \text{og} \quad \boldsymbol{b} = [b_1, \ldots, b_n]^t \quad \text{(marginvektoren)},$$

slik at likningssystemet $Y\boldsymbol{a} = \boldsymbol{b}$ skal løses med hensyn på $\boldsymbol{a}$. Dersom Y er kvadratisk (dvs. $n \times n$) og $|Y| \neq 0$ gir dette løsningen:

$$\boldsymbol{a} = Y^{-1}\boldsymbol{b}.$$

Vanligvis er imidlertid $n >> \hat{d}$, slik at likningssystemet er overbestemt og ingen eksakt løsning eksisterer.

Minste kvadraters metode (forts.)

Her søkes i stedet en minste kvadraters løsning der lengden av feilvektoren

$$\boldsymbol{e} = Y\boldsymbol{a} - \boldsymbol{b} \quad \text{er så liten som mulig.}$$

Søker derfor minste kvadraters løsning der kriteriefunksjonen

$$J_s(\boldsymbol{a}) = \|\boldsymbol{e}\|^2 = \|Y\boldsymbol{a} - \boldsymbol{b}\|^2 = \sum_{i=1}^{n} (\boldsymbol{a}^t \boldsymbol{y}_i - b_i)^2 \quad \text{skal minimaliseres.}$$

Løsningsmetoder:

- Direkte løsning (*Pseudoinvers løsningsmetode*),
- Gradientsøk (f.eks. Widrow-Hoff algoritmen).

Pseudoinvers løsningsmetode

En nødvendig betingelse for minimum av kriteriefunksjonen $J_s(\boldsymbol{a})$ er at gradienten er null:

$$\nabla J_s(\boldsymbol{a}) = 2\sum_{i=1}^{n}(\boldsymbol{a}^t\boldsymbol{y}_i - b_i)\boldsymbol{y}_i = 2Y^t(Y\boldsymbol{a} - \boldsymbol{b}) = 0,$$

slik at

$$Y^tY\boldsymbol{a} = Y^t\boldsymbol{b} \text{ der } Y^tY \text{ er kvadratisk } (\hat{d} \times \hat{d}).$$

Antar nå $|Y^tY| \neq 0$ (som oftest tilfelle). Dette gir løsningen

$$\boldsymbol{a} = (Y^tY)^{-1}Y^t\boldsymbol{b} = \underline{\underline{Y^\dagger\boldsymbol{b}}},$$

der

$$Y^\dagger = (Y^tY)^{-1}Y^t$$

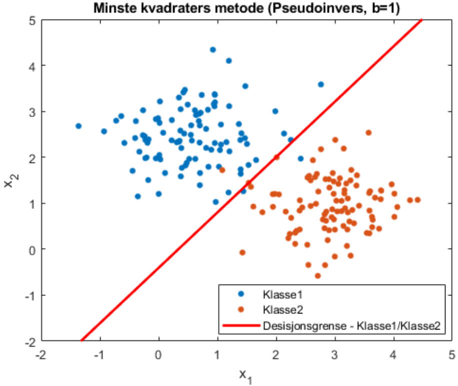
er den *pseudoinverse* til Y. Legg merke til at en minste kvadraters løsning vil imidlertid alltid eksistere, selv om Y^tY er singulær.

Pseudoinvers løsningsmetode (forts.)

Løsningen for $\boldsymbol{a}$ avhenger av hvilket valg som gjøres for $\boldsymbol{b}$, og vil ikke nødvendigvis være en separerende vektor, selv om datasettet er lineært separabelt.

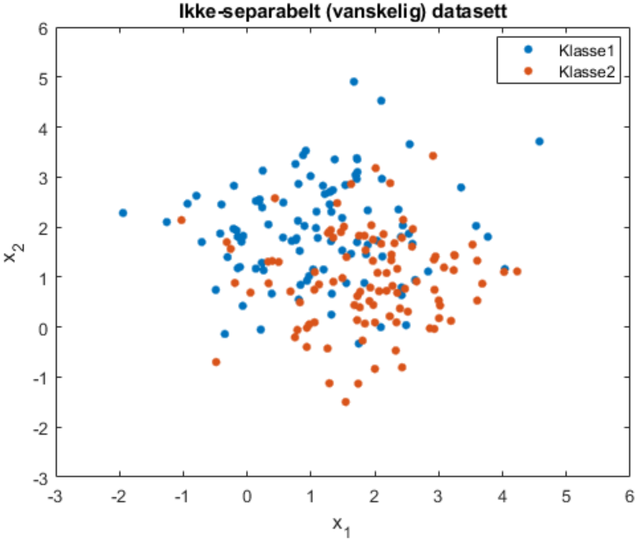
Håpet er å finne en god løsning, enten settet er separabelt eller ikke-separabelt.

Et vanlig valg for marginvektoren er $\boldsymbol{b} = [1, \ldots, 1]^t$, der poenget er at alle b'ene er like. En annen verdi enn én vil bare føre til en skalering av $\boldsymbol{a}$.

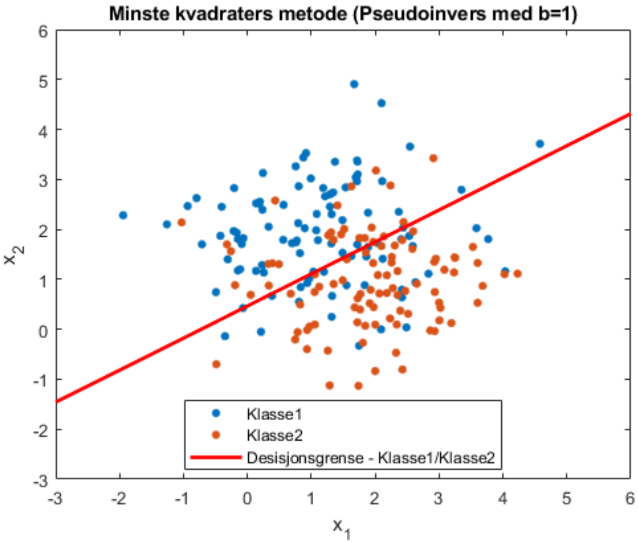


Pseudoinvers løsning på samme ikke-separable datasett som tidligere i forelesningen.

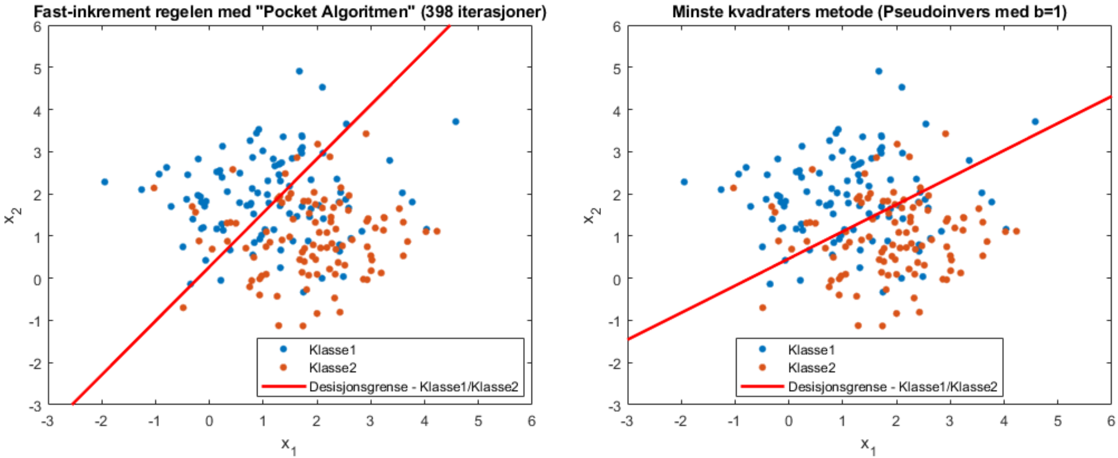
Eksempel – stor overlapp mellom klassene



Eksempel – stor overlapp mellom klassene (forts.)



Eksempel – stor overlapp mellom klassene (forts.)



Sammenlikning av resultatene for fast-inkrement regelen og pseudoinvers-metoden.

Alternativt valg av marginvektor

Skal her velge forskjellige b-verdier for klassene ut fra antall sampler i hver klasse i treningssettet. Starter med å dele treningssettet (bestående av de opprinnelige x-vektorene) i to delmengder ut fra klassetilhørighet, dvs.

$$\mathscr{X} = \{\underbrace{\mathbf{x}_1, \mathbf{x}_2, \ldots}_{n_1}\underbrace{\ldots, \mathbf{x}_n}_{n_2}\} = \mathscr{X}_1 + \mathscr{X}_2$$

dvs. et treningssett med n_1 sampler fra ω_1 og n_2 sampler fra ω_2. Datamatrisen Y kan da uttrykkes ved hjelp av de opprinnelige egenskapsvektorene på følgende måte:

$$Y = \begin{bmatrix} \mathbf{y}_1^t \\ \vdots \\ \mathbf{y}_n^t \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{x}_1^t \\ \vdots & \vdots \\ 1 & \mathbf{x}_{n_1}^t \\ -1 & -\mathbf{x}_{n_1+1}^t \\ \vdots & \vdots \\ -1 & -\mathbf{x}_n^t \end{bmatrix} = \begin{bmatrix} \mathbf{u}_1 & X_1 \\ -\mathbf{u}_2 & -X_2 \end{bmatrix} \quad \text{der} \quad \mathbf{u}_i = \left.\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}\right\} n_i, \; i = 1, 2.$$

Alternativt valg av marginvektor (forts.)

Vektvektoren kan tilsvarende uttrykkes vha. den opprinnelige vektvektoren og skalarvekten:

$$\boldsymbol{a} = \begin{bmatrix} w_0 \\ \boldsymbol{w} \end{bmatrix}.$$

Det valget for marginvektoren $\boldsymbol{b}$ som skal brukes her er

$$\boldsymbol{b} = \begin{bmatrix} \frac{n}{n_1}\boldsymbol{u}_1 \\ \frac{n}{n_2}\boldsymbol{u}_2 \end{bmatrix}$$

slik at klassen med færrest representanter i treningssettet vil vektlegges sterkere ved å få en større verdi på sine b'er. Dette vil normalt gi en bedre løsning dersom treningssettet er ubalansert mht. antall representanter fra de to klassene.

Alternativt valg av marginvektor (forts.)

Likningssystemet

$$Y^t Y \boldsymbol{a} = Y^t \boldsymbol{b}$$

kan da skrives som

$$\begin{bmatrix} \boldsymbol{u}_1^t & -\boldsymbol{u}_2^t \\ X_1^t & -X_2^t \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_1 & X_1 \\ -\boldsymbol{u}_2 & -X_2 \end{bmatrix} \begin{bmatrix} w_0 \\ \boldsymbol{w} \end{bmatrix} = \begin{bmatrix} \boldsymbol{u}_1^t & -\boldsymbol{u}_2^t \\ X_1^t & -X_2^t \end{bmatrix} \begin{bmatrix} \frac{n}{n_1}\boldsymbol{u}_1 \\ \frac{n}{n_2}\boldsymbol{u}_2 \end{bmatrix}$$

$$\Downarrow$$

$$\begin{bmatrix} n & (n_1\boldsymbol{m}_1 + n_2\boldsymbol{m}_2)^t \\ (n_1\boldsymbol{m}_1 + n_2\boldsymbol{m}_2) & S_W + n_1\boldsymbol{m}_1\boldsymbol{m}_1^t + n_2\boldsymbol{m}_2\boldsymbol{m}_2^t \end{bmatrix} \begin{bmatrix} w_0 \\ \boldsymbol{w} \end{bmatrix} = \begin{bmatrix} 0 \\ n(\boldsymbol{m}_1 + \boldsymbol{m}_2) \end{bmatrix}$$

der

$$\boldsymbol{m}_i = \frac{1}{n_i}\sum_{\boldsymbol{x}\in\mathscr{X}_i} \boldsymbol{x} \quad \text{og} \quad S_W = \sum_{i=1}^{2}\sum_{\boldsymbol{x}\in\mathscr{X}_i} (\boldsymbol{x}-\boldsymbol{m}_i)(\boldsymbol{x}-\boldsymbol{m}_i)^t.$$

Alternativt valg av marginvektor (forts.)

Løsningen av likningssystemet blir (se DHS for detaljer i utledningen):

$$\boldsymbol{a} = \begin{bmatrix} w_0 \\ \boldsymbol{w} \end{bmatrix} = \begin{bmatrix} -\boldsymbol{m}^t \boldsymbol{w} \\ \alpha n S_W^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2) \end{bmatrix}, \text{ der } \boldsymbol{w} \text{ er } \textit{Fishers lineære diskriminant.}$$

Her er

$$\boldsymbol{m}_i = \frac{1}{n_i} \sum_{\boldsymbol{x} \in \mathscr{X}_i} \boldsymbol{x} \text{ (klassemiddel)}, \quad \boldsymbol{m} = (n_1 \boldsymbol{m}_1 + n_2 \boldsymbol{m}_2)/n \text{ (middel over begge klasser), og}$$

$$S_W = \sum_{i=1}^{2} \sum_{\boldsymbol{x} \in \mathscr{X}_i} (\boldsymbol{x} - \boldsymbol{m}_i)(\boldsymbol{x} - \boldsymbol{m}_i)^t \quad \text{(spredning innen klasser).}$$

Diskriminantfunksjonen blir da

$$g(\boldsymbol{x}) = \boldsymbol{a}^t \boldsymbol{y} = [w_0, \boldsymbol{w}^t] \begin{bmatrix} 1 \\ \boldsymbol{x} \end{bmatrix} = \boldsymbol{w}^t(\boldsymbol{x} - \boldsymbol{m}),$$

som gir desisjonsregelen (*Fishers klassifikator*):

$$\boxed{\text{Velg } \omega_1 \text{ hvis } \boldsymbol{w}^t(\boldsymbol{x} - \boldsymbol{m}) > 0, \text{ ellers } \omega_2.}$$

Alternativt valg av marginvektor (forts.)

Fishers lineære diskriminant angir en retning i egenskapsrommet der separasjonen mellom klassene er maksimalisert. Separasjonsmålet er gitt som spredningen mellom klassene dividert på spredningen innen klassene

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^t S_B \boldsymbol{w}}{\boldsymbol{w}^t S_W \boldsymbol{w}}$$

der

$$S_B = (\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^t$$

angir spredningen mellom klassene og

$$S_W = \sum_{i=1}^{2} \sum_{\boldsymbol{x} \in \mathscr{X}_i} (\boldsymbol{x} - \boldsymbol{m}_i)(\boldsymbol{x} - \boldsymbol{m}_i)^t$$

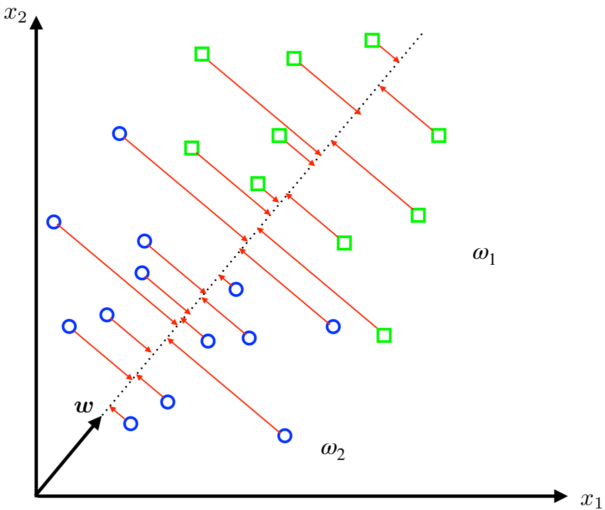
angir spredning innen klassene (som før). Det kan da vises (se utledning i DHS) at maksimalisering av $J(\boldsymbol{w})$ mhp. $\boldsymbol{w}$ gir

$$\boldsymbol{w} = S_W^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2).$$

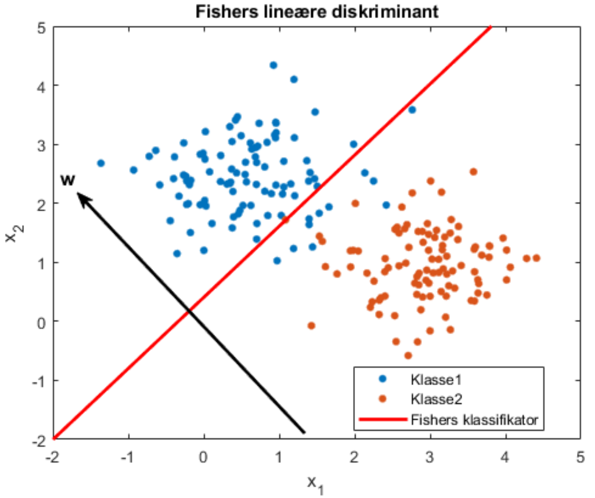
Alternativt valg av marginvektor (forts.)

Produktet $\boldsymbol{w}^t\boldsymbol{x}$ kan betraktes som en projeksjon av det opprinnelige d-dimensjonale egenskapsrommet ned i et éndimensjonalt underrom gitt ved vektoren $\boldsymbol{w}$.

Terskelen (desisjonsgrensen) mellom klassene er gitt ved $w_0 = -\boldsymbol{w}^t\boldsymbol{m}$.

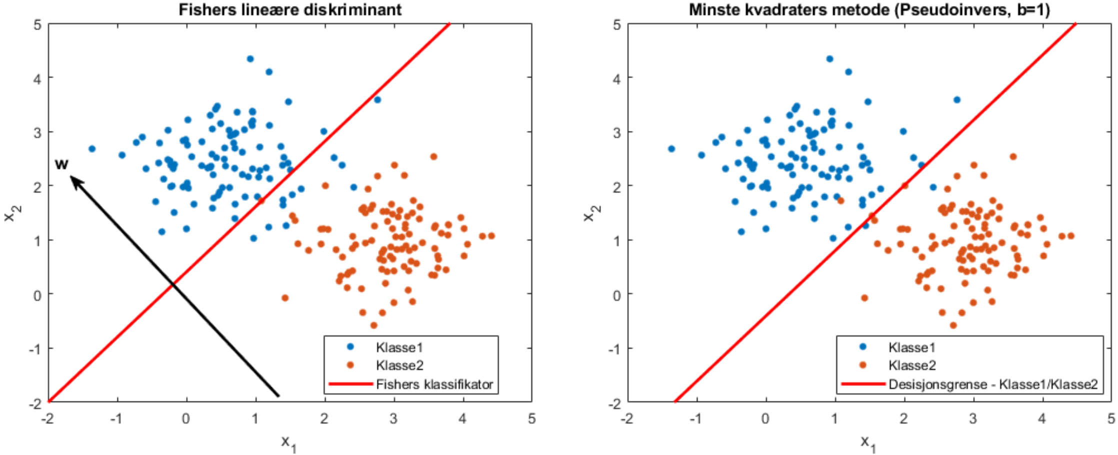


Eksempel – Fishers lineære diskriminant på ikke-separabelt datasett



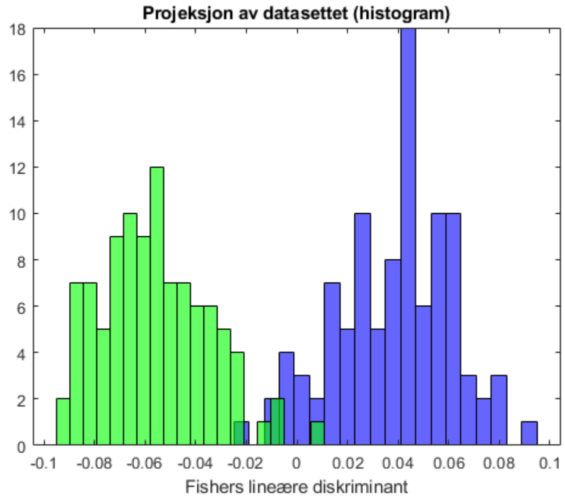
Fishers lineære diskriminant $\boldsymbol{w}$ på samme ikke-separable datasett som tidligere.

Eksempel – Fishers lineære diskriminant på ikke-separabelt datasett (forts.)



Sammenlikning av resultatene for Fishers klassifikator og pseudoinvers-metoden. Her er $n_1 = n_2 = 100$, slik at orienteringen til desisjonsgrensen er den samme i begge tilfeller.

Eksempel – Fishers lineære diskriminant på ikke-separabelt datasett (forts.)



Histogram over datasettet projisert ned på aksen definert ved vektoren $\boldsymbol{w}$.

Innhold i kurset

- Introduksjon til mønstergjenkjenning
- Beslutningsteori
- Parametriske metoder
- Ikke-parametriske metoder
- Lineære og generaliserte diskriminantfunksjoner (forts.)
- Evaluering av klassifikatorer
- Ikke-ledet læring
- Klyngeanalyse.