TEK5020/9020 Mønstergjenkjenning
Høsten 2023

Forelesning 11 – Lineære og generaliserte diskriminantfunksjoner (3)

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

UiO : Institutt for teknologisystemer

20. 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.

Minste kvadraters metode

Ønsker her en 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 størrelsene

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

Likningssystemet $Y\boldsymbol{a} = \boldsymbol{b}$ skal da løses med hensyn på $\boldsymbol{a}$. Søker minste kvadraters løsning der kriteriefunksjonen

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

Løsningsmetoder:

- Direkte løsning (*Pseudoinvers metode* - se forelesning 10),
- Gradientsøk (skal se på dette her).

Løsning ved gradientsøk

Kriteriefunksjonen

$$J_s(\boldsymbol{a}) = \|Y\boldsymbol{a} - \boldsymbol{b}\|^2 = (Y\boldsymbol{a} - \boldsymbol{b})^t(Y\boldsymbol{a} - \boldsymbol{b})$$

skal her minimaliseres ved gradientsøk. Gradienten med hensyn på $\boldsymbol{a}$ er

$$\nabla J_s(\boldsymbol{a}) = 2Y^t(Y\boldsymbol{a} - \boldsymbol{b}).$$

Dette gir da algoritmen

$$\left.\begin{aligned} \boldsymbol{a}_1 &= \text{ vilkårlig startvektor} \\ \boldsymbol{a}_{k+1} &= \boldsymbol{a}_k - \rho_k Y^t(Y\boldsymbol{a}_k - \boldsymbol{b}),\ k = 1, 2, \ldots \end{aligned}\right\} \quad \text{(oppdatering for hele treningssettet).}$$

Algoritmen kan vises å konvergere til en vektor $\boldsymbol{a}$ som tilfredsstiller $Y^t(Y\boldsymbol{a} - \boldsymbol{b}) = 0$ dersom

$$\rho_k = \rho_1/k \text{ med } \rho_1 > 0.$$

Dette gir en minste kvadraters løsning selv om matrisen Y^tY er singulær.

Enkeltsampel oppdatering

En tilsvarende enkeltsampleregel (*Widrow-Hoff* algoritmen) er gitt ved

$$\left.\begin{aligned} \boldsymbol{a}_1 &= \text{ vilkårlig startvektor} \\ \boldsymbol{a}_{k+1} &= \boldsymbol{a}_k + \rho_k(b_k - \boldsymbol{a}_k^t \boldsymbol{y}^k)\boldsymbol{y}^k, \; k = 1, 2, \ldots \end{aligned}\right\} \quad \text{(oppdatering for hvert sample)}$$

der samplene i treningssettet behandles syklisk, som i de tilsvarende feilrettings-algoritmene.

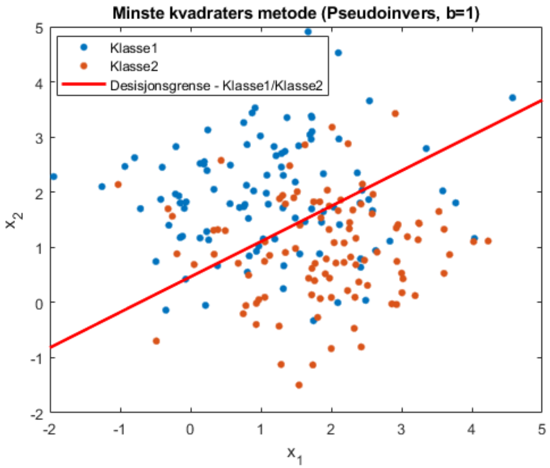
Legg merke til at her fører *alle* sampler der $\boldsymbol{a}_k^t \boldsymbol{y}^k \neq b_k$ til en justering av vektvektoren.

Avtagende ρ_k gir generelt konvergens til en vektvektor der gradienten til J_s er null, f.eks. $\rho_k = \rho_1 / k$.

Minste kvadraters metode – generelt

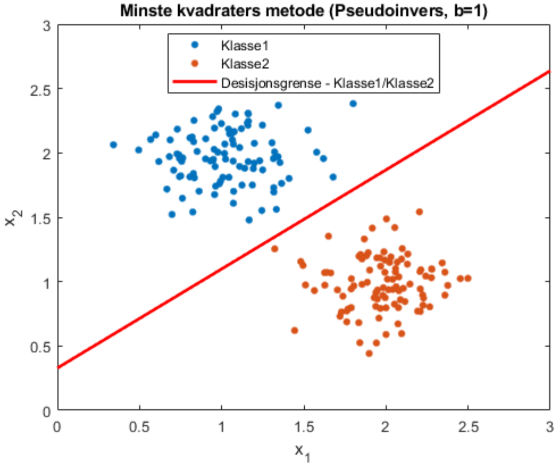
- Kriteriefunksjonen $J_s(\boldsymbol{a})$ skal minimaliseres mht. $\boldsymbol{a}$.
- En minste kvadraters løsning eksistere alltid, selv om Y^tY er singulær.
- Løsningen avhenger av $\boldsymbol{b}$ (mulig valg $\boldsymbol{b} = [1, \ldots, 1]^t$).
- Ikke garantert separerende vektor på lineært separabelt datasett.
- Håp om god løsning *både* på separable og ikke-separable sett.
- Kan generaliseres til mange klasser.

Minste kvadraters metode på ikke-separabelt datasett med to klasser



Desisjonsgrense med Pseudoinvers metode og $\boldsymbol{b} = [1, \ldots, 1]^t$.

Minste kvadraters metode på separabelt datasett med to klasser



Desisjonsgrense med Pseudoinvers metode og $\boldsymbol{b} = [1, \ldots, 1]^t$. I dette tilfellet ble en løsningsvektor funnet, men dette er ikke garantert.

Generalisering til mange klasser

Treningsalgoritmene for lineære diskriminantfunksjoner har så langt blitt beskrevet for toklasseproblemet, men metodene kan generaliseres til et vilkårlig antall klasser.

Har sett på:

- Feilrettingsmetoder (tilfredsstille et sett av ulikheter),
- Minste kvadraters metode (løse et likningssystem).

Siden disse metodene baserer seg på forskjellige prinsipper, vil også metodene for generalisering til mange klasser være forskjellige.

Fremgangsmåtene blir behandlet hver for seg i det følgende.

Generalisering av feilrettingsmetodene – Keslers konstruksjon

Keslers konstruksjon kan brukes til å generalisere feilrettingsmetodene fra to til flere klasser.

Anta først treningssettet bestående av utvidede egenskapsvektorer, som vi tenker oss sortert iht. klassetilhørighet:

$$\mathscr{Y} = \{\underbrace{\mathbf{y}_1, \mathbf{y}_2, \ldots}_{n_1} \underbrace{\ldots\ldots}_{n_2} \overset{\ldots}{\ldots} \underbrace{\ldots, \mathbf{y}_n}_{n_c}\} = \mathscr{Y}_1 + \cdots + \mathscr{Y}_c,$$

med n_i sampler fra hver klasse, der $i = 1, \ldots, c$.

Settet er *lineært separabelt* hvis det eksisterer vektvektorer $\hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_c$ slik at

$$\hat{\mathbf{a}}_i^t \mathbf{y}_k > \hat{\mathbf{a}}_j^t \mathbf{y}_k$$

for alle $\mathbf{y}_k \in \mathscr{Y}_i$ og for alle i, j der $j \neq i$.

Dette gir oss da et sett av diskriminantfunksjoner som klassifiserer alle treningssamplene perfekt.

Keslers konstruksjon (forts.)

Anta for øyeblikket $\boldsymbol{y} \in \mathscr{Y}_1$. Ønsker da at

$$\hat{\boldsymbol{a}}_1^t \boldsymbol{y} - \hat{\boldsymbol{a}}_j^t \boldsymbol{y} > 0, \quad j = 2, \ldots, c,$$

slik at $\boldsymbol{y}$ blir riktig klassifisert, dvs. $c-1$ ulikheter skal tilfredsstilles. Innfører nå en $c \cdot \hat{d}$-dimensjonal vektvektor

$$\boldsymbol{\alpha} = \left[\boldsymbol{a}_1^t, \boldsymbol{a}_2^t, \ldots, \boldsymbol{a}_c^t\right]^t$$

og $c-1$ samplevektorer med $c \cdot \hat{d}$ komponenter

$$\boldsymbol{\eta}_{12} = \begin{bmatrix} \boldsymbol{y} \\ -\boldsymbol{y} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}, \qquad \boldsymbol{\eta}_{13} = \begin{bmatrix} \boldsymbol{y} \\ \mathbf{0} \\ -\boldsymbol{y} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}, \qquad \ldots, \qquad \boldsymbol{\eta}_{1c} = \begin{bmatrix} \boldsymbol{y} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \\ -\boldsymbol{y} \end{bmatrix}.$$

Undervektorene i disse utvidete egenskapsvektorene har alle dimensjon $\hat{d}$ ($\mathbf{0}$ symboliserer en vektor bestående av $\hat{d}$ nuller, dvs. $\mathbf{0} = [0, \ldots, 0]^t$).

Keslers konstruksjon (forts.)

Med disse definisjonene kan settet av ulikheter for det valgte samplet $\boldsymbol{y} \in \mathscr{Y}_1$ skrives som

$$\hat{\boldsymbol{\alpha}}^t \boldsymbol{\eta}_{1j} = \hat{\mathbf{a}}_1^t \boldsymbol{y} - \hat{\mathbf{a}}_j^t \boldsymbol{y} > 0,\ j = 2, \dots, c.$$

For et sample $\boldsymbol{y}$ fra vilkårlig klasse ω_i ønsker vi generelt

$$\boldsymbol{a}_i^t \boldsymbol{y} > \boldsymbol{a}_j^t \boldsymbol{y} \;\text{ der }\; j \neq i \text{ når } \boldsymbol{y} \in \mathscr{Y}_i.$$

Her er

$$\boldsymbol{\alpha} = \begin{bmatrix} \boldsymbol{a}_1 \\ \vdots \\ \boldsymbol{a}_c \end{bmatrix} \quad \text{og} \quad \boldsymbol{\eta}_{ij} = \begin{array}{c} \begin{bmatrix} 0 \\ \vdots \\ \boldsymbol{y} \\ \vdots \\ -\boldsymbol{y} \\ \vdots \\ 0 \end{bmatrix} \end{array} \begin{array}{l} \leftarrow 1 \\ \vdots \\ \leftarrow i \\ \vdots \\ \leftarrow j \\ \vdots \\ \leftarrow c \end{array} \qquad j = 1, \dots, c,\ j \neq i$$

der undervektoren $\boldsymbol{y}$ står i posisjon i, mens $-\boldsymbol{y}$ står i posisjon j.

Keslers konstruksjon (forts.)

Problemet består da i å finne en vektor $\hat{\boldsymbol{\alpha}}$ som tilfredsstiller

$$\hat{\boldsymbol{\alpha}}^t \boldsymbol{\eta}_{ij} > 0 \; \forall i,j \text{ der } j \neq i.$$

Mangeklasseproblemet er da omformulert som et toklasseproblem (med én enkelt diskriminantfunksjon), der dimensjonen på egenskapsrommet er multiplisert med c og antall sampler med $c-1$.

Dette gjør det nå mulig å generalisere perceptron- og relaksasjonsalgoritmene fra toklasse- til mangeklasseproblemet.

Keslers konstruksjon gjør det også mulig å generalisere bevisene for konvergens av disse metodene fra toklasseproblemet, til problemer med vilkårlig antall klasser.

Generalisering av fast-inkrement regelen

Fast inkrement regelen for to klasser kan skrives som

$$\begin{aligned}\boldsymbol{a}(1) &= \text{vilkårlig}\\ \boldsymbol{a}(k+1) &= \boldsymbol{a}(k) + \boldsymbol{y}^k\end{aligned}$$

Denne kan nå umiddelbart generaliseres til

$$\begin{aligned}\boldsymbol{\alpha}(1) &= \text{vilkårlig}\\ \boldsymbol{\alpha}(k+1) &= \boldsymbol{\alpha}(k) + \boldsymbol{\eta}_{ij}^k\end{aligned}$$

der $\boldsymbol{\eta}_{ij}^k$ er et feilklassifisert sample med hensyn til $\boldsymbol{\alpha}(k)$, dvs. $\boldsymbol{\alpha}^t(k)\boldsymbol{\eta}_{ij}^k \leq 0$.

Generalisering av fast-inkrement regelen (forts.)

Oppdateringen i den generaliserte algoritmen kan skrives ut som

$$\begin{bmatrix} \boldsymbol{a}_1(k+1) \\ \vdots \\ \boldsymbol{a}_i(k+1) \\ \vdots \\ \boldsymbol{a}_j(k+1) \\ \vdots \\ \boldsymbol{a}_c(k+1) \end{bmatrix} = \begin{bmatrix} \boldsymbol{a}_1(k) \\ \vdots \\ \boldsymbol{a}_i(k) \\ \vdots \\ \boldsymbol{a}_j(k) \\ \vdots \\ \boldsymbol{a}_c(k) \end{bmatrix} + \begin{bmatrix} 0 \\ \vdots \\ \boldsymbol{y}^k \\ \vdots \\ -\boldsymbol{y}^k \\ \vdots \\ 0 \end{bmatrix},$$

der samplet $\boldsymbol{y}^k$ fra klasse ω_i er feilklassifisert til klasse ω_j.

Generalisering av fast-inkrement regelen (forts.)

Den generaliserte fast-inkrement regelen kan da skrives som

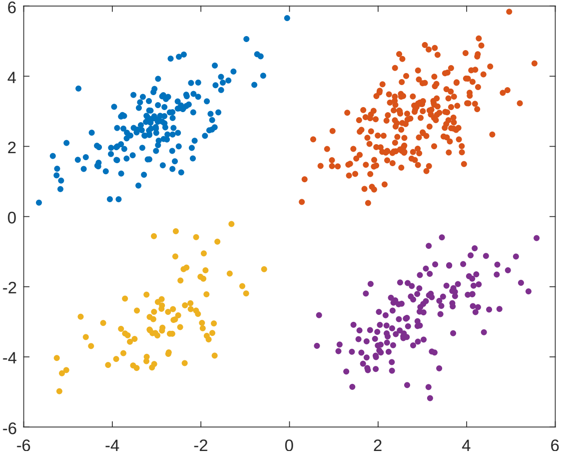
$$\boldsymbol{a}_1(1), \ldots, \boldsymbol{a}_c(1) = \text{vilkårlige startvektorer}$$

$$\left.\begin{aligned} \boldsymbol{a}_i(k+1) &= \boldsymbol{a}_i(k) + \boldsymbol{y}^k \quad \text{(sann klasse)} \\ \boldsymbol{a}_j(k+1) &= \boldsymbol{a}_j(k) - \boldsymbol{y}^k \quad \text{(feil klasse)} \\ \boldsymbol{a}_l(k+1) &= \boldsymbol{a}_l(k), \; l \neq i,j \end{aligned}\right\} \quad k = 1, 2, \ldots$$

Her endres kun vektvektoren til klassen det feilklassifiserte samplet faktisk tilhører og den klassen det er blitt feilaktig klassifisert til. Vektvektorene for de andre klassene endres ikke.

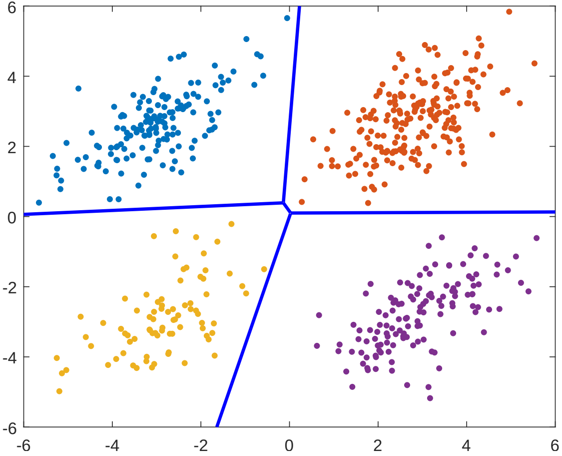
De øvrige feilrettingsmetodene kan generaliseres på tilsvarende måte.

Eksempel – lineært separabelt datasett med fire klasser



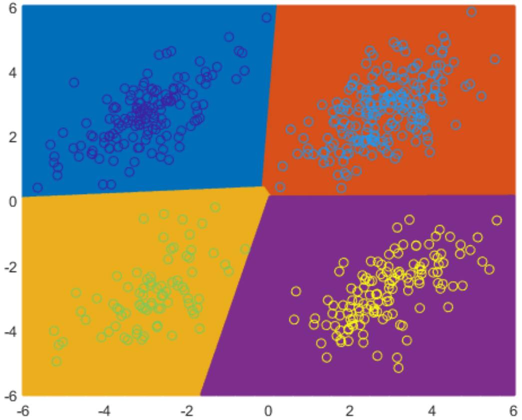
Fire klasser med hhv. 150 (øverst t.v.), 200 (øverst t.h.), 80 (nederst t.v.) og 130 (nederst t.h.) sampler.

Resultat – fast-inkrement regelen med startverdier i origo

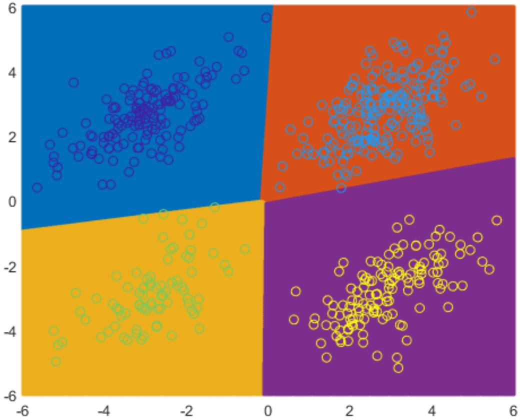


Lineært separerende desisjonsgrenser funnet etter fire iterasjoner ($\boldsymbol{a}_0 = 0$ for alle klasser).

Fast-inkrement regelen med forskjellige startverdier



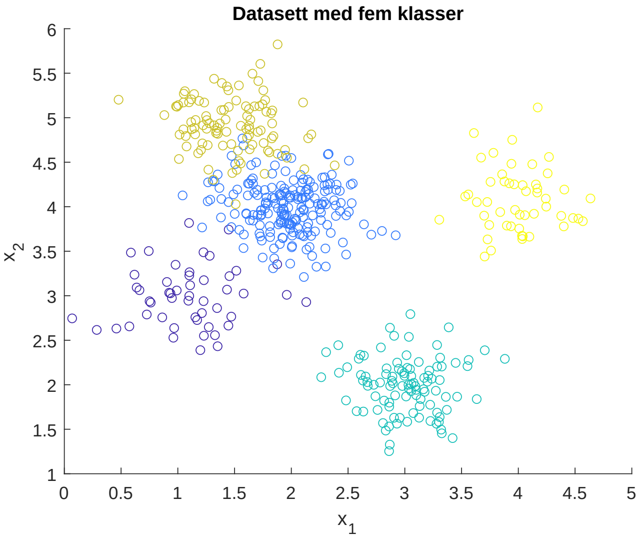
Fire iterasjoner ($\boldsymbol{a}_0 = 0$).



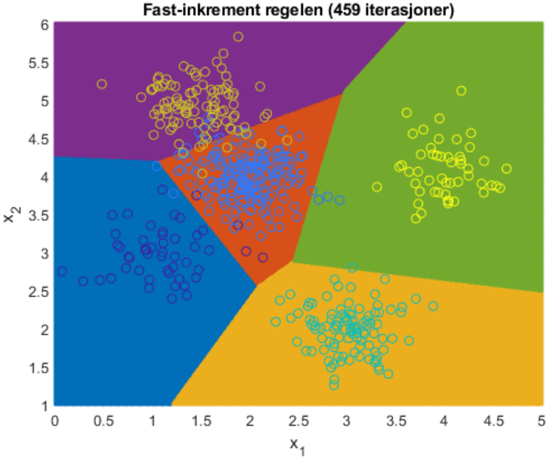
Fem iterasjoner ($\boldsymbol{a}_0 =$ tilfeldig).

Konvergens til løsningsvektorer uavhengig av startverdier.

Ikke-separabelt datasett med fem klasser

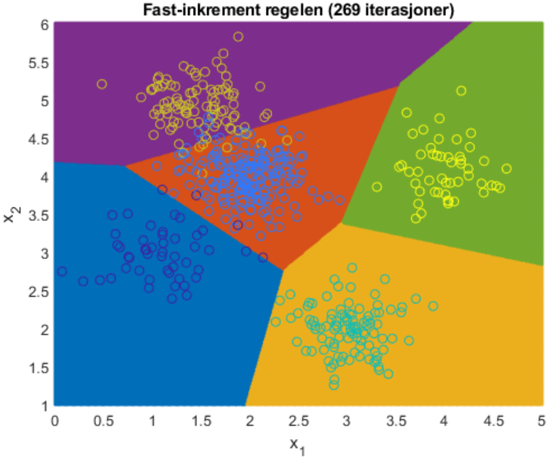


Ikke-separabelt datasett med fem klasser (forts.)



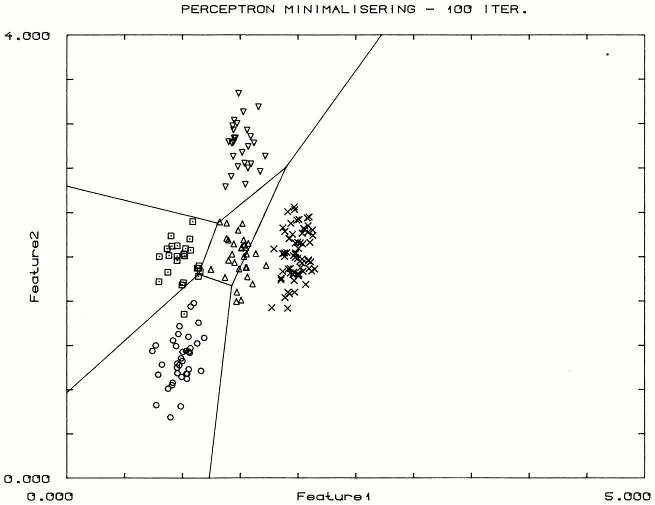
Desisjonsregioner for fast-inkrement regelen med pocket algoritmen (stopp etter 459 av 1000 iterasjoner, 24 feilklassifiseringer).

Ikke-separabelt datasett med fem klasser (forts.)



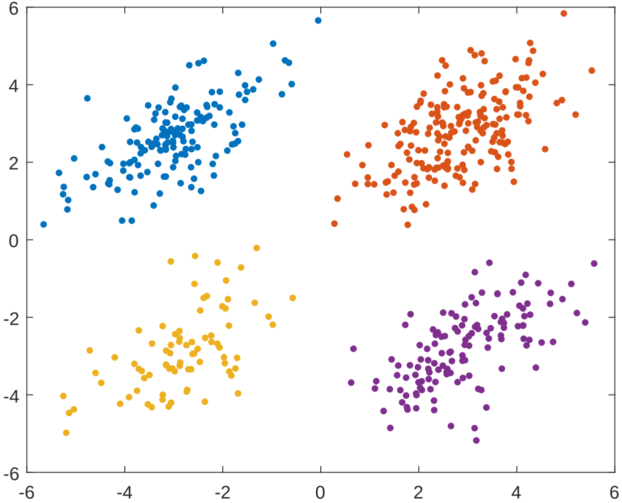
Desisjonsregioner for fast-inkrement regelen med pocket algoritmen på tilfeldig permutert treningssett (stopp etter 269 av 1000 iterasjoner, 14 feilklassifiseringer).

Perceptron-algoritmen på annet datasett med fem klasser



Stykkevis lineære desisjonsgrenser på ikke-separabelt datasett.

Fast-inkrement regelen for kvadratiske diskriminantfunksjoner



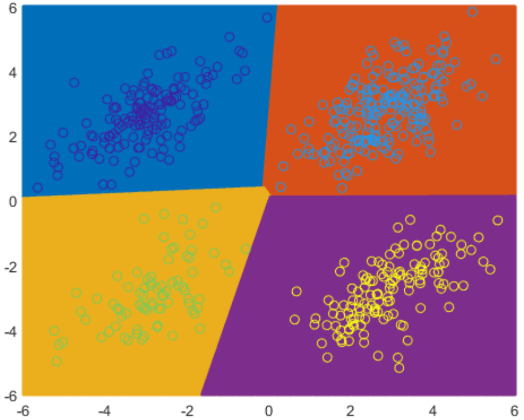
Datasett med fire klasser (d=2).

Transformasjon fra x-rom til y-rom (d=3):

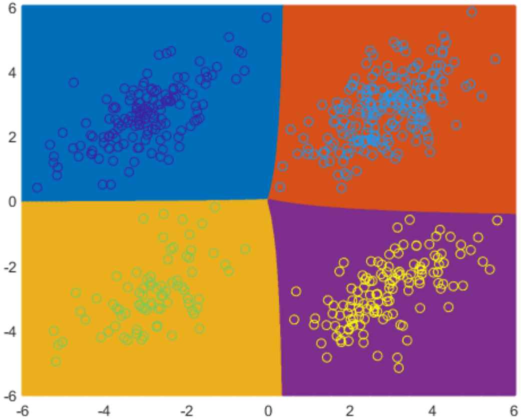
$$\boldsymbol{y} = \begin{bmatrix} x_1 \\ x_2 \\ \frac{1}{3}x_1x_2 \end{bmatrix}$$

Ny egenskapsvektor utvidet til annen orden i $\boldsymbol{x}$.

Fast-inkrement regelen for kvadratiske diskriminantfunksjoner (forts.)



Lineære diskriminantfunksjoner.



Kvadratiske diskriminantfunksjoner.

Konvergens til løsningsvektorer etter fire iterasjoner ($\boldsymbol{a}_0 = 0$).

Generalisering av minste kvadraters metode

For to klasser ønsker man her å oppfylle likningssystemet $\boldsymbol{a}^t\boldsymbol{y} = b$ for alle sampler $\boldsymbol{y}$ i treningssettet. Ved å oppheve fortegnskonvensjonen (forenklingen som ble brukt under behandlingen av toklasseproblemet) og for enkelthets skyld velge $b = 1$, kan liknings-systemet skrives som

$$\begin{aligned}\boldsymbol{a}^t\boldsymbol{y} &= 1 \text{ når } \boldsymbol{y} \in \mathscr{Y}_1 \\ \boldsymbol{a}^t\boldsymbol{y} &= -1 \text{ når } \boldsymbol{y} \in \mathscr{Y}_2\end{aligned}$$

I mangeklassetilfellet kan dette omformuleres til

$$\left.\begin{aligned}\boldsymbol{a}_i^t\boldsymbol{y} &= 1 \text{ når } \boldsymbol{y} \in \mathscr{Y}_i \\ \boldsymbol{a}_i^t\boldsymbol{y} &= 0 \text{ når } \boldsymbol{y} \notin \mathscr{Y}_i\end{aligned}\right\}, i = 1, \ldots, c$$

En eksakt løsning av dette likningssystemet ville gitt vektvektorer som tilfredsstiller

$$\boldsymbol{a}_i^t\boldsymbol{y} > \boldsymbol{a}_j^t\boldsymbol{y} \quad \forall j \neq i \text{ når } \boldsymbol{y} \in \mathscr{Y}_i$$

som er nettopp det man ønsker å oppnå.

Generalisering av minste kvadraters metode (forts.)

Et slikt sett av vektvektorer vil da klassifisere alle treningssampler riktig, men siden likningssystemet er overbestemt må man, som i toklassetilfellet, nøye seg med en minste kvadraters løsning.

Innfører nå datamatrisen Y for mangeklasseproblemet:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_c \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1^t \\ \vdots \\ \boldsymbol{y}_n^t \end{bmatrix} \qquad (n \times \hat{d})$$

og en matrise A der søylene er vektvektorene for hver av klassene:

$$A = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_c] \quad (\hat{d} \times c).$$

Generalisering av minste kvadraters metode (forts.)

I tillegg defineres en matrise B der søylene inneholder marginene for hver klasse:

$$B = \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_c \end{bmatrix} \quad (n \times c) \quad \text{der} \quad B_i = \begin{bmatrix} 0 \dots 0 & 1 & 0 \dots 0 \\ \cdots & . & \cdots \\ \cdots & . & \cdots \\ \cdots & . & \cdots \\ 0 \dots 0 & 1 & 0 \dots 0 \end{bmatrix} \quad (n_i \times c),$$

dvs. en matrise med enere i søyle nr. i og nuller ellers.

Likningssystemet kan da skrives på formen

$$YA = B \quad \text{som skal løses med hensyn til vektmatrisen } A.$$

Normalt er som nevnt dette likningssystemet overbestemt, slik at det ikke eksisterer noen eksakt løsning.

Generalisering av minste kvadraters metode (forts.)

En minste kvadraters løsning kan derimot finnes ved å minimalisere

$$Tr\{(YA - B)^t (YA - B)\} = \sum_{i=1}^{c} \|Y\boldsymbol{a}_i - \boldsymbol{b}_i\|^2 \quad \text{(kan vises)}.$$

Dette svarer til å minimalisere hvert ledd i summen, som for toklasseproblemet, dvs:

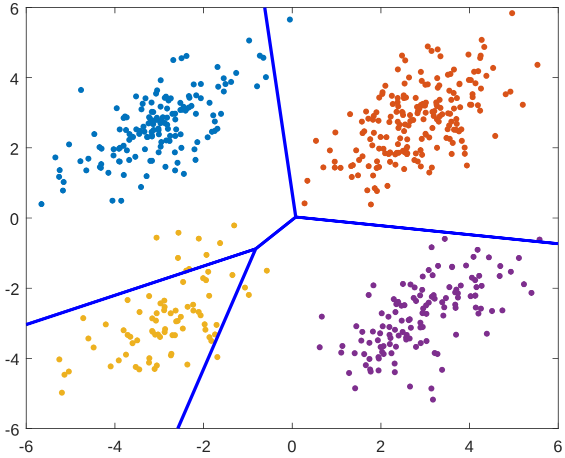
$$\underline{\underline{\boldsymbol{a}_i = Y^\dagger \boldsymbol{b}_i}}, \quad i = 1, \ldots, c \quad \text{eller} \quad \underline{\underline{A = Y^\dagger B}}.$$

Her er

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

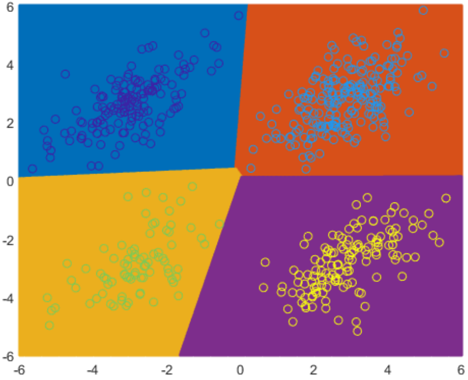
den pseudoinverse til Y, som tidligere.

Eksempel – minste kvadraters metode – fire klasser

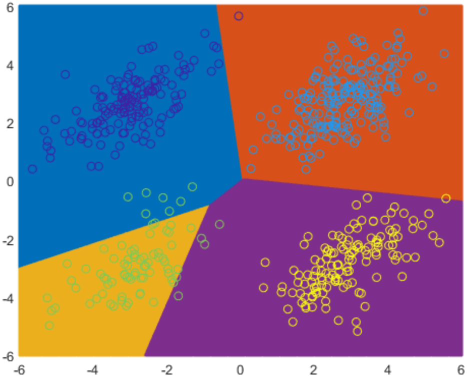


Stykkevis lineære desisjonsgrenser for Pseudoinvers-metoden ($b = 1$ for riktig klasse).

Sammenligning av feilrettingsmetoden og minste kvadraters metode



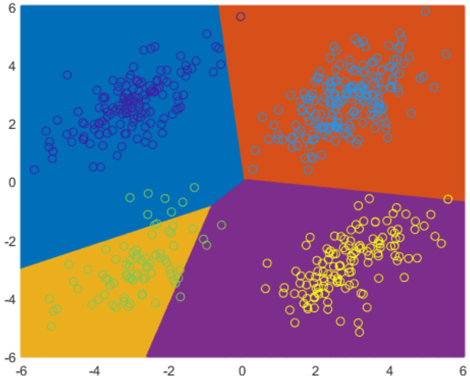
Fast-inkrement regelen.



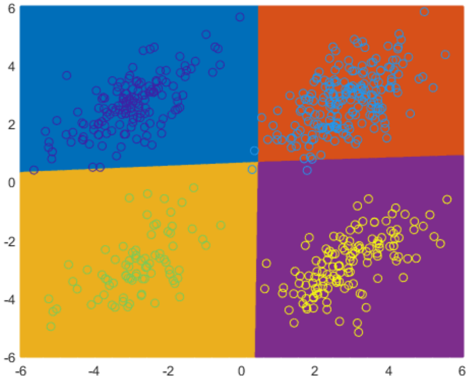
Pseudoinvers-metoden.

Feilrettingsmetoden konvergerer til lineær separabel løsning, mens minste kvadraters metode gir en (kanskje) god løsning med tanke på generalisering til nye/ukjente data.

Minste kvadraters metode med forskjellige verdier av b



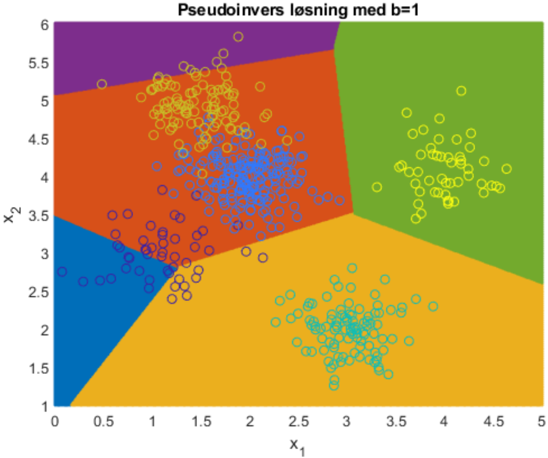
Pseudoinvers-metoden med $b = 1$.



Pseudoinvers-metoden med $b = n/n_i$.

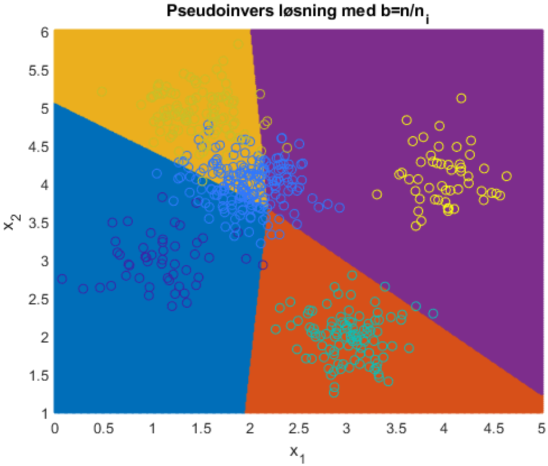
Sterkere vektlegging (større verdi av marginen b) for klasser med få treningssampler gir her bedre og mer balanserte desisjonsgrenser.

Minste kvadraters metode på datasett med fem klasser



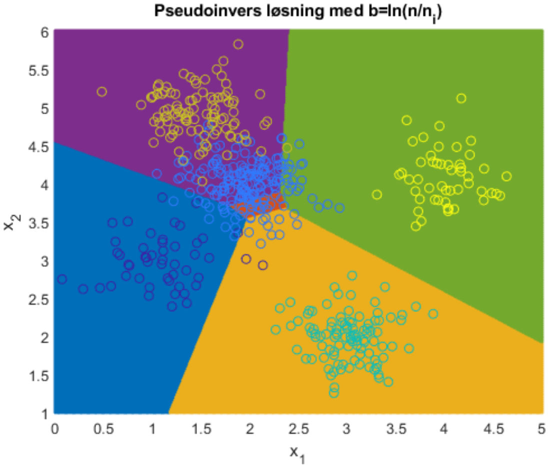
Desisjonsgrenser for ikke-separabelt datasett (Pseudoinvers metode med $b = 1$).

Minste kvadraters metode på datasett med fem klasser (forts.)



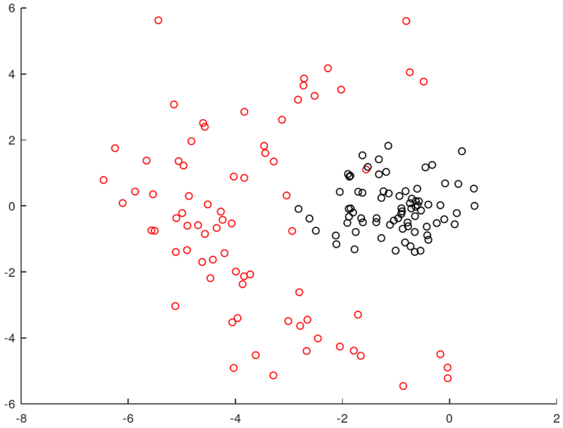
Desisjonsgrenser for ikke-separabelt datasett (Pseudoinvers metode med $b = n/n_i$).

Minste kvadraters metode på datasett med fem klasser (forts.)



Desisjonsgrenser for ikke-separabelt datasett (Pseudoinvers metode med $b = \ln(n/n_i)$).

Generaliserte diskriminantfunksjoner

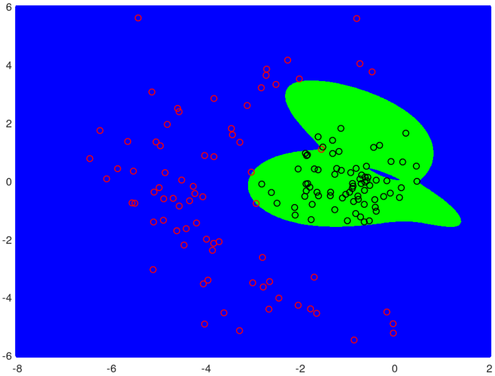


Datasett med to klasser (d=2).

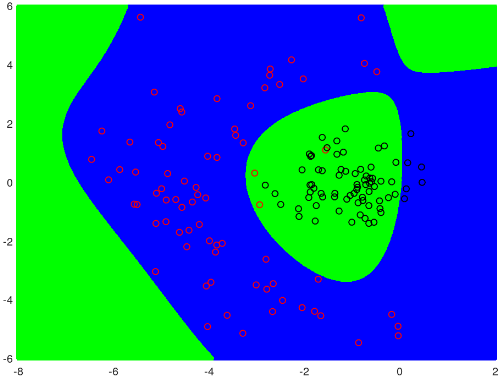
$$\boldsymbol{y} = \begin{bmatrix} x_1 \\ x_2 \\ x_1 x_2 \\ x_1^2 \\ x_2^2 \\ x_1^2 x_2 \\ x_1 x_2^2 \\ x_1^3 \\ x_2^3 \\ x_1^2 x_2^2 \\ x_1^3 x_2 \\ x_1 x_2^3 \\ x_1^4 \\ x_2^4 \end{bmatrix}$$

Egenskapsvektor av fjerde orden i $\boldsymbol{x}$.

Generaliserte diskriminantfunksjoner (forts.)



Generalisert fast-inkrement regel - løsningsvektorer etter 5652 iterasjoner.



Generalisert minste kvadraters metode.

Desisjonsregioner i det opprinnelige egenskapsrommet (x-rommet).

Innhold i kurset

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