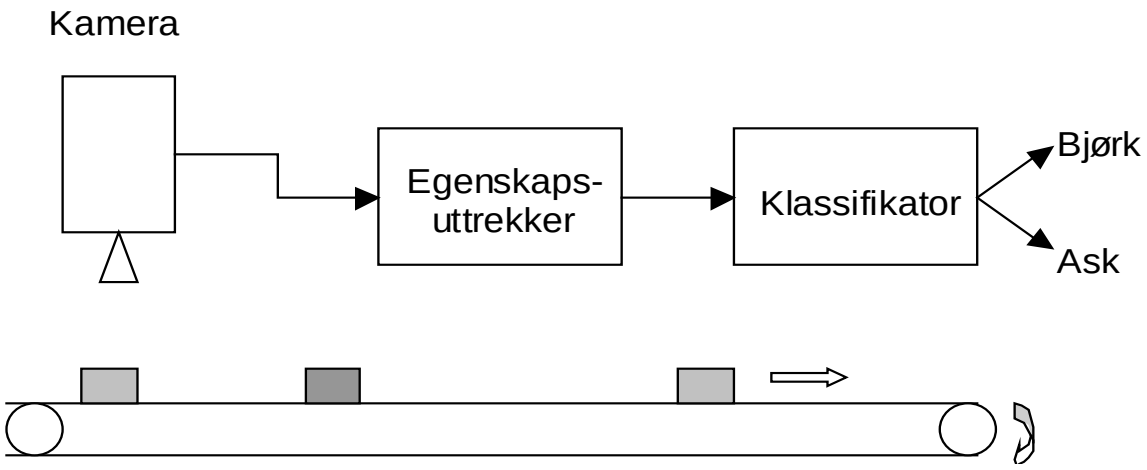
Løsningsforslag eksamen TEK5020 - 2022H

Oppgave 1

Innledning

a) Et typisk mønstergjenkjenningssystem kan bestå av følgende komponenter:

- En sensor som henter inn rådata (f.eks. et kamera, som vist i figuren nedenfor)
- En egenskapsuttrekker som bearbeider rådataene og beregner en eller flere tallstørrelser (egenskaper, målinger) som forhåpentligvis karakteriserer objektene som skal gjennkjennes (inneholder informasjon som bidrar til å skille mellom de ulike klassene)
- En klassifikator som gjør et valg av klasse, basert på de beregnede egenskapene



Figur 1: Klassifiseringssystem for å skille mellom trestykker fra klassene bjørk og ask.

b) Av eksempler på praktisk bruk av mønstergjenkjenning kan nevnes

- Automatisk lesing av bilskilt i bomstasjoner og på parkeringsplasser
- Strekkodelesing f.eks. i butikker
- Lesing av fingeravtrykk
- Automatisk lesing av trykt og håndskreven tekst
- Ansiktgjenkjenning, f.eks. i digitale kameraer

c) Den vanligste måten å estimere feilraten til en klassifikator på er den *empiriske metoden*. Den går i korthet ut på å kjøre klassifikatoren på testdata av kjent klassetilhørighet, og telle opp antall feilklassifiseringer. Forholdet mellom antall feil og det totale antall testede sampler gir da et godt estimat av feilraten. Det er viktig å benytte et uavhengig sett av testdata til feilrateestimeringen. Dersom man hadde brukt treningssettet til dette ville feilrateestimatet bli for optimistisk, og man ville ikke være i stand til å oppdage eventuell overtrening av klassifikatoren.

Leave-one-out metoden går i korthet ut på å holde ett og ett sample fra treningssettet utenfor, trene klassifikatoren på de øvrige samplene i det merkede datasettet og teste klassifikatoren på det utelatte samplet. Prosessen gjentas for hvert sample i datasettet, og feilraten beregnes som forholdet mellom antall feil og det totale antall sampler som er testet. På denne måten får man brukt hele datasettet både til trening og testing, men dette er mye mer tidkrevende enn å dele opp datasettet, trene på den ene delen og teste på den andre. En mellomløsning er k-fold kryssvalidering (kan også nevnes her).

d) Klyngeanalyse består i å dele et umerket datasett i en antall grupper (klynger), slik at sampler innen hver klynge er mest mulig like, mens det er størst mulig ulikhet mellom sampler fra forskjellige klynger. Klyngeanalysen er datadrevet og basert på et avstandsmål i egenskapsrommet (et mål på likhet/similaritet). Klyngeanalyse brukes ofte til å kartlegge strukturen i ukjente data. To hovedtyper av metoder for klyngeanalyse som kan nevnes her er

- Iterativ optimalisering
- Hierarkiske metoder

Oppgave 2

Beslutningsteori

a) *Betinget risk* for en gitt handling α_i kan uttrykkes ved hjelp av kostfunksjoner og a posteriori sannsynligheter ved uttrykket

$$R(\alpha_i|\boldsymbol{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|\boldsymbol{x}), \quad i = 1,\ldots,a.$$

Her er a antall mulige handlinger (actions) og $\lambda(\alpha_i|\omega_j) = \lambda_{ij}$ er kostnaden forbundet med å utføre handling α_i når sann klasse er ω_j. $P(\omega_j|\boldsymbol{x})$ er a posteriori sannsynlighet for klasse ω_j gitt egenskapsvektoren $\boldsymbol{x}$. Generelt kan antall handlinger være forskjellig fra antall klasser ($a \neq c$).

Det kan vises at minimum *total risk* (minimum kostnad) oppnås ved å velge handlingen med minimum betinget risk for en gitt egenskapsvektor $\boldsymbol{x}$. Minimum-risk beslutningsregelen kan da skrives som

$$\text{Velg } \alpha_i \text{ hvis } R(\alpha_i|\boldsymbol{x}) \leq R(\alpha_j|\boldsymbol{x}), \quad j = 1,\ldots,a.$$

b) I denne deloppgaven er $a = c = 2$, slik at de to handlingene α_1 og α_2 svarer til å velge henholdsvis klasse ω_1 eller ω_2. Fordelingsfunksjonene for klassene er endimensjonale, og er på formen

$$p(x|\theta) = \theta^2 x e^{-\theta x},\ \theta > 0, x \geq 0.$$

Parametrene for de to klassene er henholdsvis θ_1 og θ_2. Videre antas at $\lambda_{11} = \lambda_{22} = 0$, dvs. null kostnad for feilfri klassifisering, mens kostnadene λ_{12} og λ_{21} for feilklassifiseringer forutsettes å være større enn null.

Terskelen (desisjonsgrensen) x_0 som minimaliserer den totale risken finner vi der den betingede risken for hver av handlingene er like, dvs. $R(\alpha_1|x_0) = R(\alpha_2|x_0)$. Den betingede risken forbundet med handlingene blir, etter innsetting for kostnadene og aposteriorisannsynlighetene, i dette tilfellet

$$R(\alpha_1|x) = \lambda_{11}P(\omega_1|x) + \lambda_{12}P(\omega_2|x) = \lambda_{12}P(\omega_2|x) = \lambda_{12}\theta_2^2 x e^{-\theta_2 x}P(\omega_2)/p(x)$$
$$R(\alpha_2|x) = \lambda_{21}P(\omega_1|x) + \lambda_{22}P(\omega_2|x) = \lambda_{21}P(\omega_1|x) = \lambda_{21}\theta_1^2 x e^{-\theta_1 x}P(\omega_1)/p(x).$$

Terskelen kan da bestemmes ved å sette disse størrelsene like:

$$\begin{aligned} R(\alpha_1|x) &= R(\alpha_2|x) \\ &\Downarrow \\ \lambda_{12}\theta_2^2 x e^{-\theta_2 x}P(\omega_2) &= \lambda_{21}\theta_1^2 x e^{-\theta_1 x}P(\omega_1) \\ &\Downarrow \\ e^{(\theta_1-\theta_2)x} &= \frac{\lambda_{21}\theta_1^2 P(\omega_1)}{\lambda_{12}\theta_2^2 P(\omega_2)}. \end{aligned}$$

Ved å ta logaritmen på begge sider av likhetstegnet kan det løses ut for x, slik at terskelen blir

$$x_0 = \frac{1}{\theta_1 - \theta_2} \ln\left[\frac{\lambda_{21}\theta_1^2 P(\omega_1)}{\lambda_{12}\theta_2^2 P(\omega_2)}\right].$$

c) I denne deloppgaven antas parameterverdiene $\theta_1 = 1$ og $\theta_2 = 2$, samt at kostnadene λ_{12} og λ_{21} er like og at apriorisannsynlighetene er like. Innsetting i uttrykket for x_0 gir da resultatet

$$x_0 = -\ln[1/2^2] = \ln(4) \approx 1,3863.$$

d) Figur 2 viser fordelingsfunksjonene, terskelen og desisjonsregionene i dette tilfellet, med like kostnader og apriorisannsynligheter.

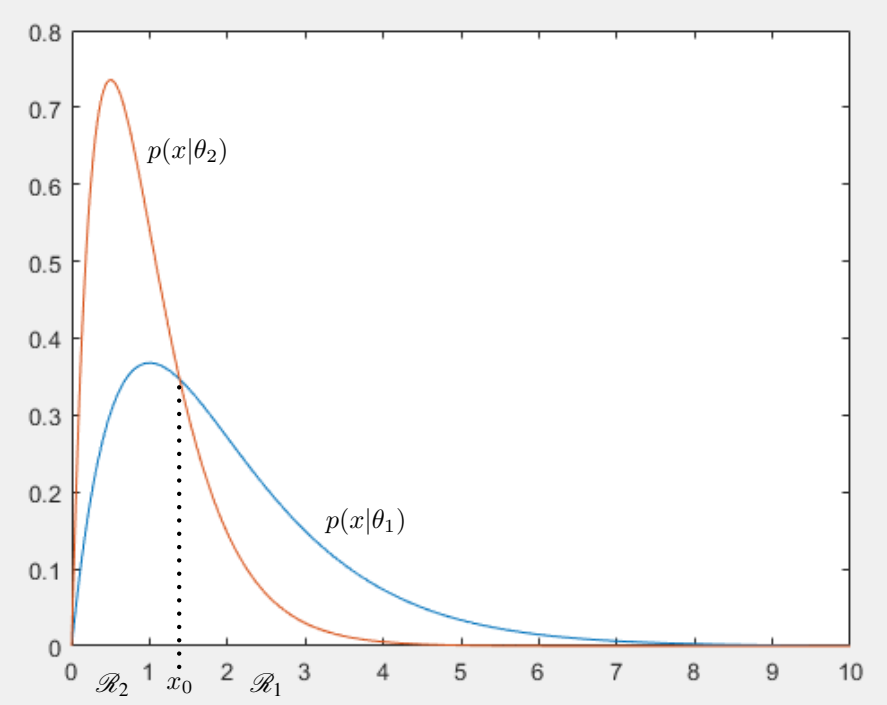
Oppgave 3

Parametriske metoder

a) Den multivariate normalfordelingen er gitt ved

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right].$$

Her er $\boldsymbol{\mu}$ forventningen til den stokastiske variabelen $\boldsymbol{x}$ (vektor av dimensjon d som angir det mest sannsynlige utfallet av variabelen $\boldsymbol{x}$) og Σ kovariansmatrisen til fordelingen (matrise av dimensjon $d \times d$, der elementene i matrisen sier noe om innbyrdes avhengighet mellom komponentene i $\boldsymbol{x}$). Parametrene i denne fordelingen er $\boldsymbol{\mu} = E\{\boldsymbol{x}\}$ og $\Sigma = E\{(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^t\}$.



Figur 2: Fordelingsfunksjonene, terskelen og de tilhørende desisjonsregionene.

b) *Maksimum-likelihood* metoden kan brukes til estimering av parametervektoren $\boldsymbol{\theta}$ i en antatt fordelingsfunksjon $p(\boldsymbol{x}|\boldsymbol{\theta})$ ved ledet læring. Den simultane sannsynlighetstettheten for de observerte treningssamplene kan uttrykkes ved *likelihoodfunksjonen*:

$$p(\mathscr{X}|\boldsymbol{\theta}) = p(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\boldsymbol{x}_k|\boldsymbol{\theta})$$

som skal maksimaliseres med hensyn til den ukjente parametervektoren $\boldsymbol{\theta}$. Det er enklere å arbeide med logaritmen til likelihoodfunksjonen, siden produktet da erstattes med en sum, og resultatet uansett blir det samme siden logaritmen er en monotont voksende funksjon. Derved vil denne *log-likelihoodfunksjonen*

$$\mathscr{L}(\boldsymbol{\theta}) = \ln p(\mathscr{X}|\boldsymbol{\theta}) = \sum_{k=1}^{n} \ln p(\boldsymbol{x}_k|\boldsymbol{\theta})$$

ha maksimum for samme verdi av $\boldsymbol{\theta}$. Maksimum av $\mathscr{L}$ finnes ved å ta gradienten til log-likelihoodfunksjonen med hensyn til $\boldsymbol{\theta}$:

$$\nabla_{\boldsymbol{\theta}} \mathscr{L}(\boldsymbol{\theta}) = \sum_{k=1}^{n} \nabla_{\boldsymbol{\theta}} \ln p(\boldsymbol{x}_k|\boldsymbol{\theta}),$$

og sette den lik null. Dette gir følgende likningssystem for parametervektoren:

$$\sum_{k=1}^{n} \nabla_{\boldsymbol{\theta}} \ln p(\boldsymbol{x}_k|\boldsymbol{\theta}) = 0.$$

c) I denne deloppgaven skal maksimum-likelihood metoden brukes til å finne et estimat for $\boldsymbol{\mu}$ i den multivariate normalfordelingen i deloppgave a, ved hjelp av et treningssett

bestående av egenskapsvektorene $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ (kovariansmatrisen til fordelingen antas kjent). Sannsynlighetstettheten i et vilkårlig samplepunkt $\boldsymbol{x}_k$ som funksjon av $\boldsymbol{\mu}$, blir her:

$$p(\boldsymbol{x}_k|\boldsymbol{\mu}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}_k-\boldsymbol{\mu})^t\Sigma^{-1}(\boldsymbol{x}_k-\boldsymbol{\mu})\right].$$

Logaritmen til tettheten blir da:

$$\ln p(\boldsymbol{x}_k|\boldsymbol{\mu}) = -\frac{1}{2}(\boldsymbol{x}_k-\boldsymbol{\mu})^t\Sigma^{-1}(\boldsymbol{x}_k-\boldsymbol{\mu}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma|,$$

og gradienten med hensyn til $\boldsymbol{\mu}$ blir:

$$\nabla_{\boldsymbol{\mu}} \ln p(\boldsymbol{x}_k|\boldsymbol{\mu}) = \Sigma^{-1}(\boldsymbol{x}_k-\boldsymbol{\mu}),$$

som innsatt i likningssystemet fra deloppgave b gir:

$$\sum_{k=1}^{n} \nabla_{\boldsymbol{\mu}} \ln p(\boldsymbol{x}_k|\boldsymbol{\mu}) = \sum_{k=1}^{n} \Sigma^{-1}(\boldsymbol{x}_k-\boldsymbol{\mu}) = \Sigma^{-1}\sum_{k=1}^{n}(\boldsymbol{x}_k-\boldsymbol{\mu}) = 0.$$

Ved å multiplisere med Σ på begges sider av likhetstegnet reduseres likningssystemet til

$$\sum_{k=1}^{n}(\boldsymbol{x}_k-\boldsymbol{\mu}) = 0,$$

som gir løsningen

$$\underline{\underline{\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n}\boldsymbol{x}_k.}}$$

For å kunne komme fram til denne løsningen må treningssamplene $\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n$ antas å være innbyrdes uavhengige, slik at det blir mulig å faktorisere likelihoodfunksjonen som vist i deloppgave b.

Oppgave 4

Diskriminantfunksjoner

a) I et todimensjonalt problem med to klasser ω_1 og ω_2, der apriorisannsynlighetene for klassene er henholdsvis $P(\omega_1) = 1/3$ og $P(\omega_2) = 2/3$, er klassene multivariat normalfordelte med felles kovariansmatrise

$$\Sigma = \begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix}$$

og forventningsvektorene

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \text{ og } \boldsymbol{\mu}_2 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}.$$

For å utlede en toklasse (felles) diskriminantfunksjon for dette problemet, kan det være naturlig å velge diskriminantfunksjoner på formen

$$g_i(\boldsymbol{x}) = \ln\left[p(\boldsymbol{x}|\omega_i)P(\omega_i)\right] = \ln p(\boldsymbol{x}|\omega_i) + \ln P(\omega_i).$$

Innsetting av den multivariate normalfordelingen (se deloppgave 3a), som i dette tilfellet blir

$$p(\boldsymbol{x}|\omega_i) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^t\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right] = N(\boldsymbol{\mu}_i,\Sigma)$$

gir da

$$g_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^t\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma| + \ln P(\omega_i).$$

Ved å multiplisere ut kvadratleddet og stryke de to neste leddene (siden de er uavhengig av klassen), blir diskriminantfunksjonene

$$\begin{aligned} g_i(\boldsymbol{x}) =& -\frac{1}{2}[-2\boldsymbol{\mu}_i^t\Sigma^{-1}\boldsymbol{x} + \boldsymbol{\mu}_i^t\Sigma^{-1}\boldsymbol{\mu}_i] + \ln P(\omega_i) \\ =& \boldsymbol{\mu}_i^t\Sigma^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_i^t\Sigma^{-1}\boldsymbol{\mu}_i + \ln P(\omega_i). \end{aligned}$$

Her er leddet $\boldsymbol{x}^t\Sigma^{-1}\boldsymbol{x}$ strøket siden det er det samme for begge klasser, og derved ikke har noen betydning for valg av klasse. Den felles diskriminantfunksjonen for de to klassene blir da

$$\begin{aligned} g(\boldsymbol{x}) =& g_1(\boldsymbol{x}) - g_2(\boldsymbol{x}) \\ =& \boldsymbol{\mu}_1^t\Sigma^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_1^t\Sigma^{-1}\boldsymbol{\mu}_1 + \ln P(\omega_1) - \boldsymbol{\mu}_2^t\Sigma^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_2^t\Sigma^{-1}\boldsymbol{\mu}_2 + \ln P(\omega_2) \\ =& (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^t\Sigma^{-1}\boldsymbol{x} + \frac{1}{2}[\boldsymbol{\mu}_2^t\Sigma^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1^t\Sigma^{-1}\boldsymbol{\mu}_1] + \ln\frac{P(\omega_1)}{P(\omega_2)}. \end{aligned}$$

Den inverse kovariansmatrisen kan finnes ved å løse likningssystemet $\Sigma\Sigma^{-1} = I$, som gir

$$\Sigma^{-1} = \frac{1}{5}\begin{bmatrix} 2 & -1 \\ -1 & 3 \end{bmatrix}.$$

Innsetting av forventningsvektorene, den inverse kovariansmatrisen og apriorisannsynlighetene i diskriminantfunksjonen gir da

$$\begin{aligned} g(\boldsymbol{x}) =& -\frac{1}{5}[3,2]\begin{bmatrix} 2 & -1 \\ -1 & 3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \frac{1}{10}[3,2]\begin{bmatrix} 2 & -1 \\ -1 & 3 \end{bmatrix}\begin{bmatrix} 3 \\ 2 \end{bmatrix} - \ln 2, \\ =& -\frac{1}{5}[4,3]\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \frac{18}{10} - \ln 2, \\ =& -\frac{4}{5}x_1 - \frac{3}{5}x_2 + \frac{9}{5} - \ln 2, \\ =& \underline{\underline{-\frac{1}{5}(4x_1 + 3x_2 - 9) - \ln 2,}} \end{aligned}$$

der det er satt inn for komponentene i egenskapsvektoren $\boldsymbol{x}$.

b) Desisjonsgrensen vil være lineær (en rett linje i planet), siden diskriminantfunksjonen er til første orden i $\boldsymbol{x}$. Grunnen til dette er at kvadratleddet i $\boldsymbol{x}$ kunne strykes under utledningen av diskriminantfunksjonen, fordi kovariansmatrisen er lik for begge klasser (se deloppgave a).

c) Her skal egenskapsvektoren

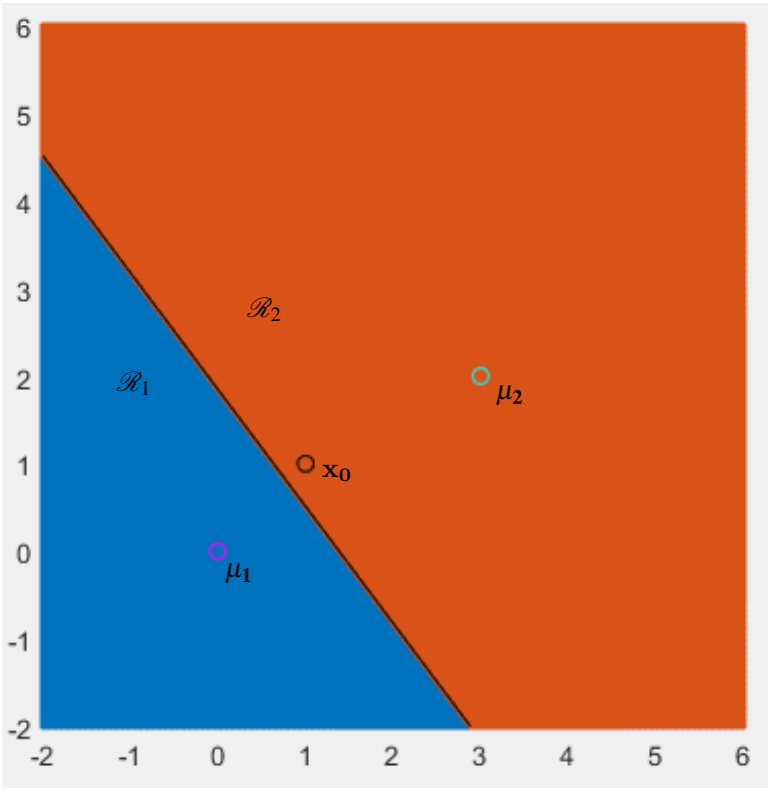
$$\boldsymbol{x}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

klassifiseres ved hjelp av diskriminantfunksjonen fra deloppgave a. Innsetting gir

$$g(\boldsymbol{x}_0) = \frac{2}{5} - \ln 2 \approx -0,293 < 0.$$

$\boldsymbol{x}_0$ blir da klassifisert til ω_2.

d) Figuren viser forventningsvektorene $\boldsymbol{\mu}_1$ og $\boldsymbol{\mu}_2$, punktet $\boldsymbol{x}_0$ og desisjonsregionene.

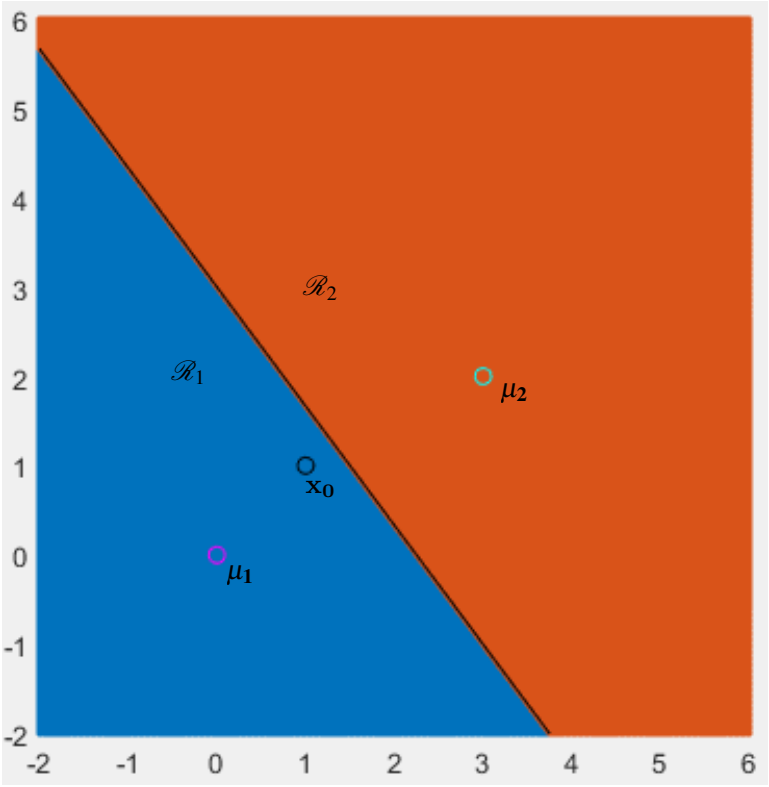


Figur 3: Illustrasjon av desisjonsgrensen og de tilhørende desisjonsregionene for diskriminantfunksjonen i deloppgave a.

e) Dersom apriorisannsynlighetene er like, vil det siste leddet i diskriminantfunksjonen fra deloppgave a falle bort, siden logaritmen til forholdet mellom apriorisannsynlighetene er null. Diskriminantfunksjonen kan derfor skrives som

$$g(\boldsymbol{x}) = -\frac{1}{5}(4x_1 + 3x_2 - 9).$$

Her er det bare skalarleddet (tersklingsvekten) som er endret. Dette fører kun til en parallelforskyvning av desisjonsgrensen, i dette tilfellet fra $\boldsymbol{\mu}_1$ mot $\boldsymbol{\mu}_2$. Innsetting av $\boldsymbol{x}_0$ gir nå $g(\boldsymbol{x}_0) = 2/5 > 0$ slik at klassifiseringsresultatet nå er ω_1 (illustrert i figur 4).



Figur 4: Desisjonsgrensen og de tilhørende desisjonsregionene med like apriorisannsynligheter (denne figuren kreves ikke).

Oppgave 5

Ikke-parametriske metoder

a) Fordeler ved ikke-parametriske metoder er at man ikke må gjøre en antakelse om formen på fordelingene, samt at metodene kan tilpasse seg vilkårlige fordelinger (det gjøres for eksempel ingen forutsetning om lineært separable klasser). Ulemper er at hele treningssettet i utgangspunktet må lagres og gjennomsøkes ved hver klassifisering av et sample. Tetthetsestimatet i et punkt $\boldsymbol{x}$ kan skrives som

$$p_n(\boldsymbol{x}) = \frac{k_n/n}{V_n},$$

basert på et treningssett av egenskapsvektorer $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$. Her er k_n antall treningssampler innenfor en vilkårlig region omkring $\boldsymbol{x}$ og V_n er volumet av regionen.

b) For å oppnå konvergens av estimatet til den sanne tettheten $p(\boldsymbol{x})$ i punktet $\boldsymbol{x}$ når $n \to \infty$, må volumet kunne gå mot null, samtidig som antall sampler innenfor regionen går mot uendelig. Konvergens av estimatet, dvs.

$$\lim_{n\to\infty} p_n(\boldsymbol{x}) = p(\boldsymbol{x})$$

krever derfor at

$$\lim_{n\to\infty} V_n = 0, \quad \lim_{n\to\infty} k_n = \infty, \quad \lim_{n\to\infty} \frac{k_n/n}{V_n} = 0.$$

c) For å kunne estimere a posteriori sannsynlighet, må hele treningssettet brukes under ett (sampler fra alle klasser må være med). La k_i være antall sampler fra ω_i blant det totale antall k sampler innenfor regionen. La videre n_i være antall sampler i treningssettet fra ω_i.Ved å sette tetthetsestimatet fra deloppgave a inn i Bayes formel, kan a posteriori sannsynlighet for klasse ω_i i punktet $\boldsymbol{x}$ estimeres ved

$$P_n(\omega_i|\boldsymbol{x}) = \frac{p_n(\boldsymbol{x}|\omega_i)P_n(\omega_i)}{\sum_{j=1}^{c} p_n(\boldsymbol{x}|\omega_j)P_n(\omega_j)} = \frac{\frac{k_i/n_i}{V} \cdot \frac{n_i}{n}}{\sum_{j=1}^{c} \frac{k_j/n_j}{V} \cdot \frac{n_j}{n}} = \frac{k_i}{\underline{\underline{k}}},$$

for et problem med c klasser. Minimum-feilrateprinsippet tilsier nå at man skal velge klassen med flest representanter innen regionen, siden det er klassen med størst a posteriori sannsynlighet som skal velges. Ved å velge regionen som en hyperkule med $\boldsymbol{x}$ i sentrum, der radien justeres slik at det alltid er k sampler innenfor regionen, leder dette til *k-nærmeste-nabo* regelen (k-NNR). Den består kort og godt i å velge klassen med flest representanter blant de k nærmeste naboene til $\boldsymbol{x}$.

d) *Nærmeste-nabo* regelen (NNR) er et spesialtilfelle der k er satt lik en. Vi ser med andre ord bare på den nærmeste naboen til $\boldsymbol{x}$. Regelen består kort og godt i å velge samme klasse som den nærmeste naboen i treningssettet. NNR har en øvre grense for den asymptotiske feilraten P som funksjon av den optimale feilraten P^*, slik at

$$P \leq P^*(2 - \frac{c}{c-1}P^*).$$

For små verdier av den optimale feilraten er tommelfingeregelen at den asymptotiske feilraten alltid er mindre enn to ganger den optimale feilraten, dvs. $P \leq 2P^*$ (det er tilstrekkelig å oppgi dette).