

# Løsningsforslag eksamen TEK5020/9020 – 2023H

## Oppgave 1

### Innledning

a) Formålet med faget mønstergjenkjenning er å lære maskiner å klassifisere *objekter* (fysiske gjenstander, tilstander mm.) i et antall kategorier eller klasser. Av eksempler på praktisk bruk av dette fagfeltet kan nevnes automatisk lesing av bilskilt, tolking av strekkode o.l., gjenkjenning av fingeravtrykk, automatisk lesing av trykt og håndskreven tekst og ansiktsgjenkjenning.

b) Et typisk mønstergjenkjenningssystem kan bestå av følgende komponenter:

- En *sensor* som henter inn rådata (f.eks. et kamera)
- En *egenskapsuttrekker* som bearbeider rådataene og beregner en eller flere tallstørrelser (egenskaper, målinger) som skal brukes til klassifisering av objektene
- En *klassifikator* som gjør et valg av klasse, basert på de beregnede egenskapene

c) Av eksempler på inngangsdata til et mønstergjenkjenningssystem kan nevnes digitale bilder og signaler (tidsrekker), avleste verdier fra et måleinstrument og besvarte spørreskjemaer.

Formålet med egenskapsuttrekking er å komprimere rådataene til et begrenset antall numeriske størrelser (egenskaper) som inneholder mest mulig informasjon som er relevant for å skille mellom klassene, mens irrelevant informasjon for klassifisering i størst mulig grad er fjernet. Bruk av egenskapsuttrekking gjør det mulig å trene opp en mer robust klassifikator med mindre treningsdata og mindre behov for prosessorkraft, enn om man prøver å trene opp klassifikatoren direkte på rådataene (her kan dyp læring eventuelt nevnes). Konstruksjon av egenskapsuttrekkeren krever tradisjonelt domenekunnskap, dvs. a priori informasjon om problemstillingen mønstergjenkjenningssystemet skal brukes i.

d) Ledet læring er trening av en klassifikator ved hjelp av treningsdata med kjent klassetilhørighet.

I parametriske metoder gjøres det en antagelse om formen på de klassebetingede tetthetsfunksjonene i problemet. Det er kun et sett av parametre som er ukjente, og som estimeres ut fra treningssettet for en og en klasse om gangen. Fordelen med parametriske metoder er at treningen er hurtig og gir kompakte og raske klassifikatorer. Ulempen er at det kreves en antakelse om formen på tetthetsfunksjonene. Dersom antakelsen er feil, blir resultatene naturligvis deretter. Dersom det ikke finnes egnede, enkle analytiske uttrykk for fordelingsfunksjonene, eller fordelingene er kompliserte (f.eks. multimodale), kan det være vanskelig å komme frem til gode parameterestimater (manglende identifiserbarhet, kompliserte beregninger), selv om formen i

og for seg er kjent.

I ikke-parametriske metoder gjøres det ingen antakelse om formen på fordelingene. Her estimeres tettheten direkte fra treningsdataene for hvert punkt i egenskapsrommet der det er behov for å kjenne tettheten. I praksis vil dette være punkter som representerer egenskapsvektorene til ukjente objekter (objekter som skal klassifiseres). Fordelen med ikke-parametriske metoder er nettopp at de er fordelingsfrie, dvs. at de kan brukes på vilkårlig kompliserte fordelinger. Her kan det også nevnes at de ikke-parametriske beslutningsreglene *nærmeste-nabo regelen* og *k-nærmeste-nabo regelen* har en øvre grense for den asymptotiske feilraten, noe som gjør disse metodene godt egnet til feilrateestimering. Ulempen er at ikke-parametriske i utgangspunktet krever lagring av hele treningssettet, og gjennom søking av hele settet ved hver klassifisering.

## Oppgave 2

### Beslutningsteori

a) *Klassebetinget sannsynlighetstetthet* er en fordelingsfunksjon for egenskapsvektoren  $\mathbf{x}$  for en gitt klasse  $\omega_i$ , og kan skrives som  $p(\mathbf{x}|\omega_i)$ .

*A priori sannsynlighet* er sannsynligheten for at en gitt klasse skal opptre (hyppigheten den opptrer med), og kan skrives som  $P(\omega_i)$  for klassen  $\omega_i$ . Denne sannsynligheten tenkes gitt på forhånd (a priori), før det er utført måling på objektet som skal klassifiseres.

*A posteriori sannsynlighet* er sannsynligheten for at en gitt klasse skal opptre *etter* at det er utført målinger på objektet (funnet en egenskapsvektor), og kan uttrykkes ved *Bayes regel* (Bayes formel)

$$P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{p(\mathbf{x})}, i = 1, \dots, c$$

der nevneren blir

$$p(\mathbf{x}) = \sum_{j=1}^c p(\mathbf{x}|\omega_j)P(\omega_j)$$

dvs. den totale sannsynlighetstettheten i punktet  $\mathbf{x}$  uavhengig av klasse. Her er  $c$  antall klasser i problemet.

b) Minimum-feilrateprinsippet er en strategi for valg av klasse, som sikrer at sannsynligheten for å feilklassifisere objekter blir minimalisert. Det kan vises at minimum feilrate oppnås ved å velge klassen med størst a posteriori sannsynlighet. Beslutningsregelen kan da formuleres som

$$\text{Velg } \omega_i \text{ hvis } P(\omega_i|\mathbf{x}) \geq P(\omega_j|\mathbf{x}), j = 1, \dots, c$$

der  $c$  er antall klasser i problemet.

c) I dette klassifiseringsproblem med de to klassene  $\omega_1$  og  $\omega_2$ , er fordelingsfunksjonene gitt ved univariate (endimensjonale) normalfordelinger

$$p(x|\omega_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(x - \mu_i)^2}{2\sigma_i^2}\right], \quad i = 1, 2$$

Her er  $x$  den målte egenskapen til et vilkårlig objekt. Parametrene til disse fordelingene er henholdsvis  $\mu_i$  (forventningsverdien) og  $\sigma_i$  (standardavviket) for klasse  $\omega_i$ ,  $i = 1, 2$ .

d) Her skal det utledes en minimum feilrate beslutningsregel for problemet i deloppgave c, der klassene har a priori sannsynligheter  $P(\omega_1)$  og  $P(\omega_2)$ . En mulighet er å sette opp diskriminantfunksjoner  $g_1(x)$  og  $g_2(x)$  for de to klassene i problemet, som har som utgangspunkt at klassen med størst a posteriori sannsynlighet skal velges. Et mulig valg for diskriminantfunksjonene er da

$$g_i(x) = \ln P(\omega_i|x) = \ln p(x|\omega_i) + \ln P(\omega_i) - \ln p(x), i = 1, 2$$

som ved innsetting gir

$$g_i(x) = -\frac{1}{2} \left( \frac{x - \mu_i}{\sigma_i} \right)^2 - \frac{1}{2} \ln(2\pi) - \ln \sigma_i + \ln P(\omega_i) - \ln p(x)$$

Etter fjerning av ledd som er like for begge klasser, forenkles dette til

$$g_i(x) = -\frac{1}{2\sigma_i^2}x^2 + \frac{\mu_i}{\sigma_i^2}x - \frac{\mu_i^2}{2\sigma_i^2} - \ln \sigma_i + \ln P(\omega_i), i = 1, 2.$$

Siden det her er bare to klasser, kan diskriminantfunksjonene slås sammen til én funksjon

$$\begin{aligned} g(x) &= g_1(x) - g_2(x) \\ &= \frac{1}{2} \left( \frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2} \right) x^2 + \left( \frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2} \right) x + \frac{1}{2} \left( \frac{\mu_2^2}{\sigma_2^2} - \frac{\mu_1^2}{\sigma_1^2} \right) + \ln \left( \frac{\sigma_2 P(\omega_1)}{\sigma_1 P(\omega_2)} \right) \end{aligned}$$

Uttrykt ved diskriminantfunksjonen kan beslutningsregelen da skrives som

$$\underline{\text{Velg } \omega_1 \text{ hvis } g(x) > 0, \omega_2 \text{ ellers.}}$$

e) Med forventningsverdiene  $\mu_1 = 1$  og  $\mu_2 = 3$ , standardavvikene  $\sigma_1 = \sigma_2 = 1$  og a priorisannsynlighetene  $P(\omega_1) = 1/3$  og  $P(\omega_2) = 2/3$ , forenkles diskriminantfunksjonen til

$$\begin{aligned} g(x) &= (\mu_1 - \mu_2)x + \frac{1}{2}(\mu_2^2 - \mu_1^2) - \ln(2) \\ &= -2x + 4 - \ln(2) \end{aligned}$$

Desisjongrensen finnes ved å løse likningen  $g(x) = 0$ , som da gir terskelen

$$\underline{\underline{x_0 = \frac{1}{2}(4 - \ln(2)) \approx 1.6534.}}$$

Beslutningsregelen vil her bestå i å velge  $\omega_1$  for  $x > x_0$  og  $\omega_2$  ellers.

f) For objektet med egenskapsverdien  $x = 2$  blir diskriminantfunksjonen

$$g(x) = g(2) = -2 \cdot 2 + 4 - \ln(2) < 0.$$

Objektet blir da klassifisert til  $\omega_2$  med beslutningsregelen fra deloppgave e.

### Oppgave 3

#### Parametriske metoder

a) I parametriske metoder antas de klassebetingede tetthetsfunksjonene å ha kjent form, dvs. tetthetsfunksjonene kan uttrykkes på formen  $p(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i)$ ,  $i = 1, \dots, c$  der  $\boldsymbol{\theta}_i$  er den såkalte *parametervektoren* til fordelingen for klasse  $\omega_i$ . Fordelingen antas altså å være en *kjent* funksjon av  $\mathbf{x}$  og  $\boldsymbol{\theta}_i$ , f.eks. en multivariat normalfordeling, men parametervektoren må være kjent for at den aktuelle fordelingen skal være fullstendig bestemt. Parametervektoren estimeres ved hjelp av treningssettet. To hovedmetoder for parameterestimering kan nevnes.

I *maksimum-likelihoodmetoden* antas parametervektoren  $\boldsymbol{\theta}$  å ha en fast, men ukjent verdi. Den simultane sannsynlighetstettheten for de observerte samplene i  $\mathcal{X}$  skal maksimaliseres med hensyn på  $\boldsymbol{\theta}$ . Estimaten  $\hat{\boldsymbol{\theta}}$  benyttes videre som erstatning for den sanne, men ukjente parametervektoren.

I *Bayesisk estimering* betraktes parametervektoren  $\boldsymbol{\theta}$  som en stokastisk variabel med  $p(\boldsymbol{\theta})$  som tilhørende fordelingsfunksjon (a priori parameterfordeling). Treningssettet brukes til å oppdatere denne a priorifordelingen til en a posteriori parameterfordeling. Tetthetsestimaten finnes deretter ved å integrere tetthetsfunksjonen veiet med aposteriorifordelingen over alle mulige verdier av  $\boldsymbol{\theta}$ .

Bayesisk estimering er mer komplisert og regnekrevende enn maksimum-likelihoodmetoden, men kan være et gunstig alternativ hvis treningssettet er lite og man har god a priori kunnskap om problemet, som det med denne metoden er mulig å dra nytte av.

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

$$p(\mathcal{X}|\boldsymbol{\theta}) = p(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n|\boldsymbol{\theta}) = \prod_{k=1}^n p(\mathbf{x}_k|\boldsymbol{\theta}) \quad (1)$$

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*

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

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

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

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

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

Faktoriseringen av likelihoodfunksjonen i likning (1) krever at samplene i treningssettet er innbyrdes uavhengige.

c) Her er oppgaven å finne maksimum-likelihoodestimatet av parameteren  $\theta$  i den univariate fordelingen gitt ved

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

der  $x \geq 0$  og  $\theta > 0$ , og treningssettet er  $\mathcal{X} = \{x_1, \dots, x_n\}$ . Med denne fordelingen blir likelihoodfunksjonen

$$P(\mathcal{X}|\theta) = \prod_{k=1}^n p(x_k|\theta) = \prod_{k=1}^n [\theta^2 x_k e^{-\theta x_k}].$$

Log-likelihoodfunksjonen blir da

$$\mathcal{L}(\theta) = \sum_{k=1}^n \ln [\theta^2 x_k e^{-\theta x_k}] = \sum_{k=1}^n [2 \ln \theta + \ln x_k - \theta x_k].$$

Maksimum-likelihoodestimatet av  $\theta$  finnes ved å sette den deriverte av log-likelihoodfunksjonen til null, dvs.

$$\frac{d\mathcal{L}(\theta)}{d\theta} = \sum_{k=1}^n \left[ \frac{2}{\theta} - x_k \right] = 0,$$

som medfører at

$$\frac{2n}{\theta} = \sum_{k=1}^n x_k.$$

Estimatet av  $\theta$  blir da

$$\hat{\theta} = \frac{2n}{\sum_{k=1}^n x_k}.$$

## Oppgave 4

### Diskriminantfunksjoner

a) Diskriminantfunksjoner er et sett av funksjoner  $g_i(\mathbf{x})$ ,  $i = 1, \dots, c$ , av en input egenskapsvektor  $\mathbf{x}$ . Vi kan definere en slik funksjon for hver av de  $c$  klassene i klassifiseringsproblemet. Funksjonsverdien er et mål på tiltro til klassen, og vi definerer diskriminantfunksjonene slik at vi skal velge klassen med størst funksjonsverdi for en gitt egenskapsvektor  $\mathbf{x}$ . Beslutningsregelen kan da skrives som:

Velg  $\omega_m$  hvis  $g_m(\mathbf{x}) = \max_i \{g_i(\mathbf{x})\}$ .

Her har vi altså like mange diskriminantfunksjoner, som det er klasser i problemet.

b) En lineære diskriminantfunksjon kan skrives om til utvidet form på følgende måte:

$$g(\mathbf{x}) = \mathbf{w}^t \mathbf{x} + w_0 = w_0 + \sum_{i=1}^d w_i x_i = [w_0, w_1, \dots, w_d] \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{bmatrix} = \mathbf{a}^t \mathbf{y}$$

der  $\mathbf{a}$  = utvidet vektvektor og  $\mathbf{y}$  = utvidet egenskapsvektor. Dette er en avbildning fra dimensjon  $d$  til dimensjon  $\hat{d} = d + 1$ . Alle sampler fra  $x$ -rommet vil ligge i et  $d$ -dimensjonalt underrom i det nye  $y$ -rommet som er av dimensjon  $d + 1$ .

c) I minste kvadraters metode er målet å finne en vektvektor  $\mathbf{a}$  som i størst mulig grad tilfredsstiller likningssystemet

$$\mathbf{a}^t \mathbf{y}_i = b_i \quad \text{der } b_i > 0, \quad i = 1, \dots, n,$$

med håp om at de fleste produkter da vil bli positive, dvs. riktig klassifisert av vektvektoren  $\mathbf{a}$ . Her er det forutsatt av fortegnet på utvidede egenskapsvektorer fra  $\omega_2$  er snudd (fortegnskonvensjonen).

Definerer en datamatrise  $Y$  og en vektor av marginverdier  $\mathbf{b}$ :

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

Hver linje i  $Y$  inneholder en utvidet egenskapsvektor. Ønsket om å finne en vektvektor som gjør alle produkter med egenskapsvektorer positive, leder da til følgende likningssystem

$$Y\mathbf{a} = \mathbf{b} \quad \text{som skal løses med hensyn på } \mathbf{a}.$$

Siden dette likningssystemet er overbestemt (siden vi normalt vil ha langt flere likninger enn ukjente) og ingen eksakt løsning eksisterer, søkes i stedet en minste kvadraters løsning der lengden av feilvektoren  $\mathbf{e} = Y\mathbf{a} - \mathbf{b}$  er så liten som mulig.

d) Med utgangspunkt i deloppgave c kan vi sett opp kriteriefunksjonen

$$J_s(\mathbf{a}) = \|\mathbf{e}\|^2 = \|Y\mathbf{a} - \mathbf{b}\|^2 = \sum_{i=1}^n (\mathbf{a}^t \mathbf{y}_i - b_i)^2,$$

som skal minimaliseres mht.  $\mathbf{a}$ . Dette vil lede til den ønskede minste kvadratiske løsningen for den utvidede vektvektoren.

Den pseudoinverse løsningsmetoden består i å sette gradienten til kriteriefunksjonen lik null, siden dette er en nødvendig betingelse for minimum, og løse det likningssystemet man da får direkte for  $\mathbf{a}$ :

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

som gir

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

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

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

der

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

er den *pseudoinverse* til  $Y$ .

## Oppgave 5

### Ikke-parametriske metoder

a) Fordeler ved ikke-parametriske metoder er at det ikke kreves noen antakelse om formen på fordelingene, men at metodene tvert imot kan tilpasse seg vilkårlige, f.eks. multimodale fordelinger. Ulemper er at hele treningssettet i utgangspunktet må lagres og gjennomføres ved hver klassifisering av et sample.

Tetthetsestimatet i et punkt  $\mathbf{x}$  kan skrives som

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

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

b) Sannsynlighetstettheten i et vilkårlig punkt i  $\mathbf{x}$  i egenskapsrommet kan estimeres ved hjelp av estimatoren i deloppgave a. For å klassifisere et ukjent objekt representert ved  $\mathbf{x}$  beregnes tetthetsestimatet for hver av klassene i det samme punktet. Bayes formel kan deretter brukes til å beregne a posteriori classesannsynligheter ved å kombinere tetthetsestimatene med a priori-sannsynlighetene for hver av klassene.

To hovedtyper av metoder for å utføre selve tetthetsestimeringen i foregående deloppgave er hhv. vindumetoden og nærmeste-nabo metoden. Forskjellen på disse teknikkene er at i vindumetoden holdes volumet til regionen omkring punktet  $\mathbf{x}$  konstant, mens antall sampler innenfor regionen holdes konstant i nærmeste-nabo metoden.

c) I dette todimensjonalt klassifiseringsproblemet med de tre klasser  $\omega_1$ ,  $\omega_2$  og  $\omega_3$  består treningssettet  $\mathcal{X} = \mathcal{X}_1 + \mathcal{X}_2 + \mathcal{X}_3$  av henholdsvis 5, 4 og 5 sampler (se figur 1).

A priorisannsynlighetene for klassene i et vilkårlig problem kan estimeres ut fra antall treningssampler i hver klasse ved uttrykket

$$\hat{P}(\omega_i) = \frac{n_i}{n}$$

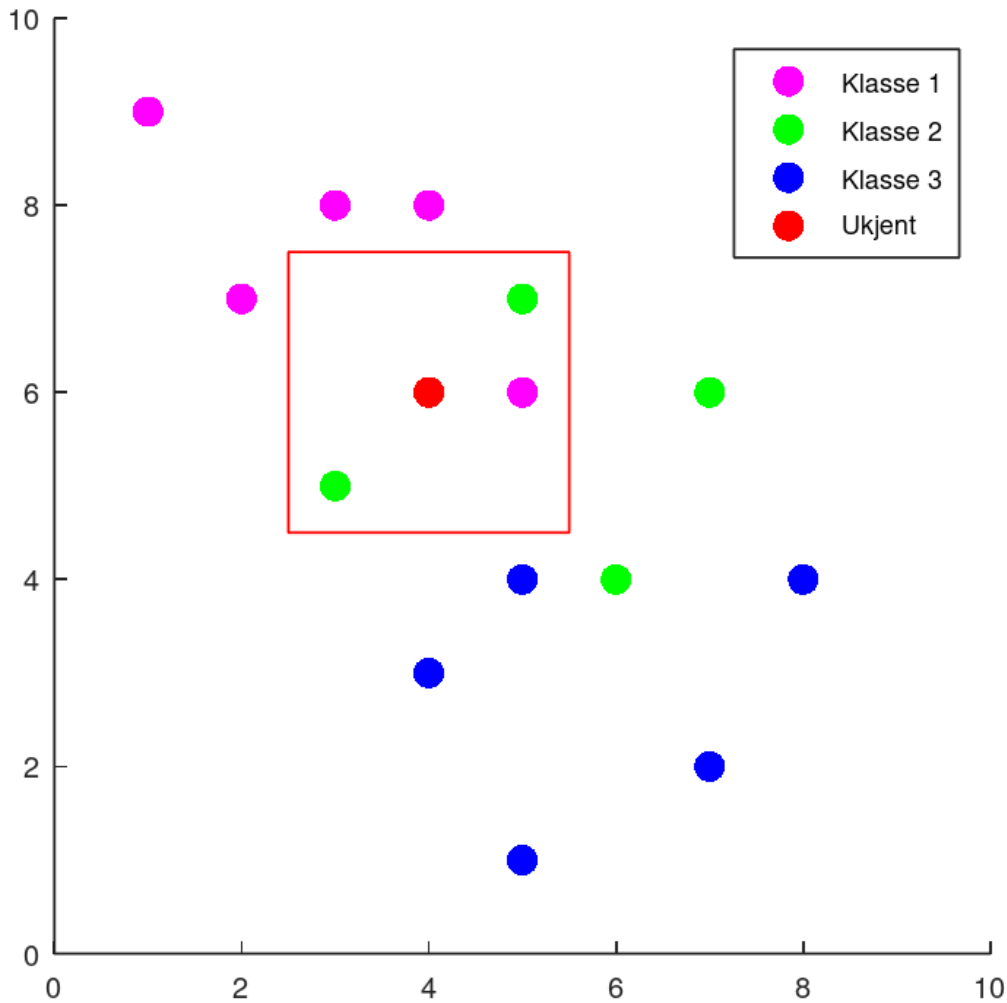
der  $n_i$  er antall sampler i klasse  $\omega_i$  og  $n$  er totalt antall sampler i treningssettet. For problemet i denne deloppgaven blir estimatene hhv.  $\hat{P}(\omega_1) = 5/14$ ,  $\hat{P}(\omega_2) = 4/14$  og  $\hat{P}(\omega_3) = 5/14$ .

Vindumetoden med hyperkubisk vindufunksjon (i dette tilfellet et kvadrat, som vist i figuren, med side  $h = 3$ ) skal her brukes til å klassifisere et ukjent objekt i punktet  $[4, 6]^t$  i egenskapsrommet.

Av figuren ser vi at regionen inneholder ett sample fra  $\omega_1$ , to sampler fra  $\omega_2$  og ingen sampler fra  $\omega_3$ .

A posteriori sannsynlighet for klasse  $\omega_i$  i punktet  $\mathbf{x}$  estimeres ved Bayes formel:

$$\hat{P}(\omega_i|\mathbf{x}) = \frac{p_n(\mathbf{x}|\omega_i)\hat{P}(\omega_i)}{\sum_{j=1}^c p_n(\mathbf{x}|\omega_j)\hat{P}(\omega_j)} = \frac{\frac{k_i/n_i}{V_n} \cdot \frac{n_i}{n}}{\sum_{j=1}^c \frac{k_j/n_j}{V_n} \cdot \frac{n_j}{n}} = \alpha \cdot \frac{k_i/n_i}{V_n} \cdot \frac{n_i}{n}$$



Figur 1: Plott av treningssamplene, det ukjente objektet og den omsluttende vindufunksjonen (denne figuren kreves ikke i besvarelsen).

for et problem med  $c$  klasser. Her er  $k_i$  antall sampler fra  $\omega_i$  innenfor regionen (kvadratet i dette tilfellet) med volum  $V_n = h^2 = 9$ , og  $\alpha$  er den inverse av nevneren, som er felles for alle klasser. Estimaten blir da:

$$\hat{P}(\omega_1|\mathbf{x}) = \alpha \cdot \frac{1}{9 \cdot 5} \cdot \frac{5}{14} = \frac{\alpha}{9 \cdot 14}$$

$$\hat{P}(\omega_2|\mathbf{x}) = \alpha \cdot \frac{2}{9 \cdot 4} \cdot \frac{4}{14} = \frac{2 \cdot \alpha}{9 \cdot 14}$$

$$\hat{P}(\omega_3|\mathbf{x}) = \alpha \cdot \frac{0}{9 \cdot 5} \cdot \frac{5}{14} = 0$$

Vi ser at a posteriorisannsynligheten er størst, slik at objektet klassifiseres til  $\omega_2$ .



**Oppgave 6 (bare TEK9020)**

Ikke-ledet læring

a) Ikke-ledet læring går ut på å trene en klassifikator ved hjelp av et treningssett med *umerkede* sampler, dvs. klassetilhørigheten til treningssamplene er ukjent. En blandingstetthet er en vektet sum av de klassebetingede tetthetsfunksjonene, der vektene er klassenes a priori sannsynligheter.

b) Blandingstettheten for et problem med to klasser kan skrives som

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{i=1}^2 p(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i)P(\omega_i) = p(\mathbf{x}|\omega_1, \boldsymbol{\theta}_1)P(\omega_1) + p(\mathbf{x}|\omega_2, \boldsymbol{\theta}_2)P(\omega_2).$$

Her er  $p(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i)$ ,  $i = 1, 2$  komponenttetthetene, mens  $P(\omega_i)$  er blandingsparametrene.

c) Se utledningen av likningssystemet for  $c$  klasser i avsnitt 7.1 (side 97) i forelesningsnotatene.

d) Innsetting av normalfordelingene, som i dette tilfellet blir

$$p(x_k|\omega_i, \mu_i) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x_k - \mu_i)^2\right], \quad i = 1, 2$$

i likningssystemet

$$\sum_{k=1}^n P(\omega_i|x_k, \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}_i} \ln p(x_k|\omega_i, \boldsymbol{\theta}_i) = 0, \quad i = 1, 2,$$

gir da

$$\sum_{k=1}^n P(\omega_i|x_k, \mu_1, \mu_2) \frac{d}{d\mu_i} \left[ -\frac{1}{2}(x_k - \mu_i)^2 - \ln \sqrt{2\pi} \right] = 0, \quad i = 1, 2,$$

som gir

$$\sum_{k=1}^n P(\omega_i|x_k, \mu_1, \mu_2)(x_k - \mu_i) = 0, \quad i = 1, 2.$$

Forventningsestimatene kan da uttrykkes ved

$$\hat{\mu}_i = \frac{\sum_{k=1}^n P(\omega_i|x_k, \hat{\mu}_1, \hat{\mu}_2)x_k}{\sum_{k=1}^n P(\omega_i|x_k, \hat{\mu}_1, \hat{\mu}_2)}, \quad i = 1, 2.$$

der

$$\begin{aligned} P(\omega_i|x_k, \hat{\mu}_1, \hat{\mu}_2) &= \frac{p(x_k|\omega_i, \hat{\mu}_i)P(\omega_i)}{\sum_{j=1}^2 p(x_k|\omega_j, \hat{\mu}_j)P(\omega_j)} \\ &= \frac{\exp\{-\frac{1}{2}(x_k - \hat{\mu}_i)^2\}}{\exp\{-\frac{1}{2}(x_k - \hat{\mu}_1)^2\} + \exp\{-\frac{1}{2}(x_k - \hat{\mu}_2)^2\}}, \quad i = 1, 2, \end{aligned}$$

etter innsetting av normalfordelingene og forkorting av like faktorer over og under brøkstreken. Her er  $x_{k,k=1\dots n}$  de umerkede samplene i treningssettet.

Dette er implisitte uttrykk som kan løses ved iterasjon, som beskrevet på side 98 i forelesningsnotatene. Man velger startverdier for  $\mu_1$  og  $\mu_2$  og oppdaterer deretter aposteriorisannsynlighetene og forventningsestimatene rekursivt.