

Løsningsforslag eksamen TEK5020/9020 - 2021H

Oppgave 1

Innledning

a) *Klassebetinget sannsynlighetstetthet* er fordelingsfunksjonen for egenskapsvektoren \mathbf{x} for en gitt klasse ω_i , og kan skrives som $p(\mathbf{x}|\omega_i)$. *A priori sannsynlighet* er sannsynligheten for at en gitt klasse skal opptre (hyppigheten den optrer med), og kan skrives som $P(\omega_i)$ for klassen ω_i . Denne sannsynligheten tenkes gitt på forhånd (a priori) uten at 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. c er antall klasser i problemet.

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

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

c) En *klassifikator* er en maskin (eller algoritme) som foretar valg av klasse basert på en beslutningsregel man har kommet frem til, f.eks. ved maskinlæring. Inngangsdata til klassifikatoren vil typisk være en egenskapsvektor, dvs. et sett av numeriske størrelser som er målt på objektene som skal klassifiseres.

d) Lag en skisse i likhet med figur 3 på side 7 i forelesningsnotatene, som viser trinnene i et klassifiseringssystem, fra rådata, via egenskaper, til valg av klasse.

Oppgave 2

Beslutningsteori

a) I den multivariate normalfordelingen

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

er i denne sammenhengen \mathbf{x} en egenskapsvektor (et punkt i egenskapsrommet), d er dimensjonen til egenskapsrommet, $\boldsymbol{\mu}$ er forventningsvektoren til fordelingen (vektor av dimensjon d) og Σ er kovariansmatrisen (matrise av dimensjon $d \times d$). Parametrene som bestemmer denne fordelingen er $\boldsymbol{\mu}$ og Σ .

b) I henhold til minimum-feilrateprinsippet skal (i dette tilfellet) egenskapsvektoren \mathbf{x}_0 klassifiseres til klassen med størst a posteriori sannsynlighet. Siden klassenes a priori sannsynligheter her er like, svarer dette til å klassifisere til klassen med størst klassebetinget tetthet $p(\mathbf{x}|\omega_i)$. I dette tilfellet der klassene er multivariat normalfordelte, oppnås dette ved å velge klassen med minst Mahalanobisavstand

$$r_i^2 = (\mathbf{x}_0 - \boldsymbol{\mu}_i)^t \Sigma^{-1} (\mathbf{x}_0 - \boldsymbol{\mu}_i)$$

fra punktet \mathbf{x}_0 . For kovariansmatrisen

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

blir den inverse

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

som kan finnes ved å løse likningssystemet $\Sigma \Sigma^{-1} = I$, f.eks. ved bruk av Cramers regel. Ved å sette forventningsvektorene

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{\mu}_2 = \begin{bmatrix} 4 \\ 2 \end{bmatrix} \text{ og } \boldsymbol{\mu}_3 = \begin{bmatrix} -1 \\ 3 \end{bmatrix}.$$

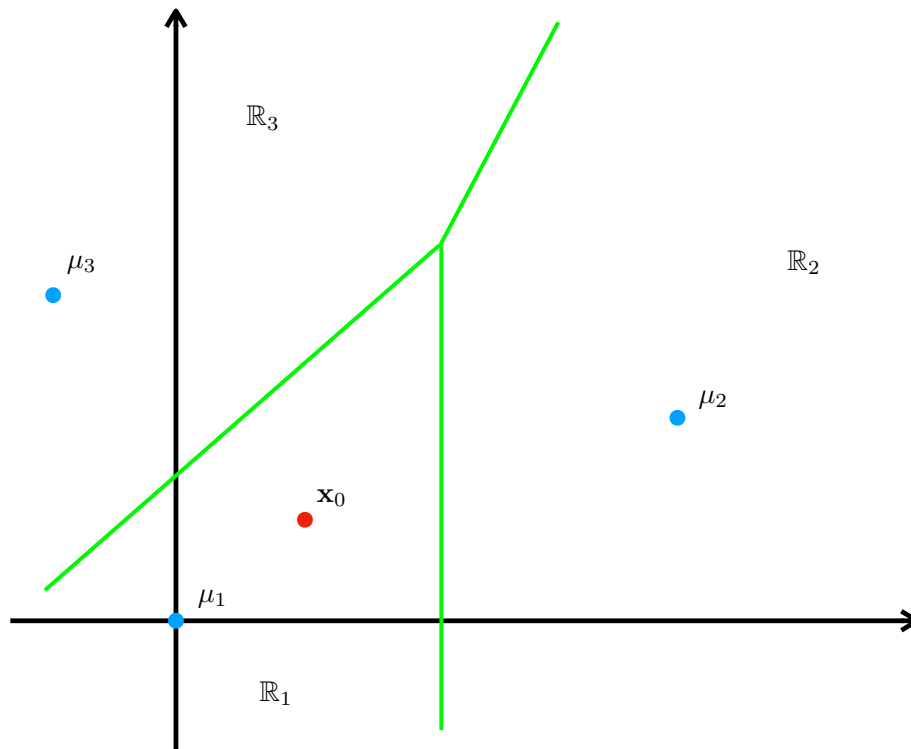
for de tre klassene inn i uttrykket for Mahalanobisavstanden, blir avstanden fra egenskapsvektoren $\mathbf{x}_0 = [1, 1]^t$:

$$r_1^2 = \frac{3}{5} = 0.6, \quad r_2^2 = \frac{23}{5} = 4.6 \quad \text{og} \quad r_3^2 = \frac{28}{5} = 5.6.$$

Siden r_1^2 er minst, blir \mathbf{x}_0 klassifisert til ω_1 . Denne deloppgaven kan også løses ved hjelp av diskriminantfunksjoner.

c) Figuren på neste side viser forventningsvektorene (blå sirkler) for hver klasse og punktet \mathbf{x}_0 (rød sirkel) i det todimensjonale egenskapsrommet.

d) Siden kovariansmatrisene er like, vil beslutningsgrensene mellom klassene være lineære (generelt hyperplan, som forklart i avsnitt 2.9.2 i forelesningsnotatene), og i dette todimensjonale tilfelle rette linjestykker. Beslutningsgrensene (i grønt) er skissert i figuren (vi ser at \mathbf{x}_0 ligger i desisjonsregionen \mathcal{R}_1 for klasse ω_1).



Oppgave 3

Parametriske metoder

a) Maksimum-likelihoodmetoden for estimering av parametervektoren θ i fordelingsfunksjon $p(\mathbf{x}|\theta)$, består i å finne den verdien av parametervektoren som maksimaliserer den simultane sannsynlighetstettheten for de observerte treningssamplene fra den aktuelle klassen (vi ser her på ledet læring). Dette svarer til å maksimalisere den såkalte likelihoodfunksjonen mhp. θ , som kan oppnås ved å sette gradienten til likelihoodfunksjonen lik null. Se avsnitt 3.1 på side 34 i forelesningsnotatene.

b) Likelihoodfunksjonen kan skrives som

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

og skal maksimaliseres med hensyn til den ukjente parametervektoren θ . 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 blir denne *log-likelihoodfunksjonen*

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

maksimum for samme verdi av θ . Maksimum av \mathcal{L} finnes ved å ta gradienten til log-likelihoodfunksjonen mht. θ :

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

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

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

Treningssamplene må forutsettes å være innbyrdes uavhengige for at faktoriseringen av likelihoodfunksjonen skal være mulig (se avsnitt 3.1 på side 34 i forelesningsnotatene).

c) Oppgaven består her i å finne maksimum-likelihood estimatet av parameteren θ i den univariate fordelingen gitt ved

$$p(x|\theta) = \frac{1}{6} \theta^4 x^3 e^{-\theta x},$$

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

Med fordelingen $p(x|\theta)$ i oppgaveteksten blir likelihoodfunksjonen

$$P(\mathcal{X}|\theta) = \prod_{k=1}^n \left[\frac{1}{6} \theta^4 x_k^3 e^{-\theta x_k} \right].$$

Log-likelihoodfunksjonen blir da

$$\mathcal{L}(\theta) = \sum_{k=1}^n \ln \left[\frac{1}{6} \theta^4 x_k^3 e^{-\theta x_k} \right] = \sum_{k=1}^n [4 \ln \theta - \ln(6) + 3 \ln x_k - \theta x_k].$$

Maksimum-likelihood estimatet av θ finnes ved å sette den deriverte av log-likelihoodfunksjonen til null, dvs.

$$\frac{d\mathcal{L}(\theta)}{d\theta} = \sum_{k=1}^n \left[\frac{4}{\theta} - x_k \right] = 0,$$

som medfører at

$$\frac{4n}{\theta} = \sum_{k=1}^n x_k.$$

Estimatet blir da

$$\hat{\theta} = \frac{4n}{\sum_{k=1}^n x_k}.$$

Oppgave 4

Lineære diskriminantfunksjoner

a) 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

$$d \rightarrow \hat{d} = d + 1$$

dimensjoner. Alle sampler fra x -rommet vil ligge i et d -dimensjonalt underrom i det nye y -rommet som er av dimensjon $d + 1$.

b) Se avsnittene 5.3.1, 5.3.2 og 5.3.3 på side 70-72 i forelesningsnotatene.

c) Se avsnitt 5.4.3 på side 74 i forelesningsnotatene. Forutsetningen for at denne algoritmen skal lede til en løsningsvektor etter et endelig antall iterasjoner er at treningssettet er lineært separabelt.

d) Treningssettet består her av fire univariate sampler fra to klasser ω_1 og ω_2 :

$$\mathcal{X} = \underbrace{\{1, 2\}}_{\omega_1}, \underbrace{\{6, 7\}}_{\omega_2}.$$

I det utvidede egenskapsrommet kan treningssettet skrives som

$$\mathcal{Y} = \left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} -1 \\ -6 \end{bmatrix}, \begin{bmatrix} -1 \\ -7 \end{bmatrix} \right\}.$$

$\mathbf{y}_1 \quad \mathbf{y}_2 \quad \mathbf{y}_3 \quad \mathbf{y}_4$

Ved å bruke fast-inkrementregelen til å finne en løsningsvektor, blir iterasjonsprosessen med utgangspunkt i startvektoren $\mathbf{a}_0 = [0, 0]^t$ som følger:

$$\begin{aligned} \mathbf{a}_0 &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} & \Rightarrow \mathbf{a}_0^t \mathbf{y}_1 = 0 & \Rightarrow \text{feil} \\ \mathbf{a}_1 &= \mathbf{a}_0 + \mathbf{y}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} & \Rightarrow \mathbf{a}_1^t \mathbf{y}_3 < 0 & \Rightarrow \text{feil} \\ \mathbf{a}_2 &= \mathbf{a}_1 + \mathbf{y}_3 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} -1 \\ -6 \end{bmatrix} = \begin{bmatrix} 0 \\ -5 \end{bmatrix} & \Rightarrow \mathbf{a}_2^t \mathbf{y}_1 < 0 & \Rightarrow \text{feil} \\ \mathbf{a}_3 &= \mathbf{a}_2 + \mathbf{y}_1 = \begin{bmatrix} 0 \\ -5 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ -4 \end{bmatrix} & \Rightarrow \mathbf{a}_3^t \mathbf{y}_2 < 0 & \Rightarrow \text{feil} \\ \mathbf{a}_4 &= \mathbf{a}_3 + \mathbf{y}_2 = \begin{bmatrix} 1 \\ -4 \end{bmatrix} + \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \\ -2 \end{bmatrix} & \Rightarrow \mathbf{a}_4^t \mathbf{y}_1 = 0 & \Rightarrow \text{feil} \\ \mathbf{a}_5 &= \mathbf{a}_4 + \mathbf{y}_1 = \begin{bmatrix} 2 \\ -2 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \underline{\underline{\begin{bmatrix} 3 \\ -1 \end{bmatrix}}} \end{aligned}$$

Alle produkter av \mathbf{a}_5 med $\mathbf{y}_i, i = 1, \dots, 4$ er positive, slik at \mathbf{a}_5 er en løsningsvektor.

e) Terskelen mellom klassene i det opprinnelige egenskapsrommet kan finnes ved å løse likningen $\mathbf{a}_5^t \mathbf{y}_0 = 0$ med hensyn på \mathbf{y}_0 , dvs.

$$[3, -1] \begin{bmatrix} 1 \\ x_0 \end{bmatrix} = 0,$$

som gir $3 - x_0 = 0$. Terskelen blir da $x_0 = 3$.

Oppgave 5

Ikke-parametriske metoder

a) Ikke-parametriske metoder skiller seg fra parametriske metoder ved at det ikke gjøres noen antagelser om formen på tetthetsfunksjonene. I stedet brukes treningssamplene direkte til å foreta et punkttestimat av tettheten i et gitt punkt i egenskapsrommet ved hjelp av de omkringliggende treningssamplene. Tetthetsestimatet i et punkt \mathbf{x} basert på et treningssett av egenskapsvektorer $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ (fra den aktuelle klassen) kan skrives som

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

der V_n er volumet av en vilkårlig region omkring punktet \mathbf{x} , k_n er antall treningssamplene innenfor regionen og n er det totale antall samplene i treningssettet (husk at vi her ser på klassene enkeltvis). Se figur 32 på side 51 i forelesningsnotatene.

b) Se avsnitt 4.2 på side 56-57 i forelesningsnotatene. Utledningen av estimatet for a posteriori sannsynlighet er vist i likning 24 på side 57.

c) Resultatet i foregående deloppgave viser at klassen med den største estimerte a posteriori sannsynlighet (den klassen som skal velges i henhold til minimum-feilrateprinsippet), er den samme som klassen med flest representanter innenfor regionen. Dette leder til *k-nærmeste-naboregelen*, som kan formuleres som følger:

Velg klassen med flest representanter blant de k nærmeste naboene til egenskapsvektoren som skal klassifiseres.

d) Ved å sette $k = 1$ får vi spesialtilfellet *nærmeste-naboregelen* som består i å *velge samme klasse som den nærmeste naboen til egenskapsvektoren som skal klassifiseres*. For sammenhengen mellom den asymptotiske feilraten P og den optimale feilraten P^* gjelder følgende:

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

Her er det tilstrekkelig å oppgi det forenklete resultatet $P \leq 2P^*$, dvs. at *den asymptotiske feilraten alltid er mindre enn to ganger den optimale feilraten*. Se forøvrig figur 40 på side 61 i forelesningsnotatene.

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* samplene, 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

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

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

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

d) Innsetting av normalfordelingene

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

i likningssystemet

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

gir da

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

som gir

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

Forventningsestimatene kan da uttrykkes ved

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

der

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

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 μ_1 og μ_2 og oppdaterer deretter aposteriorisannsynlighetene og forventningsestimatene rekursivt.