Løsningsforslag TEK5020 - 2019H

Oppgave 1

Innledning

a) Mulig svar: Egenskapsuttrekking består i å trekke ut informative tallstørrelser rådata; størrelser som inneholder mest mulig relevant informasjon for å kunne skille mellom klasser av objekter. Klassifisering består i å gjøre et valg av klasse, basert på denne komprimerte informasjonen. Lag en skisse i likhet med figur 8 i forelesningsnotatene.

b) Mulig svar: Ledet læring består i å trene opp en klassifikator ved hjelp av objekter (egenskapsvektorer) med *kjent* klassetilhørighet. Man har altså en fasit som kan brukes som en rettesnor i treningsprosessen. Ikke-ledet læring består i å trene klassifikatoren uten en slik fasit, dvs. klassetilhørigheten til objektene i treningssettet er *ukjente*. Årsaken kan være at det er for dyrt eller tidkrevende å merke treningsobjektene med riktig klassetilhørighet.

c) Mulig svar: En beslutningsregel er en oppskrift på hvilken klasse som skal velges for et gitt ukjent objekt, representert ved en uttrukket egenskapsvektor. Kort og godt beskriver beslutningsregelen hvordan klassifikatoren fungerer. Se avsnitt 1.2 i forelesningsnotatene og eksempelet på beslutningsregel midt på side 7.

d) Mulig svar: Beslutningsregler kan konstrueres ved hjelp av f.eks. parametriske- eller ikke-parametriske metoder for estimering av sannsynlighetstetthetsfunksjoner, eller optimalisering av kriteriefunksjoner for å finne vektene i diskriminantfunksjoner direkte fra objektene i treningssettet.

Oppgave 2

Bayesisk beslutningsteori

a) Sett opp Bayes regel som i avsnitt 2.1 i forelesningsnotatene, og forklar at dette uttrykket gir aposteriori sannsynlighet for hver klasse (sannsynligheten etter at egenskapsvektoren er målt), uttrykt ved den klassebetingede tetthetsfunksjonen (fordelingsfunksjonen) for samme klasse og klassens apriori sannsynlighet (sannsynligheten for at klassen skal opptre i utgangspunktet).

b) Mulig svar: Minimum-feilrate prinsippet består i å velge klassen med størst aposteriori sannsynlighet. Dette valget kan vises å sikre lavest mulig feilrate. Se avsnitt 2.3 i forelesningsnotatene.

c) Desisjonsgrensen (terskelverdien) er i dette tilfellet gitt ved den verdien av x som tilfredsstiller likningen

$$P(\omega_1|x) = P(\omega_2|x),$$

som ved innsetting i Bayes regel og fjerning av den felles nevneren $p(x)$ gir

$$p(x|\omega_1)P(\omega_1) = p(x|\omega_1)P(\omega_1).$$

Innsetting av den univariate normalfordelingen gir da

$$\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left[-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2\right]P(\omega_1) = \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left[-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2\right]P(\omega_2),$$

som med innsetting av $\sigma_1 = \sigma_2 = \sigma$, forkorting og bruk av den naturlige logaritmen på begge sider av likhetstegnet gir

$$-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2 + \ln P(\omega_1) = -\frac{1}{2}\left(\frac{x-\mu_2}{\sigma}\right)^2 + \ln P(\omega_2)$$

og derav

$$-x^2 + 2\mu_1 x - \mu_1^2 + 2\sigma^2 \ln P(\omega_1) = -x^2 + 2\mu_2 x - \mu_2^2 + 2\sigma^2 \ln P(\omega_2).$$

Ved innsetting av $\mu_1 = 0$ og $\mu_2 = 1$ forenkles dette til

$$-2x + 1 + 2\sigma^2 \ln\frac{P(\omega_1)}{P(\omega_2)} = 0.$$

Desisjonsgrensen er derved gitt ved

$$x_0 = \underline{\underline{\frac{1}{2} + \sigma^2 \ln\frac{P(\omega_1)}{P(\omega_2)}}}.$$

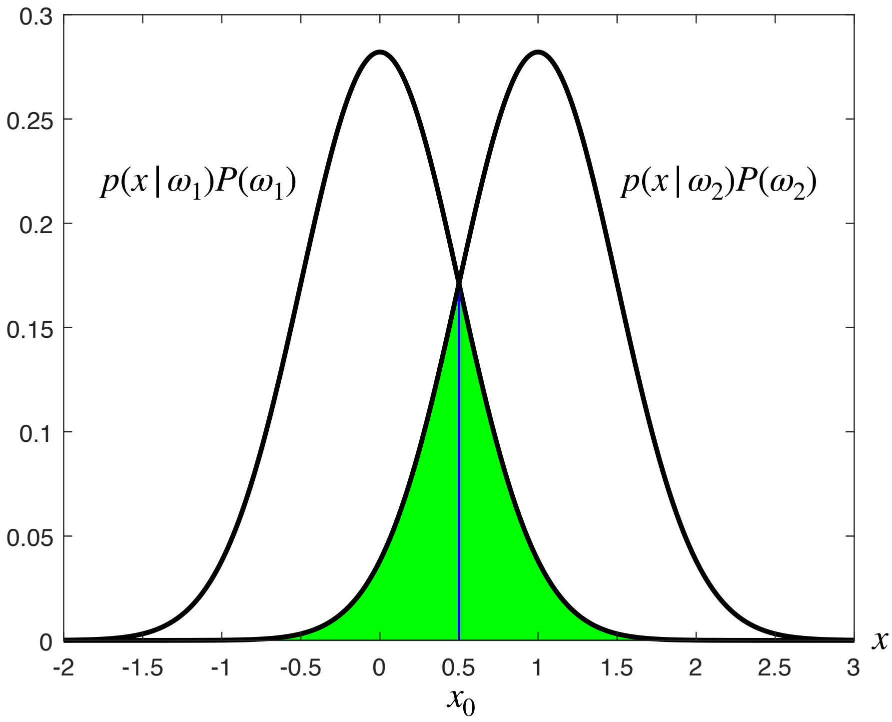
d) Med like apriori sannsynligheter blir terskelverdien

$$x_0 = \underline{\underline{1/2}},$$

siden $\ln(P(\omega_1)/P(\omega_2)) = 0$ i dette tilfellet. Figuren på neste side viser terskelen og feilraten, illustrert ved det grønne arealet (i denne figuren er $\sigma = 1$).

e) Med $P(\omega_1) = 1/3$ og $P(\omega_2) = 2/3$ blir terskelverdien

$$x_0 = \frac{1}{2} + \sigma^2 \ln\frac{P(\omega_1)}{P(\omega_2)} = \frac{1}{2} + \sigma^2 \ln\left(\frac{1/3}{2/3}\right) = \underline{\underline{1/2 - \sigma^2\ln(2)}} \approx 1/2 - 0.693\sigma^2.$$



Oppgave 3

Parametriske metoder

a) Maksimum likelihood (ML) metoden består i å finne den parametervektoren $\boldsymbol{\theta}$ i tetthetsfunksjon $p(\boldsymbol{x}|\boldsymbol{\theta})$ som maksimaliserer den simultane sannsynligheten for å observere nettopp det treningssettet man har til rådighet. Konkret skal man altså maksimalisere likelihoodfunksjonen som funksjon av parametervektoren (sett opp uttrykket). Deretter beregner man gradienten til log-likelihoodfunksjonen og setter den lik null (sett opp log-likelihood funksjonen og gradienten). Parameterestimatet kan deretter finnes ved å løse likningssystemet

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

med hensyn på $\boldsymbol{\theta}$. Dette er beskrevet i avsnitt 3.1 i forelesningsnotatene (side 34).

b) Eksponensialfordelingen er gitt ved

$$p(x|\theta) = \begin{cases} \theta\,\mathrm{e}^{-\theta x} & x \geq 0 \\ 0 & ellers \end{cases}$$

der $\theta > 0$. Denne tetthetsfunksjonen kan settes inn i likningssystemet fra punkt a. Alternativt kan løsningen finnes som følger, ved å starte med likelihoodfunksjonen

$$p(\mathscr{X}|\theta) = \prod_{k=1}^{n} \theta e^{-\theta x_k},$$

der treningssettet er gitt ved $\mathscr{X} = \{x_1, \ldots, x_n\}$. Log-likelihoodfunksjonen blir da

$$\mathscr{L}(\theta) = \sum_{k=1}^{n} (\ln\theta - \theta x_k).$$

Gradienten til $\mathscr{L}$ (i dette tilfellet den deriverte mht. θ) skal deretter settes til null for å finne maksimum, dvs

$$\sum_{k=1}^{n}\left(\frac{1}{\theta}-x_k\right)=0 \quad \Rightarrow \quad \hat{\theta}=\underline{\underline{\frac{n}{\sum_{k=1}^{n}x_k}}}.$$

Oppgave 4

Diskriminantfunksjoner

a) Mulig svar: Diskriminantfunksjoner er et sett av funksjoner, en for hver klasse, som uttrykker en grad av tiltro til at et objekt tilhører den aktuelle klassen; dess større funksjonsverdi, dess større tiltro. Ved klassifisering av objekter velges den klassen som har størst verdi på den tilhørende diskriminantfunksjon. Se avsnitt 2.5 i forelesningsnotatene.

b) Med fordelingen

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

kan man sette opp diskriminantfunksjoner på formen

$$g_i(\boldsymbol{x})=\ln p(\boldsymbol{x}|\omega_i)+\ln P(\omega_i),$$

som ved innsetting for tetthetsfunksjonen gir

$$\begin{aligned}g_i(\boldsymbol{x})&=-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^t\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)-\frac{d}{2}\ln 2\pi-\frac{1}{2}\ln|\Sigma_i|+\ln P(\omega_i)\\&=-\frac{1}{2}\boldsymbol{x}^t\Sigma_i^{-1}\boldsymbol{x}+\boldsymbol{x}^t\Sigma_i^{-1}\boldsymbol{\mu}_i-\frac{1}{2}\boldsymbol{\mu}_i^t\Sigma_i^{-1}\boldsymbol{\mu}_i-\frac{d}{2}\ln 2\pi-\frac{1}{2}\ln|\Sigma_i|+\ln P(\omega_i),\end{aligned}$$

og siden kovariansmatriser og apriorisannsynligheter er like for alle klasser

$$=-\frac{1}{2}\boldsymbol{x}^t\Sigma^{-1}\boldsymbol{x}+\boldsymbol{x}^t\Sigma^{-1}\boldsymbol{\mu}_i-\frac{1}{2}\boldsymbol{\mu}_i^t\Sigma^{-1}\boldsymbol{\mu}_i-\frac{d}{2}\ln 2\pi-\frac{1}{2}\ln|\Sigma|+\ln\frac{1}{3}.$$

Her kan ledd som er felles for alle klasser strykes, slik at diskriminantfunksjonene blir

$$g_i(\boldsymbol{x})=\boldsymbol{w}_i^t\boldsymbol{x}+w_{i0},$$

der

$$\boldsymbol{w}_i=\Sigma^{-1}\boldsymbol{\mu}_i$$

og

$$w_{i0}=-\frac{1}{2}\boldsymbol{\mu}_i^t\Sigma^{-1}\boldsymbol{\mu}_i$$

for hver klasse ω_i, $i=1,2,3$. Her er kovariansmatrisen for alle klasser, gitt ved

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

og den inverse, som kan finnes ved å løse et likningssystem, blir

$$\Sigma^{-1} = \frac{1}{10}\begin{bmatrix} 4 & -2 \\ -2 & 6 \end{bmatrix}.$$

Innsetting av Σ^{-1} og forventningsvektorene

$$\boldsymbol{\mu}_1 = \begin{bmatrix} -2 \\ 5 \end{bmatrix}, \quad \boldsymbol{\mu}_2 = \begin{bmatrix} 2 \\ 3 \end{bmatrix} \quad \text{og} \quad \boldsymbol{\mu}_3 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}$$

i uttrykket for diskriminantfunksjonen ovenfor, gir da

$$\begin{aligned} g_1(\boldsymbol{x}) &= [-1.8, 3.4]\boldsymbol{x} - 10.3 \\ g_2(\boldsymbol{x}) &= [0.2, 1.4]\boldsymbol{x} - 2.3 \\ g_3(\boldsymbol{x}) &= [1.6, -1.8]\boldsymbol{x} - 4.2. \end{aligned}$$

c) Diskriminantfunksjonene er lineære, siden $\boldsymbol{x}$ bare inngår til første orden. Desisjonsgrensene vil da bestå av hyperplan (rette linjestykker i dette tilfellet, siden $d = 2$). Siden lineære diskriminantfunksjoner gir opphav til like mange (enkeltsammenhengende) regioner som antall klasser, er svaret her *tre* regioner.

d) Ved å sette inn det ukjente objektet representert ved egenskapsvektoren

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

i diskriminantfunksjonene ovenfor, får man

$$\begin{aligned} g_1(\boldsymbol{x}_0) &= [-1.8, 3.4]\begin{bmatrix} 1 \\ -1 \end{bmatrix} - 10.3 = -15.5 \\ g_2(\boldsymbol{x}_0) &= [0.2, 1.4]\begin{bmatrix} 1 \\ -1 \end{bmatrix} - 2.3 = -3.5 \\ g_3(\boldsymbol{x}_0) &= [1.6, -1.8]\begin{bmatrix} 1 \\ -1 \end{bmatrix} - 4.2 = -0.8. \end{aligned}$$

Siden g_3 har størst verdi, blir objektet klassifisert til ω_3.

Oppgave 5

Ikke-parametriske metoder

a) Forklar prinsippet som beskrevet i avsnitt 4.1 i forelesningsnotatene (illustrer gjerne med en skisse tilsvarende figur 32) og sett opp formelen for $p_n(\boldsymbol{x})$ på side 50.

b) Mulig svar: Ved å beregne slike tetthetsestimater for hver klasse, kan disse kombineres med apriori sannsynligheter til å estimere aposteriori sannsynligheter for klassene ved hjelp av Bayes regel. Ved å bruke f.eks. minimum-feilrate prinsippet, kan man deretter velge den klassen som har størst aposteriori sannsynlighet. c) Se avsnitt 4.2. og 4.2.2 (side 57).

d) Ved å beregne avstanden mellom det ukjente objektet $[4,6]^t$ og hvert av egenskapsvektorene i treningssettet (eller ved å plotte vektorene inn i et rutenett) vil man se at $[5,6]^t$ fra ω_1 er det nærmeste objektet, mens $[5,7]^t$ og $[3,5]^t$ fra ω_2 er de nest nærmeste. NNR vil derfor velge ω_1, mens k-NNR med $k = 3$ vil velge ω_2.

e) Mulig svar: Ikke-parametriske metoder krever ingen antakelse om formen på fordelingene, men er på den annen side mer tidkrevende å bruke for klassifisering, og krever også at hele treningssettet lagres som del av klassifikatoren. Det kan også nevnes at nærmeste-nabo metodene (k-NNR og NNR) har en øvre grense for den asymptotiske feilraten, og er derfor godt egnet til å prøve ut ulike egenskapskombinasjoner, fordi man da har mulighet til å sammenlikne feilraten med hva som er optimalt oppnåelig.