Løsningsforslag Unik4590/9590/TTK4205 - 2016H

16. desember 2017

Oppgave 1

Innledning

a) Se f.eks. figur 8 i forelesningsnotatene, og den tilhørende figurteksten.

b) Se f.eks. figur 3 i forelesningsnotatene.

c) Se avsnitt 1.2 i forelesningsnotatene og eksempelet på beslutningsregel midt på side 7.

d) Mulig svar: Ledet læring består i å trene opp en klassifikator ved hjelp av objekter (egenskapsvektorer) med *kjent* klassetilhørighet. Treningen kan utføres ved bl.a. parametriske- eller ikke-parametriske metoder for estimering av sannsynlighetstetthetsfunksjoner, eller optimalisering av kriteriefunksjoner for å finne vektene i diskriminantfunksjoner direkte fra treningssampler.

Oppgave 2

Bayesisk beslutningsteori

a) Se avsnitt 2.1 i forelesningsnotatene.

b) 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 $\sigma_1=\sigma_2=\sigma$ 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

$$-x^2+2\mu_1x-\mu_1^2+2\sigma^2\ln P(\omega_1)=-x^2+2\mu_2x-\mu_2^2+2\sigma^2\ln P(\omega_2)$$

som kan forenkles til

$$2(\mu_1-\mu_2)x+(\mu_2-\mu_1)(\mu_2+\mu_1)+2\sigma^2\ln\frac{P(\omega_1)}{P(\omega_2)}=0.$$

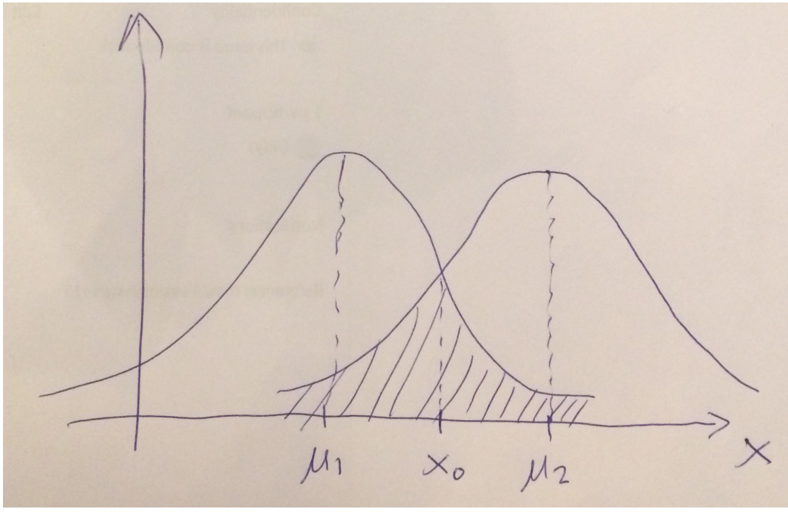
Desisjonsgrensen er derved gitt ved

$$x_0=\frac{(\mu_1-\mu_2)(\mu_1+\mu_2)-2\sigma^2\ln\frac{P(\omega_1)}{P(\omega_2)}}{2(\mu_1-\mu_2)}=\underline{\underline{\frac{1}{2}(\mu_1+\mu_2)-\frac{\sigma^2}{(\mu_1-\mu_2)}\ln\frac{P(\omega_1)}{P(\omega_2)}}}.$$

d) Med like a priori sannsynligheter blir terskelverdien

$$x_0=\underline{\underline{\frac{1}{2}(\mu_1+\mu_2)}},$$

siden $\ln(P(\omega_1/P(\omega_2))=0$ i dette tilfellet. Figuren viser terskelen og feilraten (skravert areal).



e) Med $P(\omega_1)=1/4$ og $P(\omega_2)=3/4$ blir terskelverdien

$$x_0=\frac{1}{2}(\mu_1+\mu_2)-\frac{\sigma^2}{\mu_1+\mu_2}\ln\left(\frac{1/4}{3/4}\right)=\underline{\underline{\frac{1}{2}(\mu_1+\mu_2)+\frac{\sigma^2}{\mu_1-\mu_2}\ln(3).}}$$

Oppg. 3

Diskriminantfunksjoner

a) 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

$$\Sigma = \begin{bmatrix} 4 & 1 \\ 1 & 3 \end{bmatrix},$$

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

$$\Sigma^{-1} = \frac{1}{8}\begin{bmatrix} 3 & -2 \\ -2 & 4 \end{bmatrix}.$$

Innsetting av Σ^{-1} og forventningsvektorene

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \quad \boldsymbol{\mu}_2 = \begin{bmatrix} 3 \\ 6 \end{bmatrix} \quad \text{og} \quad \boldsymbol{\mu}_3 = \begin{bmatrix} 4 \\ 1 \end{bmatrix}$$

i uttrykket for diskriminantfunksjonen ovenfor, gir da

$$\begin{aligned}
g_1(\boldsymbol{x}) &= 2[-3, 10]\boldsymbol{x} - 27 \\
g_2(\boldsymbol{x}) &= 2[-3, 18]\boldsymbol{x} - 99 \\
g_3(\boldsymbol{x}) &= 2[10, -4]\boldsymbol{x} - 36.
\end{aligned}$$

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

$$\boldsymbol{x}_0 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

i diskriminantfunksjonene ovenfor, får man

$$\begin{aligned}
g_1(\boldsymbol{x}_0) &= 2[-3, 10]\begin{bmatrix} 2 \\ 2 \end{bmatrix} - 27 = 1 \\
g_2(\boldsymbol{x}_0) &= 2[-3, 18]\begin{bmatrix} 2 \\ 2 \end{bmatrix} - 99 = -39 \\
g_3(\boldsymbol{x}_0) &= 2[10, -4]\begin{bmatrix} 2 \\ 2 \end{bmatrix} - 36 = -12.
\end{aligned}$$

Siden g_1 har størst verdi, blir objektet klassifisert til ω_1.

Oppg. 4

Parametriske metoder

a) Se avsnitt 3.1 i forelesningsnotatene (side 34).

b) For Maxwellfordelingen

$$p(x|\theta) = \begin{cases} \dfrac{4}{\sqrt{\pi}}\theta^{3/2}x^2 e^{-\theta x^2} & x \geq 0 \\ 0 & \text{ellers,} \end{cases}$$

der $\theta > 0$, blir log-likelihoodfunksjonen

$$\begin{aligned}
\mathscr{L}(\theta) &= \sum_{k=1}^{n} \ln p(x_k|\theta) \\
&= \sum_{k=1}^{n} \ln\left(\frac{4}{\sqrt{\pi}}\theta^{3/2}x_k^2 e^{-\theta x_k^2}\right) \\
&= \sum_{k=1}^{n}\left(\ln\frac{4}{\sqrt{\pi}}x_k^2 + \frac{3}{2}\ln\theta - \theta x_k^2\right).
\end{aligned}$$

Gradienten til log-likelihoodfunksjonen m.h.p. θ (i dette tilfellet den deriverte) blir da

$$\begin{aligned}\frac{\mathrm{d}\mathscr{L}(\theta)}{\mathrm{d}\theta} &= \sum_{k=1}^{n} \frac{\mathrm{d}}{\mathrm{d}\theta}\left(\ln \frac{4}{\sqrt{\pi}}x_k^2 + \frac{3}{2}\ln\theta - \theta x_k^2\right) \\ &= \sum_{k=1}^{n}\left(\frac{3}{2\theta} - x_k^2\right) = \frac{3n}{2\theta} - \sum_{k=1}^{n} x_k^2.\end{aligned}$$

Ved å sette den deriverte lik null, blir løsningen for parameterestimatet

$$\hat{\theta} = \frac{3n}{2\sum\limits_{k=1}^{n} x_k^2}.$$

c) Forslag til svar: Bayesisk estimering skiller seg fra maksimum likelihood estimering ved at man her betrakter parametervektoren i problemet som en stokastisk variabel. Dette gir mulighet til å inkorporere apriori kunnskap i parameterestimeringen gjennom en apriori parameterfordeling $p(\boldsymbol{\theta})$. Ved å introdusere treningssamplene, blir apriorifordelingen konvertert til en a posteriori parameterfordeling $p(\boldsymbol{\theta}|\mathscr{X})$, som normalt vil konvergere mot Diracs deltafunksjon omkring maksimum likelihoodløsningen når $n \to \infty$. For endelig n vil parameterestimatet typisk være et vektet middel av aprioriinformasjon og informasjon fra samplene i treningssettet.

Oppg. 5

Ikke-parametriske metoder

a) Se avsnitt 4.1 i forelesningsnotatene, først og fremst formelen for $p_n(\boldsymbol{x})$ og konvergensbetingelsene på side 50.

c) Se avsnitt 4.2.2 og resultatet i 4.2.3 (side 61).

d) Forslag: 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.