TEK5020/9020 – Mønstergjenkjenning
Høsten 2023

Løsningsforslag – Øvingsoppgaver 5

Idar Dyrdal (idar.dyrdal@its.uio.no)

UiO : Institutt for teknologisystemer

23. august 2023

Oppgave 1

I denne oppgaven er blandingstettheten

$$\left.\begin{array}{l} p(\boldsymbol{x}|\boldsymbol{\theta}) = \displaystyle\sum_{i=1}^{c} p(\boldsymbol{x}|\omega_i, \boldsymbol{\theta}_i)P(\omega_i) \\ \text{der } p(\boldsymbol{x}|\omega_i, \boldsymbol{\theta}_i) = N(\boldsymbol{\mu}_i, \sigma_i^2 I) \end{array}\right\} \text{Her er } \boldsymbol{\mu}_i, \sigma_i \text{ og } P(\omega_i),\ i = 1, \ldots, c \text{ ukjente.}$$

Likningssystemet

$$\left.\begin{array}{l} \displaystyle\sum_{k=1}^{n} \hat{P}(\omega_i|\boldsymbol{x}_k, \hat{\boldsymbol{\theta}}) \nabla_{\boldsymbol{\theta}_i} \ln p(\boldsymbol{x}_k|\omega_i, \hat{\boldsymbol{\theta}}_i) = 0 \\ \hat{P}(\omega_i|\boldsymbol{x}_k, \hat{\boldsymbol{\theta}}) = \dfrac{p(\boldsymbol{x}|\omega_i, \boldsymbol{\theta}_i)\hat{P}(\omega_i)}{\sum_{j=1}^{n} p(\boldsymbol{x}|\omega_j, \boldsymbol{\theta}_j)\hat{P}(\omega_j)} \end{array}\right\} \text{der } \hat{P}(\omega_i) = \frac{1}{n}\sum_{k=1}^{n} \hat{P}(\omega_i|\boldsymbol{x}_k, \hat{\boldsymbol{\theta}}), i = 1, \ldots, c,$$

skal løses med hensyn på σ_i, for $i = 1, \ldots, c$.

Oppgave 1 (forts.)

Komponenttetthetene er gitt ved

$$p(\boldsymbol{x}_k|\omega_i, \boldsymbol{\theta}_i) = \frac{1}{(2\pi)^{d/2}\sigma_i^d} \exp\left\{-\frac{||\boldsymbol{x}_k - \boldsymbol{\mu}_i||^2}{2\sigma_i^2}\right\}, i = 1, \ldots, c,$$

slik at

$$\ln p(\boldsymbol{x}_k|\omega_i, \boldsymbol{\theta}_i) = -\frac{||\boldsymbol{x}_k - \boldsymbol{\mu}_i||^2}{2\sigma_i^2} - \frac{d}{2}\ln(2\pi) - d\ln\sigma_i.$$

Dette gir

$$\frac{\partial}{\partial\sigma_i}\ln p(\boldsymbol{x}_k|\omega_i, \boldsymbol{\theta}_i) = \frac{||\boldsymbol{x}_k - \boldsymbol{\mu}_i||^2}{\sigma_i^3} - \frac{d}{\sigma_i}.$$

Oppgave 1 (forts.)

Innsetting i likningssystemet gir da

$$\sum_{k=1}^{n} \hat{P}\left(\omega_i | \boldsymbol{x}_k, \hat{\boldsymbol{\theta}}\right) \left[\frac{\|\boldsymbol{x}_k - \hat{\boldsymbol{\mu}}_i\|^2}{\sigma_i^3} - \frac{d}{\sigma_i} \right] = 0$$

$$\Downarrow$$

$$\sum_{k=1}^{n} \hat{P}\left(\omega_i | \boldsymbol{x}_k, \hat{\boldsymbol{\theta}}\right) \|\boldsymbol{x}_k - \hat{\boldsymbol{\mu}}_i\|^2 = \sigma_i^2 d \sum_{k=1}^{n} \hat{P}\left(\omega_i | \boldsymbol{x}_k, \hat{\boldsymbol{\theta}}\right)$$

$$\Downarrow$$

$$\hat{\sigma}_i^2 = \frac{\sum_{k=1}^{n} \hat{P}(\omega_i | \boldsymbol{x}_k, \hat{\boldsymbol{\theta}}) ||\boldsymbol{x}_k - \hat{\boldsymbol{\mu}}_i||^2}{d \sum_{k=1}^{n} \hat{P}(\omega_i | \boldsymbol{x}_k, \hat{\boldsymbol{\theta}})}$$

hvilket skulle bevises.

Oppgave 2

I denne oppgaven er den univariate fordelingsfunksjonen

$$p(x|\mu_1, \mu_2)$$

en blandingstetthet av to univariate, normalfordelte komponenter, med kjente a priori sannsynligheter $P(\omega_1) = P(\omega_2) = 1/2$ og kjente varianser $\sigma_1 = \sigma_2 = \sigma = 2$, mens bare forventningsverdiene μ_1 og μ_2 for de to underliggende klassene er ukjente.

Treningssettet er

$$\mathscr{X} = \{1.0, 2.9, 4.2, 5.1, 7.3, 5.4, 7.9, 8.8, 9.8, 11.2\}.$$

Klassetilhørigheten til treningssamplene er ukjent.

Oppgave 2a

Forventningsestimatet $\hat{\boldsymbol{\mu}}_{i, i=1,\ldots,c}$ er gitt ved

$$\hat{\boldsymbol{\mu}}_i = \frac{\sum_{k=1}^{n} P(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\mu}})\mathbf{x}_k}{\sum_{k=1}^{n} P(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\mu}})}$$

der

$$\hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}) = \frac{p(\mathbf{x}_k|\omega_i, \hat{\boldsymbol{\theta}}_i)\hat{P}(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x}_k|\omega_j, \hat{\boldsymbol{\theta}}_j)\hat{P}(\omega_j)}.$$

Her er inneholder parametervektoren $\boldsymbol{\theta}$ alle ukjente parametre i problemet.

Oppgave 2a (forts.)

I denne oppgaven er tetthetsfunksjonene univariate normalfordelinger, der bare forventningsverdiene til de to komponenttetthetene er ukjente ($c = 2$).

Estimatene kan derfor forenkles til

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

der

$$\hat{P}(\omega_i | x_k, \hat{\mu}_1, \hat{\mu}_2) = \frac{p(x_k | \omega_i, \hat{\mu}_i) P(\omega_i)}{\sum_{j=1}^{2} p(x_k | \omega_j, \hat{\mu}_j) P(\omega_j)}.$$

Oppgave 2a (forts.)

Siden tetthetsfunksjonene er normalfordelinger og apriorisannsynlighetene og variansene er like for begge komponenter, blir det siste uttrykket

$$\hat{P}(\omega_i|x_k, \hat{\mu}_1, \hat{\mu}_2) = \frac{\exp\left\{-\frac{1}{2}\left(\frac{x_k - \hat{\mu}_i}{\sigma}\right)^2\right\}}{\exp\left\{-\frac{1}{2}\left(\frac{x_k - \hat{\mu}_1}{\sigma}\right)^2\right\} + \exp\left\{-\frac{1}{2}\left(\frac{x_k - \hat{\mu}_2}{\sigma}\right)^2\right\}}, \tag{2}$$

der $x_{k,k=1\ldots n}$ de umerkede samplene i treningssettet.

Oppgave 2b

Deloppgaven består i å løse det implisitte likningssystemet bestående av (1) og (2) ved iterasjon.

Velger startverdiene $\mu_1(0) = 2$ og $\mu_2(0) = 10$, som settes inn i (2) for å beregne initielle aposteriorisannsynligheter for hvert treningssampel. Disse settes så inn i (1) for å beregne oppdaterte verdier $\mu_1(1)$ og $\mu_2(1)$ av forventningsestimatene, osv.

Forventningsestimatene blir her

$$\hat{\mu}_1 = 3.94 \text{ og } \hat{\mu}_2 = 8.69.$$

Andre startverdier gir også samme resultat.

Resultatene for hver iterasjon blir

1. $\mu_1(1) = 3.6700 \quad \mu_2(1) = 8.7390$
2. $\mu_1(2) = 3.8602 \quad \mu_2(2) = 8.6674$
3. $\mu_1(3) = 3.9092 \quad \mu_2(3) = 8.6703$
4. $\mu_1(4) = 3.9252 \quad \mu_2(4) = 8.6774$
5. $\mu_1(5) = 3.9312 \quad \mu_2(5) = 8.6817$
6. $\mu_1(6) = 3.9337 \quad \mu_2(6) = 8.6838$
7. $\mu_1(7) = 3.9348 \quad \mu_2(7) = 8.6848$
8. $\mu_1(8) = 3.9353 \quad \mu_2(8) = 8.6853$
9. $\mu_1(9) = 3.9355 \quad \mu_2(9) = 8.6855$
10. $\mu_1(10) = 3.9356 \quad \mu_2(10) = 8.6856$

Ytterligere iterasjoner gir ingen endring.