Tentamen: Introduction to Pattern Recognition for AI (NB054B)

6 juli 2007, 9:00 - 12:00

Schrijf boven elk vel je naam, studentennummer en studierichting. Schrijf duidelijk, motiveer je antwoorden en schrijf ook eventuele tussenstappen op: dit kan punten schelen!

Opgave 1

Een fabriek produceert halffabrikaten X. 40% van de geproduceerde halffabrikaten is van kwaliteit $x = 1$ en de rest is kwaliteit $x = 2$. Er is een test Z om halffabrikaten te testen. De uitslag van Z is een getal z tussen 0 en 1. Stel dat de testuitslag, afhankelijk van de kwaliteit, als volgt verdeeld is:

$$\begin{aligned} p(z|x=1) &= \frac{1}{C_1}(1-z^2) \\ p(z|x=2) &= \frac{1}{C_2}(1+z) \end{aligned}$$

voor $0 \leq z \leq 1$. C_1 en C_2 zijn constanten. $p(z|x) = 0$ voor z buiten het interval.

Vraag 1.1 *Laat zien dat $C_1 = 2/3$ en $C_2 = 3/2$.*
ANT:C_1 en C_2 zijn normering constantes. Zij volgen uit

$$\int 1 - z^2 dz = (z - \frac{1}{3}z^3)\Big|_0^1 = 2/3$$

$$\int 1 + z dz = (z + \frac{1}{2}z^2)\Big|_0^1 = 3/2$$

Vraag 1.2 *Bereken met de regel van Bayes de kans op kwaliteit $x = 1$ en $x = 2$ voor het geval dat de testuitslag $z = 0$ is. Doe hetzelfde voor het geval dat de testuitslag $z = 1$ is.*
ANT: Bayes rule:

$$p(x_1|z) = \frac{p(z|x_1)p(x_1)}{p(z|x_1)p(x_1) + p(z|x_2)p(x_2)}$$

Invullen van $z = 0$

$$p(x_1|z=0) = \frac{\frac{3}{2}\frac{2}{5}}{\frac{3}{2}\frac{2}{5} + \frac{2}{3}\frac{3}{5}} = \frac{\frac{3}{5}}{\frac{3}{5} + \frac{2}{5}} = 0.6$$

en dus $p(x_2|z = 0) = 1 - p(x_1|z = 0) = 0.4$. Invullen van $z = 1$. Omdat $p(z = 1|x = 1) = 0$ en $p(z = 1|x = 2) \neq 0$ kan de meetwaarde alleen onstaan door $x = 2$. Dus $p(x_1|z = 1) = 0$ en $p(x_2|z = 1) = 1$

Opgave 2

Beschouw de kansverdeling

$$p(x, k|\boldsymbol{\mu}, \sigma^2, \boldsymbol{\alpha}) = p(x|k, \boldsymbol{\mu}, \sigma^2)p(k|\boldsymbol{\alpha})$$

met $x \in I\!R$ en 'classes' $k \in \{1, \ldots, K\}$. De kansverdeling heeft parameters $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$, σ^2 en $\boldsymbol{\alpha} = \alpha_1, \ldots, \alpha_K$, waarbij $\alpha_k \geq 0$ en $\sum_{k=1}^K \alpha_k = 1$.

De classes zijn verdeeld volgens $p(k|\boldsymbol{\alpha}) = \alpha_k$. De class-conditional densities zijn Gaussisische verdelingen met allemaal dezelfde variantie σ^2, d.w.z.,

$$p(x|k, \boldsymbol{\mu}, \sigma^2) = \mathcal{N}(x|\mu_k, \sigma^2)$$

Stel we hebben een trainingset $\{x_n, k_n\}$ met $n = 1, \ldots, N$. De datapunten zijn onafhankelijk getrokken uit de verdeling.

Vraag 2.1 *Laat zien dat de negative log-likelihood bij deze data wordt gegeven door*

$$E = \frac{1}{2\sigma^2} \sum_{k=1}^{K} \sum_{n=1}^{N} \delta_{kk_n} (x_n - \mu_k)^2 + N \log \sigma + \frac{N}{2} \log(2\pi) - \sum_{k=1}^{K} N_k \log \alpha_k$$

met $N_k = \sum_{n=1}^N \delta_{kk_n}$, d.w.z. het aantal datapunten met $k_n = k$.

ANT: Likelihood is

$$p(\{x_n, k_n\}_{n=1}^N | \boldsymbol{\mu}, \sigma^2, \boldsymbol{\alpha}) = \prod_{n=1}^{N} p(x_n, k_n | \boldsymbol{\mu}, \sigma^2, \boldsymbol{\alpha})$$

De negatieve loglikelihood is

$$\begin{aligned}
E &= -\sum_{n=1}^{N} \log p(x_n, k_n | \boldsymbol{\mu}, \sigma^2, \boldsymbol{\alpha}) \\
&= -\sum_{n=1}^{N} \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_n - \mu_{k_n})^2}{2\sigma^2} \right) \alpha_{k_n} \right) \\
&= -\sum_{n=1}^{N} \log(\frac{1}{\sqrt{2\pi}}) - \sum_{n=1}^{N} \log \frac{1}{\sqrt{\sigma^2}} - \sum_{n=1}^{N} \log \exp\left(-\frac{(x_n - \mu_{k_n})^2}{2\sigma^2} \right) - \sum_{n=1}^{N} \log \alpha_{k_n} \\
&= N \log(\sqrt{2\pi}) + N \log \sigma - \sum_{n=1}^{N} \left(-\frac{(x_n - \mu_{k_n})^2}{2\sigma^2} \right) - \sum_{n=1}^{N} \log \alpha_{k_n} \\
&= \frac{N}{2} \log(2\pi) + N \log \sigma + \frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu_{k_n})^2 - \sum_{n=1}^{N} \log \alpha_{k_n}
\end{aligned}$$

Maak gebruik dat i.h.a. geldt $F(k_n) = \sum_{k=1}^K \delta_{kk_n} F(k)$. (en idem in sub-index notatie: $F_{k_n} = \sum_{k=1}^K \delta_{kk_n} F_k$.) Dus

$$(x_n - \mu_{k_n})^2 = \sum_{k=1}^{K} \delta_{kk_n} (x_n - \mu_k)^2$$

en dus

$$\begin{aligned} E &= \frac{N}{2}\log(2\pi) + N\log\sigma + \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu_{k_n})^2 - \sum_{n=1}^{N}\log\alpha_{k_n} \\ &= \frac{N}{2}\log(2\pi) + N\log\sigma + \frac{1}{2\sigma^2}\sum_{n=1}^{N}\sum_{k=1}^{K}\delta_{kk_n}(x_n - \mu_k)^2 - \sum_{n=1}^{N}\sum_{k=1}^{K}\delta_{kk_n}\log\alpha_k \\ &= \frac{N}{2}\log(2\pi) + N\log\sigma + \frac{1}{2\sigma^2}\sum_{n=1}^{N}\sum_{k=1}^{K}\delta_{kk_n}(x_n - \mu_k)^2 - \sum_{k=1}^{K}\sum_{n=1}^{N}\delta_{kk_n}\log\alpha_k \\ &= \frac{N}{2}\log(2\pi) + N\log\sigma + \frac{1}{2\sigma^2}\sum_{n=1}^{N}\sum_{k=1}^{K}\delta_{kk_n}(x_n - \mu_k)^2 - \sum_{k=1}^{K}N_k\log\alpha_k \end{aligned}$$

Vraag 2.2 *Laat zien door minimalizatie van E dat de maximum likelihood schattingen van μ_k, σ^2 en α_k worden gegeven door*

$$\begin{aligned} \mu_k &= \frac{1}{N_k}\sum_{n=1}^{N}\delta_{kk_n}x_n \\ \sigma^2 &= \frac{1}{N}\sum_{k=1}^{K}\sum_{n=1}^{N}\delta_{kk_n}(x_n - \mu_k)^2 \\ \alpha_k &= \frac{N_k}{N} \end{aligned}$$

ANT: μ_k, σ^2 en α_k kun je vinden door

$$\begin{aligned} \frac{\partial E}{\partial \mu_k} &= 0 \\ \frac{\partial E}{\partial \sigma^2} &= 0 \\ \frac{\partial E}{\partial \alpha_k} &= 0 \end{aligned}$$

te stellen en op te lossen. Let op dat bij α_k nog de constraint geldt dat $\sum_k \alpha_k = 0$. Verder moet je gebruiken dat iha geldt

$$\frac{\partial}{\partial x_k}\sum_{k'=1}^{K}F(x'_k) = \frac{\partial}{\partial x_k}F(x_k) \tag{1}$$

Partiele afgeleide naar μ_k:

$$\frac{\partial E}{\partial \mu_k} = \frac{1}{\sigma^2}\sum_{n=1}^{N}\delta_{kk_n}(x_n - \mu_k)$$

Nulstellen:

$$\sum_{n=1}^{N}\delta_{kk^n}(x^n - \mu_k) = 0$$

dus

$$\sum_{n=1}^{N}\delta_{kk^n}\mu_k = \sum_{n=1}^{N}\delta_{kk^n}x^n$$

met definitie van N_k is dit

$$N_k\mu_k = \sum_{n=1}^{N}\delta_{kk^n}x^n$$

dus

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \delta_{kk^n} x^n \ .$$

Nu σ

$$\frac{\partial E}{\partial \sigma} = \frac{1}{2\sigma^3} \sum_{n=1}^{N} \sum_{k=1}^{K} \delta_{kk_n} (x_n - \mu_k)^2 + \frac{N}{\sigma}$$

Nul stellen, beide kanten met σ^3 *vermenigvuldigen en delen door* N *geeft het gevraagde resultaat.*

Tenslotte α_k*: voor lagrangemultiplier* λ *voor de constraint in, stel de Lagrangiaan* $E+\lambda(\sum_k \alpha_k - 1)$ *op. Minimalizeer naar* $P(k)$ *en los* λ *op uit de constraints*

$$\frac{\partial E + \lambda(\sum_k \alpha_k - 1)}{\partial \alpha_k} = -\frac{N_k}{\alpha_k} + \lambda$$

Daaruit volgt

$$\alpha_k = \lambda N_k$$

Uit $\sum_k \alpha_k = 1$ *volgt* $\lambda = \sum_k N_k = N$*, en daaruit volgt het gevraagde resultaat.*

Opgave 3

Vraag 3.1 *Bespreek de voor en nadelen van de methode van maximum likelihood ten opzichte van Bayesiaanse inferentie .*

ANT: Maximum likelihood geeft een enkele parameter(vector), namelijk die het beste past bij de dataset. Bayesiaanse inferentie geeft een complete distributie van parameters gegeven de data, en houdt dus rekening met onzekerheid in de parameters.

Voordelen van ML

- Maximum likelihood is simpeler:
  - Het werken met een parameter vector is veel simpeler dan het moeten meesjouwen van een hele distributie
  - Het berekenen van de ML-oplossing is het minimalizeren van een error-functie. Dit is i.h.a. numeriek nog behoorlijk goed te doen. Bayesiaanse inferentie vereist het berekenen van een posterior distributie. Dit is eigenlijk alleen in de meest eenvoudige gevallen nog analytisch te doen. Numerieke integratie in hoog dimensionale parameter ruimtes is veel moeilijker en kost veel meer computertijd dan numerieke minimalizatie van een error functie.
- Maximum likelihood heeft geen prior nodig. Dit kan een voordeel zijn omdat je er dan ook niet over hoeft na te denken.
- Als er relatief veel data is, is het resultaat van ML en Bayes hetzelfde, en is ML het eenvoudigste alternatief.

Nadelen van ML

- Maximum likelihood houdt geen rekening met parameter onzekerheid en heeft a.h.w. te veel vertrouwen in de parameter die het terug geeft. Als er relatief weinig data ten opzichte van de modelcomplexiteit is heeft dit overfitting bij ML tot gevolg. Omgekeerd laat Bayes het in principe toe om met veel complexere modellen te werken, omdat Bayesiaanse inferentie automatisch regularizeert.
- Omdat Bayesiaanse inferentie een hele distributie meesjouwt, kun je ook andere statistieken bekijken, bijvoorbeeld onzekerheid in de parameters. Dit kan van belang zijn voor de vraag of er bijvoorbeeld nog meer data nodig is om te leren.
- In Bayesiaanse inferentie is het mogelijk om verschillende modelklasses van verschillende complexiteit mee te nemen. Dit is in ML veel lastiger. ML zou dan keizen voor een parameter uit het meest complexe model (die het beste fit).
- Bayesiaanse inferentie heeft een prior nodig en dwingt je om over je modelaannames na te denken. Bij ML wordt dit onder de mat geschoven.

Opgave 4

Beschouw de verdeling $p(t|\lambda) = \lambda \exp(-t\lambda)$ voor $t > 0$. De verdeling is geparametriseerd door $\lambda > 0$.

Vraag 4.1 *Laat zien dat*

$$\int_0^\infty p(t|\lambda)dt = 1$$

ANT:

$$\int_0^\infty \lambda \exp(-t\lambda)dt = -\exp(-t\lambda)\Big|_{t=0}^{\infty} = 0 - (-1) = 1$$

Stel we hebben data $D = \{t_n\}_{n=1}^N$ met gemiddelde $\frac{1}{N}\sum_{n=1}^N t_n = \langle t \rangle$.

Vraag 4.2 *Laat zien dat de maximum likelihood oplossing wordt gegeven door* $\lambda = 1/\langle t \rangle$.
ANT: Negatieve loglikelihood is

$$E(\lambda) = -\sum_n \log(\lambda \exp(-t_n\lambda)) = -N\log(\lambda) + \sum_n t_n\lambda = -N\log(\lambda) + N\langle t \rangle\lambda$$

Afgeleide nemen

$$\frac{dE}{d\lambda} = -\frac{N}{\lambda} + N\langle t \rangle$$

nul stellen geeft het antwoord

Nu gaan we Bayesiaanse inferentie doen. We kiezen als prior $p(\lambda) \propto \lambda^\nu \exp(-\nu\tau\lambda)$, met hyperparameters ν en τ.

Vraag 4.3 *Laat zien dat de posterior geschreven kan worden als*

$$p(\lambda|D) \propto \lambda^{N+\nu} \exp\Big(-(N\langle t \rangle + \nu\tau)\lambda \Big) \tag{2}$$

ANT: toepassing van Bayes rule:

$$p(\lambda|D) \propto p(D|\lambda)p(\lambda)$$

Likelihood is

$$p(D|\lambda) = \prod_n \lambda \exp(-t_n\lambda) = \lambda^N \exp(-\sum_n t_n\lambda) = \lambda^N \exp(-N\langle t\rangle\lambda)$$

Prior is

$$p(\lambda) \propto \lambda^\nu \exp(-\nu\tau\lambda)$$

Met elkaar vermenigvuldigen geeft

$$p(D|\lambda)p(\lambda) \propto \lambda^N \lambda^\nu \exp(-N\langle t\rangle\lambda) \exp(-\nu\tau\lambda)$$

Bij elkaar vegen van exponenten geeft resultaat.

De prior en de posterior zijn Gamma verdelingen (zie boek blz. 688). De Gamma verdeling heeft twee parameters $a > 0$ en $b > 0$, en is van de vorm

$$\text{Gamma}(\lambda|a, b) \propto \lambda^{a-1} \exp(-b\lambda) \tag{3}$$

Door (2) en (3) te vergelijken vind je dat de posterior een Gamma verdeling is met $a = N + \nu + 1$ en $b = N\langle t\rangle + \nu\tau$.

In dit geval kunnen we de posterior exact uitrekenen, of misschien beter gezegd: opzoeken. Meestal kan dat niet, en moeten we de posterior benaderen. Dit kan bijvoorbeeld met de Laplace benadering zoals we in de volgende vragen zien. In dit geval, waarbij we de exacte posterior hebben kunnen we zien hoe goed de Laplace benadering is door bijvoorbeeld statististische eigenschappen van de verdelingen te vergelijken.

De Laplace benadering van de posterior is een Gaussische benadering rond de MAP oplossing (ofwel de mode λ_0 van de verdeling),

$$q(\lambda|D) = \mathcal{N}(\lambda|\lambda_0, s^2)$$

Vraag 4.4 *De mode van* $\text{Gamma}(\lambda|a, b)$ *is*

$$\lambda_0 = \frac{a-1}{b}$$

Verifieer dit door $\lambda^{a-1}\exp(-b\lambda)$ *te maximalizeren naar* λ*. Waarom is de normering hier niet van belang?*

Verifieer vervolgens dat de mode van de posterior $p(\lambda|D)$ *wordt gegeven door*

$$\lambda_0 = \frac{N+\nu}{N\langle t\rangle + \nu\tau}.$$

ANT.

Afgeleide naar λ *nemen en nul stellen.*

$$\frac{d}{d\lambda}\text{Gamma}(\lambda|a, b) = (a-1)\lambda^{a-2}\exp(-b\lambda) - b\lambda^{a-1}\exp(-b\lambda) = 0$$

Omdat $\lambda^{a-2}\exp(-b\lambda)$ *positief is, geldt dan ook*

$$(a-1) - b\lambda = 0$$

Ofwel $\lambda = (a-1)/b$.

Normering is constant in λ*. Als* $\lambda*$ *het maximum is van* $f(\lambda)$*, dan ook van* $\alpha f(\lambda)$*.*

Invullen van $a = N + \nu + 1$ *en* $b = N\langle t\rangle + \nu\tau$ *in* $\lambda = (a-1)/b$ *geeft het antwoord.*

Vraag 4.5 *Beschrijf hoe de variantie in de Laplace benadering wordt bepaald, en laat zien dat deze gegeven wordt door*

$$s^2 = \frac{N+\nu}{(N\langle t\rangle + \nu\tau)^2}$$

ANT.

Als $p(\lambda) \propto f(\lambda)$ *wordt de Laplace benadering berekend door het maximum* λ_0 *van* $f(\lambda)$ *te bepalen. Vervolgens wordt de variantie bepaald door de tweede afgeleide van* $\log f$ *rond* λ_0 *te berekenen. De variantie is dan*

$$s^2 = -\frac{1}{d^2/d\lambda^2 \log f(\lambda_0)}$$

Hier is $g = \log f = (a-1)\log\lambda - b\lambda$. *Eerste afgeleide is* $g' = (a-1)/\lambda - b$. *Tweede afgeleide is* $g'' = -(a-1)/\lambda^2$ *dus* $s^2 = \lambda_0^2/(a-1)$. *Invullen van* $\lambda_0 = (a-1)/b$ *geeft* $s^2 = (a-1)/b^2$. *Invullen van* $a = N+\nu+1$ *en* $b = N\langle t\rangle + \nu\tau$ *geeft het resultaat.*

We gaan nu naar de kwaliteit van de benadering kijken. De verwachtingswaarde en variantie van Gamma$(\lambda|a,b)$ zijn:

$$\begin{aligned} I\!E[\lambda] &= \frac{a}{b} \\ \mathrm{var}[\lambda] &= \frac{a}{b^2} \end{aligned}$$

Vraag 4.6 *Neem voor het gemak* $\nu \approx 0$, *zodat* ν *verwaarloosd kan worden t.o.v.* N *en vergelijk de benaderingen van* $I\!E[\lambda], \mathrm{var}[\lambda]$ *volgens de Laplace benadering van de posterior (2) met hun exacte waardes. Doe dit door de benaderingen uit te drukken in termen van de exacte waardes, d.w.z.*

$$\begin{aligned} I\!E_{Laplace}[\lambda] &= (\ 1+\epsilon(N,\langle t\rangle)\ )\ I\!E_{Exact}[\lambda] \\ \mathrm{var}_{Laplace}[\lambda] &= (\ 1+\eta(N,\langle t\rangle)\ )\ \mathrm{var}_{exact}[\lambda] \end{aligned}$$

Wat kun je zeggen over de kwaliteit van de benadering als functie van N *(en* $\langle t\rangle$*)? Tenslotte, hoe goed wordt de mode* λ_0 *benaderd?*

ANT: We verwaarlozen termen met ν. *Invullen van* $a = N+1$ *en* $b = N\langle t\rangle$ *geeft*

$$\begin{aligned} I\!E_{Exact}[\lambda] &= \frac{N+1}{N\langle t\rangle} \\ \mathrm{var}_{Exact}[\lambda] &= \frac{N+1}{N^2\langle t\rangle^2} \end{aligned}$$

De Laplace waardes zijn λ_0 *en* s^2, *dus*

$$\begin{aligned} I\!E_{Laplace}[\lambda] &= \frac{N}{N\langle t\rangle} \\ \mathrm{var}_{Laplace}[\lambda] &= \frac{N}{N^2\langle t\rangle^2} \end{aligned}$$

Ofwel

$$\begin{aligned} I\!E_{Laplace}[\lambda] &= (1-\frac{1}{N+1})I\!E_{Exact}[\lambda] \\ \mathrm{var}_{Laplace}[\lambda] &= (1-\frac{1}{N+1})\mathrm{var}_{Exact}[\lambda] \end{aligned}$$

Dus de Laplace benadering geeft een kleine onderschatting van mean en variance. De relative fout neemt omgekeerd evenredig met N *af. De mode is exact (voor de mode van de Laplace benadering is immers de exacte mode genomen).*