T 991210 Mönstergenkänning TF4

1 a) $\hat{m}_x = \frac{1}{N}\sum_{n=1}^{N} x_n$ $\qquad \sigma_x^2 = E\{(x-m_x)^2\} = E\{x^2\} - m_x^2$

$$E\{\hat{m}_x\} = \frac{1}{N}\sum_{n=1}^{N} E\{x_n\} = \frac{1}{N} N m_x = m_x$$

b) $s_x^2 = E\{(\hat{m}_x - m_x)^2\} = E\{\hat{m}_x^2\} - m_x^2$

$$E\{\hat{m}_x^2\} = \frac{1}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N} E\{x_n x_m\}$$

$$= \frac{1}{N^2}\sum_{n=1}^{N} E\{x_n^2\} + \frac{1}{N^2}\sum_{n=1}^{N}\sum_{m\neq n} E\{x_n\}E\{x_m\}$$

(x_n och x_m oberoende)

$$= \frac{1}{N^2}\sum_{n=1}^{N} E\{x_n^2\} + m_x^2\,\frac{N(N-1)}{N^2}$$

$$= \frac{1}{N^2}\sum_{n=1}^{N} \underbrace{E\{x_n^2\}}_{\sigma^2+m_x^2} + m_x^2 - \frac{m_x^2}{N^2}$$

$$= \frac{\sigma_x^2}{N} + \frac{m_x^2}{N} + m_x^2 - \frac{m_x^2}{N}$$

$$s_x^2 = \frac{\sigma_x^2}{N} + m_x^2 - m_x^2 = \frac{\sigma_x^2}{N} \qquad \text{Q.E.D}$$

c) $s_x = 0.01 = \frac{\sigma_x}{\sqrt{N}} \Rightarrow N = \frac{\sigma_x^2}{0.01^2}$ $\qquad N = 10000\sigma_x^2$

$N = \frac{\sigma_x^2}{s_x^2}$ $\qquad (N = 100)$

2

a) Okorrelerade komponenter, varians σ_i^2 och medel m_i för komponent x_i

$$\Downarrow$$

$$C_{xx} = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_N^2 \end{pmatrix}$$ kovariansmatrisen en diagonalmatris

Egenvektorerna pekar i riktningar med störst varians.

$$\Downarrow$$

Egenvektorerna i detta fall samma som enhetsvektorerna $\bar{e}_i = [0\,0 \ldots 0\,1\,0 \ldots 0]$

↑ Position i

Detta inses också direkt från

$$C_{xx}\bar{e}_i = \lambda_i \bar{e}_i \quad \text{där } \lambda_i = \sigma_i^2$$

Svar: Egenvektorerna är enhetsbasen
Egenvärdena är $\lambda_i = \sigma_i^2$

b) Kan eliminera alla basvektorer med små egenvärden $\Rightarrow$ Behåll $\bar{e}_i$, $i = 1, 2, \ldots, 1000, 18001, \ldots, 20000$

$\Rightarrow$ 3000 dimensioner kvar

c) K-means, $K = 53$. Likt vektorkvantisering $\|\bar{x} - \hat{\bar{x}}_k\|$ $\Rightarrow$ 53 prototypvektorer täcker in data mycket väl

$\Rightarrow$ Kan koda vektor $\bar{x}$ som ett enda heltal!

3. a) Outliers förlyga: Påverkar skattning av medel och kovariansmatris (mer generellt parameter-skattningar). ⇒ Beslutsgränser kan bli helt fel. Detta demonstreras på lab.

b) Outliers kan uppstå tex p.g.a trasig (felaktig) sensor eller på manuell felavläsning.

c) Man kan "filtrera" bort outliers före träning genom att tex utföra PCA och titta på data i 2D/3D och eller genom att införa villkor

$$x_{min}^{(i)} < x_i < x_{max}^{(i)}$$

för varje komponent.

4 a) Icke-parametriska metoder: "Universella", kan justeras (via parametrar) så att de utför godtyckliga klassificeringar och funktionsapproximationer. Kallas ibland för data-drivna metoder

b)

Parametrisk metoder: En familj av tänkbara lösningar parametriseras och en sökmetod väljer ut den bästa. Dessa metoder ej universella dvs de kan inte justeras till att passa godtyckliga problem.

Enlagersperceptron: Parametrisk, dess beslutsytor består av hyperplan genom val av parametrar.

Flerlagersperceptron: Icke-parametrisk, dess beslutsytor kan bli godtyckligt komplicerade genom att öka antalet gömda noder

b) K-NN: Snabb "träning" Långsam respons
Konceptuellt enkel att förstå Mycket minne kr

Enlagersperceptron: Snabb respons Långsam träning
Svår att tolka
Lokala minimum
Lik minne krävs

$$P(\text{klass } i \mid \bar{x}) = \frac{f(\bar{x} \mid \text{klass } i)\, P(\text{klass } i)}{f(\bar{x})} \quad \text{Bayes}$$

$$P(\text{klass } i \mid \bar{x}) > P(\text{class } j \mid \bar{x}) = g_j(\bar{x}) \quad ?$$

$$g_j(\bar{x}) = f(\bar{x} \mid \text{klass } i)\, P(\text{klass } i)$$

Beslutsyta:

$$\{ \bar{x} \mid g_1(\bar{x}) = g_2(\bar{x}) \}$$

$$\left\{ \bar{x} \;\middle|\; \frac{e^{-\frac{1}{2}(\bar{x}-\bar{m}_1)^T C_1^{-1}(\bar{x}-\bar{m}_1)} P(\text{klass } 1)}{(2\pi)^{d/2} |C_1|^{1/2}} = \frac{e^{-\frac{1}{2}(\bar{x}-\bar{m}_2)^T C_2^{-1}(\bar{x}-\bar{m}_2)}}{(2\pi)^{d/2} |C_2|^{1/2}} \right\}$$

$$\left\{ \bar{x} \;\middle|\; -\tfrac{1}{2}(\bar{x}-\bar{m}_1)^T C_1^{-1}(\bar{x}-\bar{m}_1) + \tfrac{1}{2}(\bar{x}-\bar{m}_2)^T C_2^{-1}(\bar{x}-\bar{m}_2) = \tfrac{1}{2}\log|C_1| - \tfrac{1}{2}\log|C_2| \right\}$$

Detta är en kvadratisk form

$$\{ \bar{x} \mid \bar{x}^T A \bar{x} + b^T \bar{x} + c = 0 \}$$

Välkänd i linjär algebra. Olika kvadratytor erhålles beroende på hur A, b, c ser ut.

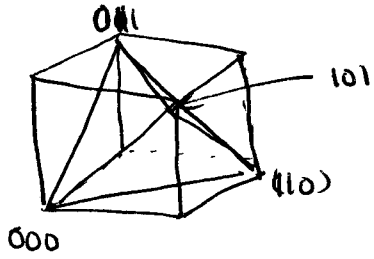
$C_1 = C_2 \Rightarrow$ Kvadratiska termer tar ut varandra

$$\{ \bar{x} \mid \tilde{b}^T \bar{x} + \tilde{c} = 0 \}$$ Hyperplan

6 Önskar samma avstånd mellan alla bokstäver

Ledning:

fx.



Hörn på en kub

samma avstånd ($\sqrt{2}$) mellan alla.

Välj tex
000 ↔ A
011 ↔ C
101 ↔ G
110 ↔ T

Anm.
Man kan tänka sig att flytta denne tetraeder så att den har sitt tyngdpunkt i origo men detta är inte nödvändigt. Enhälles direkt med kodningen
-1-1-1 ↔ A
-1+1+1 ↔ C
+1-1+1 ↔ G
+1+1-1 ↔ T

7

a) K-NN: 1. För ett givet $\bar{x}$, beräkna avstånd till de K närmaste grannarna i träningsmängden

2. Beräkna antalet medlemmar K_c från klass c bland de K närmaste grannarna

3. Välj klass c om $K_c > K_i$ $c \neq i$.

b) Korsvalidering: 1. Dela upp tillgänglig data i träning och testset på flera olika sätt, extremfallet är att plocka ut endast ett testexempel varje gång

2. Testa prestanda m.h.a testmängderna för olika antal K. Upprepa

3. Beräkna medelprestanda för alla olika antal K över de testmängder som studerats

4. Välj det K som ger högst medelprestanda

c) K stor → Liten varians i $\frac{K_c}{N_c}$

→ Bra estimat men inte lokalt

→ Globalt eftersom alla punkter i rummet med i skattning

K liten → Lokalt estimat men stor varians

8 a) K-NN: Alltför långsam respons
Måste utföra många beräkningar
Vid alla jämförelser

b) Flerlagesperceptron: Alltför långsam att
träna.

c) RBF: Träna gömda lagret med K-means
Snabbt och enkelt

Linjär regression: Mycket snabb träning och
respons. Gäller även RBF
där gömda lagret fixerat.

Linjär regression absolut snabbast men
om sambanden olinjära så ger (troligen)
RBF bättre resultat.

9) a) Olinearieter

$$\varphi(x) = \frac{1}{1+e^{-\alpha x}}$$

$$\varphi(x) = \tanh(x)$$

$$\varphi(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases}$$

$$\varphi(x) = e^{-x^2}$$

b) Bias: Gör det möjligt att få beslutslinjen/ytan som går utanför origo.