Introduktion till maskininlärning

Kurskod: EEN175

Tentamen 2022-10-27

Tid: 8:30-12:30

Lärare: Gabriel Arslan Waltersson, tel: 0737-71 04 16

Tentamen omfattar 25 poäng, där betyg tre fordrar 10 poäng, betyg fyra 15 poäng och betyg fem 20 poäng.

Granskning av rättning sker den 14 och 15 november kl 12:30-13:00 på avdelningen.

Tillåtna hjälpmedel:

- Matematiska och fysikaliska tabeller, t ex Beta och Physics handbook.
- Valfri kalkylator med tömt minne.
- De två formelblad som ingår i tentamenstesen får också tas med på tentamen och då inkluderande egna handskrivna anteckningar på fram och baksida på de två formelbladen, dvs sammanlagt fyra A4-sidor. Datorutskrifter förutom de ingående formlerna och figurerna på de två formelbladen är ej tillåtna.

Institutionen för elektroteknik
Avdelningen för system- och reglerteknik
Chalmers tekniska högskola

1

Visa hur de okända parametrarna a, b och c i följande modeller kan skattas med hjälp linjär regression. Det kan antas att variablerna x, y, u, v, φ, och ω är kända för ett antal tidpunkter $t = 0, h, 2h, 3h, \ldots, Nh$, där h är lika med samplingsintervallets längd.

a) $y = e^{-2a}x^2 + x^4/b + c$

b) $y = a + b\cos(c - u)$

c) $a\dot{v} + ab\dot{\omega}\cos\varphi = u + bc\dot{\omega}^2\sin\varphi$

Ange speciellt hur derivatorna i uppgift c) bestäms då h antas vara kort i förhållande till dynamiken för den aktuella modellen. (4 p)

2

En tidsdiskret modell för ett första ordningens system med varierande fördröjning av insignalen u ska skattas. Samplingsintervallets längd h antas vara tidsenhet, vilket innebär att $h = 1$. Insignalens fördröjning i förhållande till utsignalen y kan antas ligga i ett intervall mellan $d_{\min}$ och $d_{\max}$, vilket innebär att systemet antas kunna beskrivas med hjälp av differensekvationen

$$y(k) = ay(k-1) + b_1u(k-d_{\min}) + \ldots + b_nu(k-d_{\max}) + e(k), \quad k = 1, 2, 3, \ldots$$

där antalet insignalparametrar $n = d_{\max} - d_{\min} + 1$, och den icke mätbara insignalen $e(k)$ antas vara vitt brus med variansen r_e. Eftersom denna systemmodell kan formuleras som en linjär regressionsmodell undviks bias då parameterskattningen sker med hjälp av minsta kvadratmetoden.

Antag nu också att den mätbara insignalen u är vitt brus. På vilket sätt påverkas medelkvadratfelet (mean square error)

$$\text{tr}\,\mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}))(\hat{\theta} - \mathbb{E}(\hat{\theta}))^T]$$

av antalet datapunkter N samt hur stor skillnaden antas vara mellan största och minsta tidsfördröjningen $d_{\max} - d_{\min}$. Notera att tr betyder trace, dvs summan av diagonalelementen. (4 p)

3

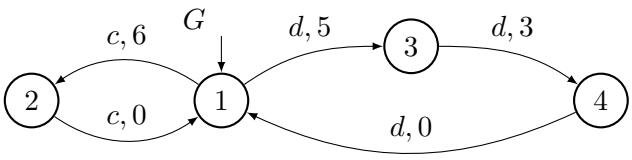
Följande tabell visar ett utfall då personer har valt att spela eller inte spela tennis givet väder, temperatur och vind.

Väder	Temperatur	Vind	Spela tennis
soligt	varmt	blåsigt	nej
soligt	varmt	stilla	ja
molnigt	medel	stilla	ja
molnigt	kallt	blåsigt	nej
soligt	medel	blåsigt	ja
soligt	kallt	blåsigt	ja
regn	medel	stilla	ja
regn	kallt	stilla	nej
molnigt	varmt	stilla	ja
molnigt	medel	blåsigt	ja
regn	kallt	blåsigt	nej

Skatta med hjälp av naiv bayesiansk klassificering om en person kommer att spela tennis i följande situationer: a) det regnar, är varmt och blåser, b) det regnar, är varmt och är vindstilla. (4 p)

4

Betrakta följande diskreta tillståndsmodell (automat) G, inkluderande olika belöningar för de olika tillståndsövergångarna.



a) Iterera $\widehat{Q}$-inlärningsalgoritmen

$$\widehat{Q}_{k+1}(x,a) = (1-\alpha_k)\widehat{Q}_k(x,a) + \alpha_k\big(r' + \gamma \max_{b\in\Sigma(x')} \widehat{Q}_k(x',b)\big)$$

de första 10 stegen för $\alpha_k = 1$ and $\gamma = 0.9$. Beslutet (action) för det största skattade $\widehat{Q}$-värdet väljs (Greedy) i tillstånd 1, förutom då $\widehat{Q} = 0$ vilket har högre prioritet. Denna strategi väljs för att förbättra den initiala utforskningen av tillståndsrummet. (2 p)

b) Bestäm värdet som Q-funktionen kommer att konvergera mot för de alternativa besluten c och d i tillståndet $x = 1$ genom att studera $J^*(x)$. Jämför dessa Q-värden med de värden som erhölls efter 10 steg i uppgift a). Vilket är det optimala beslutet i begynnelsetillståndet? (2 p)

5

Ett två-lagers neuralt nätverk för multiklassificering (mer än två klasser) kan med ReLU (Rectified Linear Unit) som aktiveringsfunktion uttryckas som

$$z_m = b_m^{(z)} + \sum_{j=1}^{U} w_{mj}^{(z)} \max\Big(\sum_{i=1}^{n} w_{ji}^{(x)} x_i + b_j^{(x)}, 0\Big), \quad m \in \{1, \ldots, M\}$$

där $z = [\, z_1 \cdots z_M]^T$ och $g = \text{softmax}(z)$. Antag att $n = 1$, $U = 2$ och $M = 3$, samt att $w_{11}^{(x)} = 2$, $b_1^{(x)} = -2$, $w_{21}^{(x)} = -2$, $b_2^{(x)} = 4$, $w_{m1}^{(z)} = m - 1$, $w_{m2}^{(z)} = -m + 2$, medan samtliga bias termer $b_m^{(z)} = 0$.

a) Skissera detta neurala nätverk inklusive vikter. (2 p)

b) Vilka klasser erhålls för testdata $x = 1$ och $x = 2$? (2 p)

6

a) Hur förhåller sig bias och varians till varandra vid regularisering? (1 p)

b) Ange den fundamentala förenkling som tillämpas vid naiv bayesiansk klassificering. (1 p)

c) Ange skillnader och likheter mellan support vektor maskiner för regression och klassificering. (1 p)

d) Vilken relation råder mellan ett två-lagers neuralt nätverk för multiklassificering (mer än två klasser) och logistisk regression? (1 p)

e) Ange två fundamentala skillnader mellan ett fullständigt (densed, fully connected) neuralt nätverk och ett ”convolutional” neuralt nätverk. (1 p)

Lösning tentamen Intro maskininlärning 221027

BL221025

1. a) $y(kh) = [\theta_1 \quad \theta_2 \quad \theta_3] \begin{bmatrix} x^2(kh) \\ x^4(kh) \\ 1 \end{bmatrix} + \varepsilon(kh)$

$\theta_1 = e^{-2a}$, $\theta_2 = 1/b$

$a = -\frac{\ln \theta_1}{2}$

$b = 1/\theta_2$

$c = \theta_3$

b) $y(kh) = a + b\cos c \cos(u(kh)) + \sin c \cos u(kh) + \varepsilon(kh)$

$= [\theta_1 \quad \theta_2 \quad \theta_3] \begin{bmatrix} 1 \\ \cos u(kh) \\ \sin u(kh) \end{bmatrix} + \varepsilon(kh)$

$\theta_1 = a$, $\theta_2 = b\cos(c)$, $\theta_3 = b\sin(c)$

$a = \theta_1$

$c = \arctan \frac{\theta_3}{\theta_2}$

$b = \frac{\theta_3}{\sin(c)}$

c) $u(kh) = a\dot{v}(kh) + ab\dot{w}(kh)\cos\varphi(kh) - bc\dot{w}^2(kh)\sin\varphi(kh) + \varepsilon(kh)$

$= [\theta_1 \quad \theta_2 \quad \theta_3] \begin{bmatrix} \dot{v}(kh) \\ \dot{w}(kh)\cos\varphi(kh) \\ \dot{w}^2(kh)\sin\varphi(kh) \end{bmatrix} + \varepsilon(kh)$

$\theta_1 = a$, $\theta_2 = ab$, $\theta_3 = bc$

$a = \theta_1$

$b = \theta_2/a$

$c = \theta_3/b$

$\dot{v}(kh) \approx (v(kh) - v(kh-h))/h$

$\dot{w}(kh) \approx (w(kh) - w(kh-h))/h$

2.

$$y(k) = \underbrace{[a \quad b_1 \quad \dots \quad b_n]}_{\theta} \underbrace{\begin{bmatrix} y(k-1) \\ u(k-d_{min}) \\ \vdots \\ u(k-d_{max}) \end{bmatrix}}_{x(k)} + \varepsilon(k)$$

$E\hat{\theta} = \theta \Rightarrow$

$$Cov(\hat{\theta}) = E\left[(\hat{\theta}-\theta)(\hat{\theta}-\theta)^T\right] =$$

$$= \left(\sum_{k=1}^{N} x(k)x^T(k)\right)^{-1} r_e = \left(\frac{1}{N}\sum_{k=1}^{N} x(k)x^T(k)\right)^{-1}\frac{r_e}{N}$$

$$\approx \left(E\,x(k)x^T(k)\right)^{-1}\frac{r_e}{N} =$$

$$= \begin{bmatrix} Ey^2(k-1) & & \bigcirc \\ & Eu^2(k-d_{min}) & \\ \bigcirc & & \ddots \\ & & Eu^2(k-d_{max}) \end{bmatrix}^{-1}\frac{r_e}{N} = \begin{bmatrix} \frac{1}{r_y} & & \bigcirc \\ & \frac{1}{r_u} & \\ \bigcirc & & \ddots \\ & & \frac{1}{r_u} \end{bmatrix}\frac{r_e}{N}$$

($Ey^2(k-1) = \frac{1}{r_y}$)

Alla icke diagonala element blir noll eftersom u = vitt brus och därmed även okorrelerad med $y(k-1)$ eftersom $d_{min} \geq 1$

$$MSE = tr\,Cov(\hat{\theta}) \approx \left(\frac{1}{r_y} + \frac{n}{r_u}\right)\frac{r_e}{N}$$

MSE minskar då antalet datapunkter N ökar men ökar då skillnaden $d_{max} - d_{min} = n + 1$, dvs antalet b-parametrar i skattningen, ökar.

3. Introducera följande förkortningar

Väder: s = soligt, m_V = molnigt, r = regn

Temperatur: v = varmt, m_T = medel, k = kallt

Vind: b = blåsigt, $\neg b$ = ej blåsigt (stilla)

ST = spela tennis $\neg ST$ = ej spela tennis

$p(ST) = 7/11$ $p(\neg ST) = 4/11$

a)

$$p(ST \mid r, v, b) = \frac{p(r \mid ST)\, p(v \mid ST)\, p(b \mid ST)\, p(ST)}{p(r, v, b)}$$

$$= \frac{1/7 \cdot 2/7 \cdot 3/7 \cdot 7/11}{p(r, v, b)} = \frac{0.0111}{p(r, v, b)}$$

$$p(\neg ST \mid r, v, b) = \frac{p(r \mid \neg ST)\, p(v \mid \neg ST)\, p(b \mid \neg ST)\, p(\neg ST)}{p(r, v, b)}$$

$p(\neg ST \mid r, v, b) > p(ST \mid r, v, b) \Rightarrow$ ej spela tennis

$$= \frac{2/4 \cdot 1/4 \cdot 3/4 \cdot 4/11}{p(r, v, b)} = \frac{0.0341}{p(r, v, b)}$$

b)

$$p(ST \mid r, v, \neg b) = \frac{\ldots\, p(\neg b \mid ST) \ldots}{p(r, v, \neg b)} = \frac{\ldots\, 4/7 \ldots}{p(r, v, \neg b)} = \frac{0.0148}{p(r, v, \neg b)}$$

$$p(\neg ST \mid r, v, \neg b) = \frac{\ldots\, p(\neg b \mid \neg ST) \ldots}{p(r, v, \neg b)} = \frac{\ldots\, 1/4 \ldots}{p(r, v, \neg b)} = \frac{0.0114}{p(r, v, \neg b}$$

$p(ST \mid r, v, \neg b) > p(\neg ST \mid r, v, \neg b) \Rightarrow$ spela tennis

4 a)

k	x	a	x'	r	$\hat{Q}_k(x',b)$	$\hat{Q}_{k+1}(x,a)$
1	1	c	2	6	$\hat{Q}(2,c)=0$	$\hat{Q}(1,c)=6$
2	2	c	1	0	$\hat{Q}(1,c)=6$, $\hat{Q}(1,d)=0$	$\hat{Q}(2,c)=6\gamma$
3	1	d	3	5	$\hat{Q}(3,d)=0$	$\hat{Q}(1,d)=5$
4	3	d	4	3	$\hat{Q}(4,d)=0$	$\hat{Q}(3,d)=3$
5	4	d	1	0	$\hat{Q}(1,c)=6$, $\hat{Q}(1,d)=5$	$\hat{Q}(4,d)=6\gamma$
6	1	c	2	6	$\hat{Q}(2,c)=6\gamma$	$\hat{Q}(1,c)=6(1+\gamma^2)$
7	2	c	1	0	$\hat{Q}(1,c)=6(1+\gamma^2)$, $\hat{Q}(1,d)=5$	$\hat{Q}(2,c)=6(\gamma+\gamma^3)$
8	1	c	2	6	$\hat{Q}(2,c)=6(\gamma+\gamma^3)$	$\hat{Q}(1,c)=6(1+\gamma^2+\gamma^4)$
9	2	c	1	0	$\hat{Q}(1,c)=6(1+\gamma^2+\gamma^4)$, $\hat{Q}(1,d)=5$	$\hat{Q}(2,c)=6(\gamma+\gamma^3+\gamma^5)$
10	1	c	2	6	$\hat{Q}(2,c)=6(\gamma+\gamma^3+\gamma^5)$	$\hat{Q}(1,c)=6(1+\gamma^2+\gamma^4+\gamma^6)$

b) $J^*(2)=\gamma J^*(1)$ $J^*(4)=\gamma J^*(1)$

$J^*(3)=3+\gamma J^*(4)=3+\gamma^2 J^*(1)$

$J^*(1)=\max_{c,d}\left(\underbrace{6+\gamma J^*(2)}_{c},\underbrace{5+\gamma J^*(3)}_{d}\right)=$

$=\max(6+\gamma^2 J^*(1),\ 5+3\gamma+\gamma^3 J^*(1))$

Fixpunkt för c $\Rightarrow (1-\gamma^2)J^*(1)=6$

$J^*(1)=\frac{6}{1-\gamma^2}=\frac{6}{1-0.81}=31.6$

Fixpunkt för d $\Rightarrow (1-\gamma^3)J^*(1)=5+3\gamma$

$J^*(1)=\frac{5+3\gamma}{1-\gamma^3}=28.4$

Action c ger störst reward

för $x=1$ $\Rightarrow J^*(1)=31.6$

$J^*(2) = J^*(4) = \gamma J^*(1) = 28.4$

$J^*(3) = 3 + \gamma^2 J^*(1) = 28.6$

$\hat{Q}_\infty(x,a) = g(x,a) + \gamma J^*(\delta(x,a))$

$\hat{Q}_\infty(1,c) = 6 + \gamma J^*(2) = J^*(1) = 31.6$

$\hat{Q}_\infty(1,d) = 5 + \gamma J^*(3) = 30.7$

Då endast en aktion, säg a, är tillgänglig gäller att $\hat{Q}_\infty(x,a) = J^*(x) \Rightarrow$

$\hat{Q}_\infty(2,c) = J^*(2) = 28.4$, $\hat{Q}_\infty(3,d) = J^*(3) = 28.6$

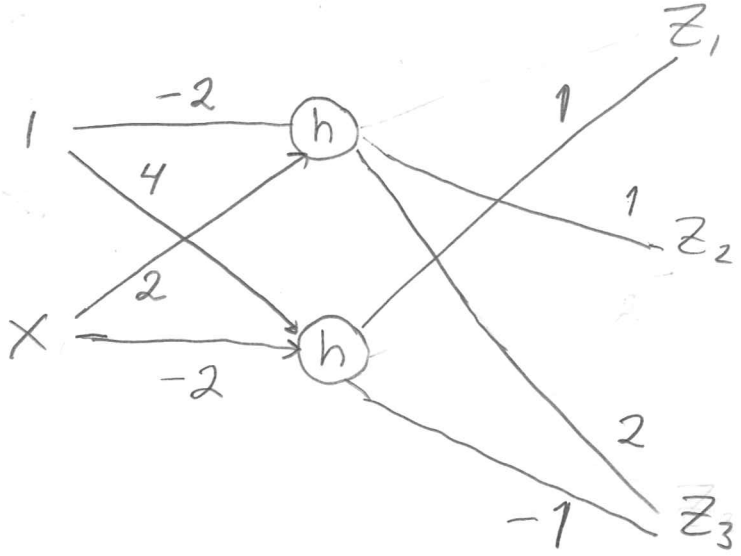
$\hat{Q}_\infty(4,d) = J^*(4) = 28.4$

Optimal aktion i tillstånd 1

$= \arg\max_{a \in \{c,d\}} \hat{Q}_\infty(1,a) = c$ eftersom

$\hat{Q}_\infty(1,c) > \hat{Q}_\infty(1,d)$

5. a)



b)

$$z_1 = \max(4-2x, 0) = \begin{cases} 2, & x=1 \\ 0, & x=2 \end{cases}$$

$$z_2 = \max(-2+2x, 0) = \begin{cases} 0, & x=1 \\ 2, & x=2 \end{cases}$$

$$z_3 = 2\max(-2+2x, 0) - \max(4-2x, 0)$$

$$= \begin{cases} -2 & x=1 \\ 4 & x=2 \end{cases}$$

$x=1$

$$g = \frac{1}{e^{z_1}+e^{z_2}+e^{z_3}} \begin{bmatrix} e^{z_1} \\ e^{z_2} \\ e^{z_3} \end{bmatrix} = \begin{bmatrix} 0.867 \\ 0.117 \\ 0.016 \end{bmatrix}$$

$x=2$

$$g = \begin{bmatrix} 0.016 \\ 0.117 \\ 0.867 \end{bmatrix}$$

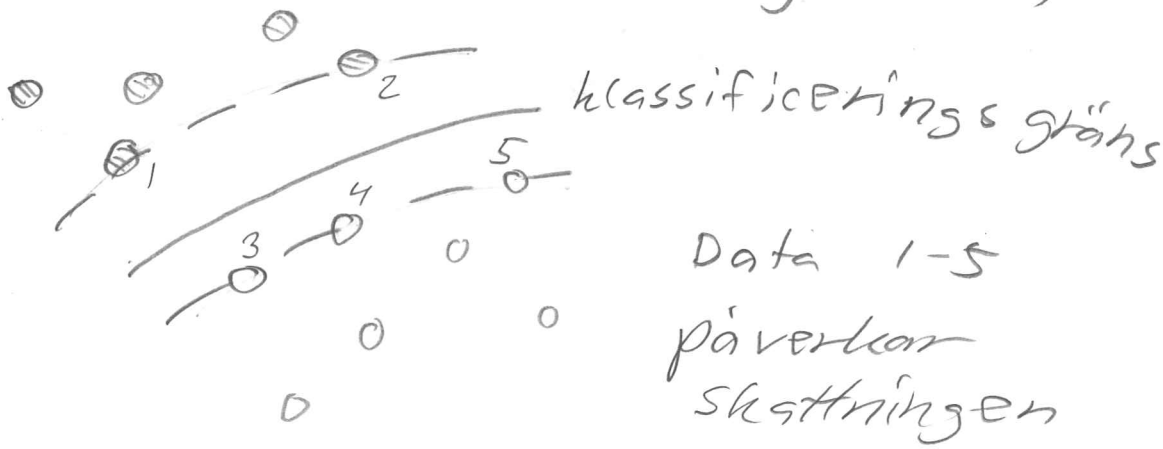
$x=1 \Rightarrow m=1 \qquad x=2 \Rightarrow m=3$

6. a) Regularisering innebär att variansen i skattningen minskar på bekostnad av viss ökning av bias.

b) Givet en klass y antas de olika elementen i featurevektorn x vara oberoende av varandra

$$\Rightarrow p(x|y) = \prod_{i=1}^{n} p(x_i|y)$$

c) Vid SVM regression påverkar samtliga data sådana att $|y - \hat{y}| \geq \varepsilon$ skattningen. Under förutsättningen att data är separerbara i två klasser är det bara de data som ligger närmast klassficerings-gränsen som påverkar skattningen vid SVM klassificering, se fig



d) Logistisk regression erhålls då det gömda lagret i ett två-lagers neuralt nätverk tas bort.

e) I ett "convolutional" neuralt nätverk ingår bara vikter från ett begränsat antal noder i det föregående lagret närmare bestämt de närmaste noderna. Vikterna är dessutom lika från dessa noder.