Tentamen – Bayesiansk inferens och maskininlärning (TIF385)

Tid och plats:	21 augusti, 2024, fm, Johanneberg.
Hjälpmedel:	Physics Handbook, Beta Mathematics Handbook, typgodkänd kalkylator.
Lösningsskiss:	Christian Forssén.

Tentamen består av sex uppgifter som kan ge maximalt 60 poäng totalt. För att bli godkänd med betyg 3 krävs 24 poäng, för betyg 4 krävs 36 poäng och för betyg 5 krävs 48 poäng.

Rättningsprinciper: Alla svar skall motiveras (om ej annat anges) och införda storheter skall förklaras. Lösningarna förväntas vara välstrukturerade och begripligt presenterade. Skriv och rita tydligt. Vid tentamensrättning gäller följande allmänna principer:

- För maximal (10) poäng krävs fullständigt korrekt och välmotiverad lösning.
- Mindre fel ger 1-3 poängs avdrag. Gäller även mindre brister i presentationen.
- Lösningar som inte går att följa (t.ex. avsaknad av relevant figur, ej definierade variabler, svårtydd, etc) renderar poängavdrag även om svaret verkar vara korrekt.
- Allvarliga principiella fel ger fullt poängavdrag.
- Allvarliga fel som leder till orimliga resultat kan ge lägre poängavdrag om orimligheten pekas ut.
- Även skisserade lösningar kan ge delpoäng.

Detta är enbart en skiss av den fullständiga lösningen. Det kan innebära att vissa mellansteg i uträkningarna, som egentligen är nödvändiga för en komplett lösning, inte redovisas.

---

1. Svara på följande delfrågor:

   (a) Använd produktregeln för sannolikheter för att härleda Bayes teorem

   $$\mathbb{P}\left(M \mid \mathcal{D}, I\right) = \frac{\mathbb{P}\left(\mathcal{D} \mid M, I\right)\mathbb{P}\left(M \mid I\right)}{\mathbb{P}\left(\mathcal{D} \mid I\right)}$$

   (2 poäng)

   (b) Parameterestimering innebär att utföra inferens av modellparametrarna $\boldsymbol{\theta}$ för en modell $M(\boldsymbol{\theta})$ givet data $\mathcal{D}$. Vid parameterestimering slipper man att evaluera en av termerna i Bayes teorem. Vilken och varför? *(2 poäng)*

(c) Finn ett uttryck för $p(z \mid I)$ när motsvarande slumpvariabel $Z = \sqrt{X}$ och $p(x \mid I) = \frac{1}{x_{\max} - x_{\min}}$ för $0 < x_{\min} \leq x \leq x_{\max}$ och 0 annars. Verifiera att $p(z \mid I)$ är korrekt normaliserad. *(6 poäng)*

Lösning:

(a) Se till exempel ekvation 7.1 och 7.3 i kurskompendiet.

(b) Nämnaren $\mathbb{P}(\mathcal{D} \mid I)$ beror ej på modellen och därmed inte heller på dess parametrar. Eftersom målet med parameterestimering blir att bestämma en PDF för parametrarna kommer denna term bara att ge en normaliseringskonstant.

(c) Med $z = f(x) = \sqrt{x}$ har vi $x = f^{-1}(z) = z^2$ så att $|dx/dz| = 2|z|$. Vi noterar att z är positiv så att $|z| = z$. Därför har vi att

$$p(z \mid I) = 2zp(x \mid I) = 2z\frac{1}{x_{\max} - x_{\min}},$$

för $\sqrt{x_{\min}} \leq z \leq \sqrt{x_{\max}}$ och 0 annars
Vi kontrollerar normaliseringen genom att utföra integralen

$$\int_0^\infty p(z \mid I)\, dz = \int_{\sqrt{x_{\min}}}^{\sqrt{x_{\max}}} \frac{2z}{x_{\max} - x_{\min}} dz = \frac{1}{x_{\max} - x_{\min}} \left[z^2\right]_{\sqrt{x_{\min}}}^{\sqrt{x_{\max}}} = 1.$$

---

2. Prestandan för en maskininlärningsmodell uppskattas ofta med en felfunktion (error function) som vi allmänt betecknar $E(\hat{y}, y)$, där $\hat{y}$ är en modellförutsägelse och y motsvarande instans måldata.

(a) Antag att det finns tillgång till träningsdata ($\mathcal{D}_{\text{train}}$) och valideringsdata ($\mathcal{D}_{\text{val}}$). Introducera (definiera) E_{train} och E_{val} och förklara vid vilket steg av träningen som dessa mått beräknas.

(b) Skissa hur E_{train} och E_{val} typiskt beror på antalet gradientstegsiterationer vid träning av ett neuralt nätverk. Skillnaden $E_{\text{train}} - E_{\text{val}}$ kallas för generaliseringsgapet. Beskriv hur generaliseringsgapet kan användas under träning av ett neuralt nätverk.

(c) Ge tre förslag på hur generaliseringsgapet kan minskas. Du behöver inte ta hänsyn till hur förslagen påverkar modellens träffsäkerhet utan bara att de kan minska generaliseringsgapet.

(3 poäng per deluppgift, plus 1 poäng om alla är helt korrekta.)

Lösning:

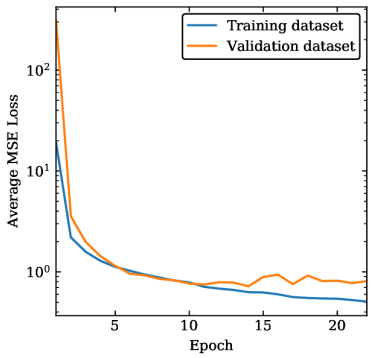
(a) Vi introducerar felfunktionen för tränings- respektive valideringsdata

$$E_{\text{train}} = \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} E(\hat{y}(x_i), y_i), \quad \text{för } (x_i, y_i) \in \mathcal{D}_{\text{train}},$$

$$E_{\text{val}} = \frac{1}{N_{\text{val}}} \sum_{i=1}^{N_{\text{val}}} E(\hat{y}(x_i), y_i), \quad \text{för } (x_i, y_i) \in \mathcal{D}_{\text{val}}.$$

Vid slutet av varje träningsiteration (eller varje epok) beräknas dessa mått för den aktuella versionen av modellen. Notera alltså att de inte används för gradientstegsoptimering (det är kostnadsfunktionen som används till det).

(b) Se figur nedan. Skillnaden mellan de heldragna kurvorna är generaliseringsgapet.



Om felfunktionen är densamma som kostnadsfunktionen så kommer E_{train} att minska som en funktion av antalet iterationer. Även E_{val} kan förväntas att minska till en början, men när man närmar sig överoptimering så avstannar minskningen. Detta kan användas som en signal att upphöra med träningen.

(c) (1) Minska modellkomplexiteten (vilket ger mindre risk för överanpassning); (2) Viktregularisering; (3) Avbryt träningen när E_{val} inte längre minskar; (4) Använd mer (och/eller mer representativ) träningsdata.

---

3. Definiera ordinär linjär regression (eller minsta kvadratmetoden) och visa att den leder till normalekvationen $\boldsymbol{X}^T\mathcal{D} = \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta}$ med lösningen $\boldsymbol{\theta}^* = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\mathcal{D}$. I uppgiften ingår att definiera designmatrisen

$\boldsymbol{X}$ samt att introducera vektorerna $\boldsymbol{\theta}$ och $\mathcal{D}$. *(10 poäng)*

Lösning:
Detta är bevis 2 från bevislistan (se kurshemsida och hänvisning till kompendiet).

---

4. Antag att vi skall skapa en maskininlärningsmodell för att klassificera bilder av handritade siffror mellan 0–9. Varje bild definieras av 8×8 pixlar med ett värde på intensiteten för varje pixel. Vi har bestämt att använda ett fullständigt kopplat neuralt nätverk med fem gömda lager som har tjugo noder vardera samt ett utlager med tio noder. Varje nod karakteriseras av vikter som inkluderar en bias-term w_0. Vi använder sigmoidfunktionen, $1/(1 + e^{-z})$, som aktivitetsfunktion i de gömda lagren.

   (a) Inför relevant notation och ge ett matematiskt uttryck för utsignalerna från noderna i det första gömda lagret. Definiera storlekar på eventuella matriser och vektorer.

   (b) Hur många parametrar har vi totalt i vårt neurala nätverk?

   (c) Givet att vi vill skapa en mjuk klassificerare som ger sannolikheter för alla tio möjliga klasserna (siffrorna 0–9). Hur skulle vi kunna definiera utlagret och utsignalen från modellen?

   (3 poäng per deluppgift, plus 1 poäng om alla korrekta.)

   Lösning:

   (a) Utsignalerna ges av vektorn $\boldsymbol{y} = f(\boldsymbol{z})$, vilket alltså innebär att utsignalen från nod i i det första gömda lagret är $y_i = f(z_i)$ där $f(z_i) = 1/(1 + e^{-z_i})$.

   Både vektorn $\boldsymbol{y}$ och $\boldsymbol{z}$ har längden N där N är bredden på det första gömda lagret.

   Aktiveringarna $\boldsymbol{z}$ ges av

$$\boldsymbol{z} = \boldsymbol{y}_{\text{in}}\boldsymbol{W} + \boldsymbol{w}_0,$$

   där vektorn med insignaler, $\boldsymbol{y}_{\text{in}}$, har längden N_{in}, viktmatrisen för det första gömda lagret, $\boldsymbol{W}$, har dimensionen $N_{\text{in}} \times N$ samt vektorn med dess basvikter, $\boldsymbol{w}_0$, har längden N. För att vara extra tydlig bör man klargöra att formen ovan bygger på att samtliga vektorer är radvektorer, dvs $1 \times N$ eller $1 \times N_{\text{in}}$.

(b) Inlagret har bredden 64. Dessa signaler går till samtliga noder i det första gömda lagret som alltså måste ha $64+1$ vikter vardera. De 20 noderna i varje gömt lager ger varsin utsignal som går till varje nod i det efterföljande lagret som alltså behöver $20+1$ vikter vardera. Detta ger

$$20 \times (64 + 1) + (20 + 20 + 20 + 20 + 10) \times (20 + 1) = 3190$$

(c) Vi önskar en mjuk klassificerare vilket innebär att modellen skall ge en (diskret) sannolikhetsfördelning, dvs en vektor där elementen kan tolkas som sannolikheten att måldata tillhör den specifika klassen $p(t = i) = \hat{y}_i$.

För att åstadkomma detta kan vi till exempel använda sigmoidfunktionen även för noderna i utlagret. Dessa leder då till en utvektor $\tilde{\boldsymbol{y}}$ med tio element $\tilde{y}_i = f(z_i)$ som uppfyller $0 \leq \tilde{y}_i \leq 1$. Denna vektor måste dock normaliseras för att omvandlas till en diskret sannolikhetsfördelning $\hat{\boldsymbol{y}}$. Alltså ges utsignalen av

$$\hat{y}_i = \frac{f(z_i)}{\sum_{i=0}^{9} f(z_i)}.$$

---

5. Avgör om Markovkedjan som beskrivs av övergångsmatrisen

$$T = \begin{bmatrix} 0 & 2/5 & 3/5 \\ 1/2 & 1/4 & 1/4 \\ 1/2 & 1/6 & 1/3 \end{bmatrix}$$

är (i) stationär och (ii) reversibel. Du skall också förklara (iii) vilka egenskaper en övergångsmatris skall uppfylla (för ett diskret och ändligt utfallsrum).

(10 poäng)

Lösning:

(i) Ja, den är stationär (eftersom Markovkedjan beskrivs av en övergångsmatris; detta gäller ju enbart för stationära kedjor).

(ii) Lös ekvationssystemet

$$\sum_i \pi_i T_{ij} = \pi_j, \quad \forall j$$

under tvångsvillkoret $\pi_1+\pi_2+\pi_3 = 1$ för att finna stationärfördelningen. Lösningen är (mellansteg saknas)

$$\pi = \left(\frac{1}{3}, \frac{4}{15}, \frac{2}{5}\right).$$

Kontrollera detaljerad balans

$$\pi_i T_{ij} = \pi_j T_{ji}, \quad \forall i, j$$

Samtliga dessa är uppfyllda (skall visas) och kedjan är alltså reversibel.

(iii) Övergångsmatrisen T har icke-negativa element $T(i,j) \geq 0$ för alla i, j samt normaliserade radsummor $\sum_j T_{ij} = 1$ för alla i.

---

6. Antag att vi har tre mynt som både ser och känns likadana. Vi får dock veta att det ena myntet har 0.75 sannolikhet att ge krona när det singlas, det andra är rättvist (dvs 0.5 sannolikhet för krona), och det tredje har 0.25 sannolikhet att ge krona.

   Antag att du väljer ett mynt, utan att veta vilket, och singlar det. Vad är chansen att du får krona?

   Antag att du fick krona. Vad är då sannolikheten att du får krona en andra gång om du singlar det en gång till? *(10 poäng)*

   Lösning:
   Låt H/T beteckna resultatet av ett myntkast som visar krona/klave. Vi söker en a posteriori prediktiv sannolikhet

$$\mathbb{P}(\mathcal{F} = H|\mathcal{D} = H, I)$$

   där I är informationen som ges i problemformuleringen. Vi vet inte vilken av de tre mynten vi kastar, så det lämpliga är att marginalisera med avseende på mynt-typ. Låt oss numrera mynten med den diskreta stokastiska variabeln $C = 1, 2, 3$. Detta ger oss följande uttryck för den posteriora prediktiva sannolikheten.

$$\begin{aligned}\mathbb{P}(\mathcal{F} = H|\mathcal{D} = H, I) &= \sum_{C=1}^{3} \mathbb{P}(\mathcal{F} = H, C|\mathcal{D} = H, I) \\ &= \sum_{C=1}^{3} \mathbb{P}(\mathcal{F} = H|C, I)\mathbb{P}(C|\mathcal{D} = H, I),\end{aligned}$$

där vi använde produktregeln för sannolikheter i det sista steget och att kunskap om C ger att sannolikheten $\mathbb{P}(\mathcal{F} = H|C, I)$ är betingat oberoende av tidigare slantsinglingar.

Den vänstra faktorn i det sista steget är en trolighet, och vi vet att $\mathbb{P}(\mathcal{F} = H|C = 1, I) = 0.75$, $\mathbb{P}(\mathcal{F} = H|C = 2, I) = 0.5$, samt $\mathbb{P}(\mathcal{F} = H|C = 3, I) = 0.25$. Den högra faktorn är en a posteriori sannolikhet för att ha mynt C givet att man observerar krona (i det första kastet). För att beräkna detta måste vi använda Bayes sats

$$\mathbb{P}(C|\mathcal{D} = H, I) = \frac{\mathbb{P}(\mathcal{D} = H|C, I)\mathbb{P}(C|I)}{\mathbb{P}(\mathcal{D} = H|I)}.$$

Nämnaren blir (använd marginalisering och produktregeln)

$$\mathbb{P}(\mathcal{D} = H|I) = \sum_{C=1}^{3} \mathbb{P}(\mathcal{D} = H|C, I)\mathbb{P}(C|I) = \frac{3}{4}\frac{1}{3} + \frac{1}{2}\frac{1}{3} + \frac{1}{4}\frac{1}{3} = \frac{1}{2}.$$

Givet detta har vi de tre a posteriori sannolikheterna

$$\begin{aligned}
\mathbb{P}(C = 1|\mathcal{D} = H, I) &= 2 \cdot \frac{3}{4} \cdot \frac{1}{3} = \frac{1}{2}, \\
\mathbb{P}(C = 2|\mathcal{D} = H, I) &= 2 \cdot \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{3}, \\
\mathbb{P}(C = 3|\mathcal{D} = H, I) &= 2 \cdot \frac{1}{4} \cdot \frac{1}{3} = \frac{1}{6}.
\end{aligned}$$

Vi kan slutligen beräkna den a posteriori sannolikhet som vi söker

$$\begin{aligned}
\mathbb{P}(\mathcal{F} = H|\mathcal{D} = H, I) &= \sum_{C=1}^{3} \mathbb{P}(\mathcal{F} = H|C, I)\mathbb{P}(C|\mathcal{D} = H, I) \\
&= \frac{3}{4}\frac{1}{2} + \frac{1}{2}\frac{1}{3} + \frac{1}{4}\frac{1}{6} = \frac{7}{12} \approx 0.58.
\end{aligned}$$
