Tentamen – Bayesiansk inferens och maskininlärning (TIF385)

Tid och plats:	3 april, 2023, fm, Johanneberg.
Hjälpmedel:	Physics Handbook, Beta Mathematics Handbook, typgodkänd kalkylator.
Examinator:	Christian Forssén.
Jourhavande lärare:	Christian Forssén (031–772 3261).

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.

Lycka till!

---

1. Den gemensamma frekvensfunktionen (="probability mass function") för två diskreta variabler X, Y ges av

$$p(x, y) = \frac{x + y}{N}, \quad \text{för } x, y \in \{0, 1, 2\}.$$

   (a) Beräkna normaliseringskonstanten N.

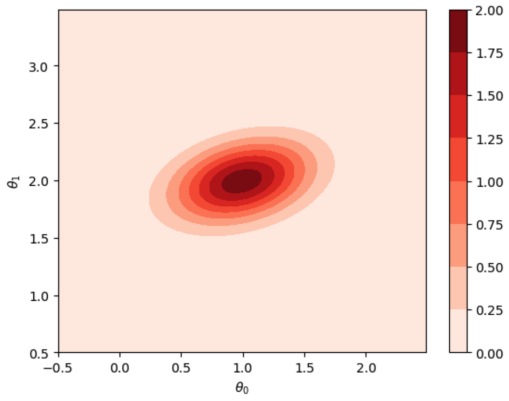
   (b) Finn den marginaliserade sannolikheten $p(x)$.

   (c) Finn den betingade sannolikheten $p(x|y)$.

   (3 poäng per deluppgift, plus 1 poäng om åtminstone en är helt korrekt.)

2. Finn ett uttryck för $p(z|I)$ när slumpvariablen $Z = \sqrt{X}$ och X är uniformt fördelad på det positiva intervallet $[x_{\min}, x_{\max}]$ (dvs $0 < x_{\min} < x_{\max}$). Verifiera att $p(z|I)$ är korrekt normaliserad. *(10 poäng)*

3. Figuren visar en a posteriori-fördelning som är resultatet från en inferens av parametrarna (θ_0, θ_1) i en linjär modell $y = \theta_0 + \theta_1 x$. Skissa modellens prediktion för y i intervallet $x \in [-1, 1]$.



4. Betrakta en stationär Markovkedja med endast två utfall (utfallsrummet $S = \{1, 2\}$) som beskrivs av övergångsmatrisen

$$T = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix},$$

där $p, q \in [0, 1]$.

   (a) Betrakta först specialfallet $p = q = 1$. Finn en stationärfördelning.

   (b) Är stationärfördelningen från (a) också en gränsfördelning? (motivera ditt svar)

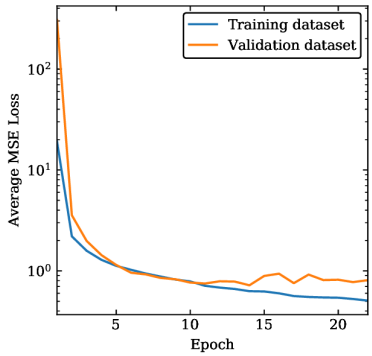
   (c) Finn stationärfördelningen för godtyckliga värden $p, q \in [0, 1]$.

   (3 poäng per deluppgift, plus 1 poäng om åtminstone en är helt korrekt.)

5. Betrakta en maskininlärningsmodell för regression: $\hat{\boldsymbol{y}} = \hat{\boldsymbol{y}}(\boldsymbol{\theta}; \boldsymbol{x})$ där $\boldsymbol{x}$ är oberoende variabler och $\boldsymbol{\theta}$ är modellparametrar. Antag vidare att det finns en uppsättning med data $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N}$ som delas upp i två delar: $\mathcal{T}$ med N_{train} träningsdata och $\mathcal{V}$ med N_{val} valideringsdata ($N = N_{\text{train}} + N_{\text{val}}$).

(a) Medelkvadratsfelet MSE (mean-squared error) är en vanlig felmåttsfunktion för att mäta modellens prestanda. Skriv ner uttryck för MSE($\boldsymbol{\theta}$) beräknat för tränings- respektive valideringsdata. (*4 poäng*)

(b) Figuren nedan visar en inlärningskurva som tagits fram vid iterativ träning av modellen givet en kostnadsfunktion $C(\boldsymbol{\theta})$. Förklara konkret hur denna kurva har tagits fram och vad det är som beräknas. Varje epok (epoch) motsvarar en träningsiteration där all träningsdata har matats igenom nätverket en gång. (*6 poäng*)



6. Betrakta ett neuralt nätverk med P noder i inlagret, M noder i det enda gömda lagret (lager 1) samt en enda nod i utlagret (lager 2). Dvs modellen avbildar $\mathbb{R}^P \to \mathbb{R}^1$.

   I det gömda lagret (1) används en sigmoidfunktion $f^1(z) = e^z/(1+e^z)$ som aktiveringsfunktion för alla noder (där z är aktiveringen för respektive nod) medan utlagret (2) är linjärt $f^2(z) = z$.

   Ge ett förenklat uttryck för utsignalen givet att samtliga vikter i det gömda lagret är små. Uttrycket skall visa att modellen blir approximativt linjär i prediktorvariablerna x_i $(i = 1, \ldots, P)$. I uppgiften ingår också att kvantifiera vad som menas med villkoret att vikterna är "små". *(10 poäng)*

Tentamen – Bayesiansk inferens och maskininlärning (TIF385)

Tid och plats:	3 april, 2023, 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. Den gemensamma frekvensfunktionen (="probability mass function") för två diskreta variabler X, Y ges av

$$p(x, y) = \frac{x + y}{N}, \quad \text{för } x, y \in \{0, 1, 2\}.$$

   (a) Beräkna normaliseringskonstanten N.

   (b) Finn den marginaliserade sannolikheten $p(x)$.

   (c) Finn den betingade sannolikheten $p(x|y)$.

   (3 poäng per deluppgift, plus 1 poäng om åtminstone en är helt korrekt.)

   Lösning:

(a) Vi kan räkna ut normaliseringen genom att summera över alla möjliga utfall

$$\sum_{x=0}^{2}\sum_{y=0}^{2} p(x,y) = \frac{x+y}{N} = \frac{3+6+9}{N}.$$

Detta innebär att $N = 18$.

(b) Den marginaliserade sannolikheten fås genom att summera över möjliga utfall på y

$$p(x) = \sum_{y=0}^{2} p(x,y) = \frac{x}{18} + \frac{x+1}{18} + \frac{x+2}{18} = \frac{x+1}{6}.$$

(c) Den betingade sannolikheten fås från

$$p\,(x\,|\,y) = \frac{p(x,y)}{p(y)}.$$

Från (b) (och symmetrin i x, y) inser vi att $p(y) = (y+1)/6$ så att

$$p\,(x\,|\,y) = \frac{(x+y)/18}{(y+1)/6} = \frac{x+y}{3(y+1)} \quad \text{för } x \in \{0,1,2\}.$$

---

2. Finn ett uttryck för $p(z|I)$ när slumpvariablen $Z = \sqrt{X}$ och X är uniformt fördelad på det positiva intervallet $[x_{\min}, x_{\max}]$ (dvs $0 < x_{\min} < x_{\max}$). Verifiera att $p(z|I)$ är korrekt normaliserad. *(10 poäng)*

Lösning:

- Med $z = f(x) = \sqrt{x}$ har vi den inversa transformationen $x = f^{-1}(z) = z^2$ så att $|dx/dz| = 2|z|$. Vi noterar att z är positivt vilket ger $|z| = z$ och därmed

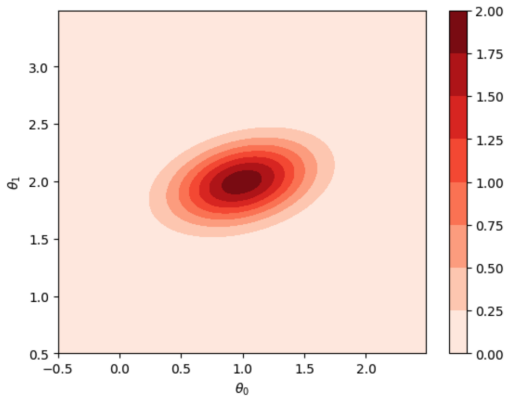
$$p(z|I) = 2zp(x|I) = \frac{2z}{x_{\max} - x_{\min}} \quad \text{för } \sqrt{x_{\min}} \le z \le \sqrt{x_{\max}},$$

och 0 annars.

- Vi kontrollerar normaliseringen genom att marginalisera

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

3. Figuren visar en a posteriori-fördelning som är resultatet från en inferens av parametrarna (θ_0, θ_1) i en linjär modell $y = \theta_0 + \theta_1 x$. Skissa modellens prediktion för y i intervallet $x \in [-1, 1]$.



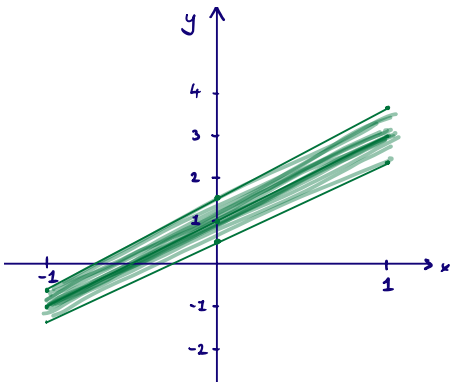
(10 poäng)

Lösning:

- En a posteriori-fördelning för modellens prediktion y kallas för en PPD och fås genom att marginalisera över a posteriori-fördelningen för modellparametrarna. Det enklaste sättet att få denna är genom att evaluera modellen för många stickprov av modellparametrar

$$\{\theta_0 + \theta_1 x \ : \ (\theta_0, \theta_1) \sim p(\theta_0, \theta_1 | \mathcal{D}, I)\} \, ,$$

  där $p(\theta_0, \theta_1 | \mathcal{D}, I)$ är a posteriori-fördelningen som visas i figuren.
- Skissen skall visa att modellens prediktion ges av en sannolikhetsfördelning. Man kan t.ex. plocka en handfull representativa stickprov ut fördelningen och skissa motsvarande räta linjer.
- Maxpunkten $(\theta_0, \theta_1) \approx (1.0, 2.0)$ ger en prediktion $y \approx 1 + 2x$ med hög sannolikhet, medan t.ex. $(\theta_0, \theta_1) \approx (1.5, 2.2)$ och $(\theta_0, \theta_1) \approx (0.5, 1.8)$ ger prediktioner $y \approx 1.5 + 2.2x$ respektive $y \approx 0.5 + 1.8x$ med lägre sannolikhet.
- Skissen borde därför visa ett band för y där korrelationen mellan den linjära modellens lutning och skärningspunkt framgår.



4. Betrakta en stationär Markovkedja med endast två utfall (utfallsrummet $S = \{1, 2\}$) som beskrivs av övergångsmatrisen

$$T = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix},$$

där $p, q \in [0, 1]$.

(a) Betrakta först specialfallet $p = q = 1$. Finn en stationärfördelning.

(b) Är stationärfördelningen från (a) också en gränsfördelning? (motivera ditt svar)

(c) Finn stationärfördelningen för godtyckliga värden $p, q \in [0, 1]$.

(3 poäng per deluppgift, plus 1 poäng om åtminstone en är helt korrekt.)

Lösning:

(a) I detta specialfall har vi

$$T = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Fördelningen $\pi = (0.5, 0.5)$ är stationär eftersom $\pi T = \pi$.

(b) Fördelningen ovan är dock inte en gränsfördelning eftersom den inte nås från alla slumpmässigt valda startfördelningar. Inför beteckningen $\pi_0 = (\pi_{0,1}, \pi_{0,2})$ för en godtycklig startfördelning. Då har vi allmänt att

$$\lim_{n\to\infty} \pi_0 T^n \neq \pi.$$

Testa t.ex. med $\pi_0 = (0.9, 0.1)$.

(c) Stationärfördelningen $\pi = (\pi_1, \pi_2)$ för godtyckliga värden $p, q \in [0, 1]$ fås genom att lösa ekvationssystemet

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

med normaliseringsvillkoret $\pi_1 + \pi_2 = 1$. Dvs

$$\begin{aligned} \pi_1(1-p) + \pi_2 q &= \pi_1, \\ \pi_1 p + \pi_2(1-q) &= \pi_2, \\ \pi_1 + \pi_2 &= 1 \end{aligned}$$

Lösningen är

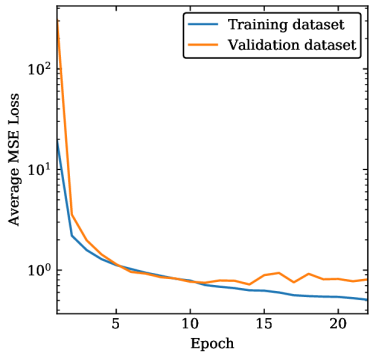
$$\pi = \left(\frac{q}{p+q}, \frac{p}{p+q} \right).$$

---

5. Betrakta en maskininlärningsmodell för regression: $\hat{\boldsymbol{y}} = \hat{\boldsymbol{y}}(\boldsymbol{\theta}; \boldsymbol{x})$ där $\boldsymbol{x}$ är oberoende variabler och $\boldsymbol{\theta}$ är modellparametrar. Antag vidare att det finns en uppsättning med data $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N$ som delas upp i två delar: $\mathcal{T}$ med N_{train} träningsdata och $\mathcal{V}$ med N_{val} valideringsdata ($N = N_{\text{train}} + N_{\text{val}}$).

   (a) Medelkvadratsfelet MSE (mean-squared error) är en vanlig felmåttsfunktion för att mäta modellens prestanda. Skriv ner uttryck för MSE($\boldsymbol{\theta}$) beräknat för tränings- respektive valideringsdata. (*4 poäng*)

   (b) Figuren nedan visar en inlärningskurva som tagits fram vid iterativ träning av modellen givet en kostnadsfunktion $C(\boldsymbol{\theta})$. Förklara konkret hur denna kurva har tagits fram och vad det är som beräknas. Varje epok (epoch) motsvarar en träningsiteration där all träningsdata har matats igenom nätverket en gång. (*6 poäng*)



Lösning:

(a) Medelkvadratsfelet för tränings- respektive valideringsdata är

$$\mathrm{MSE}_{\mathrm{train}}(\boldsymbol{\theta}) = \frac{1}{N_{\mathrm{train}}} \sum_{i=1}^{N_{\mathrm{train}}} \Big(\hat{\boldsymbol{y}}\left(\boldsymbol{\theta}\,;\, \boldsymbol{x}_i\right) - \boldsymbol{y}_i \Big)^2, \quad \text{för } (\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathcal{T},$$

$$\mathrm{MSE}_{\mathrm{val}}(\boldsymbol{\theta}) = \frac{1}{N_{\mathrm{val}}} \sum_{i=1}^{N_{\mathrm{val}}} \Big(\hat{\boldsymbol{y}}\left(\boldsymbol{\theta}\,;\, \boldsymbol{x}_i\right) - \boldsymbol{y}_i \Big)^2, \quad \text{för } (\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathcal{V}.$$

(b) Träningen syftar till att finna den uppsättning med modellparametrar som minimerar kostnadsfunktionen (vilken evalueras enbart med träningsdata). Dvs man söker

$$\boldsymbol{\theta}^\star = \mathrm{argmin}_{\boldsymbol{\theta}} C_{\mathrm{train}}(\boldsymbol{\theta}).$$

Träningen genomförs iterativt (vanligtvis med gradientstegsoptimering) där varje epok innebär att modellparametrarna uppdateras i en riktning som minskar kostnadsfunktionen. Låt oss säga att modellparametrarna är $\boldsymbol{\theta}^{\star,(n)}$ vid epok n.

Figuren visar då $\mathrm{MSE}_{\mathrm{train}}(\boldsymbol{\theta}^{\star,(n)})$ samt $\mathrm{MSE}_{\mathrm{val}}(\boldsymbol{\theta}^{\star,(n)})$ som en funktion av antalet epoker med inlärning.

---

6. Betrakta ett neuralt nätverk med P noder i inlagret, M noder i det enda gömda lagret (lager 1) samt en enda nod i utlagret (lager 2). Dvs modellen avbildar $\mathbb{R}^P \to \mathbb{R}^1$.

   I det gömda lagret (1) används en sigmoidfunktion $f^1(z) = e^z/(1 + e^z)$ som aktiveringsfunktion för alla noder (där z är aktiveringen för respektive nod) medan utlagret (2) är linjärt $f^2(z) = z$.

   Ge ett förenklat uttryck för utsignalen givet att samtliga vikter i det gömda lagret är små. Uttrycket skall visa att modellen blir approximativt linjär i prediktorvariablerna x_i $(i = 1, \ldots, P)$. I uppgiften ingår också att kvantifiera vad som menas med villkoret att vikterna är "små". *(10 poäng)*

   Lösning:

- Låt z_j^1 beteckna aktiveringen av nod j i det gömda lagret. Notera att superskriptet betecknar lager (inte upphöjt till). I termer av insignalerna x_i och nodens vikter $w_{j,i}^1$ blir aktiveringen

$$z_j^1 = \sum_{i=1}^{P} w_{j,i}^1 x_i + w_{j,0}^1.$$

- Vi förutsätter att alla vikter är små så att $z_j^1 \ll 1$. Mer kvantitativt skulle vi t.ex. kunna kräva att "små" vikter uppfyller $\max\left(w_{j,i}^1\right) \ll 1/(P\bar{x})$, där $\bar{x}$ är en typisk storleksordning på insignalerna. Då skall även bias uppfylla $w_{j,0}^1 \ll 1/P$.
- Utsignalen från nod j i det gömda lagret ges enligt uppgift av sigmoidfunktionen

$$y_j^1 = \frac{e^{z_j^1}}{1 + e^{z_j^1}}.$$

  För små vikter (dvs $z_j^1 \ll 1$) kan vi Taylorutveckla exponenten runt noll och försumma termer som är kvadratiska i z_j^1. Detta ger

$$y_j^1 \approx \frac{1 + z_j^1}{2 + z_j^1} \approx \frac{1}{2}\left(1 + z_j^1\right)\left(1 - \frac{z_j^1}{2}\right) \approx \frac{1}{2}\left(1 + \frac{z_j^1}{2}\right).$$

- Utlagret innehåller en enda nod. Dess aktivering är

$$z_1^2 = \sum_{j=1}^{M} w_{1,j}^2 y_j^1 + w_{1,0}^2 \approx \sum_{j=1}^{M}\sum_{i=1}^{P} \frac{1}{2} w_{1,j}^2 \left(1 + \frac{w_{j,i}^1 x_i + w_{j,0}^1}{2}\right) + w_{1,0}^2,$$

  där vi har använt approximationen ovan i det andra steget.
- Enligt uppgift är utsignalen linjär i nodens aktivering. Dvs $y_1^2 = z_1^2$. Genom att studera uttrycket ovan för z_1^2 ser vi att vi kan skriva utsignalen som en linjär funktion av prediktorvariablerna

$$y_1^2 = \sum_{i=1}^{P} a_i x_i + b,$$

  där $a_i = a_i(\boldsymbol{w}^1, \boldsymbol{w}^2)$ och $b = b(\boldsymbol{w}^1, \boldsymbol{w}^2)$ är koefficienter som är (mer eller mindre komplicerade) funktioner av vikterna.