Tentamen – Bayesiansk inferens och maskininlärning (TIF385)

Tid och plats:	20 augusti, 2025, 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. Betrakta ett neuralt nätverk med två insignaler, två gömda lager bestående av tio noder vardera och ett utlager med en enda nod. Samtliga noder använder samma aktiveringsfunktion $f(z)$ och inkluderar en biasvikt vid beräkning av z.

   (a) Hur många fria parametrar har det neurala nätverket?

   (b) Inför lämpliga beteckningar och ge ett uttryck för utsignalen.

   (c) Diskutera syftet med viktregularisering för ett neuralt nätverk och ge ett konkret exempel på hur detta skulle kunna implementeras.

   (3 poäng per korrekt besvarad deluppgift, 10 poäng för alla tre.)

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. Betrakta en prediktion $\hat{y}^{\odot} \equiv \hat{y}(\boldsymbol{x}^{\odot})$ med en maskininlärningsmodell som har tränats med en uppsättning träningsdata. Denna prediktion skall jämföras med en framtida observation $y^{\odot} \equiv y(\boldsymbol{x}^{\odot})$. Alla observationer är behäftade med ett irreducibelt fel, dvs $y(\boldsymbol{x}) = f(\boldsymbol{x}) + \epsilon$ där $f(\boldsymbol{x})$ är den underliggande datagenererande processen ("sanningen"). Vi kan anta att observationsfelen är oberoende och identiskt fördelade med $\mathbb{E}[\epsilon] = 0$ och $\mathrm{Var}[\epsilon] = \sigma_\epsilon^2$.

   Visa att (kvadraten av) det förväntade prediktionsfelet kan skrivas som en summa av tre termer

$$\mathbb{E}\left[\left(y^{\odot} - \hat{y}^{\odot}\right)^2\right] = \left(f^{\odot} - \mathbb{E}\left[\hat{y}^{\odot}\right]\right)^2 + \mathrm{Var}\left[\hat{y}^{\odot}\right] + \sigma_\epsilon^2$$

   där $f^{\odot} = f(\boldsymbol{x}^{\odot})$. Förklara kort innebörden av varje term. *(10 poäng)*

4. 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) Träningen genomförs ofta iterativt med gradientsteg. Förklara vad som menas med begreppet "epok" i detta sammanhang och förklara hur en epok genomförs i praktiken med en grad av slumpmässighet när datamängden är väldigt stor.

   (c) Ge tre exempel på användbara felfunktioner (med matematisk definition) varav minst en är specifik för klassificeringsalgoritmer. Dina tre exempel skall vara olika på så sätt att de inte garanteras att ha samma optimum vid träning mot en given uppsättning träningsdata.

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

5. Betrakta en stationär Markovkedja med tre möjliga utfall.

   (a) Visualisera Markovkedjan grafiskt med ett så kallat övergångsdiagram där tillstånden i kedjan visas som noder och övergångssannolikheterna som riktade kanter mellan noderna. Ge lämpliga beteckningar för alla övergångssannolikheter. *(2 poäng)*

(b) Använd matris-vektor-formalism för att beskriva övergången mellan två på varandra följande tillstånd i Markovkedjan. *(3 poäng)*

(c) Ge ett konkret exempel på en situation som uppfyller detaljerad balans. För full poäng skall samtliga sannolikheter i ditt exempel vara nollskilda och ges av rationella tal (kvoter av två heltal). Vilken egenskap uppfyller det tillstånd som du har valt i ditt exempel? *(5 poäng)*

6. Du skall göra en Bayesiansk parameteruppskattning av livstiden τ för ett radioaktivt sönderfall. Sannolikheten att en atom med livstiden τ sönderfaller vid tidpunkten t ges av

$$p\left(t \mid \tau, I\right) = \frac{1}{\tau} e^{-t/\tau}.$$

En mätning av N stycken atomers sönderfallstider ger $\mathcal{D} = \{t_1, \ldots, t_N\}$.

(a) Vad är trolighetsfunktionen för denna datamängd? *(2 poäng)*

(b) Vad blir a-posteriori-fördelningen givet en a-priori-fördelning $p\left(\tau \mid I\right) \propto 1/\tau$ för $\tau \in [\tau_{\min}, \tau_{\max}]$ och noll annars? Ge ett matematiskt uttryck för normaliseringskonstanten, men du behöver inte lösa integralen. *(2 poäng)*

(c) Ge uttryck för förväntansvärdet och standardavvikelsen för τ givet datamängden. Antag för enkelhets skull att vi kan approximera $\tau_{\min} \to 0$ och $\tau_{\max} \to \infty$ i relevanta integraler. Hur stor datamängd behöver vi för att kunna uppskatta τ med en standardavvikelse som är 10% av förväntansvärdet? *(6 poäng)*

Ledtråd. För deluppgift (c) kan följande integral vara av intresse: $I(\alpha, \beta) \equiv \int_0^\infty \tau^{-\alpha-1} e^{-\beta/\tau}\, d\tau = \beta^{-\alpha}\, \Gamma(\alpha)$ för $\alpha > 0, \beta > 0$. Gammafunktionen $\Gamma(\alpha) = (\alpha - 1)!$ om argumentet är ett positivt heltal.

Tentamen – Bayesiansk inferens och maskininlärning (TIF385)

Tid och plats:	20 augusti, 2025, 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. Betrakta ett neuralt nätverk med två insignaler, två gömda lager bestående av tio noder vardera och ett utlager med en enda nod. Samtliga noder använder samma aktiveringsfunktion $f(z)$ och inkluderar en biasvikt vid beräkning av z.

   (a) Hur många fria parametrar har det neurala nätverket?

   (b) Inför lämpliga beteckningar och ge ett uttryck för utsignalen.

   (c) Diskutera syftet med viktregularisering för ett neuralt nätverk och ge ett konkret exempel på hur detta skulle kunna implementeras.

(3 poäng per korrekt besvarad deluppgift, 10 poäng för alla tre.)

Lösning:

(a) Två insignaler samt en bias till varje nod i det första gömda lagret ger $3 \cdot 10 = 30$ vikter. Tio insignaler samt en bias till varje nod i det andra gömda lagret ger $11 \cdot 10 = 110$ vikter. Tio insignaler och en bias till den enda noden i utlagret ger 11 vikter. Totalt: 151 parametrar.

(b) Introducera matriserna $\boldsymbol{W}^{(l)}$ för vikterna i lager l (de gömda lagren motsvarar $l = 1$, $l = 2$ och utlagret $l = 3$). Dessa matriser har element $W^{(l)}_{i,j}$ där varje nod representeras av en kolumn, dvs index j motsvarar nodnummer. Vi kan notera att viktmatriserna för de gömda lagren har storlekarna 2×10 och 10×10 samt för utlagret 10×1.

Vidare introduceras radvektorerna $\boldsymbol{b}^{(l)}$ som innehåller alla biasvikter för nod l.

Insignaler samlas i radvektorn $\boldsymbol{x} = (x_1, x_2)$. Utsignalerna från respektive nod betecknas av radvektorn $\boldsymbol{y}^{(l)}$ medan aktiveringarna för respektive nod betecknas av radvektorn $\boldsymbol{z}^{(l)}$. Notera att bägge dessa har längden 1 för utlagret med en enda nod och vi använder därför skalär notation för dessa.

Aktiveringsfunktionen betecknas $f(z)$ och ger en vektor om argumentet är en vektor. Då fås slutligen utsignalen $y^{(3)} = f(z^{(3)})$ där

$$
\begin{aligned}
z^{(3)} &= \boldsymbol{y}^{(2)} \cdot \boldsymbol{W}^{(3)} + b^{(3)} \\
\boldsymbol{y}^{(2)} &= f(\boldsymbol{z}^{(2)}) \\
\boldsymbol{z}^{(2)} &= \boldsymbol{y}^{(1)} \cdot \boldsymbol{W}^{(2)} + \boldsymbol{b}^{(2)}. \\
\boldsymbol{y}^{(1)} &= f(\boldsymbol{z}^{(1)}) \\
\boldsymbol{z}^{(1)} &= \boldsymbol{x} \cdot \boldsymbol{W}^{(1)} + \boldsymbol{b}^{(1)}.
\end{aligned}
$$

(c) Viktregularisering syftar till att motverka överanpassning mot träningsdata genom att eliminera lokala optima där modellparametrarna tar stora värden. Detta kan åstadkommas genom att lägga till en term i kostnadsfunktionen, till exempel av formen $\lambda \sum_i w_i^2$ där summan löper över alla modellparametrar (vikter och biasvikter) och λ är en hyperparameter.

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

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

  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. Betrakta en prediktion $\hat{y}^\odot \equiv \hat{y}(\boldsymbol{x}^\odot)$ med en maskininlärningsmodell som har tränats med en uppsättning träningsdata. Denna prediktion skall jämföras med en framtida observation $y^\odot \equiv y(\boldsymbol{x}^\odot)$. Alla observationer är behäftade med ett irreducibelt fel, dvs $y(\boldsymbol{x}) = f(\boldsymbol{x}) + \epsilon$ där $f(\boldsymbol{x})$ är den underliggande datagenererande processen ("sanningen"). Vi kan anta att observationsfelen är oberoende och identiskt fördelade med $\mathbb{E}[\epsilon] = 0$ och $\mathrm{Var}[\epsilon] = \sigma_\epsilon^2$.

Visa att (kvadraten av) det förväntade prediktionsfelet kan skrivas som en summa av tre termer

$$\mathbb{E}\left[\left(y^\odot - \hat{y}^\odot\right)^2\right] = \left(f^\odot - \mathbb{E}\left[\hat{y}^\odot\right]\right)^2 + \mathrm{Var}\left[\hat{y}^\odot\right] + \sigma_\epsilon^2$$

där $f^\odot = f(\boldsymbol{x}^\odot)$. Förklara kort innebörden av varje term. *(10 poäng)*

Lösning:
Detta är bevis 1 från bevislistan (se kurshemsida och hänvisning till kompendiet).

---

4. 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) Träningen genomförs ofta iterativt med gradientsteg. Förklara vad som menas med begreppet "epok" i detta sammanhang och förklara hur en epok genomförs i praktiken med en grad av slumpmässighet när datamängden är väldigt stor.

(c) Ge tre exempel på användbara felfunktioner (med matematisk definition) varav minst en är specifik för klassificeringsalgoritmer. Dina tre exempel skall vara olika på så sätt att de inte garanteras att ha samma optimum vid träning mot en given uppsättning träningsdata.

(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) En träningsepok innebär att all data har använts exakt en gång för justering av parametrarna. För stora datamängder genomförs stokastisk optimering vilket innebär upprepade gradientsteg av modellparametrarna med avseende på (slumpmässigt uppdelade) delmängder av träningsdatan $\mathcal{D}_{\text{train},i} \subset \mathcal{D}_{\text{train}}$. En epok innebär då flera gradientsteg där samtliga sådana delmängder har använts exakt en gång.

(c) För regression används ofta kvadratfelet $E(\hat{y}, y) = (\hat{y} - y)^2$ eller absolutfelet $E(\hat{y}, y) = |\hat{y} - y|$. För klassificering används ofta felklassificering som ett mått $E(\hat{y}, y) = 1 - \delta_{\hat{y},y}$.

5. Betrakta en stationär Markovkedja med tre möjliga utfall.

   (a) Visualisera Markovkedjan grafiskt med ett så kallat övergångsdiagram där tillstånden i kedjan visas som noder och övergångssannolikheterna som riktade kanter mellan noderna. Ge lämpliga beteckningar för alla övergångssannolikheter. *(2 poäng)*

   (b) Använd matris-vektor-formalism för att beskriva övergången mellan två på varandra följande tillstånd i Markovkedjan. *(3 poäng)*

   (c) Ge ett konkret exempel på en situation som uppfyller detaljerad balans. För full poäng skall samtliga sannolikheter i ditt exempel vara nollskilda och ges av rationella tal (kvoter av två heltal). Vilken egenskap uppfyller det tillstånd som du har valt i ditt exempel? *(5 poäng)*

Lösning:

(a) Bild med tre noder och pilar mellan alla. Glöm inte pilarna för övergångarna $1 \to 1$, $2 \to 2$, $3 \to 3$. Lämpliga beteckning för övergångssannolikhet är T_{ij} för sannolikheten att gå från tillstånd i till j.

(b) Med övergångsmatrisen $\boldsymbol{T}$ och tillståndet $\boldsymbol{\pi}$ definierade som

$$\boldsymbol{T} = \begin{pmatrix} T_{11} & T_{12} & T_{13} \\ T_{21} & T_{22} & T_{23} \\ T_{31} & T_{32} & T_{33} \end{pmatrix} \qquad \boldsymbol{\pi} = \begin{pmatrix} \pi_1 & \pi_2 & \pi_3 \end{pmatrix}$$

får vi övergången mellan tillståndet $\boldsymbol{\pi}_N$ och $\boldsymbol{\pi}_{N+1}$ enligt

$$\boldsymbol{\pi}_{N+1} = \boldsymbol{\pi}_N \boldsymbol{T}.$$

(c) Detaljerad balans innebär att $\pi_i T_{ij} = \pi_j T_{ji}$ för alla i, j, där vi skall förstå π_i som element i tillståndsvektorn. Vidare har vi normaliseringarna $\sum_i \pi_i = 1$ samt $\sum_j T_{ij} = 1$ (för alla i).

Ett enkelt val som uppfyller detaljerad balans och där alla sannolikheter är nollskilda är $\pi_1 = \pi_2 = \pi_3 = 1/3$ samt $T_{ij} = 1/3$ för alla i, j. Dessa val uppfyller alla villkor.

Det inses att detta tillstånd är ett stationärtillstånd för den Markovkedja som ges av detta val av $\boldsymbol{T}$. Dvs $\boldsymbol{\pi} = \boldsymbol{\pi}\boldsymbol{T}$.

6. Du skall göra en Bayesiansk parameteruppskattning av livstiden τ för ett radioaktivt sönderfall. Sannolikheten att en atom med livstiden τ sönderfaller vid tidpunkten t ges av

$$p\left(t \mid \tau, I\right) = \frac{1}{\tau} e^{-t/\tau}.$$

En mätning av N stycken atomers sönderfallstider ger $\mathcal{D} = \{t_1, \ldots, t_N\}$.

(a) Vad är trolighetsfunktionen för denna datamängd? *(2 poäng)*

(b) Vad blir a-posteriori-fördelningen givet en a-priori-fördelning $p\left(\tau \mid I\right) \propto 1/\tau$ för $\tau \in [\tau_{\min}, \tau_{\max}]$ och noll annars? Ge ett matematiskt uttryck för normaliseringskonstanten, men du behöver inte lösa integralen. *(2 poäng)*

(c) Ge uttryck för förväntansvärdet och standardavvikelsen för τ givet datamängden. Antag för enkelhets skull att vi kan approximera $\tau_{\min} \to 0$ och $\tau_{\max} \to \infty$ i relevanta integraler. Hur stor datamängd behöver vi för att kunna uppskatta τ med en standardavvikelse som är 10% av förväntansvärdet? *(6 poäng)*

Ledtråd. För deluppgift (c) kan följande integral vara av intresse: $I(\alpha, \beta) \equiv \int_0^\infty \tau^{-\alpha-1} e^{-\beta/\tau}\, d\tau = \beta^{-\alpha}\, \Gamma(\alpha)$ för $\alpha > 0, \beta > 0$. Gammafunktionen $\Gamma(\alpha) = (\alpha - 1)!$ om argumentet är ett positivt heltal.

Lösning:

(a) Eftersom de uppmätta sönderfallstiderna är oberoende får vi trolighetsfunktionen som en produkt

$$p\left(\mathcal{D} \mid \tau, I\right) = \tau^{-N} e^{-N\bar{t}/\tau},$$

där $\bar{t} \equiv \sum_{i=1}^{N} t_i/N$, dvs genomsnittstiden.

(b) A-posteriori-fördelningen blir proportionell mot produkten av a-priori och trolighetsfunktion. Notera gränserna.

$$p\left(\tau \mid \mathcal{D}, I\right) = \frac{1}{Z} \tau^{-N-1} e^{-N\bar{t}/\tau} \text{ för } \tau_{\min} < \tau < \tau_{\max}$$

och noll annars. Normaliseringskonstanten Z ges av

$$Z = \int_{\tau_{\min}}^{\tau_{\max}} \tau^{-N-1} e^{-N\bar{t}/\tau} d\tau.$$

(c) Man kan visa att de relevanta integralerna har följande form:

$$\int_0^\infty \tau^{-N+\nu-1} e^{-N\bar{t}/\tau}\, d\tau = I(N-\nu, N\bar{t}) \equiv I_\nu$$

med definitionen $I(\alpha, \beta)$ från ledtråden och med ytterligare en kort notation I_ν. Specifikt blir

- normaliseringen $Z = I_0$ (se b-uppgiften),
- förväntansvärdet $\mathbb{E}[\tau] \equiv \bar{\tau} = I_1/I_0 = \frac{N}{N-1}\bar{t}$,
- variansen $\mathrm{Var}(\tau) \equiv \sigma_\tau^2 = \frac{I_2}{I_0} - \bar{\tau}^2 = \frac{1}{N-2}\bar{\tau}^2$.

För att standardavvikelsen σ_τ skall vara 10% av förväntansvärdet $\bar{\tau}$ behöver vi $N \approx 100$ datapunkter.
