Tentamen – Bayesiansk inferens och maskininlärning (TIF385)

Tid och plats:	17 januari, 2025, fm, Johanneberg.
Hjälpmedel:	Physics Handbook, Beta Mathematics Handbook, typgodkänd kalkylator.
Lösningsskiss:	Andreas Ekström.

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. Vi ska undersöka begreppen *parvis oberoende*, *ömsesidigt oberoende*, och *betingad sannolikhet*.

   (a) Vi har en urna med 4 bollar numrerade 1,2,3,4. Antag att du drar en boll slumpmässigt. Betrakta händelserna

   - X_1: boll 1 eller boll 2 dras
   - X_2: boll 1 eller boll 3 dras
   - X_3: boll 2 eller boll 3 dras

   Beräkna sannolikheterna $P(X_1), P(X_2), P(X_3)$, $P(X_1, X_2), P(X_1, X_3), P(X_2, X_3)$, och $P(X_1, X_2, X_3)$, där $P(X_i, X_j) = P(X_i \cap X_j)$ och $P(X_i, X_j, X_k) = P(X_i \cap X_j \cap X_k)$. *(4 poäng)*

(b) Utnyttja (a) för att visa, genom ett motexempel, att parvis oberoende ej nödvändigtvis medför ömsesidigt oberoende. Två slumpvariabler X_1 och X_2 är *parvis oberoende* om $P(X_1|X_2) = P(X_1)$. Vi kallar n slumpvariabler *ömsesidigt oberoende* om $P(X_i|X_1, \ldots, X_n) = P(X_i)$ för alla delmängder av $\{X_1, \ldots, X_n\} \setminus \{X_i\}$; exempelvis $P(X_1|X_2, X_3) = P(X_1)$ och $P(X_i|X_j, X_k, X_l) = P(X_i)$, osv.[20250117: Korrigerat tryckfel (ömsesidigt oberoende) i första meningen. Meddelat i skrivningssal.] *(6 poäng)*

Lösning:

(a) Räkning av antalet utfall ger $P(X_1) = \frac{|\{1,2\}|}{|\{1,2,3,4\}|} = \frac{2}{4} = \frac{1}{2}$, $P(X_2) = \frac{1}{2}$, och $P(X_3) = \frac{1}{2}$. För $P(X1, X2) = P(X_1 \cap X_2) = \frac{|\{1,2\} \cap \{1,3\}|}{|\{1,2,3,4\}|} = \frac{|\{1\}|}{4} = \frac{1}{4}$, $P(X_1, X_3) = \frac{1}{4}$, och $P(X_2, X_3) = \frac{1}{4}$. För $P(X_1, X_2, X_3)$ finner vi $X_1 \cap X_2 \cap X_3 = \emptyset$ vilket ger $P(X_1, X_2, X_3) = 0$

(b) Med hjälp av produktregeln har vi att $P(X_1, X_2) = P(X_1|X_2)P(X_2)$. För parvis oberoende finner vi att $P(X1, X2) = P(X_1)P(X_2)$. Direkt beräkning visar att sannolikheterna i (a)-uppgiften är parvis oberoende ($\frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2}$). På samma sätt visar en direkt beräkning att sannolikheterna i (a)-uppgiften ej är ömsesidigt oberoende ty $0 = P(X_1, X_2, X_3) \neq \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}$. Vi har alltså konstaterat ett motexempel för påståendet att parvis oberoende ej nödvändigtvis medför ömsesidigt oberoende.

---

2. Vi betraktar en så kallad Bernoulli-Beta modell, med en trolighetsfunktion för en enskild datapunkt x_i:

$$p(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}, \quad x_i \in \{0, 1\},\ \theta \in [0, 1],$$

(dvs vi har diskret data x_i som kan anta värden 0 eller 1 givet en kontinuerlig parameter θ i intervallet $[0, 1]$.), och en (Beta-)prior:

$$p(\theta|\alpha_0, \beta_0) = \mathcal{B}(\theta; \alpha_0, \beta_0) = \frac{1}{N_{\alpha_0,\beta_0}}\theta^{\alpha_0-1}(1-\theta)^{\beta_0-1}.$$

Här är N_{α_0,β_0} en normeringskonstant så att $\int_0^1 \mathcal{B}(\theta; \alpha_0, \beta_0)\, d\theta = 1$ och $\alpha_0 > 0$ och $\beta_0 > 0$ är så kallade hyperparametrar som definierar vår à-priorifördelning.

(a) Visa att à-posteriorfördelningen $p(\theta|D)$ för θ, givet en datamängd $D = \{x_1, x_2, \ldots, x_n\}$ med oberoende och identiskt fördelade datapunkter x_i ges av en Betafördelning $\mathcal{B}(\theta; \alpha_n, \beta_n)$, där $\alpha_n = \alpha_0 + \gamma_{x=1}, \quad \beta_n = \beta_0 + \gamma_{x=0}$, och $\gamma_{x=1} = \sum_i x_i$ och $\gamma_{x=0} = \sum_i (1 - x_i)$.

(b) Härled ett uttryck för sannolikheten $p(x_{n+1} = 1|D)$, dvs att nästa datapunkt $x_{n+1} = 1$ givet datamängden D från (a)-uppgiften. Formulera ditt slutliga uttryck med hjälp av hyperparametrarna α_0,β_0, $\gamma_{x=1}$, och n. (Ledning: Marginalisera över modellparametern θ).

(5 poäng per deluppgift)

Lösning:

(a) Vi söker $p(\theta|D) = \frac{p(D|\theta)p(\theta|\alpha_0,\beta_0)}{p(D)}$. Trolighetsfunktionen kan i det här fallet skrivas

$$p(D|\theta) = \prod_{i=1}^{n} p(x_i|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_i x_i}(1-\theta)^{\sum_i (1-x_i)}.$$

Tillsammans med den givna (Beta-)priorfördelningen har vi

$$p(\theta|D) \propto \theta^{\sum_i x_i}(1-\theta)^{\sum_i (1-x_i)}\theta^{\alpha_0-1}(1-\theta)^{\beta_0-1} = \theta^{\gamma_{x=1}+\alpha_0-1}(1-\theta)^{\gamma_{x=0}+\beta_0-1},$$

där vi har definierat $\gamma_{x=1} = \sum_i x_i$ och $\gamma_{x=0} = \sum_i (1 - x_i)$. Om vi vidare definierar $\alpha_n = \alpha_0 + \gamma_{x=1}, \quad \beta_n = \beta_0 + \gamma_{x=0}$ har vi att $p(\theta|D) \propto \theta^{\alpha_n-1}(1-\theta)^{\beta_n-1}$ vilket i sin tur, efter normering N_{α_n,β_n}, motsvarar en Betafördelning $\mathcal{B}(\theta; \alpha_n, \beta_n)$.

(b) Vi söker

$$p(x_{n+1} = 1|D) = \int p(x_{n+1} = 1, \theta|D)\, d\theta = \int p(x_{n+1} = 1|\theta)p(\theta|D)d\theta$$

där vi utnyttjat att x_{n+1} är betingat oberoende av D. Först skriver vi ut $p(x_{n+1} = 1|\theta) = \theta^1(1-\theta)^{1-1} = \theta$, från vilket vi ser att $p(x_{n+1} = 1|D) = \int_0^1 \theta p(\theta|D)\, d\theta$ är väntevärdet av en Betafördelad $(\mathcal{B}(\theta; \alpha_n, \beta_n))$ variabel θ, som ges av

$$\mathbb{E}(\theta) = \frac{\alpha_n}{\alpha_n + \beta_n} = \frac{\gamma_{x=1} + \alpha_0}{\gamma_{x=1} + \gamma_{x=0} + \alpha_0 + \beta_0} = \frac{\gamma_{x=1} + \alpha_0}{n + \alpha_0 + \beta_0}.$$

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 har en ändlig övergångsmatris T för en Markovkedja som har en stationär fördelning π.

   (a) För vilka k är $U = k(I + T)$ en väldefinierad övergångsmatris? I betecknar en enhetsmatris. *(5 poäng)*

   (b) Vad är den stationära fördelningen för de $U = k(I + T)$ som är väldefinierade övergångsmatriser? *(5 poäng)*

Lösning:

(a) Vi vet att T är en övergångsmatris, dvs alla radsummor är lika med 1 och alla matriselement är större än eller lika med noll. Den andra egenskapen medför att vi måste ha $k > 0$. Den första egenskapen innebär att vi även måste ha $k = 1/2$ ty radsummorna av T är lika med 1, och addition av en identitetsmatris innebär att radsumman annars blir lika med 2.

(b) Från uppgiften vet vi att $\pi = \pi T$. Vi söker den stationära fördelningen π' så att $\pi' = \pi' U$. Från definitionen av U har vi att $2\pi' = \pi'(1 + T) = \pi' + \pi' T$. Vilket medför att $\pi' = \pi' T$, dvs $\pi' = \pi$.

---

5. Betrakta en maskininlärningsmodell som ger förutsägelsen $\hat{y}_i$ att jämföra med träningsdata y_i (för $i = 1 \ldots N$). Vi definierar träningsfelet

$$E_{\text{train}} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i - y_i\right)^2 .$$

(a) Förklara varför E_{train} är ett dåligt mått för att uppskatta prediktionsfelet i en maskininlärningsmodell med ett stort antal parametrar. *(3 poäng)*

(b) Ge ett exempel på ett bättre mått om man vill få en uppskattning av prediktionsfelet (du behöver inte beskriva den detaljerad algoritmen utan det räcker med en till två beskrivande meningar). *(3 poäng)*

(c) Betrakta ett scenario där E_{train} är betydligt större än den noggrannhet som du siktar på för din maskininlärningsmodell. I denna situation är det bortkastat att lägga beräkningsresurser på att faktiskt uppskatta prediktionsfelet. Varför då? *(4 poäng)*

Lösning:

(a) E_{train} kommer alltid att bli lägre när vi ökar modellens komplexitet (genom att större flexibilitet ger större möjlighet att anpassa till träningsdata). Måttet säger alltså ingenting om modellens träffsäkerhet när det gäller ny data.

(b) Vi kan dela upp data i träningsdata och valideringsdata och använda den senare till att uppskatta hur väl modellen generaliserar till ny data. Modellen tränas enbart på träningsdata och vi definierar sedan ett felmått (t.ex. MSE) och beräknar det för valideringsdata.

(c) I allmänhet har vi att modellfelet är större för ny data än för träningsdata som modellen har anpassats till. Vår uppskattning av prediktionsfelet lär därför bli större än träningsfelet, som redan hade konstaterats vara för stort. Det är bättre att träna modellen mer, eller att förbättra modellen och träna om den på nytt.

---

6. Med MCMC kan vi generera kedjor med värden $\theta_1, \theta_2, \theta_3, ...\theta_S$ fördelade enligt komplicerade täthetsfunktioner $p(\theta)$. På grund av begränsningar i MCMC-algoritmen som används kan det uppstå korrelationer mellan MCMC-dragningar separerade med flera steg.

   För att illustrera effekten av korrelationer på den effektiva längden av en MCMC-kedja ska vi nu betrakta två olika, och mycket korta, MCMC-kedjor med endast två dragningar: (θ_1, θ_2) och (λ_1, λ_2). Vi antar

att (λ_1, λ_2) är korrelerade enligt en kovariansmatris

$$\sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

där $\rho > 0$, medan dragningarna (θ_1, θ_2) antas vara okorrelerade, det vill säga att $\rho = 0$.

(a) Visa att
- variansen för summan av två slumpvariabler ges av $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$ *(2 poäng)*
- variansen för en skalad slumpvariabel ges av $\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$ *(2 poäng)*

(b) Beräkna variansen för medelvärdet $\bar{\theta} = \frac{1}{2}(\theta_1 + \theta_2)$ och för medelvärdet $\bar{\lambda} = \frac{1}{2}(\lambda_1 + \lambda_2)$ i respektive MCMC-kedja. Diskutera, med stöd av dina analytiska resultat, effekten av korrelationer ($\rho > 0$ räcker) på informationsinnehållet i en MCMC-kedja. *(6 poäng)*

Lösning:

(a) Vi utnyttjar att väntevärdet är linjärt, dvs $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$ och $\mathbb{E}(aX) = a\mathbb{E}(X)$

–

$$\begin{aligned}&\mathrm{Var}(X + Y) = \mathbb{E}[(X + Y - \mathbb{E}(X + Y))^2] = \\ &\mathbb{E}[(X - \mathbb{E}(X) + Y - \mathbb{E}(Y))^2] = \\ &\mathbb{E}[(X - \mathbb{E}(X))^2 + (Y - \mathbb{E}(Y))^2 + 2(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] = \\ &\mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)\end{aligned}$$

–

$$\begin{aligned}\mathrm{Var}(aX) &= \mathbb{E}[(aX - \mathbb{E}(aX))^2] = \\ &a^2\mathbb{E}[(X - \mathbb{E}(X))^2] = a^2\mathrm{Var}(X)\end{aligned}$$

(b) Vi kan nu visa att

$$\mathrm{Var}(\bar{\theta}) = \frac{1}{4}(\sigma^2 + \sigma^2) = \frac{\sigma^2}{2}$$

och

$$\mathrm{Var}(\bar{\lambda}) = \frac{1}{4}(\sigma^2 + \sigma^2 + 2\rho\sigma^2) = \frac{\sigma^2}{2}(1 + \rho)$$

För okorrelerade dragningar ser vi att variansen av medelvärdet avtar omvänt proportionellt mot antalet dragningar. En korrelation reducerar effektivt antalet MCMC-dragningar i vårt estimat av variansen för medelvärdet, och i fallet med $\rho = 1$ har vi effektivt en MCMC-kedja bestående av en enda dragning. (Fallet då $\rho < 0$, dvs antikorrelerade dragningar, reducerar variansen och kan i extrema fall leda till estimat med noll varians. Diskussion av fallet $\rho < 0$ krävs ej för full poäng.) _