Tentamen – Bayesiansk inferens och maskininlärning (TIF385)

Tid och plats:	3 april, 2024, 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 tio noder i inlagret och en nod i utlagret i kombination med ett antal dolda lager. För träning finns en uppsättning data $\mathcal{D}$ som består av 10^6 instanser med prediktorvariabler och uppmätt respons.

   (a) Ge tre exempel på hyperparametrar för det neurala nätverket.

   (b) Beskriv kortfattat ett tillvägagångssätt för att optimera dessa hyperparametrar.

   (c) Är ditt föreslagna tillvägagångssätt robust under scenariet att antalet hyperparametrar som skall optimeras är stort? Motivera.

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

2. Betrakta en binär klassificerare som beror på två prediktorvariabler (x_1 och x_2) och vars beslutsgräns är en rät linje i (x_1, x_2)-planet.
   (a) Beskriv matematiskt en sådan klassificerare givet att den skall fungera som en hård klassificerare.
   (b) Beskriv matematiskt en sådan klassificerare givet att den skall fungera som en mjuk klassificerare.
   (c) Räcker det att känna den räta linjen som beskriver beslutsgränsen (dvs numeriska värden för k och m i uttrycket $x_2 = kx_1 + m$) för att fullständigt ha bestämt den hårda klassificeraren?
   (d) Räcker det att känna den räta linjen som beskriver beslutsgränsen (dvs numeriska värden för k och m i uttrycket $x_2 = kx_1 + m$) för att fullständigt ha bestämt den mjuka klassificeraren?

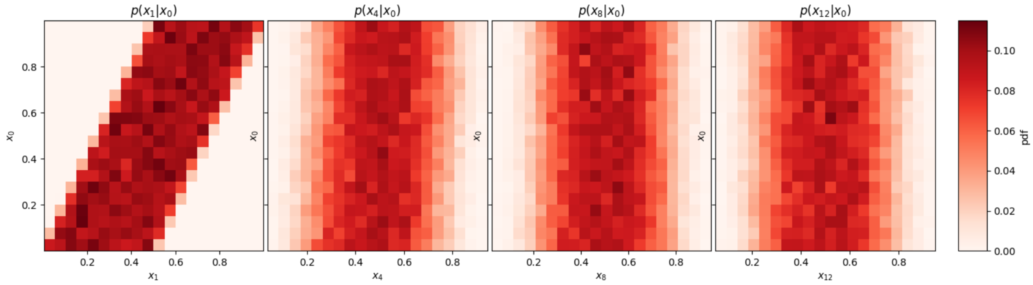
   (3 poäng per deluppgift a och b, plus 2 poäng per deluppgift c och d.)

3. Betrakta en statistisk modell $\mathcal{D} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$, som relaterar N_d observationer i kolumnvektorn $\mathcal{D}$ till en linjär modell som beskrivs av designmatrisen $\boldsymbol{X}$ och parametervektorn $\boldsymbol{\theta}$ samt slumpvariabler $\boldsymbol{\epsilon}$. De sistnämnda beskriver osäkerheter och vi kan anta att elementen är oberoende och identiskt normalfördelade $p(\epsilon_i|I) = \mathcal{N}(0, \sigma_\epsilon^2)$.
   (a) Visa att trolighetsfunktionen blir
   $$p\left(\mathcal{D}|\boldsymbol{\theta}, I\right) = \left(\frac{1}{2\pi\sigma_\epsilon^2}\right)^{N_d/2} \exp\left[-\frac{1}{2}\frac{(\mathcal{D} - \boldsymbol{X}\boldsymbol{\theta})^T(\mathcal{D} - \boldsymbol{X}\boldsymbol{\theta})}{\sigma_\epsilon^2}\right].$$
   (b) Visa därefter att
   $$p\left(\mathcal{D}|\boldsymbol{\theta}, I\right) \propto \exp\left[-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T\Sigma_\theta^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\right],$$
   där $\boldsymbol{\theta}^* = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\mathcal{D}$ är lösningen till normalekvationen och $\Sigma_\theta^{-1} = \boldsymbol{X}^T\boldsymbol{X}/\sigma_\epsilon^2$.

   Ledtråd: Definiera $L(\boldsymbol{\theta}) = (\mathcal{D} - \boldsymbol{X}\boldsymbol{\theta})^T(\mathcal{D} - \boldsymbol{X}\boldsymbol{\theta})/(2\sigma_\epsilon^2)$ och notera att gradientvektorn ges av $\boldsymbol{\nabla}_{\boldsymbol{\theta}} L = \boldsymbol{X}^T\left(\mathcal{D} - \boldsymbol{X}\boldsymbol{\theta}\right)/\sigma_\epsilon^2$ samt att Hessianmatrisen med andraderivator är $\boldsymbol{H} = \boldsymbol{X}^T\boldsymbol{X}/\sigma_\epsilon^2$.

   (10 poäng)

4. Betrakta en diskret Markovkedja $\{X_n : n \in \mathbb{Z}^{0+}\}$ med utfallsrummet $x_n \in [0,1]$. Figuren visar histogram över betingade sannolikheter för utfallen för några av de första slumpvariablerna.



(a) Motivera hur man kan se att Markovkedjan verkar ha en gränsfördelning.

(b) Vad är en matematisk definition på en gränsfördelning för en Markovkedja med ett diskret utfallsrum?

(c) Så kallat reversibla Markovkedjor är alltid stationära, vilket är ett grundkrav för att de skall kunna ha en gränsfördelning. För att konstruera en reversibel Markovkedja är det oftast enklare att fokusera på att den uppfyller detaljerad balans, vilket är ett tillräckligt (men inte nödvändigt) krav. Beskriv detaljerad balans matematiskt samt med ord (för ett diskret utfallsrum).

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

5. Betrakta den diskreta Poissonfördelningen (med $\lambda = 1$)

$$p(k) = \frac{e^{-1}}{k!}, \quad k = 0, 1, 2, \ldots$$

Konstruera och beskriv en Metropolis-Hastings-algoritm som vid konvergens kommer att generera stickprov ur denna fördelning. Utgå från en slumpvandring med steglängden 1 och beräkna de resulterande elementen $T(i,j)$ i övergångsmatrisen för $i, j \in \{0, 1, 2\}$. Notera dock att utfallsrummet är $\mathbb{Z}^{0+}$ (dvs alla icke-negativa heltal). *(10 poäng)*

6. Betrakta en modell $M(\boldsymbol{\theta}, \boldsymbol{x})$, med parametrar $\boldsymbol{\theta} \in \Omega$ och oberoende variabler $\boldsymbol{x}$. Modellen beskriver ett fysikaliskt samband (eller mer allmänt en datagenererande process). Vidare tänker vi oss att vi har tillgång till observerad data $\mathcal{D}$ som motsvarar ett experimentellt tillgängligt mätområde $\boldsymbol{x} \in X_D$ men att vi huvudsakligen är intresserade av data $\mathcal{F}$ som motsvarar ett otillgängligt mätområde $\boldsymbol{x} \in X_F$. I detta läge kan vi använda vår modell och vår tillgängliga data för att konstruera en a

posteriori sannolikhetsfördelning för modellförutsägelser (det vill säga en PPD=posterior predictive distribution)

$$p\left(\mathcal{F} \,|\, \mathcal{D}, I\right).$$

(a) Ge tre exempel på situationer gällande $X_F, X_D, \mathcal{D}$ och/eller $M(\boldsymbol{\theta}, \boldsymbol{x})$, som leder till en dålig precision i modellförutsägelsen för $\mathcal{F}$.

(b) Härled följande integraluttryck för PPD:n

$$p\left(\mathcal{F} \,|\, \mathcal{D}, I\right) = \int_\Omega p\left(\mathcal{F} \,|\, \boldsymbol{\theta}, I\right) p\left(\boldsymbol{\theta} \,|\, \mathcal{D}, I\right) d\boldsymbol{\theta}.$$

(10 poäng)

Tentamen – Bayesiansk inferens och maskininlärning (TIF385)

Tid och plats: 3 april, 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. Betrakta ett neuralt nätverk med tio noder i inlagret och en nod i utlagret i kombination med ett antal dolda lager. För träning finns en uppsättning data $\mathcal{D}$ som består av 10^6 instanser med prediktorvariabler och uppmätt respons.

   (a) Ge tre exempel på hyperparametrar för det neurala nätverket.

   (b) Beskriv kortfattat ett tillvägagångssätt för att optimera dessa hyperparametrar.

   (c) Är ditt föreslagna tillvägagångssätt robust under scenariet att antalet hyperparametrar som skall optimeras är stort? Motivera.

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

Lösning:

(a) (1) Antalet gömda lager; (2) Bredden på de gömda lagren; (3) Viktregularisering λ; (4) Val (och parametrisering) av aktiveringsfunktion; (5) Andel av data som används för träning; etc.

(b) Man kan göra en rutnätsoptimering (grid search) där man väljer ett ändligt antal alternativ för varje hyperparameter och sedan skapar alla möjliga kombinationer av dessa. För varje kombination tränar man sitt neurala nätverk på samma sätt och med samma träningsdata. Därefter beräknar man ett valideringsresultat från valideringsdata. Den optimala uppsättningen med hyperparametrar är den som har bäst valideringsresultat. Notera att man även vill behålla en uppsättning med testdata som inte används för vare sig träning eller validering. Denna kan på slutet användas för att uppskatta generaliseringsfelet för den slitliga modellen.

(c) Nej, detta tillvägagångssätt är inte robust. Antalet möjliga kombinationer växer exponentiellt med antalet hyperparametrar, och en punktbaserad optimeringsmetod kommer inte att vara effektiv.

---

2. Betrakta en binär klassificerare som beror på två prediktorvariabler (x_1 och x_2) och vars beslutsgräns är en rät linje i (x_1, x_2)-planet.

   (a) Beskriv matematiskt en sådan klassificerare givet att den skall fungera som en hård klassificerare.

   (b) Beskriv matematiskt en sådan klassificerare givet att den skall fungera som en mjuk klassificerare.

   (c) Räcker det att känna den räta linjen som beskriver beslutsgränsen (dvs numeriska värden för k och m i uttrycket $x_2 = kx_1 + m$) för att fullständigt ha bestämt den hårda klassificeraren?

   (d) Räcker det att känna den räta linjen som beskriver beslutsgränsen (dvs numeriska värden för k och m i uttrycket $x_2 = kx_1 + m$) för att fullständigt ha bestämt den mjuka klassificeraren?

(3 poäng per deluppgift a och b, plus 2 poäng per deluppgift c och d.)

Lösning:

I det följande använder vi en aktivering $z = w_0 + w_1x_1 + w_2x_2$

(a) $\hat{y} = \frac{\mathrm{sign}(z)+1}{2}$ vilken ger utsignalen 0 eller 1, dvs den kategoriserar utan osäkerhet. Beslutsgränsen, där utfallet är obestämt, fås då $z = 0$ vilket blir en rät linje i (x_1, x_2)-planet.

(b) En mjuk klassificerare ger istället en sannolikhet att utfallet tillhör den ena klassen. Det vanligaste är en sigmoid form:

$$\hat{y} = \mathbb{P}(y(\boldsymbol{x}) = 1) = \frac{1}{1+e^{-z}}.$$

(c) Ja det gör det. Den räta linjen ger ett samband mellan w_1 och w_2 och kan bestämma w_0 (den ger helt enkelt två villkor). Ett tredje villkor behövs för att bestämma samtliga tre parametrar w_0, w_1, w_2 och därmed fullständigt återskapa klassificeraren. Detta villkor kan formuleras som en övergripande normaliseringskonstant. Eftersom sign() funktionen är oberoende av en övergripande konstant så är denna skalning av parametrarna irrelevant.

(d) Nej det gör det inte. Resonemanget är samma som tidigare, men den övergripande normaliseringen av z kommer att bestämma nivåkurvorna för utsignalen. Den kommer exempelvis att bestämma exakt var i (x_1, x_2)-planet som sannolikhetslinjerna $\mathbb{P}(y(\boldsymbol{x}) = 1) = 0.25$ och $\mathbb{P}(y(\boldsymbol{x}) = 1) = 0.75$ går. Vi behöver alltså någon sådan information för att återskapa klassificeraren.

---

3. Betrakta en statistisk modell $\mathcal{D} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$, som relaterar N_d observationer i kolumnvektorn $\mathcal{D}$ till en linjär modell som beskrivs av designmatrisen $\boldsymbol{X}$ och parametervektorn $\boldsymbol{\theta}$ samt slumpvariabler $\boldsymbol{\epsilon}$. De sistnämnda beskriver osäkerheter och vi kan anta att elementen är oberoende och identiskt normalfördelade $p(\epsilon_i|I) = \mathcal{N}(0, \sigma_\epsilon^2)$.

(a) Visa att trolighetsfunktionen blir

$$p\left(\mathcal{D}|\boldsymbol{\theta}, I\right) = \left(\frac{1}{2\pi\sigma_\epsilon^2}\right)^{N_d/2} \exp\left[-\frac{1}{2}\frac{(\mathcal{D}-\boldsymbol{X}\boldsymbol{\theta})^T(\mathcal{D}-\boldsymbol{X}\boldsymbol{\theta})}{\sigma_\epsilon^2}\right].$$

(b) Visa därefter att

$$p\left(\mathcal{D}|\boldsymbol{\theta}, I\right) \propto \exp\left[-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^T\Sigma_\theta^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\right],$$

där $\boldsymbol{\theta}^* = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\mathcal{D}$ är lösningen till normalekvationen och $\Sigma_\theta^{-1} = \boldsymbol{X}^T\boldsymbol{X}/\sigma_\epsilon^2$.

Ledtråd: Definiera $L(\boldsymbol{\theta}) = (\mathcal{D} - \boldsymbol{X}\boldsymbol{\theta})^T(\mathcal{D} - \boldsymbol{X}\boldsymbol{\theta})/(2\sigma_\epsilon^2)$ och notera att gradientvektorn ges av $\boldsymbol{\nabla}_{\boldsymbol{\theta}} L = \boldsymbol{X}^T(\mathcal{D} - \boldsymbol{X}\boldsymbol{\theta})/\sigma_\epsilon^2$ samt att Hessianmatrisen med andraderivator är $\boldsymbol{H} = \boldsymbol{X}^T\boldsymbol{X}/\sigma_\epsilon^2$.

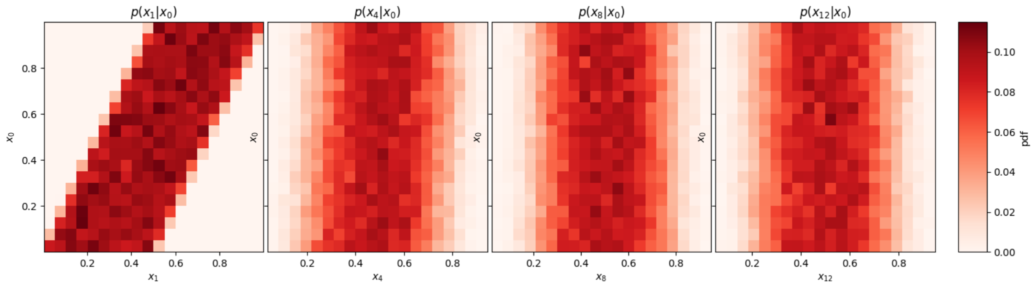
(10 poäng)

Lösning:

Detta är bevis 3 från bevislistan (se kurshemsida och hänvisning till kompendiet).

---

4. Betrakta en diskret Markovkedja $\{X_n : n \in \mathbb{Z}^{0+}\}$ med utfallsrummet $x_n \in [0, 1]$. Figuren visar histogram över betingade sannolikheter för utfallen för några av de första slumpvariablerna.



(a) Motivera hur man kan se att Markovkedjan verkar ha en gränsfördelning.

(b) Vad är en matematisk definition på en gränsfördelning för en Markovkedja med ett diskret utfallsrum?

(c) Så kallat reversibla Markovkedjor är alltid stationära, vilket är ett grundkrav för att de skall kunna ha en gränsfördelning. För att konstruera en reversibel Markovkedja är det oftast enklare att fokusera på att den uppfyller detaljerad balans, vilket är ett tillräckligt (men inte nödvändigt) krav. Beskriv detaljerad balans matematiskt samt med ord (för ett diskret utfallsrum).

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

Lösning:

(a) Vi kan se på den sista panelen att $p(x_{12} \mid x_0)$ verkar vara oberoende av x_0; dvs allt minne av utfallet i början av kedjan verkar ha försvunnit och fördelningen av utfall för senare slumpvariabler verkar nå en stabil gränsfördelning. Notera dock att vi inte kan vara helt säkra eftersom resultatet måste vara oberoende av hur startfördelningen för X_0 ser ut.

(b) Sannolikhetsfördelningen π är en gränsfördelning till Markovkedjan som beskrivs av övergångsmatrisen T om

$$\pi = \lim_{n\to\infty} \alpha T^n,$$

oberoende av startfördelning α.

(c) En kedja i detaljerad balans uppfyller

$$\pi_k T(k,j) = \pi_j T(j,k),$$

för alla utfall j, k i utfallsrummet. Det är därför ett lokalt samband och beskriver en balans av sannolikhetsflöde från j till k och från k till j.

---

5. Betrakta den diskreta Poissonfördelningen (med $\lambda = 1$)

$$p(k) = \frac{e^{-1}}{k!}, \quad k = 0, 1, 2, \ldots$$

Konstruera och beskriv en Metropolis-Hastings-algoritm som vid konvergens kommer att generera stickprov ur denna fördelning. Utgå från en slumpvandring med steglängden 1 och beräkna de resulterande elementen $T(i,j)$ i övergångsmatrisen för $i, j \in \{0, 1, 2\}$. Notera dock att utfallsrummet är $\mathbb{Z}^{0+}$ (dvs alla icke-negativa heltal). *(10 poäng)*

Lösning:

- Vi utgår från den symmetriska stegförslagsfunktionen $S(i,j) = 0.5$ för $j = i \pm 1$ (och noll annars). Dvs en slumpmässig vandring med steglängden 1.
- De relevanta elementen i acceptansfunktionen

$$A(i,j) = \max\left(1, \frac{p(j)}{p(i)}\right)$$

är: $A(i, i-1) = 1$, $A(i, i+1) = 1/(i+1)$.
Vi definierar också elementet $A(0, -1) = 0$ eftersom stegfunktionen kan föreslå ett steg från $i = 0$ till $j = -1$ vilket dock inte är ett tillåtet utfall.

- Icke-diagonala element i övergångsmatrisen $T(i,j) = S(i,j)A(i,j)$. De diagonala ges av $T(i,i) = 1 - \sum_{j \neq i} T(i,j)$.
- Låt oss betrakta den första raden som beskriver övergångar från tillstånd 0: $T(0,1) = 0.5 \cdot 1 = 1/2$, $T(0,2) = 0 \cdot A(0,2) = 0$, $T(0,0) = 1 - 1/2 - 0 - 0 - \ldots = 1/2$.
- Även om följande element inte efterfrågades så behövs det för att få det diagonala $T(2,2)$-elementet: $T(2,3) = 0.5 \cdot 1/3 = 1/6$.

Ovanstående ger

$$T = \begin{pmatrix} 1/2 & 1/2 & 0 & \ldots \\ 1/2 & 1/4 & 1/4 & \\ 0 & 1/2 & 1/3 & \\ \vdots & & & \ddots \end{pmatrix}$$

Vi beskriver algoritmen med pseudokod:

(a) Initialisera kedjan med något tillstånd. För att vara konkreta väljer vi $x_0 = 1$. Detta är alltså iteration $n = 0$.

(b) Vid iteration $n \geq 1$ med $x_{n-1} = i$:

   i. Om $i \geq 1$: Singla en rättvis slant.
      - Om utfallet är krona, $x_n = i - 1$
      - Annars om utfallet är klave, dra ett slumptal $a \sim \mathcal{U}([0,1])$: $x_n = i + 1$ om $a \leq 1/(i+1)$, $x_n = 1$ annars.

   ii. Annars om $i = 0$: Singla en rättvis slant.
      - Om utfallet är krona, $x_n = 0$, annars $x_n = 1$.

Det går även att lösa uppgiften med en stegfunktion som är asymmetrisk för $i = 0$, nämligen $S(0,1) = 1$ och $S(0,j) = 0$ för $j \neq 1$. Då måste man dock komma ihåg att Metropolis-Hastings-kvoten för $(i,j) = (0,1)$ blir

$$r = \frac{p(1)}{p(0)} \frac{S(1,0)}{S(0,1)} = 1/2.$$

Övergångsmatrisen blir därmed likadan som i lösningen ovan.

---

6. Betrakta en modell $M(\boldsymbol{\theta}, \boldsymbol{x})$, med parametrar $\boldsymbol{\theta} \in \Omega$ och oberoende variabler $\boldsymbol{x}$. Modellen beskriver ett fysikaliskt samband (eller mer allmänt en datagenererande process). Vidare tänker vi oss att vi har tillgång till observerad data $\mathcal{D}$ som motsvarar ett experimentellt tillgängligt mätområde $\boldsymbol{x} \in X_D$ men att vi huvudsakligen är intresserade av data

$\mathcal{F}$ som motsvarar ett otillgängligt mätområde $\boldsymbol{x} \in X_F$. I detta läge kan vi använda vår modell och vår tillgängliga data för att konstruera en a posteriori sannolikhetsfördelning för modellförutsägelser (det vill säga en PPD=posterior predictive distribution)

$$p\left(\mathcal{F} \mid \mathcal{D}, I\right).$$

(a) Ge tre exempel på situationer gällande $X_F, X_D, \mathcal{D}$ och/eller $M(\boldsymbol{\theta}, \boldsymbol{x})$, som leder till en dålig precision i modellförutsägelsen för $\mathcal{F}$.

(b) Härled följande integraluttryck för PPD:n

$$p\left(\mathcal{F} \mid \mathcal{D}, I\right) = \int_\Omega p\left(\mathcal{F} \mid \boldsymbol{\theta}, I\right) p\left(\boldsymbol{\theta} \mid \mathcal{D}, I\right) d\boldsymbol{\theta}.$$

(10 poäng)

Lösning:

(a) (1) Kalibreringsdata $\mathcal{D}$ skulle kunna vara behäftad med stora experimentella osäkerheter vilket leder till en stor osäkerhet hos infererade modellparametrar; (2) Avståndet mellan områdena X_D och X_F kan vara stort vilket innebär att små osäkerheter i modellen i det kalibrerande området kan leda till stora osäkerheter i det förutsagda, dvs vi utför en extrapolation av modellen; (3) Ett stort antal modellparametrar i förhållande till mängden kalibreringsdata leder sannolikt till en mer osäker inferens och större osäkerheter i förutsägelser.

(b) Vi utnyttjar först marginalisering av en gemensam fördelning och sedan produktregeln för betingade sannolikheter

$$\begin{aligned}
p\left(\mathcal{F} \mid \mathcal{D}, I\right) &= \int_\Omega p\left(\mathcal{F}, \boldsymbol{\theta} \mid \mathcal{D}, I\right) \\
&= \int_\Omega p\left(\mathcal{F} \mid \boldsymbol{\theta}, \mathcal{D}, I\right) p\left(\boldsymbol{\theta} \mid \mathcal{D}, I\right) d\boldsymbol{\theta} \\
&= \int_\Omega p\left(\mathcal{F} \mid \boldsymbol{\theta}, I\right) p\left(\boldsymbol{\theta} \mid \mathcal{D}, I\right) d\boldsymbol{\theta}.
\end{aligned}$$

Det sista sambandet erhålls genom att $\mathcal{F}$ förutses av modellen givet modellparametrarna. Det betyder att $p\left(\mathcal{F} \mid \boldsymbol{\theta}, \mathcal{D}, I\right)$ är betingat oberoende av $\mathcal{D}$ och därmed kan skrivas på formen som används i det sökta uttrycket.
