(Demo)tentamen – Bayesiansk inferens och maskininlärning (TIF385)

Tid och plats:	13 januari, 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. Antag att det finns en ovanlig sjukdom (OS) som är asymptomatisk under ett tidigt skede. Enbart en på 10,000 personer kan förväntas ha sjukdomen vid ett givet tillfälle. Det finns ett test för vilket andelen falska positiva testresultat är 0.023 och andelen falska negativa testresultat är 0.014.

   Du tar testet och får ett positivt svar. Vad är sannolikheten att du faktiskt har sjukdomen? *(10 poäng)*

2. Betrakta linjär regression $\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\theta}$ med antagandet att residualerna är normalfördelade: $y_i - \hat{y}_i \equiv \epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$.

(a) Motivera varför OLS lösningen

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\arg\min}\, \frac{1}{N}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2,$$

för N träningsdata kallas för "Maximum Likelihood Estimate" (MLE). Notera att $\|\boldsymbol{\epsilon}\|_2^2$ betecknar kvadraten av 2-normen. (*5 poäng*)

(b) Ge en Bayesiansk motivering till "ridge" regularisering för vilken lösningen är

$$\boldsymbol{\theta}^*_{\lambda,2} = \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\arg\min}\, C_{\lambda,2}(\boldsymbol{\theta}),$$

där kostnadsfunktionen

$$C_{\lambda,2}(\boldsymbol{\theta}) = \frac{1}{N}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2^2.$$

Tolka parametern λ. (*5 poäng*)

3. Är Markovkedjan med följande övergångsmatris reversibel?

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

(Utfallsrummet består alltså av tre tillstånd.) *(10 poäng)*

4. Sätt upp en Metropolis-Hastings algoritm för att ta stickprov från följande potenslagsdistribution

$$\pi_i = \frac{i^{-3/2}}{\sum_{k=1}^{\infty} k^{-3/2}},$$

för $i = 1, 2, \ldots$. Du skall beskriva algoritmen, helst med pseudokod eller annars med förklarande text. *(10 poäng)*

5. Betrakta en kNN modell.

(a) Skissa en figur med modellkomplexitet på x axeln som visar områden med under- respektive överanpassning.

(b) Indikera i samma figur systematiska-fel-eller-varians-dilemmat.

(c) Var hamnar extremvalen $k = 1$ och $k = N$ på skalan med modellkomplexitet (där N är antalet träningspunkter). Motivera.

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

6. Motivera behovet av tre olika datamängder:

   (a) Träningsdata

   (b) Valideringsdata

   (c) Testdata

   vid konstruktion av en maskininlärningsmodell. Förklara också varför det kan vara fördelaktigt med k-faldig korsvalidering. *(10 poäng)*