PRACTICE FINAL EXAM

STA414/2104 Winter 2023
Statistical Methods for Machine Learning II

University of Toronto
Faculty of Arts & Science

Duration - 3 hours
Aids allowed: Two double-sided handwritten $8.5'' \times 11''$ or A4 aid sheets.

Exam reminders:

- Fill out your name and student number on the top of this page.
- Do not begin writing the actual exam until the announcements have ended and the Exam Facilitator has started the exam.
- Write all answers in the provided answer booklets.
- Blank scrap paper is provided at the back of the exam.
- If you possess an unauthorized aid during an exam, you may be charged with an academic offence.
- Turn off and place all cell phones, smart watches, electronic devices, and unauthorized study materials in your bag under your desk. If it is left in your pocket, it may be an academic offence.
- When you are done your exam, raise your hand for someone to come and collect your exam. Do not collect your bag and jacket before your exam is handed in.
- If you are feeling ill and unable to finish your exam, please bring it to the attention of an Exam Facilitator so it can be recorded before leaving the exam hall.
- In the event of a fire alarm, do not check your cell phone when escorted outside.

Hand in all examination materials at the end
DO NOT WRITE ANY ANSWERS ON THIS PAPER

1. Decision theory (10 points). Imagine you are writing a quiz that has a true or false section. To discourage random guessing, the quiz awards x points for a correct answer, y points for a false answer, and z points for no answer.

1. (8 points) You think you know the correct answer with probability θ. How high must θ be, as a function of x, y, and z, before the expected number of points is higher for choosing the most likely answer, versus leaving the question blank?
2. (2 points) How high must θ be, before the expected number of points is higher for guessing the correct answer, when $x = 2$, $y = -2$, and $z = 0$?

1. $x\theta + y(1-\theta) > z \Rightarrow \theta > \frac{z-y}{x-y}$

2. $\theta > \frac{2}{4} = 0.5$

2. Variational Inference (10 points). Hint for this section: Jensen's inequality states that when f is concave, $f(\mathbb{E}[z]) \geq \mathbb{E}[f(z)]$.

1. (5 points) For the joint distribution $p(x, z)$, suppose we are trying to approximate a conditional distribution $p(z|x)$ using distribution $q(z|x)$. Show that for any distribution q, the "evidence lower bound"

$$\mathcal{L}(\phi) = \mathbb{E}_{q(z|x)}[\log p(x, z) - \log q(z|x)]$$

will be less than or equal to the log marginal likelihood $\log p(x)$. You can assume p and q are positive everywhere.

$$\log p(x) = \log \int p(x,z)\,dz = \log \int \frac{p(x,z)}{q_\phi(z|x)} q_\phi(z|x)\,dz$$

$$= \log \mathbb{E}_{q_\phi} \frac{p(x,z)}{q_\phi(z|x)}$$

$$\geq \mathbb{E}_{q_\phi} \log \frac{p(x,z)}{q_\phi(z|x)}$$

2. (5 points) If a training set $x_1, x_2, \ldots, x_N$ are drawn i.i.d. from $p(x|\theta)$ and the parameter $\hat{\theta}$ is estimated from the data, show that the expected log-probability of the data under $\hat{\theta}$ will be smaller in expectation on a validation set of data drawn from the same distribution $p(x|\theta)$ than it will be on the training set. That is, show that, for all $\hat{\theta}$,

$$\mathbb{E}_{p(x|\theta)}\left[\log p(x|\hat{\theta})\right] \leq \mathbb{E}_{p(x|\theta)}[\log p(x|\theta)].$$

You can assume p and q are positive everywhere.

$$** = \mathbb{E}_{p(x|\theta)}\left[\log p(x|\hat{\theta})\right] = \mathbb{E}_{p(x|\theta)}\left[\log p(x|\theta)\cdot\frac{p(x|\hat{\theta})}{p(x|\theta)}\right]$$

$$= \mathbb{E}_{p(x|\theta)}\left[\log p(x|\theta)\right] + \underbrace{\mathbb{E}_{p(x|\theta)} \log \frac{p(x|\hat{\theta})}{p(x|\theta)}}_{*}$$

$$* \leq \log \mathbb{E}_{p(x|\theta)}\left[\frac{p(x|\hat{\theta})}{p(x|\theta)}\right] = \log 1 = 0$$

Thus, $** \leq \mathbb{E}_{p(x|\theta)}\left[\log p(x|\theta)\right]$ $\blacksquare$.

3. Monte Carlo Estimators (10 points). Recall the Simple Monte Carlo estimator:

$$\hat{e}(x_1, x_2, \ldots, x_S) = \frac{1}{S}\sum_{i=1}^{S} f(x^{(i)}), \qquad \text{where each } x^{(i)} \sim p(x) \text{ independently.}$$

1. (2 points) Show that this is an unbiased estimator of $\mathbb{E}_{p(x)}[f(x)]$.

$$\mathbb{E}\,\hat{e} = \mathbb{E}\,\frac{1}{S}\sum_i f(x^{(i)}) = \frac{1}{S}\sum_i \mathbb{E} f(x^{(i)})$$

$$= \mathbb{E}[f(x^{(1)})]$$

2. (4 points) Find the variance of this estimator as a function of S.

$$\text{Var}(\hat{e}) = \text{Var}\left(\frac{1}{S}\sum_{i} f(x^{(i)})\right) = \frac{1}{S^2}\sum_{i} \text{Var}(f(x^{(i)}))$$

$$= \frac{1}{S}\text{Var}(f(x^{(i)})$$

3. (4 points) Imagine you have a distribution $p(x)$ whose normalized density you can evaluate, but which it is difficult to sample from. You also have another distribution $q(x)$, that you can sample from, and also evaluate its density. Using these two distributions, write an unbiased estimator of $\mathbb{E}_{p(x)}[f(x)]$ that can be computed without access to samples from $p(x)$.

Sample $x^{(i)} \sim q \qquad i = 1 \ldots n$

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} f(x^{(i)}) \cdot \frac{p(x^{(i)})}{q(x^{(i)})}$$

$$\mathbb{E}\hat{\theta} = \mathbb{E}_{p} f(x)$$

4. Bayesian Linear Regression (10 points). In a linear regression problem, suppose that you are given a dataset $\mathbf{t} \in \mathbb{R}^n$ and $\mathbf{X} \in \mathbb{R}^{n \times d}$ where $n > d$. We assume that target has the following distribution

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \mathcal{N}(\mathbf{t}|\mathbf{X}\mathbf{w}, \beta^{-1}\mathbf{I}).$$

We use the following prior for the weights

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mu, \mathbf{\Sigma}).$$

Derive the posterior distribution $p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \beta)$ by explicitly showing each step.

$$p(w|t, X, \beta) \propto p(t|X, w, \beta) \cdot p(w)$$

$$\propto \exp\left\{-\frac{\beta}{2}\|t - Xw\|^2 - \frac{1}{2}(w-\mu)^T \Sigma^{-1}(w-\mu)\right\}$$

$$\propto \exp\left\{-\frac{1}{2}w^T\left(\beta X^T X + \Sigma^{-1}\right) + \left(\beta X^T t + \Sigma^{-1}\mu\right)^T w\right\}$$

$$\Rightarrow \quad \mathcal{N}(w|m, S) \propto \exp\left\{-\frac{1}{2}w^T S^{-1} w + m^T S^{-1} w\right\}$$

$$S^{-1} = \beta X^T X + \Sigma^{-1}$$

$$S^{-1} m = \beta X^T t + \Sigma^{-1}\mu$$

$$w | t, X, \beta \sim \mathcal{N}\left((\beta X^T X + \Sigma^{-1})^{-1}(\beta X^T t + \Sigma^{-1}\mu), (\beta X^T X + \Sigma^{-1})^{-1}\right)$$

5. Principle Component Analysis (20 points). Suppose that you are given a centered dataset of n samples, i.e., $x_i \in \mathbb{R}^d$ for $i = 1, 2, ..., n$ such that $\sum_{i=1}^n x_i = 0$. For a given unit direction u ($\|u\|_2 = 1$), we denote by $\mathcal{P}_u(x)$ the Euclidean projection of x on u. That is,

$$\mathcal{P}_u(x) = \operatorname*{argmin}_{v=\alpha u\,:\,\alpha\in\mathbb{R}} \|x - v\|_2^2. \tag{5.1}$$

1. (2 points) *Projected data mean:* Show that the projected data in any unit direction u is still centered. That is show,

$$\sum_{i=1}^n \mathcal{P}_u(x_i) = 0. \tag{5.2}$$

$$\frac{d}{d\alpha}\|x-\alpha u\|^2 = \frac{d}{d\alpha}\{\alpha^2 - 2\alpha u^T x\} = 2\alpha - 2u^T x$$

$$\Rightarrow \alpha = u^T x$$

$$\Rightarrow P_u(x) = \alpha \cdot u = u^T x \cdot u$$

$$\sum_{i=1}^n P_u(x_i) = \sum_{i=1}^n u^T x_i \cdot u = u^T \underbrace{\left(\sum_{i=1}^n x_i\right)}_{0} \cdot u = 0$$

2. (4 points) *Maximum variance:* Show that the unit direction u that maximizes the variance of the projected data corresponds to the first principle component for the data. That is show,

$$\operatorname*{argmax}_{u\,:\,\|u\|_2=1} \sum_{i=1}^n \left\| \mathcal{P}_u(x_i) - \underbrace{\frac{1}{n}\sum_{j=1}^n \mathcal{P}_u(x_j)}_{=0} \right\|_2^2 \tag{5.3}$$

corresponds to the first principle component.

$$\hat{u} = \operatorname*{argmax}_{\|u\|=1} \sum_{i=1}^n \|P_u(x_i)\|^2$$

$$= \operatorname*{argmax}_{\|u\|=1} \sum_{i=1}^n \|u^T x_i \cdot u\|^2 = \sum_{i=1}^n u^T x_i^2$$

$$= \operatorname*{argmax}_{\|u\|=1} u^T \left(\sum_{i=1}^n x_i x_i^T\right) u$$

which is the first principle component.

3. (4 points) *Minimum error:* Show that the unit direction u that minimizes the mean squared error between projected data points and the original points corresponds to the first principal component for the data. That is show,

$$\operatorname*{argmin}_{u:\|u\|_2=1} \sum_{i=1}^{n} \|x_i - \mathcal{P}_u(x_i)\|_2^2 \tag{5.4}$$

corresponds to the first principle component.

$$\hat{u} = \operatorname*{argmin}_{\|u\|=1} \sum_{i=1}^{n} \|x_i - u^T x_i \cdot u\|^2 = \sum_{i=1}^{n} \|x_i\|^2 + {u^T x_i}^2 - 2\, {u^T x_i}^2$$

$$= \operatorname*{argmin}_{\|u\|=1} -\sum_{i=1}^{n} {u^T x_i}^2 = \operatorname*{argmax}_{\|u\|=1} u^T \left(\sum_{i=1}^{n} x_i x_i^T \right) u$$

4. (5 points) *Probabilistic PCA:* Now, assume the following model

$$z \sim N(0, \Sigma)$$
$$x|z \sim N(Wz + \mu, I).$$

Find the marginal distribution of x.

$$x = \mu + Wz + \varepsilon \qquad \varepsilon \sim N(0, I) \perp\!\!\!\perp z$$

$$= N(\mu, W\Sigma W^T + I)$$

5. (5 points) Find the posterior mean $\mathbb{E}[z|x]$. When does the above formulation reduce to classical PCA? Show your derivation.

$$p(z|x) \propto p(x|z)\,p(z) = \exp\left\{-\frac{1}{2}\|x - Wz - \mu\|^2 - \frac{1}{2}z^T\Sigma^{-1}z\right\}$$

$$= \exp\left\{-\frac{1}{2}z^T\left(\Sigma^{-1} + W^TW\right)z + (x-\mu)^TWz\right\}$$

$$\mathbb{E}[z|x] = \left(\Sigma^{-1} + W^TW\right)^{-1}W^T(x-\mu)$$

Say $\Sigma = \sigma^2 I$ and let $\sigma^2 \to \infty$,

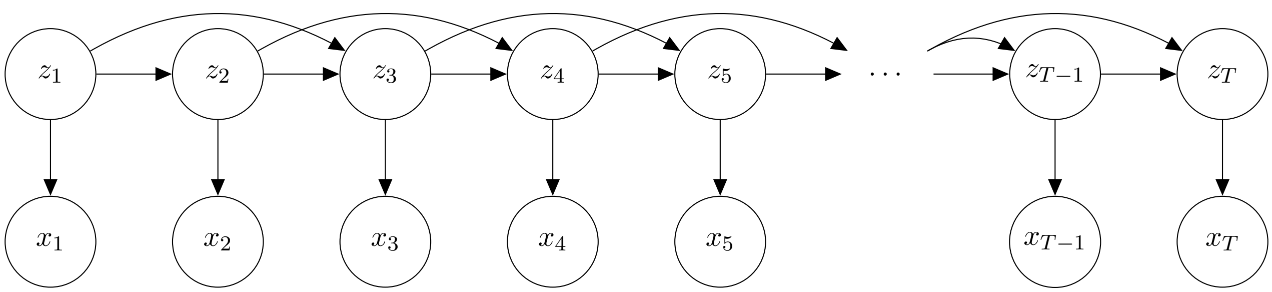
$$\mathbb{E}[z|x] = \underbrace{(W^TW)^{-1}W^T}_{\text{projects onto low dim space}}\underbrace{(x-\mu)}_{\text{centers data}}$$

↳ centers data

↳ projects onto low dim space: when $W = Q$ for $X^TX = Q\Lambda Q^T$.

6. Graphical model analysis (20 points).

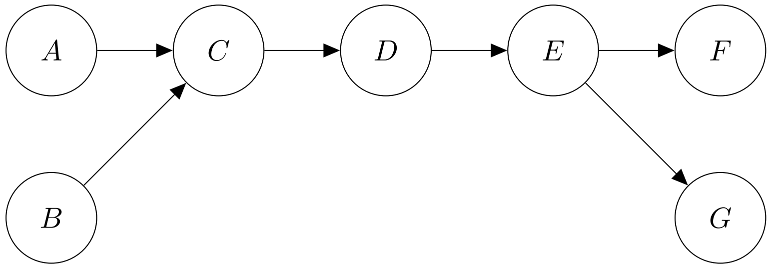
1. (5 points) Consider the graphical model shown below, a 2nd-order hidden Markov model:



Write the factorization of the joint distribution over $p(z_1, z_2, \ldots, z_T, x_1, x_2, \ldots, x_T)$ implied by this model.

$$\prod_{i=1}^{T} p(x_i|z_i) \prod_{i=3}^{T} p(z_i|z_{i-1}, z_{i-2})\; p(z_2|z_1)\,p(z_1)$$

2. (10 points) Consider another graphical model:



Answer true or false, no need to show your work:

(a) $A \perp\!\!\!\perp B$ T

(b) $B \perp\!\!\!\perp G$ F

(c) $F \perp\!\!\!\perp G$ F

(d) $A \perp\!\!\!\perp B|C$ F

(e) $A \perp\!\!\!\perp B|D$ F

(f) $A \perp\!\!\!\perp B|G$ F

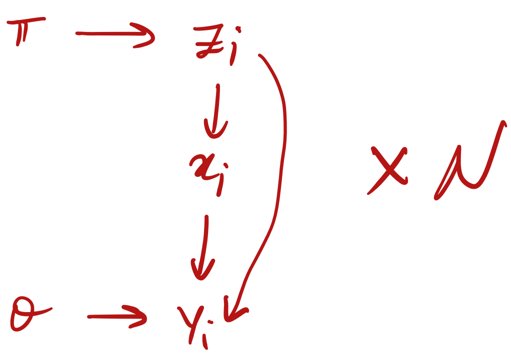
(g) $F \perp\!\!\!\perp G|E$ T

(h) $F \perp\!\!\!\perp G|A$ F

3. (5 points) Draw the graphical model for

$$p(x_1, x_2, \ldots, x_N, y_1, y_2, \ldots, y_N, z_1, z_2, \ldots, z_N, \theta, \pi) = p(\theta)p(\pi)\prod_{i=1}^{N} p(y_i|x_i, z_i, \theta)p(x_i|z_i)p(z_i|\pi)$$

.



(you need to draw this explicitly)

4. Gaussian Processes - 15 pts. We recall the following properties of the multivariate Gaussian vectors:

1. For a multivariate Gaussian vector $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and a matrix $\boldsymbol{A}$, we have

$$\boldsymbol{A}\mathbf{y} \sim \mathcal{N}\big(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T\big)$$

2. For any split,

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}\right) \tag{4.1}$$

   we have the conditional distribution again Gaussian

$$\mathbf{y}_2|(\mathbf{y}_1 = \boldsymbol{a}) \sim \mathcal{N}\big(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{a} - \boldsymbol{\mu}_1),\ \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{21}\big). \tag{4.2}$$

Suppose we have a linear model

$$t \,|\, \boldsymbol{x} \sim \mathcal{N}(y(\mathbf{x}),\ \sigma^2) \qquad y(\mathbf{x}) = \boldsymbol{w}^T\boldsymbol{\psi}(\boldsymbol{x})$$

and an isotropic prior on the weights $\boldsymbol{w} \sim \mathcal{N}(0, \alpha^{-1}\boldsymbol{I})$. We observe N data points and write them in vector form $\boldsymbol{t}_N = (t^{(1)}, t^{(2)}, ..., t^{(N)})^T$ and $\mathbf{y} = \boldsymbol{\Psi}\boldsymbol{w}$ where each row of $\boldsymbol{\Psi}$ is $\boldsymbol{\psi}(\boldsymbol{x}^{(i)})^T$.

(a) (2 pts) Find the distribution of the vector $\mathbf{y}$. Simplify notation by defining the scaled Gram matrix $\boldsymbol{K}_N = \frac{1}{\alpha}\boldsymbol{\Psi}\boldsymbol{\Psi}^T$.

$$y = \Psi w \sim \mathcal{N}\left(0, \frac{1}{\alpha}\Psi\Psi^T\right) = \mathcal{N}(0, K_N)$$

(b) (5 pts) Find the marginal distribution of $\boldsymbol{t}_N$. Simplify notation by defining the matrix $\boldsymbol{C}_N = \boldsymbol{K}_N + \sigma^2\boldsymbol{I}$.

$$t_N = y + \sigma\varepsilon \qquad \varepsilon \sim \mathcal{N}(0, I) \perp\!\!\!\perp y$$

$$\Rightarrow t_N \sim \mathcal{N}\Big(0, \underbrace{\frac{1}{\alpha}\Psi\Psi^T + \sigma^2 I}_{= K_N + \sigma^2 I = C_N}\Big)$$

(c) (8 pts) After observing a new test input $\boldsymbol{x}^{(N+1)}$, and using the above result for $N+1$, find the distribution of $p(t^{(N+1)} \mid \boldsymbol{t}_N)$.

$$t_{N+1} = \begin{bmatrix} t_N \\ t^{(N+1)} \end{bmatrix} \sim N\left(0, \begin{bmatrix} C_N & k \\ k^T & c \end{bmatrix}\right)$$

where $k = \begin{bmatrix} \vdots \\ \frac{1}{\alpha}\psi(x^{(i)})^T \psi(x^{(N+1)}) \\ \vdots \end{bmatrix}$ and $c = \sigma^2 + \frac{1}{\alpha}\psi(x^{(N+1)})^T \psi(x^{(N+1)})$

$$\Rightarrow t^{(N+1)} \mid t_N \sim N\left(k^T C_N^{-1} t_N,\ c - k^T C_N^{-1} k\right).$$