PRACTICE FINAL EXAM

STA414 Winter 2025
Probabilistic Machine Learning

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.

This practice exam contains more questions than the actual exam.

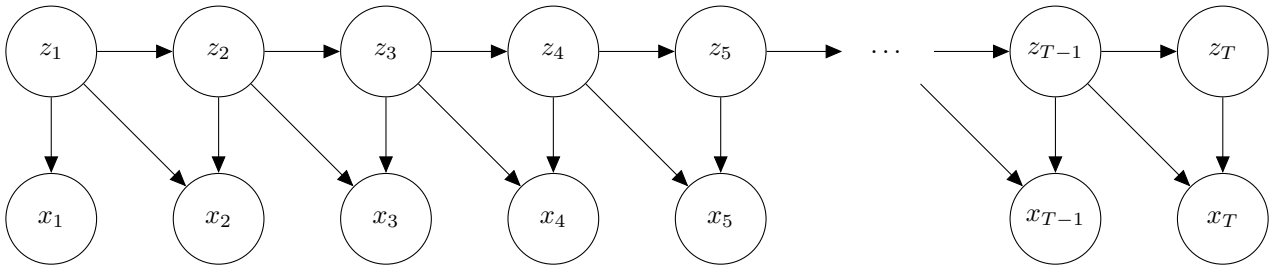
1. Decision theory (10 points). Imagine you are writing a quiz that has a true or false section. To discourage random guessing, the quiz awards x points for a correct answer, y points for a false answer, and z points for no answer.

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 $z = 0$?

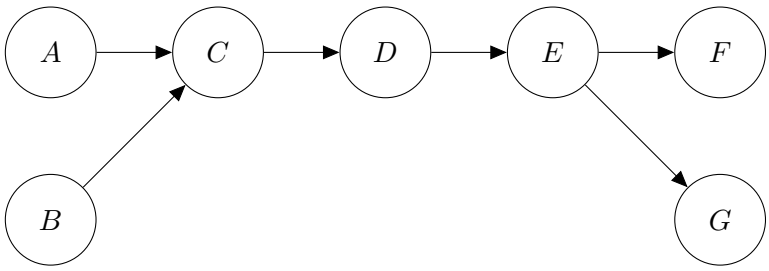
2. Graphical model analysis (20 points).

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.

2. (10 points) Consider another graphical model:



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

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

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

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

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

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

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

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

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

3. (5 points) Draw the graphical model for

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

.

3. 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.
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.

4. 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 $\mathbb{E}_{p(x)}[f(x)]$.
2. (4 points) Find the variance of this estimator as a function of S.
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 $p(x)$.

5. Bayesian Linear Regression (15 pts). Recall the multivariate Gaussian density

$$\mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \;\propto\; \exp\Big\{-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{w}-\boldsymbol{\mu})\Big\}.$$

In a linear regression problem, suppose that you are given a dataset $\mathbf{y} \in \mathbb{R}^N$ and $\mathbf{X} \in \mathbb{R}^{N\times D}$ where $N > D$ and assume $\mathbf{X}^\top\mathbf{X}$ is invertible. We assume that target has the following distribution

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

(a) (5 pts) Find a closed form solution for ordinary least squares solution defined as

$$\widehat{\mathbf{w}}_{\text{LS}} = \arg\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2.$$

For which class of covariance matrices $\boldsymbol{\Sigma}$, the MLE $\widehat{\mathbf{w}}$ for the above distribution would coincide with $\widehat{\mathbf{w}}_{\text{LS}}$?

(b) (5 pts) Now assume $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ for some scalar σ, and we use the following prior for the weights

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

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

(c) (5 pts) If the features are orthogonal, i.e. $\mathbf{X}^\top\mathbf{X} = \mathbf{I}$, show that the posterior mean is a weighted average between the prior mean $\boldsymbol{\mu}$ and the ordinary least squares solution $\widehat{\mathbf{w}}_{\text{LS}}$.

6. 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,

$$\mathcal{P}_u(x) = \operatorname*{argmin}_{v = \alpha u \,:\, \alpha \in \mathbb{R}} \|x - v\|_2^2. \tag{6.1}$$

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

$$\sum_{i=1}^n \mathcal{P}_u(x_i) = 0. \tag{6.2}$$

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,

$$\operatorname*{argmax}_{u \,:\, \|u\|_2 = 1} \sum_{i=1}^n \left\| \mathcal{P}_u(x_i) - \frac{1}{n}\sum_{j=1}^n \mathcal{P}_u(x_j) \right\|_2^2 \tag{6.3}$$

corresponds to 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,

$$\operatorname*{argmin}_{u \,:\, \|u\|_2 = 1} \sum_{i=1}^n \|x_i - \mathcal{P}_u(x_i)\|_2^2 \tag{6.4}$$

corresponds to the first principle component.

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.

5. (5 points) When does the above formulation reduce to classical PCA? Show your derivation.

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

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

We use the following prior for the weights

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

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

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

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

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

2. For any split,

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix} \right) \tag{8.1}$$

we have the conditional distribution again Gaussian

$$\mathbf{y}_2|(\mathbf{y}_1 = \boldsymbol{a}) \ \sim \mathcal{N}\big(\boldsymbol{\mu}_2 + \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}(\boldsymbol{a} - \boldsymbol{\mu}_1),\ \mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{21}\big). \tag{8.2}$$

Suppose we have a linear model

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

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 $\mathbf{y}_N = (y^{(1)}, y^{(2)}, ..., y^{(N)})^T$ and $\hat{\mathbf{y}} = \mathbf{\Psi}\boldsymbol{w}$ where each row of $\mathbf{\Psi}$ 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 $\boldsymbol{K}_N = \frac{1}{\alpha}\mathbf{\Psi}\mathbf{\Psi}^T$.

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

(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 $p(y^{(N+1)} \,|\, \mathbf{y}_N)$.

9. Decision theory - 15 pts. Recall the density of the normal distribution $\mathcal{N}(\mu, \sigma^2)$

$$p(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \Big\}$$

Suppose we have a classification problem with two classes $t \in \{0, 1\}$ and input x is 1-dimensional satisfying

$$x|t = 0 \sim \mathcal{N}(\mu_0, \sigma_0^2)$$
$$x|t = 1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$$

We assume that, a priori, both classes are equally likely. In each of the below scenarios, mathematically derive

1. the optimal decision rule that minimizes the misclassification rate,
2. the resulting value of the misclassification rate.

Decision rule will be specified by two disjoint regions $\mathcal{R}_0$ and $\mathcal{R}_1$ with $\mathcal{R}_0 \cup \mathcal{R}_1 = \mathbb{R}$. If $x \in \mathcal{R}_0$ we classify x as class 0, otherwise class 1. The misclassification rate is given by

$$p(x \in \mathcal{R}_0, t = 1) + p(x \in \mathcal{R}_1, t = 0).$$

(a) (5 pts) Suppose $\mu_0 \neq \mu_1$ and $\sigma_0 = \sigma_1$.
(b) (5 pts) Suppose $\mu_0 = \mu_1$ and $\sigma_0 = \sigma_1$.
(c) (5 pts) Suppose $\mu_0 = \mu_1$ and $\sigma_0 \neq \sigma_1$.

10. Word2vec (15 points). You are working with a dataset of M molecules built from some combination of any number of 35 atoms. You are interested in creating vector representations of the atoms to be used in downstream tasks. The data is represented as graphs with atoms being nodes, and edges corresponding to there being a bond between the two atoms. Describe how you could train a model to produce embeddings for atoms using this dataset, incorporating the idea that "atoms A and B are similar if they often bond to the same atoms". In your answer include the following:

1. (5 points) What is your model?
2. (4 points) What is the loss function?
3. (4 points) How is the data sampled in the training process?
4. (2 points) Is negative sampling necessary in this case?

End of exam
