PRACTICE FINAL EXAM

CSC412 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.

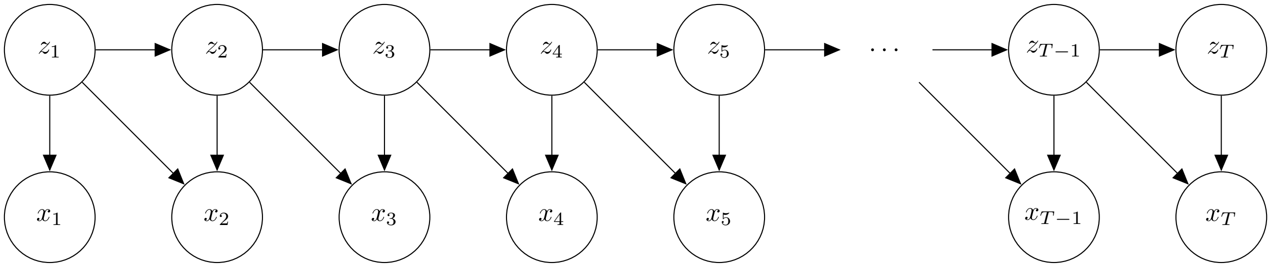
Hand in all examination materials at the end

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.

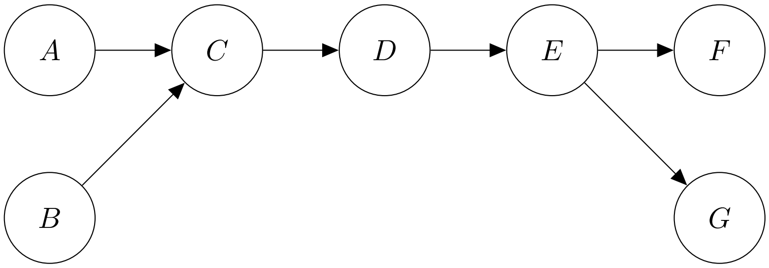
1. (8 points) You think you know the correct answer with probability θ. How high must θ be, as a function of x, y, and z, before the expected number of points is higher for choosing the most likely answer, versus leaving the question blank?
   Answer: If the question is answered, the expected reward is $\theta x + (1-\theta)y$ and if not then it is z. So the condition is $\theta > \frac{z-y}{x-y}$.
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 $z = 0$?
   Answer: 1/2

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 $p(z_1, z_2, \dots, z_T, x_1, x_2, \dots, x_T)$ implied by this model. Answer: $p(z_1)\prod_{t=2}^{T} p(z_t|z_{t-1})p(x_1|z_1)\prod_{t=2}^{T} p(x_t|z_t, z_{t-1})$
2. (10 points) Consider another graphical model:



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

   (a) $A \perp\!\!\!\perp B$ Answer: yes

   (b) $B \perp\!\!\!\perp G$ Answer: no

   (c) $F \perp\!\!\!\perp G$ Answer: no

   (d) $A \perp\!\!\!\perp B|C$ Answer: no

   (e) $A \perp\!\!\!\perp B|D$ Answer: no

   (f) $A \perp\!\!\!\perp B|G$ Answer: no

   (g) $F \perp\!\!\!\perp G|E$ Answer: yes

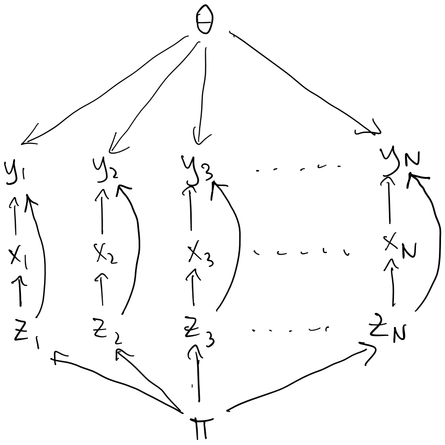
(h) $F \perp\!\!\!\perp G|A$ Answer: no

3. (5 points) Draw the graphical model for

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

.

Answer: This is the graph:



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. Answer: This was done in the lecture.

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. Answer: Note that

$$\mathbb{E}_{p(x|\theta)}[\log p(x|\theta)] - \mathbb{E}_{p(x|\theta)}\left[\log p(x|\hat{\theta})\right] = \mathbb{E}_{p(x|\theta)} \log\left[\frac{p(x|\theta)}{p(x|\hat{\theta})}\right] = \mathrm{KL}(p(x|\theta), p(x|\hat{\theta})) \geq 0$$

which implies the desired inequality.

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 $\mathbb{E}_{p(x)}[f(x)]$. Answer: See the lecture.
2. (4 points) Find the variance of this estimator as a function of S.Answer: See the lecture.
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 $p(x)$. Answer: Since we do not know how q relates to p, we cannot use rejection sampling. We can use however the importance sampling. In the lecture we discussed how to get an unbiased estimator of $\mathbb{E}_{p(x)}[f(x)]$ in this case.

5. Mixture of Experts (20 points).

1. (10 points) Draw a three layered MoE based on the switch transformer architecture. Answer: https://diophontine.github.io/csc412/slides/w11/sld11.pdf, slide 29

2. (10 points) Based on the switch transformer architecture, define the weight matrices and show how they would be multiplied together for a one layered MoE . You may choose a small layer width for simplicity Answer: Follows from the diagram

Input: $X \in \mathbb{R}^{N \times d_{\text{model}}}$

Self-Attention:

$Q = XW_Q, \quad K = XW_K, \quad V = XW_V$

$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$

$X_{\text{SA}} = \text{Attention}(Q, K, V)$

$X_{\text{SA\_out}} = \text{LayerNorm}(X + X_{\text{SA}})$

Switch MoE:

Routing Scores: $S = X_{\text{SA\_out}}W_g, \quad S \in \mathbb{R}^{N \times n_{\text{experts}}}$

Expert Assignment: $e_i = \text{argmax}(S_i) \quad \text{for } i = 1, \ldots, N$

Gating Value: $g_i = \text{softmax}(S_i)_{e_i}$

Expert Computation (for token i assigned to expert $j = e_i$):

$H_i = \text{ReLU}(X_{\text{SA\_out},i}W_{j,1} + b_{j,1})$

$E_i = H_iW_{j,2} + b_{j,2}$

MoE Output: $Y_i = g_iE_i$

Output of MoE Layer:

$X_{\text{Output}} = \text{LayerNorm}(X_{\text{SA\_out}} + Y)$

6. Diffusion (10 points).

1. (5 points) Derive the loss function for a basic diffusion model based on the forward and backwards process. Answer: https://diophontine.github.io/csc412/slides/w12/sld12.pdf Slide 22-35
2. (5 points) Explain how your loss function would change if you were training a prompt-able image diffusion model Answer: Use conditional generation conditioned on some captions, can use something like clip to map between image and text (Ramesh, 2022). A prior $P(z,|y)$ that produces CLIP image embeddings zi conditioned on captions y.
   A decoder $P(x|z_i, y)$ that produces images x conditioned on CLIP image embeddings z_i (and optionally text captions y).
   Slide 42-43.

7. 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? Answer: Tokenizer is 0 to 34 for each atom parts and potentially. Training data would be sequences of atoms in some 1d projection. You then predict either the surrounding atoms or the missing atom depending on CBOW or Skipgram approach. Embedding W project to hidden embedding layer and W' matrices back to one hot encoded vectors.
2. (4 points) What is the loss function? Answer: Cross entropy, Softmax on the output one hot encoded and substract from the predicted value. Answer: Predicting a sequence of atoms, where each one hot encoded
3. (4 points) How is the data sampled in the training process? Answer: Real molecules that exists feed them through in the 1d neighbor projection. If there are multiple bonds pass in all the pairs, not just across 1d.
4. (2 points) Is negative sampling necessary in this case? Answer: Yes, there are many combinations of atoms in molecules so we want to reduce some computational cost

8. 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 $\sigma_0 = \sigma_1$.
Answer: We know that in general the optimal decision is to classify x as 1 if $N(x; \mu_1, \sigma_1) \geq N(x; \mu_0, \sigma_0)$. If $\sigma_0 = \sigma_1$ this is equivalent to $|x - \mu_1| \leq |x - \mu_0|$.

(b) (5 pts) Suppose $\mu_0 = \mu_1$ and $\sigma_0 = \sigma_1$.
Answer: In this case the misclassification rate is $\frac{1}{2}$ irrespective of how we define the decision regions (as long as they are disjoint and cover the whole $\mathbb{R}$).

(c) (5 pts) Suppose $\mu_0 = \mu_1$ and $\sigma_0 \neq \sigma_1$.
Answer: We have $N(x; \mu, \sigma_1) \geq N(x; \mu, \sigma_0)$ if and only if

$$\log\sigma_1 + \frac{1}{2\sigma_1^2}(x-\mu)^2 \;\leq\; \log\sigma_0 + \frac{1}{2\sigma_0^2}(x-\mu)^2$$

equivalently

$$(x-\mu)^2\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right) \;\leq\; \log\frac{\sigma_0^2}{\sigma_1^2}.$$

Suppose that $\sigma_1 < \sigma_0$ then we classify x as 1 if $|x - \mu|$ is less than some threshold, given explicitly as

$$\sqrt{\frac{\log\frac{1}{\sigma_1^2} - \log\frac{1}{\sigma_0^2}}{\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}}}.$$

End of exam