PRACTICE FINAL EXAM

STA414/2104 Winter 2026

Stat Machine Learning II

University of Toronto

Faculty of Arts & Science

Duration - 180 minutes

Aids allowed: One double-sided handwritten $8.5'' \times 11''$ or A4 aid sheets.

- There are 7 written exam questions and the total possible number of points is 100. You need to show all your work to receive full credit on each question.
- Do not begin writing the actual exam until the announcements have ended and the Exam Facilitator has started the exam.
- Write all answers only in the space provided after each question.
- If you possess unauthorized aid during the exam, you may be charged with academic offence.
- Turn off and place all cell phones, smart watches, electronic devices, and unauthorized study materials in your bag under your desk. If 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. All answers should be written on the exam paper itself.

1. Decision Theory – 10 pts

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.

(a) **(8 pts)** 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?

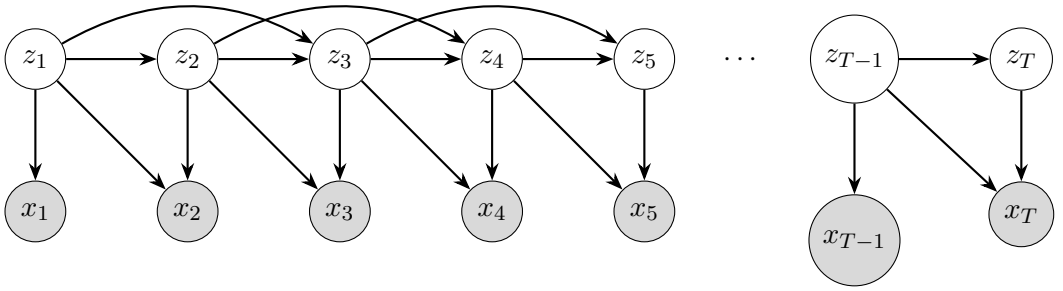
Answer: If the question is answered, the expected reward is $\theta x + (1-\theta)y$ and if not then it is z. So the condition is $\theta > \frac{z-y}{x-y}$.

(b) **(2 pts)** 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 pts

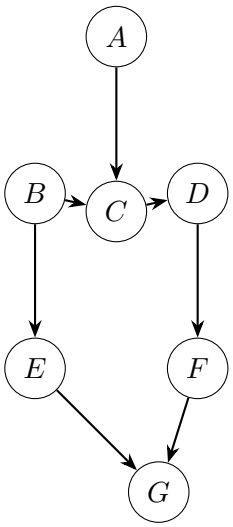
(a) **(5 pts)** Consider the graphical model shown below, a 2nd-order hidden Markov model:



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.

Answer: $p(z_1)p(z_2|z_1)\prod_{t=3}^{T} p(z_t|z_{t-1}, z_{t-2})p(x_1|z_1)\prod_{t=2}^{T} p(x_t|z_t, z_{t-1})$

(b) **(10 pts)** Consider another graphical model:



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

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

Answer: yes

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

Answer: no

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

Answer: no

(iv) $A \perp\!\!\!\perp B \mid C$

Answer: no

(v) $A \perp\!\!\!\perp B \mid D$

Answer: no

(vi) $A \perp\!\!\!\perp B \mid G$

Answer: no

(vii) $C \perp\!\!\!\perp E \mid B$

Answer: yes

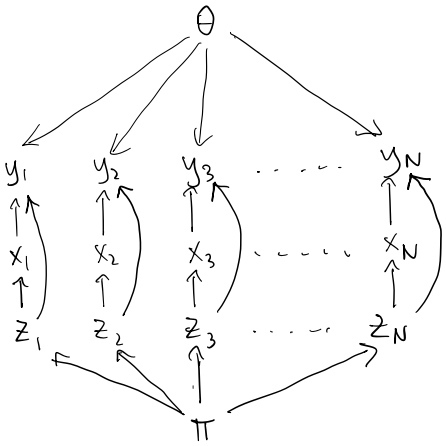
(viii) $F \perp\!\!\!\perp G \mid A$

Answer: no

(c) **(5 pts)** 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:



3. Variational Inference – 10 pts

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

(a) **(5 pts)** 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.

(b) **(5 pts)** 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 $\hat{\theta}$,

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

You can assume p and q are positive everywhere.

Answer: Note that

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

which implies the desired inequality.

4. Monte Carlo Estimators – 10 pts

Recall the Simple Monte Carlo estimator:

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

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

Answer: See the lecture.

(b) **(4 pts)** Find the variance of this estimator as a function of S.

Answer: See the lecture.

(c) **(4 pts)** 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. Bayesian Linear Regression – 15 pts

Recall the multivariate Gaussian density

$$\mathcal{N}(w|\mu,\Sigma) \propto \exp\left\{-\frac{1}{2}(w-\mu)^\top \Sigma^{-1}(w-\mu)\right\}.$$

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

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

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

$$\hat{w}_{\text{LS}} = \arg\min_w \|y - Xw\|^2.$$

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

Answer: This is a standard calculation. Suppose $\Sigma = \sigma^2 I$. The likelihood is

$$p(y|X,w,\sigma^2) = \frac{1}{(2\pi)^{N/2}}\sqrt{\det\left(\frac{1}{\sigma^2}I_N\right)}\exp\left\{-\frac{1}{2\sigma^2}(y-Xw)^\top(y-Xw)\right\}.$$

Thus, the log-likelihood function, up to the irrelevant additive constants, is

$$-\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\|y-Xw\|^2.$$

Irrespective of what σ is, the optimal w is the one minimizing $\|y-Xw\|^2$. This is the least squares solution. The explicit formula can be found by vector differentiation. If X has full column rank it is given by

$$\hat{w}_{\text{LS}} = (X^\top X)^{-1}X^\top y$$

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

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

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

Answer: Since $p(w|y) \propto p(w)p(y|w)$ we can recycle calculations from above to write that

$$\log p(w|y) = \text{const} - \frac{1}{2}\|w-\mu\|^2 - \frac{1}{2\sigma^2}\|y-Xw\|^2,$$

where all terms that do not depend on w are in the first term. We can expand this and get

$$\log p(w|y) = \text{const} - \frac{1}{2}w^\top\left(I + \frac{1}{\sigma^2}X^\top X\right)w + \left(\mu + \frac{1}{\sigma^2}X^\top y\right)^\top w.$$

Using the multivariate completion of squares introduced in the lecture, we get that the posterior is Gaussian with covariance matrix

$$\left(I + \frac{1}{\sigma^2} X^\top X\right)^{-1}$$

and mean

$$\left(I + \frac{1}{\sigma^2} X^\top X\right)^{-1} \left(\mu + \frac{1}{\sigma^2} X^\top y\right) = (\sigma^2 I + X^\top X)^{-1}(\sigma^2 \mu + X^\top y)$$

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

Answer: By the previous exercise the posterior mean is

$$(\sigma^2 I + X^\top X)^{-1}(\sigma^2 \mu + X^\top y)$$

If $X^\top X = I$ then $\hat{w}_{\text{LS}} = X^\top y$ and this expression simplifies to

$$\frac{1}{1+\sigma^2}(\sigma^2 \mu + \hat{w}_{\text{LS}}) = (1-\lambda)\mu + \lambda \hat{w}_{\text{LS}},$$

where $\lambda = \frac{1}{1+\sigma^2}$.

6. Autoencoders – 10 pts

Suppose you are given a high-dimensional dataset and wish to learn a low-dimensional latent representation to be used for downstream tasks, such as clustering. You observe that a standard, deep deterministic autoencoder produces a latent space with severe discontinuities between clusters.

(a) **(5 pts)** Explain why standard deterministic autoencoders often suffer from this issue, where proximity in the data space does not guarantee proximity in the latent (feature) space.

Answer: A deterministic autoencoder maps each input $\mathbf{x}$ to a single point $\mathbf{z} = f_{\text{enc}}(\mathbf{x})$ in the latent space. The training objective is purely reconstruction-based:

$$\mathcal{L} = \|\mathbf{x} - \hat{\mathbf{x}}\|^2 = \|\mathbf{x} - f_{\text{dec}}(f_{\text{enc}}(\mathbf{x}))\|^2$$

This objective has no constraint on the structure of the latent space. The encoder only needs to produce codes that the decoder can invert, so it is not penalized for:

- Placing similar inputs far apart in latent space
- Creating "gaps" or "holes" between clusters
- Using an irregular, non-smooth mapping

As a result, the encoder can learn a highly discontinuous mapping where small perturbations in input space cause large jumps in latent space. The latent representations are optimized solely for reconstruction, not for any geometric or probabilistic regularity. This makes the latent space unsuitable for downstream tasks like interpolation, generation, or clustering that rely on meaningful distances.

(b) **(5 pts)** Why is a Variational Autoencoder (VAE) a more suitable alternative for creating a continuous, well-structured latent space? In 3–4 sentences, describe the output of the VAE encoder and how the two components of its loss function (the ELBO) act to regularize the latent space.

Answer: Unlike a deterministic autoencoder, the VAE encoder outputs the parameters of a probability distribution over the latent space, not a single point. Specifically, for each input $\mathbf{x}$, the encoder outputs $\boldsymbol{\mu}(\mathbf{x})$ and $\boldsymbol{\sigma}^2(\mathbf{x})$ (or $\log \boldsymbol{\sigma}^2(\mathbf{x})$), defining a Gaussian posterior $q(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}(\mathbf{x}), \text{diag}(\boldsymbol{\sigma}^2(\mathbf{x})))$. The latent code $\mathbf{z}$ is then *sampled* from this distribution (using the reparameterization trick: $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$). The VAE maximizes the ELBO:

$$\mathcal{L}_{\text{ELBO}} = \underbrace{\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})]}_{\text{Reconstruction term}} - \underbrace{D_{\text{KL}}(q(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z}))}_{\text{Regularization term}}$$

The first term encourages the decoder to accurately reconstruct $\mathbf{x}$, the second prevents the encoder from placing latent codes arbitrarily far from the origin, encourages overlap between the posterior distributions of different data points and forces the latent space to be continuous and smoothly organized. The combination of these

terms ensures that the latent space is both informative (good reconstruction) and organized around the prior, making it suitable for interpolation, generation, and clustering.

7. Bayesian Linear Regression – 10 pts

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

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

We use the following prior for the weights

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

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

Answer: The logarithm of the posterior satisfies

$$\log p(w|D) = \log p(w) + \log p(D|w)$$

The likelihood term was computed as follows

$$\begin{aligned}
\sum_{i=1}^{N} \log p(y^{(i)}|x^{(i)}; w, \beta) &= \sum_{i=1}^{N} \log \mathcal{N}(y^{(i)}; w^\top x^{(i)}, \sigma^{-1}) \\
&= \sum_{i=1}^{N} \log \left[\frac{\sigma}{\sqrt{2\pi}} \exp\left\{ -\frac{\sigma^2}{2}(y^{(i)} - w^\top x^{(i)})^2 \right\} \right] \\
&= \text{const} - \frac{\sigma^2}{2} \sum_{i=1}^{N} (y^{(i)} - w^\top x^{(i)})^2 \\
&= \text{const} - \frac{\sigma^2}{2} \|y - Xw\|^2
\end{aligned}$$

For the given prior we have

$$\log p(w) = \log \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}(w-\mu)^\top \Sigma^{-1}(w-\mu) \right\} = -\frac{1}{2}(w-\mu)^\top \Sigma^{-1}(w-\mu) + \text{const}$$

Putting this together we get

$$\log p(w|D) = -\frac{1}{2}(w-\mu)^\top \Sigma^{-1}(w-\mu) - \frac{\sigma^2}{2}\|y - Xw\|^2 + \text{const}.$$

It is clear that the posterior will be Gaussian. To find its parameters explicitly we try to complete the squares

$$\log p(w|D) = -\frac{1}{2} w^\top (\Sigma^{-1} + \frac{\sigma^2}{2} X^\top X) w + \left(\mu^\top \Sigma^{-1} + y^\top X \right) w.$$

Thus, the posterior is Gaussian with covariance

$$\left(\Sigma^{-1} + \sigma^2 X^\top X \right)^{-1}$$

and mean

$$\left(\Sigma^{-1} + \sigma^2 X^\top X \right)^{-1} \left(\mu^\top \Sigma^{-1} + y^\top X \right)^\top = \left(\Sigma^{-1} + \sigma^2 X^\top X \right)^{-1} \left(\Sigma^{-1}\mu + X^\top y \right)$$

8. Gaussian Processes – 15 pts

We recall the following properties of multivariate Gaussian vectors:

1. For a multivariate Gaussian vector $y \sim \mathcal{N}(\mu, \Sigma)$ and a matrix A, we have
$$Ay \sim \mathcal{N}(A\mu, A\Sigma A^\top)$$
2. For any split,
$$y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$$
we have the conditional distribution again Gaussian
$$y_2 | (y_1 = a) \sim \mathcal{N}\left(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(a - \mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{21}\right).$$

Suppose we have a linear model
$$y|x \sim \mathcal{N}(\hat{y}(x), \sigma^2) \quad \hat{y}(x) = w^\top \psi(x)$$
and an isotropic prior on the weights $w \sim \mathcal{N}(0, \alpha^{-1}I)$. We observe N data points and write them in vector form $y_N = (y^{(1)}, y^{(2)}, \ldots, y^{(N)})^\top$ and $\hat{y} = \Psi w$ where each row of Ψ is $\psi(x^{(i)})^\top$.

(a) **(2 pts)** Find the distribution of the vector y. Simplify notation by defining the scaled Gram matrix $K_N = \frac{1}{\alpha}\Psi\Psi^\top$.

Answer: All these derivations appeared in the lecture.

(b) **(5 pts)** Find the marginal distribution of y_N. Simplify notation by defining the matrix $C_N = K_N + \sigma^2 I$.

Answer: All these derivations appeared in the lecture.

(c) **(8 pts)** After observing a new test input $x^{(N+1)}$, and using the above result for $N+1$, find the distribution of $p(y^{(N+1)}|y_N)$.

Answer: All these derivations appeared in the lecture.

9. Decision Theory – 15 pts

Recall the density of the normal distribution $\mathcal{N}(\mu, \sigma^2)$

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

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 $\mathcal{N}(x; \mu_1, \sigma_1) \geqslant \mathcal{N}(x; \mu_0, \sigma_0)$. If $\sigma_0 = \sigma_1$ this is equivalent to $|x - \mu_1| \leqslant |x - \mu_0|$. Assuming $\mu_0 < \mu_1$, the missclassification rate is

$$\frac{1}{2}\int_{-\infty}^{(\mu_0+\mu_1)/2} \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} dx + \frac{1}{2}\int_{(\mu_0+\mu_1)/2}^{\infty} \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(x-\mu_0)^2}{2\sigma_0^2}} dx$$

and using $\sigma_0 = \sigma_1$ and Φ of the standard centered Gaussian, $\frac{1}{2}\Phi\left((\mu_0 - \mu_1)/2\sigma_1\right) + \frac{1}{2}[1 - \Phi\left((\mu_1 - \mu_0)/2\sigma_0\right)] = \Phi\left((\mu_0 - \mu_1)/2\sigma_0\right)$.

(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 $\mathcal{N}(x; \mu, \sigma_1) \geqslant \mathcal{N}(x; \mu, \sigma_0)$ if and only if

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

equivalently

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

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

$$\tau = \sqrt{\frac{\log(\sigma_0^2/\sigma_1^2)}{1/\sigma_1^2 - 1/\sigma_0^2}} = \sqrt{\frac{\sigma_0^2\sigma_1^2 \log(\sigma_0^2/\sigma_1^2)}{\sigma_0^2 - \sigma_1^2}}.$$

The decision regions are $\mathcal{R}_1 = [\mu - \tau, \mu + \tau]$ and $\mathcal{R}_0 = (-\infty, \mu - \tau) \cup (\mu + \tau, \infty)$. The misclassification rate is:

$$\begin{aligned}
P(\text{error}) &= P(x \in \mathcal{R}_0, t = 1) + P(x \in \mathcal{R}_1, t = 0) \\
&= \frac{1}{2} P(|x-\mu| > \tau \mid t = 1) + \frac{1}{2} P(|x-\mu| \leqslant \tau \mid t = 0) \\
&= \frac{1}{2} \cdot 2 \left(1 - \Phi\left(\frac{\tau}{\sigma_1}\right)\right) + \frac{1}{2}\left(2\Phi\left(\frac{\tau}{\sigma_0}\right) - 1\right) \\
&= \frac{1}{2} + \Phi\left(\frac{\tau}{\sigma_0}\right) - \Phi\left(\frac{\tau}{\sigma_1}\right)
\end{aligned}$$

where Φ denotes the standard normal CDF.

10. Word2Vec for Molecular Embeddings – 15 pts

You are working with a dataset of M molecules built from some combination of 35 different atom types. You want to learn vector representations (embeddings) of atoms for use in downstream tasks such as predicting molecular properties.

The data is represented as graphs where atoms are nodes and edges represent chemical bonds between atoms. Your goal is to adapt the Word2Vec approach to this molecular setting, using the key insight that **"atoms A and B are similar if they often bond to the same atoms"** (analogous to the distributional hypothesis in NLP: "words that occur in similar contexts have similar meanings").

(a) **(5 pts)** What is your model architecture? Describe the vocabulary, input representation, and the embedding matrices involved.

Answer: **Vocabulary:** The vocabulary V consists of the 35 atom types, so $|V| = 35$. Each atom type a can be represented as a one-hot vector $\mathbf{x}_a \in \{0,1\}^{35}$. **Model (Skip-gram approach):** Given a center atom, predict its bonded neighbors (context atoms). We use two learned embedding matrices:

- $W \in \mathbb{R}^{e \times 35}$: the "input" embedding matrix, where e is the embedding dimension
- $W' \in \mathbb{R}^{35 \times e}$: the "output" embedding matrix for predicting context atoms

For a center atom a with one-hot encoding $\mathbf{x}_a$:

(a) Compute the embedding: $\mathbf{h} = W \cdot \mathbf{x}_a$ (this selects the a-th column of W)

(b) Compute output scores: $\mathbf{z} = W' \cdot \mathbf{h}$

(c) Apply softmax to get probabilities over all atom types: $p(\text{neighbor} = b \mid \text{center} = a) = \frac{\exp(z_b)}{\sum_{j=1}^{35} \exp(z_j)}$

After training, the columns of W (or equivalently, rows of $W^\top$) are the learned atom embeddings.

(b) **(4 pts)** What is the loss function?

Answer: We maximize the average log-probability of observing the true context atoms. For the Skip-gram model, given training pairs (a, b) where atom b is bonded to center atom a:

$$\mathcal{L} = \frac{1}{T} \sum_{t=1}^{T} \sum_{b \in \text{neighbors}(a_t)} \log p(b \mid a_t)$$

where $p(b \mid a) = \frac{\exp(\mathbf{u}_a^\top \mathbf{v}_b)}{\sum_{j=1}^{35} \exp(\mathbf{u}_a^\top \mathbf{v}_j)}$. Here $\mathbf{u}_a$ is the embedding of center atom a and $\mathbf{v}_b$ is the context embedding of atom b.

SCRAP PAPER – WILL NOT BE GRADED.

SCRAP PAPER – WILL NOT BE GRADED.