PLEASE HAND IN

PLEASE HAND IN

UNIVERSITY OF TORONTO
Faculty of Arts and Science

WINTER 2019 EXAMINATIONS

CSC 411 H1S

Duration: 3 hours

Aids Allowed: None

Student Number: |_|_|_|_|_|_|_|_|_|_|

Last (Family) Name(s): ______________________________

First (Given) Name(s): ______________________________

---

*Do **not** turn this page until you have received the signal to start.*
In the meantime, please read *carefully* every reminder on this page.

MARKING GUIDE

Nº 1:	______/ 2
Nº 2:	______/ 5
Nº 3:	______/ 5
Nº 4:	______/ 5
Nº 5:	______/ 5
Nº 6:	______/ 5
Nº 7:	______/ 5
Nº 8:	______/ 4
Nº 9:	______/ 2
Nº 10:	______/ 7
TOTAL:	______/45

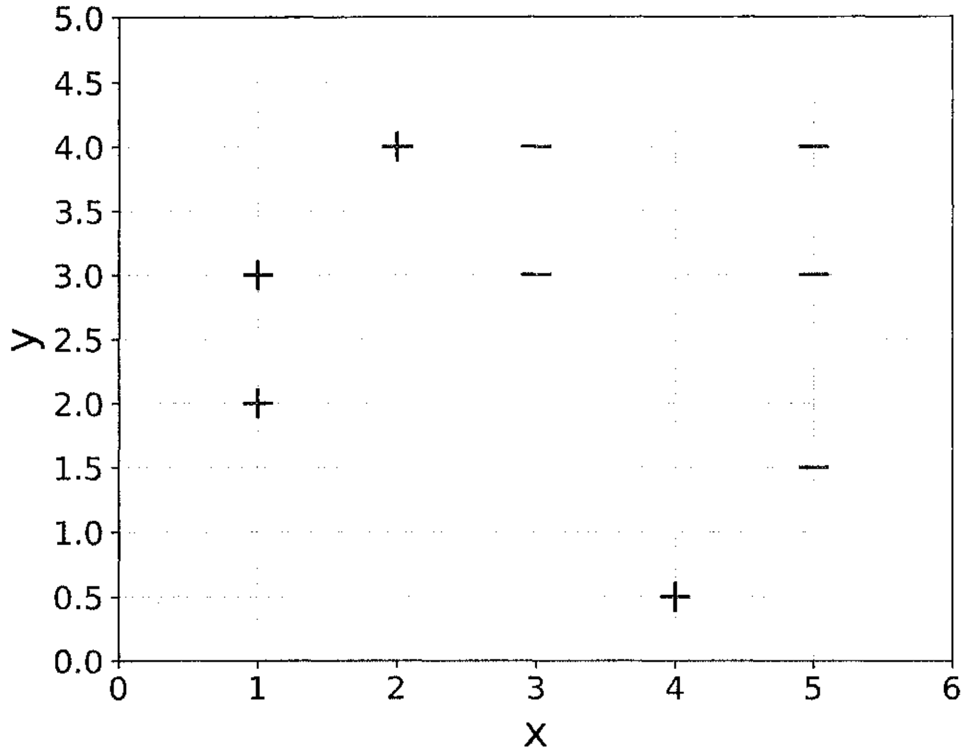
- Fill out your name and student number above—do it now, don't wait!
- This final examination consists of 10 questions on 23 pages (including this one), printed on both sides of the paper. *When you receive the signal to start, please make sure that your copy of the examination is complete.*
- Answer each question directly on the examination paper, in the space provided, and use a "blank" page for rough work. If you need more space for one of your solutions, use one of the "blank" pages and *indicate clearly the part of your work that should be marked.*
- *Remember that, in order to pass the course, you must achieve a grade of at least **30%** on this final examination.*
- As a student, you help create a fair and inclusive writing environment. If you possess an unauthorized aid during an exam, you may be charged with an academic offence.

STUDENTS MUST HAND IN ALL EXAMINATION MATERIALS AT THE END

Good Luck!

Question 1. Boosting [2 MARKS]

The figure below shows a dataset. Each example in the dataset has two input features x and y and may be classified as a positive example (labelled +) or a negative example (labelled −). We wish to apply Adaboost with axis-aligned decision stumps to solve the classification problem.



Part (a) [1 MARK]

In the figure above, draw the decision boundary corresponding to the first decision stump that the boosting algorithm would choose. Lightly shade the side of the boundary corresponding to a positive (+) classification.

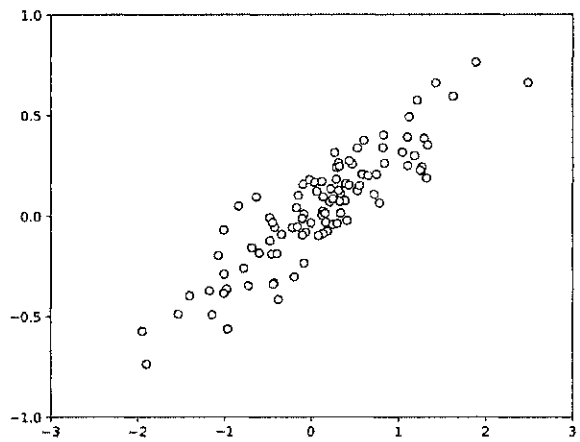
Part (b) [1 MARK]

In the same figure, circle the point(s) which have the highest weight after the first boosting iteration.

Question 2. Principal Component Analysis [5 MARKS]

Part (a) [1 MARK]

The figure below shows a two-dimensional dataset. Draw the vector corresponding to the first principal component.



Part (b) [2 MARKS]

The principal components of a dataset can be found by either minimizing an objective or, equivalently, maximizing a different objective. In words, describe the objective in each case using a single sentence.

Minimizing:

Maximizing:

Part (c) [1 MARK]

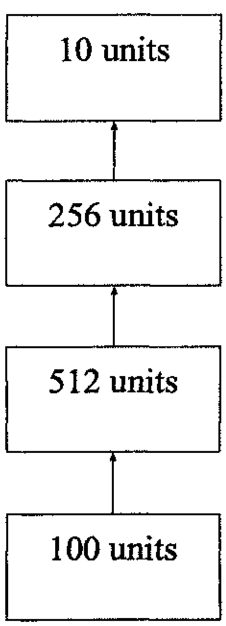
Suppose we wish to find the first two principal components of a dataset $\mathcal{D} = \{\mathbf{x}^{(n)}\}_{n=1}^{N} \subset \mathbb{R}^{50}$ using an autoencoder. Let θ denote the parameters of the autoencoder (i.e. the weights of the network) and let $f_\theta(\mathbf{x})$ denote the function the autoencoder computes. First, state the objective we would use to train the autoencoder and the activation function of the autoencoder.

Objective:

Activation Function:

Part (d) [1 MARK]

Draw the architecture of the autoencoder from the previous question. Use a similar style to the figure below, which depicts a network which takes a 100 dimensional input, processes it using hidden layers of 512 and 256 units, and produces a 10 dimensional output.



Question 3. Maximum Likelihood vs. Maximum A Posteriori vs. Fully Bayesian [5 MARKS]

Angela is at the ML Casino, where she is playing the Random Game. On each round of the game, a machine generates a real number $x \in \mathbb{R}$. If the number is positive, Angela wins x dollars. If the number is negative, Angela must pay the casino x dollars. So far, she has played 3 times and observed the following dataset:

$$\mathcal{D} = \{-5, 3, -10\}$$

Angela believes the machine is generating its numbers from a normal distribution with mean μ and variance 10:

$$x \sim \mathcal{N}(\mu, 10)$$

For this question, you may find the probability density function of the normal distribution useful:

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

Part (a) [1 MARK]

Write the log-likelihood function $\ell(\mu) = \log p(\mathcal{D}|\mu)$.

Part (b) [1 MARK]

Find the maximum likelihood estimate of the mean μ.

Part (c) [1 MARK]

Angela doesn't think the casino would set up a game where they lose money on average. She formulates this belief as a prior distribution on μ: $p(\mu) = \mathcal{N}(\mu| - 1, 5)$. Find the maximum a posteriori (MAP) estimate of the mean μ under this prior distribution.

Part (d) For a general probabilistic model, name one advantage of MAP estimation over a fully Bayesian approach. [1 MARK]

Part (e) For a general probabilistic model, name one advantage of a fully Bayesian approach over MAP estimation. [1 MARK]

Question 4. Reinforcement learning [5 MARKS]

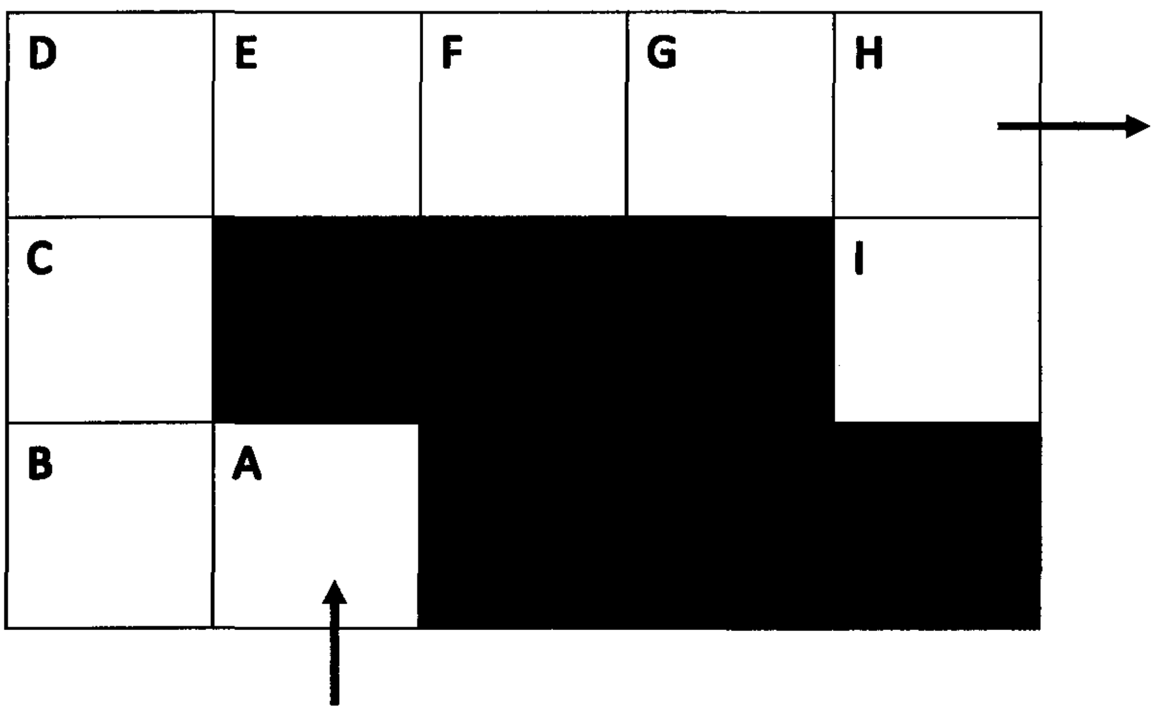
Suppose a robot is going through a maze, starting from location A. The robot is allowed to move between two adjacent cells. The robot can exit the maze and get a reward of 100 by visiting H, and the game will end. The discount factor γ is 0.8. Round your answer to the nearest integer. You do not need to keep track of numbers after the decimal place between successive multiplications.

Part (a) Value-function [1 MARK]

What is the definition of the state-value function $V^{\pi}(s)$? Write down the equation or describe in a sentence.

Part (b) Value-iteration [2 MARKS]

What is the optimal state-value function? (write the optimal value inside each cell in the diagram).

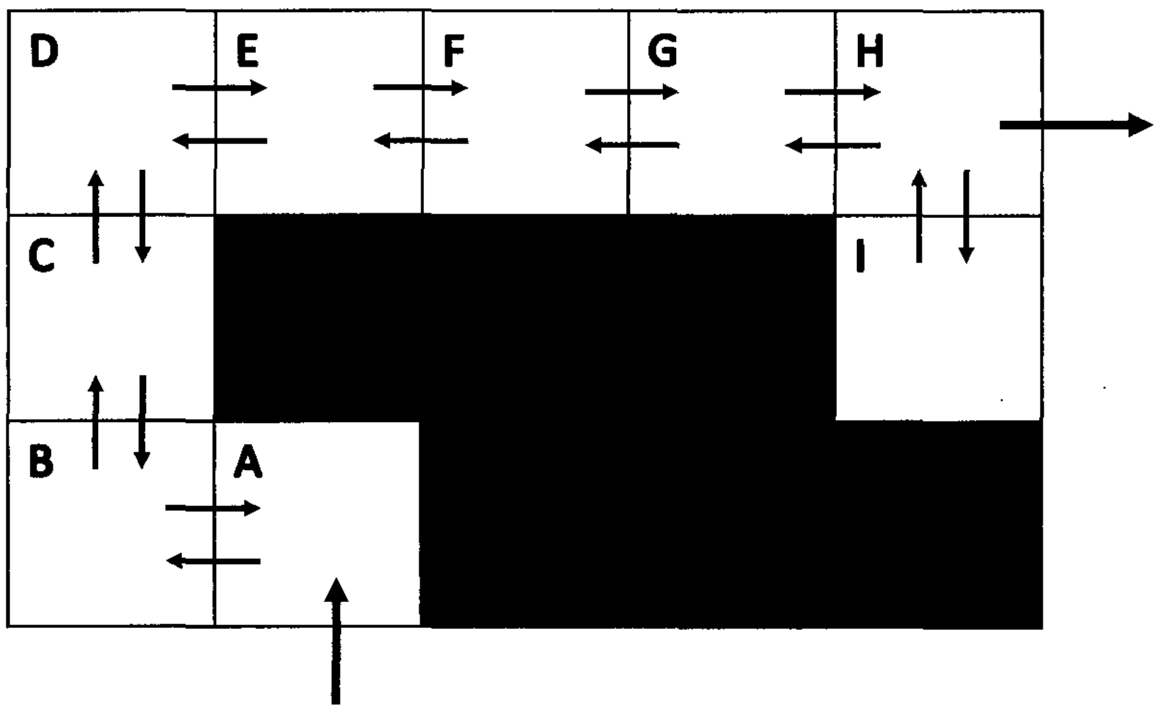


Part (c) Q-learning [2 MARKS]

Suppose that the robot is performing the Q-learning algorithm. The trajectories that the robot has explored are: EFGH, IH, FEDC, ABCDEFGFGH. What is the estimated action-value function $\hat{Q}(s,a)$ after executing these trajectories? Write down the values of $\hat{Q}(s,a)$ in between every pair of adjacent cells with an arrow to indicate the direction.

Recall the Q-learning algorithm is updated using the following rule:

$$\hat{Q}(s,a) \leftarrow r(s,a) + \gamma \max_{a'} \hat{Q}(s',a')$$



Question 5. Generative vs. Discriminative classifiers [5 MARKS]

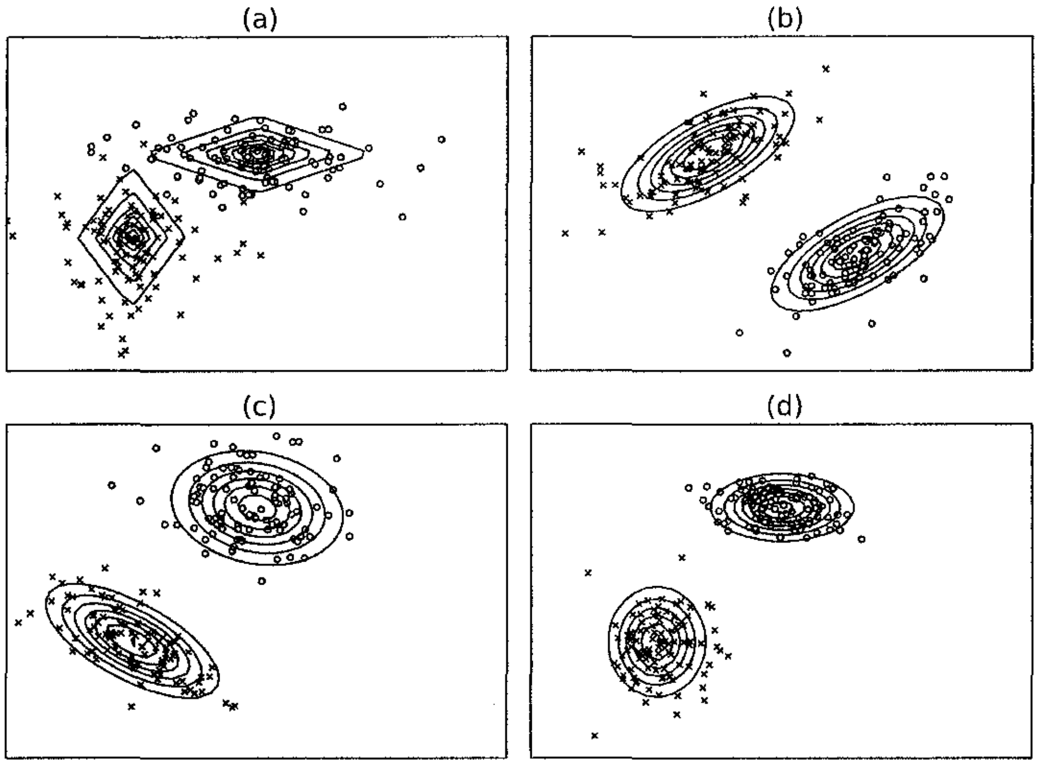
Part (a) Decision boundary [2 MARKS]

Imagine that you are training a Gaussian Bayes classifier on two classes. What does the decision boundary look like when the covariance matrix is shared between two classes? Why? (show your derivation)

Note: the multivariate Gaussian distribution is $\frac{1}{\sqrt{(2\pi)^k|\Sigma|}}\exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}))$, where $\boldsymbol{\mu}$ is the mean vector and Σ is the covariance matrix.

Part (b) Naive Bayes classifier [1 MARK]

Which of the following diagrams could NOT be a visualization of a Naive Bayes classifier? Select all that applies.



For each of the parts below, clearly circle one of the given options and justify your answer in a single sentence.

Part (c) [1 MARK]

Brian works for a software company, Facegram, which receives a large number of job applicants. Brian's task is to use machine learning to sort the applications into accept/reject categories and detect outliers, which may indicate a falsified application. Should Brian use a generative or discriminative classifier?

GENERATIVE DISCRIMINATIVE

Part (d) [1 MARK]

Catherine works for a start-up which aims to create software classifying whether or not an image contains a hot-dog. There is plenty of training data available using images from the internet. Should Catherine use a generative or discriminative classifier?

GENERATIVE DISCRIMINATIVE

Question 6. Neural Networks [5 MARKS]

Part (a) [1 MARK]

Describe two benefits of convolutional layers over fully-connected layers in neural networks applied to images.

Part (b) [2 MARKS]

Suppose we have a convolution layer which takes as input an array $\mathbf{x} = \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix}$ and convolves $\mathbf{x}$ with $\begin{pmatrix} 3 & 4 \end{pmatrix}$. This layer uses the identity activation function. The output is an array of length 4.

Now let's design a fully connected layer which computes the same function. It has an identity activation function and no bias, so it computes $\mathbf{y} = \mathbf{W}\mathbf{x}$, where the output $\mathbf{y}$ is a vector of length 4. Give the 4×3 weight matrix $\mathbf{W}$ which makes this fully connected layer equivalent to the convolution layer above.

(HINT: Begin by writing the values of each output as a linear function of the inputs.)

Part (c) [2 MARKS]

Consider one layer of a multilayer perceptron (MLP), which takes in a vector of hidden units $\mathbf{h} \in \mathbb{R}^D$ and outputs another vector $\mathbf{y} \in \mathbb{R}^D$ of the same dimensionality. The layer uses a weight matrix $\mathbf{W} \in \mathbb{R}^{D \times D}$ and bias $\mathbf{b} \in \mathbb{R}^D$ and its computations are defined as follows:

$$z_d = \sum_{j=1}^{D} w_{dj} h_j + b_d$$
$$y_d = \phi(z_d) + h_d,$$

where ϕ is a nonlinear activation function.

Recall we use the notation $\overline{v}$ to denote the derivative of the loss $\mathcal{L}$ with respect to v: $\overline{v} = \frac{d\mathcal{L}}{dv}$. Give the backprop rules for $\overline{z_d}$, $\overline{h_j}$ and $\overline{w_{dj}}$ in terms of the error signal $\overline{y_d}$. You can use ϕ' to denote the derivative of ϕ.

$\overline{z_d} =$

$\overline{h_j} =$

$\overline{w_{dj}} =$

Question 7. SVM [5 MARKS]

Consider the soft-margin SVM objective:

$$\min_{\mathbf{w},\xi^{(n)}\geqslant 0} \quad \|\mathbf{w}\|^2 + C\sum_{n=1}^{N}\xi^{(n)}$$
$$\text{subject to} \quad y^{(n)}(\mathbf{w}^\top \mathbf{x}^{(n)} + b) \geqslant 1 - \xi^{(n)} \quad \forall n. \tag{1}$$

Part (a) [1 MARK]

Explain each term in the objective above.

Part (b) [2 MARKS]

Consider the case when $C > 0$ and the dataset is linearly separable. Is the decision boundary learned by a soft-margin SVM guaranteed to separate the classes? Why?

Part (c) [2 MARKS]

Show the objective can be rewritten as an unconstrained objective with a combination of hinge loss and L2 weight regularization (HINT: For a fixed $\mathbf{w}$, there are unique values of $\xi^{(n)}$ minimizing the objective.)

Question 8. Gaussian Processes and Bayesian Linear Regression [4 MARKS]

Part (a) Name one advantage of using a Gaussian process over Bayesian linear regression [1 MARK]

Part (b) Name one advantage of using Bayesian linear regression over a Gaussian process [1 MARK]

Part (c) [2 MARKS]

Suppose we are performing Bayesian linear regression on a dataset $\{(\mathbf{x}^{(n)}, t^{(n)})\}_{n=1}^{N}$ where we apply the following feature mapping $\boldsymbol{\phi} : \mathbb{R}^D \to \mathbb{R}^N$ to our inputs $\mathbf{x} \in \mathbb{R}^D$:

$$\phi_n(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}^{(n)}\|^2}{2s^2}\right)$$

for some fixed $s \in \mathbb{R}$. Furthermore, suppose we are using the noise model $p(t|\mathbf{x}) = \mathcal{N}(t|\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}), \beta^2)$ over the targets $t \in \mathbb{R}$. Recall the posterior distribution over the weights $\mathbf{w}$ is given by $p(\mathbf{w}|\mathcal{D}) = \mathcal{N}(\mathbf{w}|\mathbf{m}, \Sigma)$ for some $\mathbf{m}, \Sigma$. This gives rise to the posterior predictive distribution:

$$p(t|\mathbf{x}, \mathcal{D}) = \mathcal{N}(t|\mu_{\text{pred}}(\mathbf{x}), \sigma^2_{\text{pred}}(\mathbf{x}))$$

where:

$$\mu_{\text{pred}}(\mathbf{x}) = \mathbf{m}^\top \boldsymbol{\phi}(\mathbf{x}), \qquad \sigma^2_{\text{pred}}(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^\top \Sigma \boldsymbol{\phi}(\mathbf{x}) + \beta^2$$

Suppose $\mathbf{x}$ lies far away from any of the points in our dataset. Use no more than two sentences to answer each of the parts below.

Describe what happens to $\mu_{\text{pred}}(\mathbf{x})$. Explain why this is desired or undesired.

Describe what happens to $\sigma^2_{\text{pred}}(\mathbf{x})$. Explain why this is desired or undesired.

Question 9. Bayesian optimization [2 MARKS]

Suppose we are using Bayesian optimization to minimize a function $f(\theta)$.

Part (a) Give one reason why Probability of Improvement is a better acquisition function than the negative predictive mean (i.e. $-\mathbb{E}[f(\theta)]$) [1 MARK]

Part (b) Give one reason why Expected Improvement is a better acquisition function than Probability of Improvement. [1 MARK]

Question 10. Expectation Maximization [7 MARKS]

Recall that in the discriminative setting we wish to predict targets $t \in \mathbb{R}$ from inputs $\mathbf{x} \in \mathbb{R}^D$. Suppose we believe the targets are a linear function of the input on some region of input space and a different linear function on another region. We can encode these beliefs using a latent variable. The resulting model is called a *Mixture of Experts* model.

In this model, an "expert" $z \in \{0, 1\}$ is selected based on the input $\mathbf{x}$, then the target t is generated as a linear function of $\mathbf{x}$, where the weights depend on which expert was chosen. More formally, we use the following probabilistic model:

$$p(z = 1|\mathbf{x};\boldsymbol{\theta}) = \sigma(\mathbf{c}^\top \mathbf{x})$$

$$p(t|\mathbf{x}, z;\boldsymbol{\theta}) = \begin{cases} \mathcal{N}(t|\mathbf{w}_0^\top \mathbf{x}, \sigma_0^2), & z = 0 \\ \mathcal{N}(t|\mathbf{w}_1^\top \mathbf{x}, \sigma_1^2), & z = 1 \end{cases}$$

The parameters of this model are $\boldsymbol{\theta} = \{\mathbf{c}, \mathbf{w}_0, \mathbf{w}_1, \sigma_0, \sigma_1\}$. Suppose we observe a dataset $\mathcal{D} = \{(\mathbf{x}^{(n)}, t^{(n)})\}_{n=1}^N$. In this question, you will derive some of the necessary steps for the EM algorithm applied to this model.

Part (a) [1 MARK]

As a warm-up, answer by clearly circling one of the options below: when applying the EM updates to the parameters, they may get trapped in a local minima.

TRUE FALSE

Part (b) [1 MARK]

Write the complete data log-likelihood for this model. Do not replace the sigmoid with its definition $\sigma(x) = \frac{1}{1+\exp(-x)}$ or substitute the pdf of the normal for $\mathcal{N}$. (HINT: Use that $p(z|\mathbf{x};\boldsymbol{\theta}) = [\sigma(\mathbf{c}^\top \mathbf{x})]^z [(1 - \sigma(\mathbf{c}^\top \mathbf{x}))]^{1-z}$ and $p(t|\mathbf{x}, z;\boldsymbol{\theta}) = \left[\mathcal{N}(t|\mathbf{w}_1^\top \mathbf{x}, \sigma_1^2)\right]^z \left[\mathcal{N}(t|\mathbf{w}_0^\top \mathbf{x}, \sigma_0^2)\right]^{1-z}$.)

Part (c) [1 MARK]

Give an expression for the posterior probability $p(z = 1|\mathbf{x}, t; \boldsymbol{\theta})$. Do not replace the sigmoid with its definition $\sigma(x) = \frac{1}{1+\exp(-x)}$ or substitute the pdf of the normal for $\mathcal{N}$.(HINT: You will need to use Bayes rule.)

Part (d) [2 MARKS]

Let $p_n = p(z^{(n)} = 1|\mathbf{x}^{(n)}, t^{(n)}; \boldsymbol{\theta}^{\text{old}})$. Give the expected complete-data log-likelihood, substituting p_n where appropriate. Do not replace the sigmoid with its definition $\sigma(x) = \frac{1}{1+\exp(-x)}$ or substitute the pdf of the normal for $\mathcal{N}$. (HINT: $z^{(n)}$ should not appear in the resulting expression)

Part (e) [2 MARKS]

Using the expected complete data log-likelihood, find the M-step update for σ_1^2. Here, you will need to replace $\mathcal{N}$ with the normal pdf:

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

(HINT: You only need to consider the parts of the expected complete data log-likelihood which contain σ_1^2. Maximize with respect to σ_1^2 and not σ_1.)

Use the space on this "blank" page for scratch work, or for any solution that did not fit elsewhere.
Clearly label each such solution with the appropriate question and part number.

Use the space on this "blank" page for scratch work, or for any solution that did not fit elsewhere.
Clearly label each such solution with the appropriate question and part number.