Last (Family) Name:

First (Given) Name:

Student Number:

Section *(circle one)*: L0101 = Mon, L0201 = Wed, L0301 = Th

UNIVERSITY OF TORONTO
FACULTY OF ARTS & SCIENCE

DECEMBER 2019 EXAMINATIONS
CSC311H1F

Introduction to Machine Learning

Duration - 3 hours
Aids allowed: Two double-sided handwritten or typed $8.5'' \times 11''$ 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 only in the space provided after each question. Last few pages are provided for scratch work. They won't be graded.
- 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
WRITE ALL ANSWERS ON THIS PAPER

1. True/False [12 pts]. For each statement below, write whether it is true or false, and give a **one or two sentence** justification of your answer.

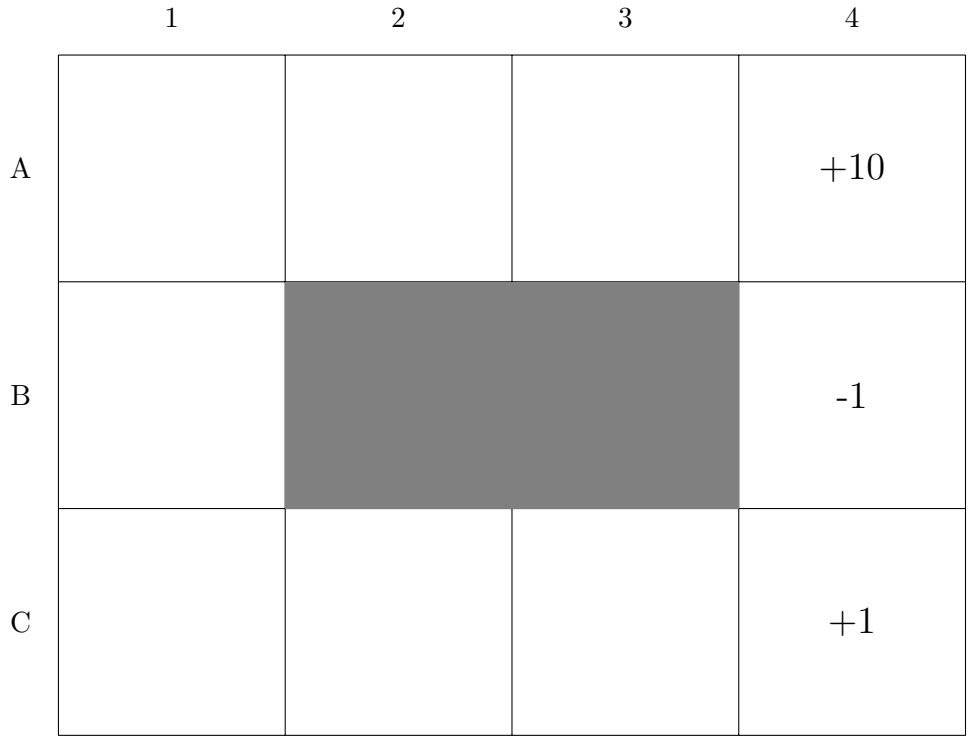
a) [3 pts] As we increase the number of parameters in a neural network model, both the bias and variance of our predictions on the test set should decrease since we can better fit the data.

b) [3 pts] We can always improve the performance of a clustering algorithm like k-means by removing features and reducing the dimensionality of our data.

c) [3 pts] Bagging and boosting are both ensemble methods that reduce the variance of our models.

d) [3 pts] Generative and discriminative approaches both model the probability distribution of inputs given target.

2. Reinforcement Learning [15 pts].



Consider the familiar robot navigation task within the gridworld shown above. You can move in any of the four directions (left/right/up/down) unless blocked by one of the gray obstacles at B2 and B3. The rewards are +1 for entering state C4, -1 for entering state B4, and +10 for entering state A4. A4 is an absorbing state. The rewards for every other state are 0.

a) [5 pts] Assume that the state transitions are deterministic. Recall that under the simple Q-learning algorithm, the estimated Q values are updated using the following rule:

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

Consider applying this algorithm when all the $\hat{Q}$ values are initialized to zero and $\gamma = 0.8$. Write the Q estimates on the figure as labeled arrows after the robot has executed the following state sequences:

- C1 → C2 → C3 → C4 → B4
- A2 → A3 → A4
- A1 → B1 → A1 → A2 → A3 → A4

Extra copies of the gridworld have been provided on the scratch paper at the back of this exam.

.

b) [5 pts] After executing these trajectories, the robot now uses the policy of always performing the action having the greatest Q value. Is this the optimal policy? Why or why not?

c) [5 pts] Suppose the robot is now allowed to explore the environment using a stochastic policy for 10,000 episodes. Is it guaranteed to find the optimal Q function? Why or why not?

3. Back propagation in Neural Networks [15 pts].
Consider a neural network $y = f(\mathbf{x})$ with a single hidden layer with 2 hidden units that uses the relu activation function ($\sigma(x) = \max(0, x)$) and has no biases. The network predicts 1d targets $t \in \mathbb{R}$ using 2d input data $\mathbf{x} = [x_1, x_2]^\top \in \mathbb{R}^2$. To train this model we can use mean squared error from the targets $\mathcal{L}(y, t) = (y - t)^2/2$. We have one training data $\mathbf{x}_0 = [1, 2]^\top$ with target $t_0 = 1$. This model has

1. input layer weights $\mathbf{W} = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \in \mathbb{R}^{2\times 2}$ and output layer weights $\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \in \mathbb{R}^2$.
2. input to the hidden units $\mathbf{z} = [z_1, z_2]^\top$ and the output from the hidden units $\mathbf{h} = [h_1, h_2]^\top$
3. At the current training step the forward pass has weights : $\mathbf{W} = \begin{bmatrix} -1 & 1 \\ 1 & -2 \end{bmatrix}$, $\mathbf{w} = \begin{bmatrix} 3 \\ -1 \end{bmatrix}$

.
a) [2 pts] Write down the forward pass in terms of W_{ij}, w_j, z_j, h_j. Then compute $f(\mathbf{x}_0)$.

$$z_1 =$$
$$z_2 =$$
$$h_1 =$$
$$h_2 =$$
$$y =$$

b) [2 pts] Compute $\frac{\partial \mathcal{L}}{\partial y}$ and the training loss at $\mathbf{x_0}$. Do we need to compute the training loss for the backpropagation update?

c) [2 pts] Write down the derivatives of the loss with respect to the outputs of the hidden layer (in terms of $\frac{\partial \mathcal{L}}{\partial y}$) and compute their exact values at this training step.

$$\frac{\partial \mathcal{L}}{\partial h_1} =$$

$$\frac{\partial \mathcal{L}}{\partial h_2} =$$

d) [2 pts] Write down the derivatives of the loss with respect to the input to the hidden layer (in terms of $\frac{\partial \mathcal{L}}{\partial h_1}, \frac{\partial \mathcal{L}}{\partial h_2}$) and compute their exact values at this training step.

$$\frac{\partial \mathcal{L}}{\partial z_1} =$$

$$\frac{\partial \mathcal{L}}{\partial z_2} =$$

e) [4 pts] Write down the derivatives of the loss with respect to the parameters of the input layer $\frac{\partial \mathcal{L}}{\partial W_{ij}}$ and compute their exact values at this training step. Do you need the derivatives of loss with respect to the input to the model $\frac{\partial \mathcal{L}}{\partial x_j}$ to calculate this?

$$\frac{\partial \mathcal{L}}{\partial W_{11}} =$$

$$\frac{\partial \mathcal{L}}{\partial W_{12}} =$$

$$\frac{\partial \mathcal{L}}{\partial W_{21}} =$$

$$\frac{\partial \mathcal{L}}{\partial W_{22}} =$$

f) [3 pts] In f) you should find some of the derivatives become zero, while others will be larger than any of the activations or weights. Explain why this is.

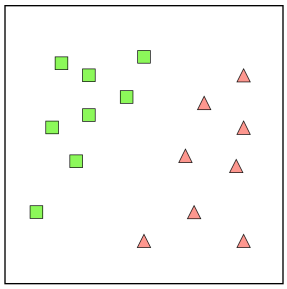
4. Support Vector Machines [12 pts].
a) [6 pts] Given a true label $t^{(i)} \in \{-1, +1\}$ and an output value $z^{(i)}(\mathbf{w}, b) = \mathbf{w}^T\mathbf{x}^{(i)} + b$, a linear SVM learns to minimize the hinge loss :

$$\min_{\mathbf{w},b} \sum_{i=1}^{N} \max\{0, 1 - t^{(i)} z^{(i)}(\mathbf{w}, b)\}$$

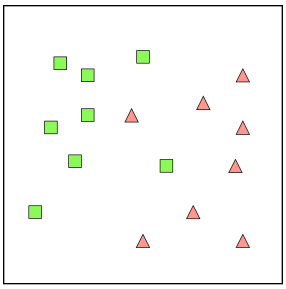
Suppose we replace this loss with a zero-one loss and instead minimize the following:

$$\min_{\mathbf{w},b} \sum_{i=1}^{N} \mathbb{1}_{\{t^{(i)} \neq \text{sign}(z^{(i)}(\mathbf{w},b))\}}$$

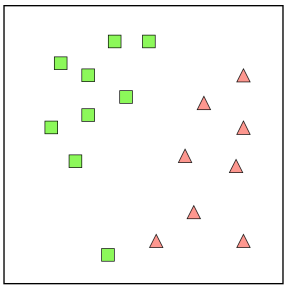
How does this change the decision boundary learned? Why do we use the hinge loss instead of the zero-one loss when learning our decision boundary?



(a)



(b)



(c)

b) [6 pts] For each distribution above, draw the decision boundary that an SVM classifier as given in the previous part (a linear SVM with trained with hinge loss) would learn.

5. Naive Bayes and Probabilistic Models [15 pts]. Assume we have two classes : spam and non-spam. We have a dictionary of D words, and binary features $\mathbf{x} = [x_1, ..., x_D]$ saying whether each word appears in the e-mail. We want to write down a joint distribution of the model.

a) [3 pts] What is the problem in calculating this joint distribution? How many parameters would be required to construct this model? How does the Naive Bayes model handle this problem?

b) [3 pts] Write down the joint distribution under the Naive Bayes assumption. How many parameters are specified under the model now?

c) [3 pts] Why does the Naive Bayes assumption allow for the model parameters to be learned efficiently? Write down the log-likelihood to explain why. Here, assume that you observed N data points $(\mathbf{x}^{(i)}, t^{(i)})$ for $i = 1, 2, ..., N$.

d) [3 pts] Show how you can use Bayes rule to predict a class given a data point.

e) [3 pts] Explain how placing a Beta prior on parameters is helpful for the Naive Bayes model.

6. k-means and EM [15 pts].

a) [3 pts] Is the k-means algorithm guaranteed to converge to the global minimum? Explain why or why not. Draw a 2d example of a dataset and an initialization that k-means will have difficulty clustering.

b) [3 pts] What is the benefit of making **soft assignments** in the k-means algorithm. What function do we use to make soft assignments (write explicitly)? Under what conditions does soft k-means become hard k-means?

c) [3 pts] What are some remaining issues with soft k-means? How does a Gaussian Mixture model address them?

d) [3 pts] For some data $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}$ the GMM assumes the following generative model

$$p(z = k) = \pi_k$$
$$p(\mathbf{x}|z = k) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \propto \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)\right\}.$$

Show that when $\boldsymbol{\Sigma}_k = \boldsymbol{I}/(2\beta)$, posterior probability $p(z = k|\mathbf{x})$ is given by

$$p(z = k|\mathbf{x}) = \frac{\pi_k \exp\left(-\beta||\mathbf{x} - \boldsymbol{\mu}_k||^2\right)}{\sum_{j=1}^{K} \pi_j \exp\left(-\beta||\mathbf{x} - \boldsymbol{\mu}_j||^2\right)}.$$

e) [3 pts] Under what conditions does the EM algorithm reduce to the soft k-means algorithm? Under what conditions does the EM algorithm reduce to the hard k-means algorithm?

7. Principal Component Analysis [16 pts]. Suppose that you are given a **centered** dataset of N samples, i.e., $\mathbf{x}^{(i)} \in \mathbb{R}^D$ for $i = 1, 2, ..., N$. We say that *the data is centered* if the mean of the samples is equal to 0, i.e., $\frac{1}{N}\sum_{i=1}^{N} \mathbf{x}^{(i)} = 0$.
For a given unit direction $\mathbf{u} \in \mathbb{R}^D$ such that $\|\mathbf{u}\|_2 = 1$, we denote by $\mathcal{P}_{\mathbf{u}}(\mathbf{x})$, the Euclidean projection of $\mathbf{x}$ on $\mathbf{u}$. Recall that the projection is given by

$$\mathcal{P}_{\mathbf{u}}(\mathbf{x}) = \mathbf{u}^\top \mathbf{x}\mathbf{u} \in \mathbb{R}^D.$$

(a) **[3 pts]** *Mean of data after projecting on* $\mathbf{u}$*:* Show that the projected data with samples $\mathcal{P}_{\mathbf{u}}(\mathbf{x}^{(i)})$ in any unit direction $\mathbf{u}$ is still centered. That is, show

$$\frac{1}{N}\sum_{i=1}^{n} \mathcal{P}_{\mathbf{u}}(\mathbf{x}^{(i)}) = 0.$$

(b) **[6 pts]** *Maximum variance:* Recall that the first principal component $\mathbf{u}_*$ is given by the largest eigenvector of the sample covariance (eigenvector associated to the largest eigenvalue). That is,

$$\mathbf{u}_* = \underset{\mathbf{u}\,:\,\|\mathbf{u}\|_2=1}{\operatorname{argmax}} \mathbf{u}^\top \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}^{(i)} [\mathbf{x}^{(i)}]^\top \mathbf{u}.$$

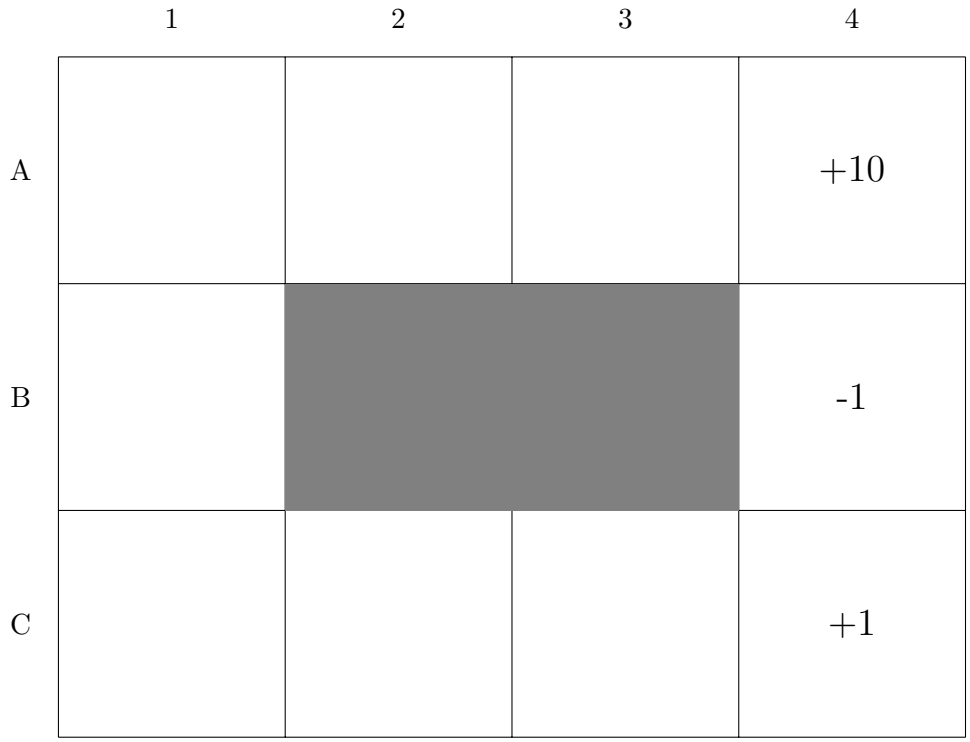
Using this, show that the unit direction $\mathbf{u}$ that maximizes the variance of the projected data corresponds to the first principal component of the data. That is, show

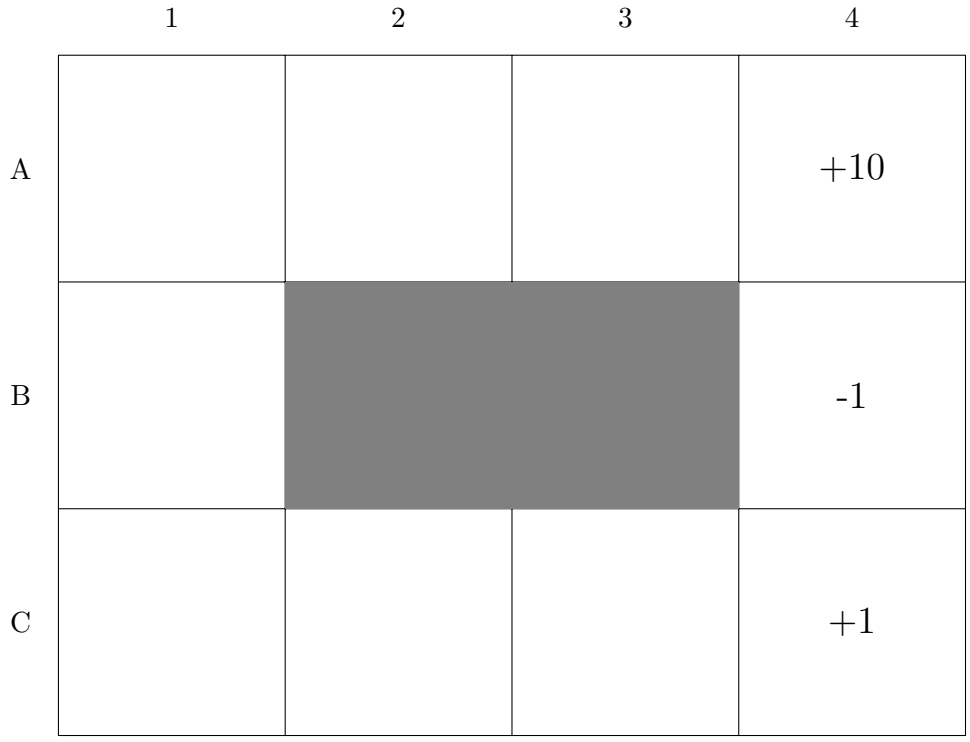
$$\mathbf{u}_* = \underset{\mathbf{u}\,:\,\|\mathbf{u}\|_2=1}{\operatorname{argmax}} \sum_{i=1}^{N} \left\| \mathcal{P}_{\mathbf{u}}(\mathbf{x}^{(i)}) - \frac{1}{N} \sum_{j=1}^{N} \mathcal{P}_{\mathbf{u}}(\mathbf{x}^{(j)}) \right\|_2^2.$$

(c) **[7 pts]** *Minimum error:* Show that the unit direction $\mathbf{u}$ that minimizes the mean squared error between projected data points $\mathcal{P}_{\mathbf{u}}(\mathbf{x}^{(i)})$ and the original data points $\mathbf{x}^{(i)}$ corresponds to the first principal component $\mathbf{u}_*$. That is, show

$$\mathbf{u}_* = \underset{\mathbf{u}\,:\,\|\mathbf{u}\|_2=1}{\operatorname{argmin}} \sum_{i=1}^{N} \|\mathbf{x}^{(i)} - \mathcal{P}_{\mathbf{u}}(\mathbf{x}^{(i)})\|_2^2. \tag{7.1}$$

SCRATCH WORK ONLY: THIS PAGE WILL NOT BE GRADED





SCRATCH WORK ONLY: THIS PAGE WILL NOT BE GRADED

SCRATCH WORK ONLY: THIS PAGE WILL NOT BE GRADED

SCRATCH WORK ONLY: THIS PAGE WILL NOT BE GRADED

SCRATCH WORK ONLY: THIS PAGE WILL NOT BE GRADED

SCRATCH WORK ONLY: THIS PAGE WILL NOT BE GRADED

SCRATCH WORK ONLY: THIS PAGE WILL NOT BE GRADED