Last (Family) Name:

First (Given) Name:

Student Number:

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

PRACTICE FINAL EXAM

CSC311 Fall 2019
Introduction to 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 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
DO NOT WRITE ANY ANSWERS ON THIS PAPER

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

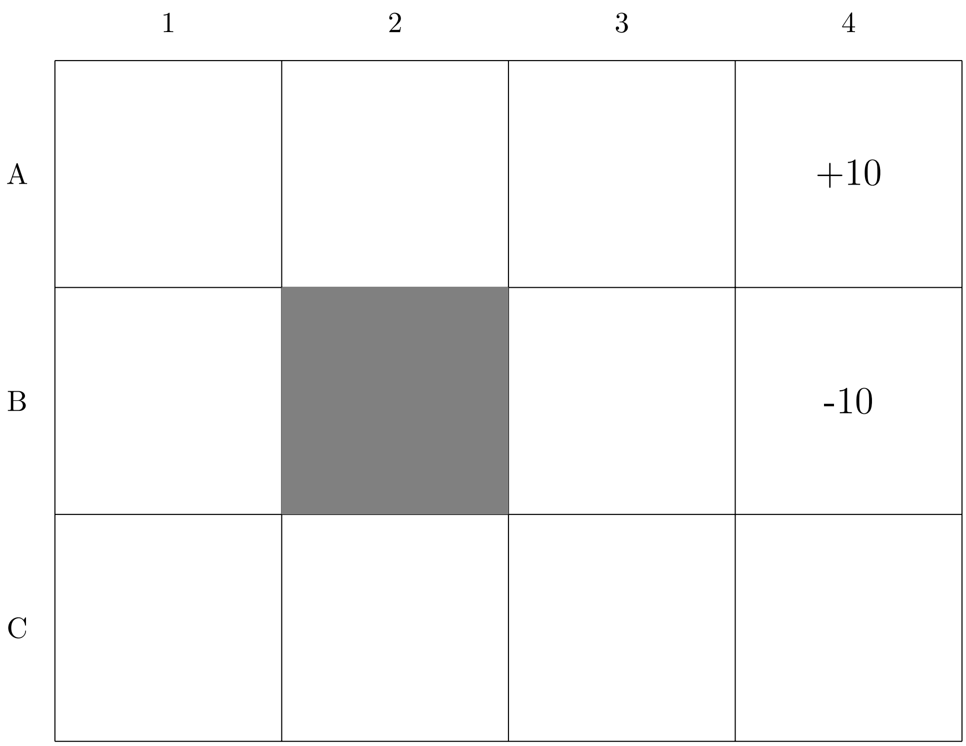
a) Adding more training data always reduces overfitting.

b) For small k, the k-means algorithm is equivalent to the k-nearest neighbors algorithm.

c) An ensemble of models always has more capacity than a single model.

d) A linear SVM will find the same decision boundary as logistic regression.

2. Reinforcement Learning.



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 +10 for state C4, and -10 for state B4. A4 and B4 are both absorbing states. The reward for every other state is 0.

a) Assume that the state transitions are deterministic. Recall that under the simple Q-learning algorithm, the estimate Q values are update 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:

- B1 $\rightarrow$ A2 $\rightarrow$ A2 $\rightarrow$ A3 $\rightarrow$ B3 $\rightarrow$ B4
- A2 $\rightarrow$ A3 $\rightarrow$ A4
- C1 $\rightarrow$ C2 $\rightarrow$ C3 $\rightarrow$ B3 $\rightarrow$ A3 $\rightarrow$ A4

.

b) Assume the robot will now use the policy of always performing the action having the greatest Q value. Is this the optimal policy? Why or why not?

c) Suppose state A3 also has a reward of -10. How can we ensure that our agent is still able to find the optimal policy in this new environment?

3. Backpropagation. Consider a L_2 regularized single layer neural network model that predicts continuous 1d targets $y = \sigma(z) \in \mathbb{R}$ where $z = wh + b$ and $h = \sigma(w'x + b')$, and σ is an activation function. To train, we use mean squared error from the targets with L_2 penalty on $t \in \mathbb{R}: \ \mathcal{L} = (y - t)^2/2 + w^2 + b'^2$

a) Write the loss as a function of the parameters w, b and compute directly $\frac{\partial \mathcal{L}}{\partial w}, \frac{\partial \mathcal{L}}{\partial b'}$.

b) Now compute $\frac{\partial \mathcal{L}}{\partial w}, \frac{\partial \mathcal{L}}{\partial b'}$ using the backprogation algorithm.

c) What are the disadvantages of doing a) versus backpropagation? Why do we use backpropagation in machine learning as opposed to direct differentiation?

4. Principal Component Analysis. Recall that the optimal PCA subspace can be determined from the eigendecomposition of the empirical covariance matrix $\hat{\boldsymbol{\Sigma}}$. Also recall that the eigendecomposition can be expressed in terms of the following spectral decomposition of $\hat{\boldsymbol{\Sigma}}$:

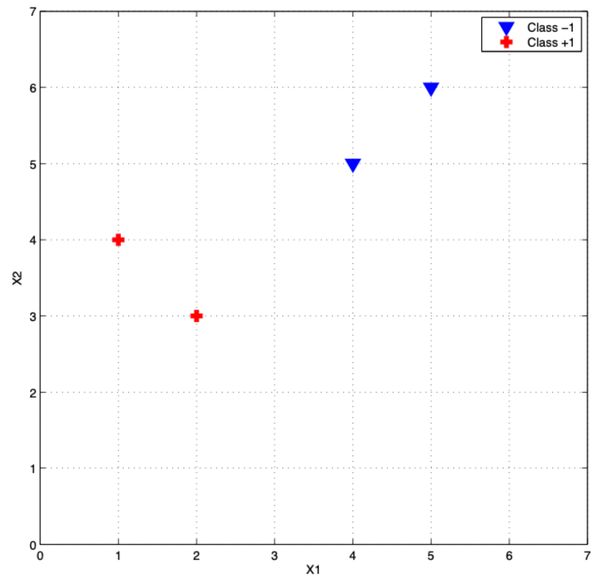
$$\hat{\boldsymbol{\Sigma}} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\top,$$

where $\mathbf{Q}$ is an orthogonal matrix and $\boldsymbol{\Lambda}$ is a diagonal matrix. Assume the eigenvalues are sorted from largest to smallest. You may assume all of the eigenvalues are distinct.

1. If you've already computed the eigendecomposition (i.e. $\mathbf{Q}$ and $\boldsymbol{\Lambda}$), how do you obtain the orthogonal basis $\mathbf{U}$ for the optimal PCA subspace? (You do not need to justify your answer.)

2. The PCA code vector for a data point $\mathbf{x}$ is given by $\mathbf{z} = \mathbf{U}^\top(\mathbf{x} - \hat{\boldsymbol{\mu}})$ where $\hat{\boldsymbol{\mu}}$ is the data mean. Show that the entries of $\mathbf{z}$ are uncorrelated.

5. Support Vector Machines.
Support vector machines learn a decision boundary leading to the largest margin from both classes. You are training SVM on a tiny dataset with 4 points shown below. This dataset consists of two examples with class label -1 (denoted with plus), and two examples with class label +1 (denoted with triangles).



a) Write down the SVM loss function for this data and state how to find the weight vector $\mathbf{w}$ and bias b.

b) Draw the (approximate) decision boundary.

6. Probabilistic Models .

The Laplace distribution, parameterized by μ and β, is defined as follows:

$$\text{Laplace}(w; \mu, \beta) = \frac{1}{2\beta} \exp\left(-\frac{|w - \mu|}{\beta}\right).$$

Consider a variant of the homework2 question where we assume that the prior over the weights $\mathbf{w}$ consists of an independent zero-centered Laplace distribution for each dimension, with shared parameter β:

$$w_j \sim \text{Laplace}(0, \beta)$$
$$t \,|\, \mathbf{w} \sim \mathcal{N}(t; \mathbf{w}^\top \mathbf{x}, \sigma^2)$$

For reference, the Gaussian PDF is:

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

1. Suppose you have a labeled training set $\{(\mathbf{x}^{(i)}, t^{(i)})\}_{i=1}^N$. Give the cost function you would minimize to find the MAP estimate of $\mathbf{w}$.

2. Based on your answer to part (a), how might the MAP solution for a Laplace prior differ from the MAP solution if you use a Gaussian prior (which is exactly homework2)?

7. EM Algorithm.

1. Is EM algorithm a supervised or an unsupervised learning method? Explain your answer.

2. How does EM algorithm and k-means compare? Write 3 similarities and 3 differences.

3. Explain why we call these steps expectation and maximization steps. What is it that we take expectation of and what is it that we maximize?

SCRATCH WORK ONLY: THIS PAGE WILL NOT BE GRADED

	1	2	3	4
A				+10
B				-10
C				

	1	2	3	4
A				+10
B				-10
C				

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