UNIVERSITY OF TORONTO
Faculty of Arts & Science

PRACTICE EXAM EXAMINATIONS

CSC 311 H1F

Duration: 3 hours

Aids Allowed: One *double*-sided *8.5"×11"* reference sheet and one non-programmable calculator.

Do ***not*** *turn this page until you have received the signal to start.*
In the meantime, write your name, and student number below (please do this now!) and *carefully* read *all* the information on the rest of this page.

- 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.
- You must obtain a grade of at least 40% on the final exam to pass the course.
- When you are done with 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.
- This final examination consists of 5 questions worth 74 marks on 16 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.*
- 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

Question 1. Short Answers [42 MARKS]

Part (a) [2 MARKS]

Suppose you have a choice between two neural net architectures for a classification task. In the first layer, Architecture A has a convolution layer with M input units and N output units. Architecture B has a fully connected layer with M input units and N output units. Both architectures use the same activation function, and subsequent layers are identical between the two architectures.

Which architecture has more parameters? Why?

Part (b) [2 MARKS]

Consider logistic regression with L2 regularization. The weight for the L2 regularizer (λ) is a hyperparameter for this model. Suppose that we are optimizing the weights using gradient descent. Is it a good idea to choose a value for λ by minimizing accuracy on a validation set?

Circle the correct answer: Yes or No

Explain briefly:

Part (c) [2 MARKS]

What is *underfitting* in Machine Learning and how do we prevent it?

Part (d) [5 MARKS]

What is *overfitting* in Machine Learning? List four methods to prevent overfitting and explain when each method should be preferably used.

Part (e) [2 marks]

Why is it called *Logistic Regression* if it is a classification model? What are the outputs of a logistic regression model?

Part (f) [4 marks]

Explain the *Gradient Descent* algorithm. Write down the update formula in each iteration. Visualize and explain the effects of high and low learning rates.

Part (g) [4 marks]

What are the most common activation functions for the hidden and output layers of a Neural Network. For the output layer, name the activation functions used for each model type (i.e. regression, multi-class classification, and multi-label classification). Write a short sentence for the reason behind using each function.

Part (h) [2 marks]

Explain *Auto-Differentiation* and its relationship with the Backpropagation algorithm.

Part (i) [2 marks]

What are the benefits of *Convolutional Networks* versus *Fully Connected Networks* when we are dealing with images as the model input? List two benefits.

Part (j) [4 marks]

Explain *Equivariance* and *Invariance* properties of Convolutional Networks and their relationship with each layer type of the network (i.e. convolution layer and pooling layer).

Part (k) [3 marks]

Explain the difference between *Discriminative* and *Generative* classifiers. Mathematically explain how the classification task is done in a generative classifier.

Part (l) [2 marks]

What is the Naive assumption in the *Naive Bayes Models*. At what stage will you use Bayes rule in this algorithm?

Part (m) [3 marks]

Explain *Maximum Likelihood Estimation* and *Maximum A-Posteriori (MAP) Estimation* methods. What assumption do we make for MLE that makes it different from MAP.

Part (n) [2 marks]

Explain the *Local Minima* issue in the *K-Means* clustering algorithm. Visualize the problem with a simple example.

Part (o) [3 marks]

Explain, in words, the *Exploration-exploitation tradeoff* in Reinforcement Learning. Provide one example of Exploration-Exploitation in the real world.

Question 2. Sensitivity and Specificity [8 marks]

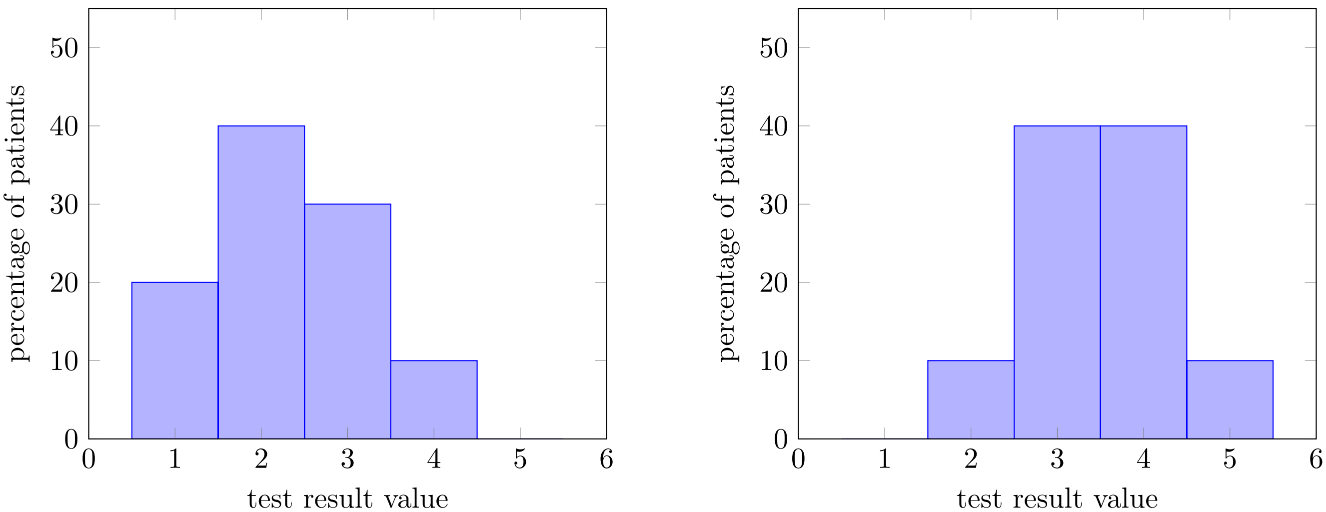
We want to build a model to diagnose a disease. Whether a patient has the disease is correlated with the result of Test A.

- Figure 1a shows the distribution of test result values given that the patient **does not have the disease**.
- Figure 1b shows the distribution of test result values given that the patient **has the disease**.

Our model makes the prediction based on a threshold value for the test.

For example, if we choose the threshold value to be 2.5, then the predictive rules are follows:

- If the patient's test result value is **greater than or equal to** 2.5, then our model predicts that the patient **has the disease**.
- Otherwise, if the patient's test result value is **less than** 2.5, then our model predicts that the patient **does not have the disease**.



(a) Test result if patient doesn't have disease D

(b) Test result if the patient has disease D

Figure 1: Test result distributions

Complete the table below. For each test threshold value, calculate the sensitivity and specificity of our predictive model. Recall that sensitivity is the true positive rate and specificity is the true negative rate. Express every value as a percentage.

Test Threshold Value	1.5	2.5	3.5	4.5
Sensitivity				
Specificity				

Table 1: Sensitivity and Specificity for Different Test Threshold Values

This page intentionally left blank

Question 3. Linear Regression [10 marks]

Consider a data set with three examples in Figure 2.

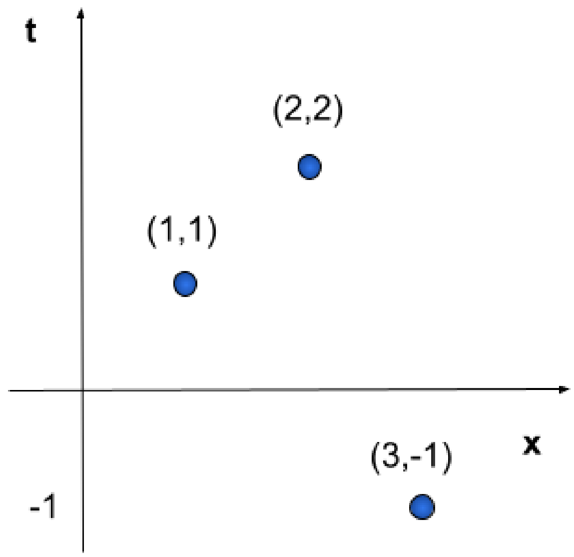


Figure 2: Data set for linear regression

We will fit a linear regression model to this data set. The closed-form solution for linear regression is given below.

$$w = (X^T X)^{-1} X^T t$$

Calculate the values of w for the linear regression model by using the formula above. We will give partial marks for correct intermediate steps.

Below is the formula to invert a 2x2 matrix.

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Question 4. Neural Networks [6 marks]

Part (a) [2 marks]

Your friend is training a feed-forward neural network model for a regression problem. They produced the following curve of training loss over iterations. Did your friend use stochastic gradient descent or batch gradient descent?

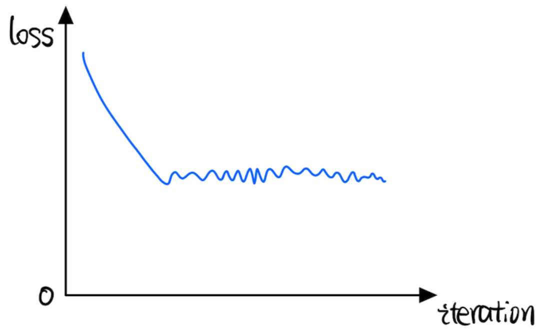


Figure 3: Training loss.

Circle the correct answer: Stochastic gradient descent OR Batch gradient descent

Explain briefly:

Part (b) [2 marks]

Your friend is having had a hard time reducing the loss further. Could you come up with two solutions to help them further reduce the loss?

Part (c) [1 mark]

What is one advantage of using neural networks over linear regression?

Part (d) [1 mark]

What is one advantage of linear regression over neural networks?

This page intentionally left blank

Question 5. Naive Bayes [8 MARKS]

Let's continue with the candy example from the previous question. Suppose that the manufacture wants to give customers a little hint by wrapping the candies in **y**ellow or blue wrappers. The wrapper for each candy depends on the flavour probabilistically, but the conditional distribution is unknown to us.

Let $P_d \in [0, 1]$ denote the probability that a candy is dark chocolate. Let $P_{dy} \in [0, 1]$ denote the probability that a candy has a yellow wrapper given that the candy is dark chocolate. Let $P_{my} \in [0, 1]$ denote the probability that a candy has a yellow wrapper given that the candy is milk chocolate.

Part (a) [2 MARKS]

What is the likelihood of observing a milk chocolate candy in a blue wrapper?

Part (b) [3 MARKS]

Your friend Avery unwrapped N candies. The flavour and wrapper counts are as follows.

- d of the N candies are dark chocolate.
  Among these d candies, d_y of the d wrappers are yellow and the rest are blue.
- $N - d$ of the N candies are milk chocolate.
  Among these $N - d$ candies, m_y of the $N - d$ wrappers are yellow and the rest are blue.

What is the likelihood of the data (the N candies)?

Part (c) [3 MARKS]

Derive the Maximum Likelihood estimates for P_d, P_{dy}, and P_{my}.

This page intentionally left blank

End of Exam