Last Name (PRINT): _____ First Name: _____

Student Number: _____ UTORid: _____

Signature: _____

UNIVERSITY OF TORONTO MISSISSAUGA
DECEMBER 2022 FINAL EXAMINATION

CSC311H5F
Introduction to Machine Learning

Sonya Allin, Lisa Zhang

Duration - 3 Hours

Examination Aids: One page double sided letter

The University of Toronto Mississauga and you, as a student, share a commitment to academic integrity. You are reminded that you may be charged with an academic offence for possessing any unauthorized aids during the writing of an exam. Clear, sealable, plastic bags have been provided for all electronic devices with storage, including but not limited to: cell phones, smart watches, SMART devices, tablets, laptops, and calculators. Please turn off all devices, seal them in the bag provided, and place the bag under your desk for the duration of the examination. You will not be able to touch the bag or its contents until the exam is over.

If, during an exam, any of these items are found on your person or in the area of your desk other than in the clear, sealable, plastic bag, you may be charged with an academic offence. A typical penalty for an academic offence may cause you to fail the course.

*Please note, once this exam has begun, you **CANNOT** re-write it.*

- Please fill out the information at the top of this cover page.
- **DO NOT** write anything in the QR code area.
- This examination has **8** questions. There are a total of **14 pages**, **DOUBLE-SIDED**.
- **DO NOT** open or turn over the exam paper until the exam has started.
- Answer questions clearly and completely; illegible answers will not be marked.
- Show all your work and provide justifications to your answers unless explicitly asked not to do so.
- Remember that you must earn a grade of at least **40%** on this exam to pass this course.

Question	Out of
Q1	7
Q2	16
Q3	10
Q4	5
Q5	6
Q6	6
Q7	12
Q8	8
Total	70

Use this page for rough work. If you want work on this page to be marked, please indicate this clearly *at the location of the original question.*

1. **[7 marks]** For each of the following quantities, circle whether **increasing** the quantity would **increase the bias** of our model , **increase the variance** of our model, or **neither**. No explanation is required.

   (a) The maximum depth of a decision tree classifier.

   Circle one: Increase the Bias or **Increase the Variance** or **Neither**

   (b) The number of PCA features to use as input to a decision tree classifier.

   Circle one: Increase the Bias or **Increase the Variance** or **Neither**

   (c) The number of individual decision trees in a random forest classifier.

   Circle one: Increase the Bias or **Increase the Variance** or **Neither**

   (d) The weight of the regularizer λ in the regularized cost function $\mathcal{E}_{\text{reg}}(\mathbf{w}) = \mathcal{E}(\mathbf{w}) + \frac{\lambda}{2}\sum_j w_j^2$.

   Circle one: Increase the Bias or **Increase the Variance** or **Neither**

   (e) The value of k in a k-Nearest Neighbour model.

   Circle one: Increase the Bias or **Increase the Variance** or **Neither**

   (f) The number of data points in our training set.

   Circle one: Increase the Bias or **Increase the Variance** or **Neither**

   (g) The number of input features to use in a Naive Bayes classifier.

   Circle one: Increase the Bias or **Increase the Variance** or **Neither**

2. **[16 marks]** For each of the following questions, circle either **True** or **False**, and briefly justify your choice.

(a) **[2 marks]** In a Naive Bayes model, we make the conditional independence assumption that

$$p(\mathbf{x}|c) = p(x_1, x_2, \ldots, x_D|c)$$

Circle one: True or **False**

Explanation:

(b) **[2 marks]** A decision tree divides the input space into overlapping regions, one for each leaf of the tree.
Circle one: True or **False**

Explanation:

(c) **[2 marks]** In the "E" step of the Expectation-Maximization algorithm to fit a Gaussian Mixture Model, we compute the expected value of the mean and covariance matrix of each Gaussian.
Circle one: True or **False**

Explanation:

(d) **[2 marks]** An ensemble of models will typically achieve a higher accuracy compared to a single model.
Circle one: True or **False**

Explanation:

(e) **[2 marks]** PCA features are uncorrelated.
Circle one: True or **False**

Explanation:

(f) **[2 marks]** The entropy of an unfair coin is lower than the entropy of a fair coin.
Circle one: True or **False**

Explanation:

(g) **[2 marks]** The more data we have, the more important the prior is for MAP inference.
Circle one: True or **False**

Explanation:

(h) **[2 marks]** PCA directions of a given data set correspond to the eigenvectors of the empirical covariance matrix.
Circle one: True or **False**

Explanation:

3. [**10 marks**] Answer each question below.

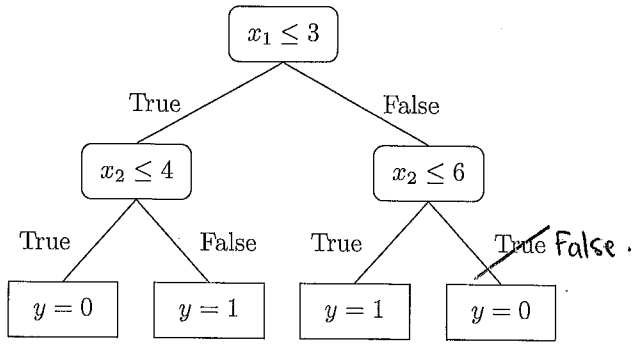
(a) [**4 marks**] Explain why, when training a decision tree classifier, we do *not* choose splits that ~~minimize~~ maximize the classification accuracy. Provide an example to illustrate your point.

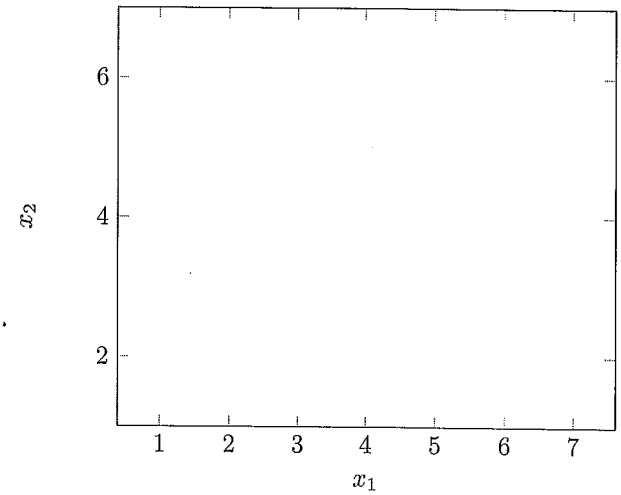
(b) [**4 marks**] A random forest classifier consists of many individual trees. However, each tree is trained differently. What are two differences in how each tree is trained? Why are these differences important?

(c) [**2 marks**] In Stochastic Gradient Descent, what happens if the batch size is too large? Explain briefly.

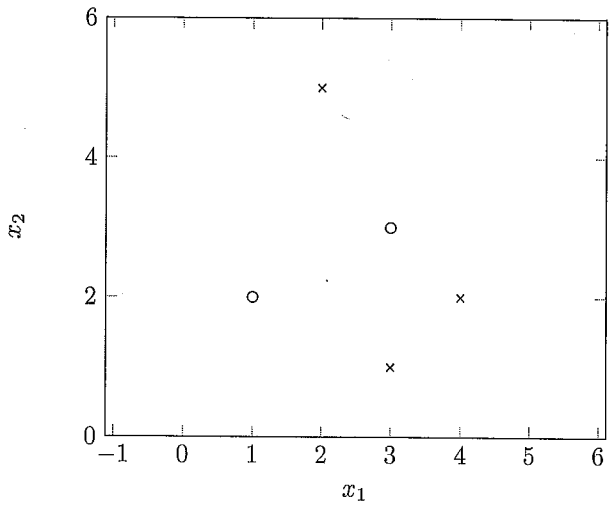
4. [**5 marks**] Draw the decision boundaries for the following models.

   (a) [**3 marks**] Draw the decision boundary for the following decision tree.

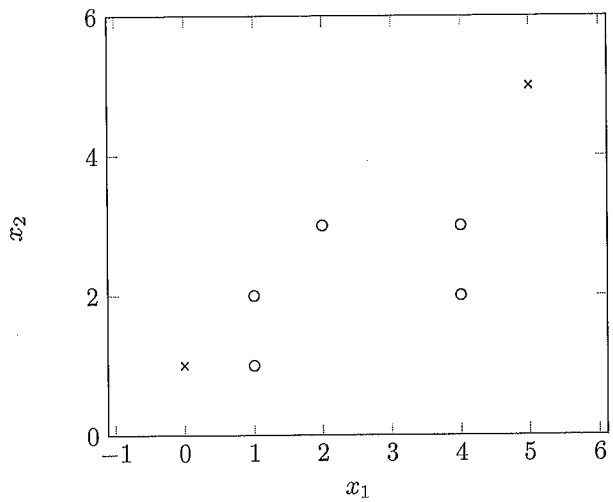




   (b) [**2 marks**] Draw the decision boundary for the 1-nearest neighbour model trained on the following data set.



5. **[6 marks]** In this question, we would like to use k-Means clustering to group the following 5 data points (labeled "o") into two clusters. We initialize the cluster centers to $(0, 1)$ and $(5, 5)$ (labeled "x").



Perform one iteration of k-means clustering to update the location of the cluster centers. Describe each step and show your work.

6. **[6 marks]** Consider neural network models A and B below, both of which can be trained to solve binary classification problems. The difference between the two models is that Model A models the target as a scalar $t \in \{0, 1\}$ while Model B models the target as a one-hot vector $\mathbf{t} = \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}$ with $t_i \in \{0, 1\}$ with either $t_1 = 1$ or $t_2 = 1$ but not both. Model A therefore uses the sigmoid activation function to compute $y = \sigma(z)$, while Model B uses the softmax activation function to compute $\mathbf{y} = \text{softmax}(\mathbf{z})$.

Model A	Model B
$\mathbf{h} = \text{ReLU}(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}$	$\mathbf{h} = \text{ReLU}(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}$
$z = \mathbf{W}^{(2)}\mathbf{h} + b^{(2)}$	$\mathbf{z} = \mathbf{W}^{(2)}\mathbf{h} + \mathbf{b}^{(2)}$
$y = \sigma(z)$	$\mathbf{y} = \text{softmax}(\mathbf{z})$

Both models uses the same input features $\mathbf{x} \in \mathbb{R}^D$, and both models have the same hidden activation size $\mathbf{h} \in \mathbb{R}^M$.

(a) **[2 marks]** Compute the number of parameters in Model A.

(b) **[2 marks]** Compute the number of parameters in Model B.

(c) **[2 marks]** Which model is more likely to overfit? Explain.

7. **[12 marks]** In this question, we will fit the parameters to a Naive Bayes model that classifies whether an email is spam ($c = 1$) or not-spam ($c = 0$).

   (a) **[2 marks]** Write down the data matrix $\mathbf{X}$ containing the bag-of-words features for the following emails, the first two of which are labeled spam ($c = 1$) and the last email is not spam ($c = 0$). Assume that the vocabluary consists of only the words appearing in these emails.

   - buy now
   - buy it now
   - it is now

   (b) **[2 marks]** What parameters do we need to fit in order to fit a Naive Bayes model to this data? Clearly define each parameter.

(c) **[4 marks]** Use MLE to fit the parameters from part (b). Do not derive the maximum likelihood estimate formula for the parameters, but make it clear what formula you are using.

(d) **[4 marks]** Using those parameters, produce a discrete estimate for whether a new email consisting of the words "buy it" is spam or not spam. Show all your work. You do not need to multiply together the individual probabilities.

8. **[8 marks]** Suppose we would like to model the distribution of data $\mathcal{D} = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$ using a univariate Gaussian distribution:

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

(a) **[2 marks]** Show that the log likelihood of the data can be written

$$\ell(\mu, \sigma^2) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x^{(i)} - \mu)^2$$

(b) **[2 marks]** Show that the maximum likelihood estimate of μ is

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x^{(i)}$$

(c) [**4 marks**] Show that the maximum likelihood estimate of σ^2 is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x^{(i)} - \mu)^2$$

Use this page for rough work. If you want work on this page to be marked, please indicate this clearly *at the location of the original question.*