UNIVERSITY OF TORONTO
Faculty of Arts and Science

APRIL 2020 EXAMINATIONS

CSC 311 H1S

Duration: Please submit to MarkUs by Monday April 20 at 16:59.

Aids Allowed: You may consult the course slides and your notes.

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

Last (Family) Name(s): ________________________________

First (Given) Name(s): ________________________________

Please read *carefully* every reminder on this page.

- Fill out your name and student number above—do it now, don't wait!
- This take-home test consists of 10 questions on 17 pages (including this one), printed on both sides of the paper.
- You may either (1) print these pages, answer each question directly on the examination paper, and then scan it and upload it as a PDF file, or (2) type your answers using your favourite word processor using the same order of questions as here. In the latter case, make sure you use the right question numbers and part (e.g., 4(a), 7(c), etc.). If you don't, you may not get the mark.
- As a student, you should help us having a fair take-home test. You may consult the slides and your notes. We in fact encourage you to do so. But do not discuss the questions with anyone else. And do not search for answer to these questions on the Internet.
- Do not share this take-home test with anyone else, even after the semester ends, as we may reuse some of these questions in the future.
- There will not be an auto-fail in this take-home test. But try hard to do a good job. This will be a good opportunity to practice your ML skills once more, and get feedback.

Marking Guide

Nº 1:	______/ 11
Nº 2:	______/ 18
Nº 3:	______/ 9
Nº 4:	______/ 8
Nº 5:	______/ 5
Nº 6:	______/ 8
Nº 7:	______/ 10
Nº 8:	______/ 8
Nº 9:	______/ 8
Nº 10:	______/ 15
TOTAL:	______/100

Question 1. True or False Questions [11 MARKS]

For each of these questions, a correct answer is +1 point, an empty answer is 0 point, and a wrong answer is −1 point.

	Statement	**True**	**False**
(1)	Decision trees can achieve zero classification error on any training data (assuming each training data point is unique)		
(2)	A linear regression model trained with the squared error loss is more robust to outliers than the same model trained with the absolute error loss		
(3)	A lower entropy implies lower uncertainty		
(4)	Scaling features between [0,1] is a mandatory pre-processing step before using K-NN algorithm		
(5)	PCA is a feature selection method		
(6)	Covariance matrix can have negative values		
(7)	The ℓ_1 regularization cannot shrink parameters to zero, hence it can be used for the purpose of feature selection		
(8)	Linear regression is invariant to scaling		
(9)	AdaBoost cannot overfit		
(10)	Projection to the highest variance direction is the same as projection to the direction that minimizes the total squared norm of each point to its projection		
(11)	ε-greedy is an action selection mechanism that allows some control over the exploration-exploitation tradeoff		

Question 2. Short Answer Questions [18 MARKS]

Answer these questions concisely. In most cases, one or two sentences are enough.

Part (a) Fill in the blanks, using the following terms Posterior, Likelihood, Prior, Evidence. [3 MARKS]

$$\ldots\ldots\ldots\ldots\ldots\ldots = \frac{(\ldots\ldots\ldots\ldots)\times(\ldots\ldots\ldots\ldots)}{(\ldots\ldots\ldots\ldots)}$$

Part (b) Explain the relationship between posterior probability and prior probability given the likelihood. [2 MARKS]

Part (c) Suppose that your classifier achieves poor accuracy on both the training and test sets. Which would be a better choice to try to improve the performance: Bagging or Boosting? Briefly justify your answer. [3 MARKS]

Part (d) What is the effect of Boosting and Bagging on the bias/variance tradeoff? [4 MARKS]

Part (e) What is the main modelling assumption in the Naïve Bayes classifier? [2 MARKS]

Part (f) What objective function is minimized by the k-Means clustering? Write down the formula [2 MARKS]

Part (g) Write down the Bellman optimality equation in the expanded form (so not $Q^* = T^*Q^*$, but the one that has an integral in it.) [2 MARKS]

Question 3. Hyper-Parameter Identification [9 MARKS]

Consider the following ML methods. Write down one or more hyper-parameters used in each method.

Part (a) KNN (1 answer) [1 MARK]

Part (b) Deep Neural Networks (2 answers) [2 MARKS]

Part (c) Support Vector Machines (1 answer) [1 MARK]

Part (d) PCA (1 answer) [1 MARK]

Part (e) K-Means Clustering (1 answer) [1 MARK]

Part (f) Gaussian Mixture Models (1 answer) [1 MARK]

Part (g) Q-Learning (2 answers) [2 MARKS]

Question 4. Diagnosis of Rare Diseases [8 MARKS]

Part (a) Someone you know is tested positive for a rare disease that affects only 0.1% of the population. Assuming that the test is 99% accurate. What is the probability that this person has the disease? You need to write down your calculations to get the mark. [4 MARKS]

Part (b) The same person asks another laboratory to re-do the test in order to increase the confidence of diagnosis. The new laboratory uses the same technology, so its test accuracy is the same 99% accurate. The result is positive again. What is the probability that this person actually has the disease. [4 MARKS]

Question 5. Multi-Class Classification for Regression Problems [5 MARKS]

Suppose that you have a regression problem with data $(\mathbf{x}, t)$ with the target values t being between $[0, 1]$. We usually solve this type of problems using one of a regression method by minimizing the squared error. But assume that you only need to predict t with the resolution of 0.1, e.g., it does not matter whether you predict 0.72 or 0.79; predicting 0.7 would be enough. Briefly describe how this problem can be formulated as a multi-class classification problem.

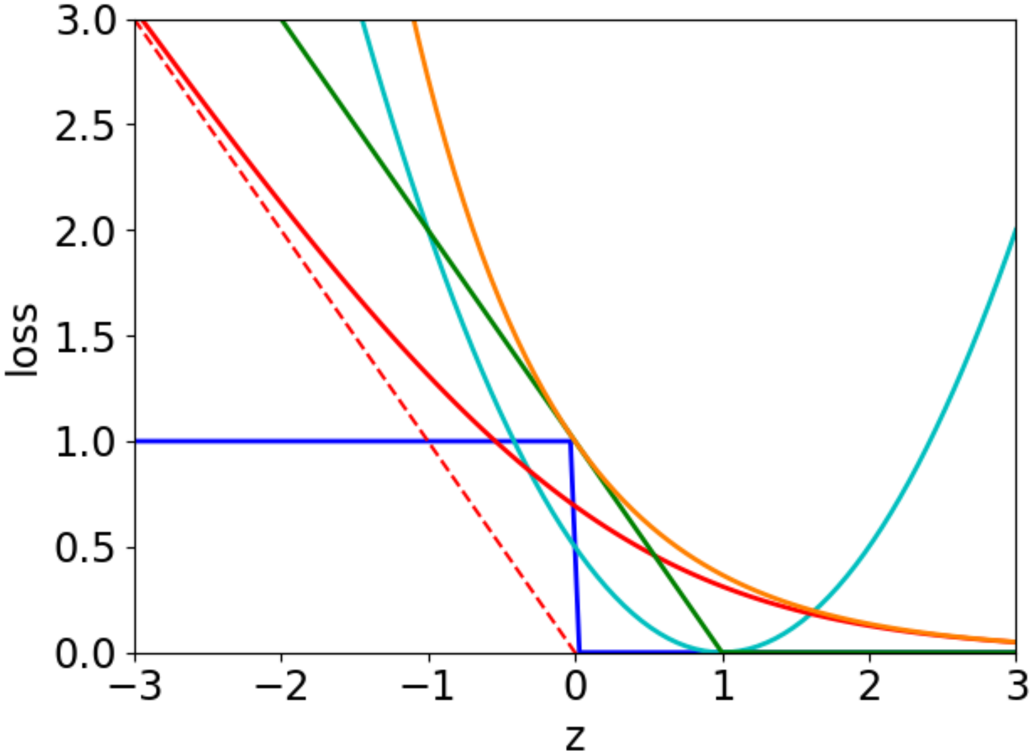
Question 6. Loss Functions for Classification [8 MARKS]

Consider a model whose output is z, e.g., the output of a linear classifier is $z = w^\top x + b$. We have been introduced to a variety of classification loss functions $\mathcal{L}(z, t)$, where t is the target (either $\{0, 1\}$ or $\{-1, +1\}$ for binary classification). This question asks you about these loss functions.

Part (a) Identifying Loss Functions [5 MARKS]

Identify the following loss functions on the figure (or write down their formulate, if you cannot write down on the figure). In this figure, you should assume that the target is $t = 1$. Make sure you identify them clearly without any ambiguity.

- $0 - 1$ Loss
- Exponential Loss
- Hinge Loss
- Cross-Entropy with a Logistic Model
- Squared Error Loss



Part (b) Why don't we minimize the $0-1$ loss function with a linear model? [1 MARK]

Part (c) What ML method uses the Hinge loss? [1 MARK]

Part (d) What ML method can be interpreted as using the exponential loss? [1 MARK]

Question 7. Optimization of Deep Linear Neural Networks [10 marks]

Consider the simplest deep linear neural network that is described by the following equations:

$$z = w_1 x,$$
$$y = w_2 z,$$

with $x, w_1, w_2, y \in \mathbb{R}$. This is indeed a very simple DNN that receives a one dimensional input, has one hidden layer with one unit, and one output. This simplicity is to ensure that the calculations are all easy. Consider the squared error loss function $l(y,t) = \frac{1}{2}(y-t)^2$.

Part (a) Show that one can replace this 2-layer NN with a 1-layer NN (show the relation of the input x to the output y). [2 marks]

Part (b) Compute the gradient of the loss of the 2-layer NN with respect to w_1 and w_2. [4 marks]

Part (c) Is the loss function of the 2-layer NN convex with respect to w_1 and w_2 or not? Prove your claim. [4 marks]

Question 8. Bayesian Classifier [8 MARKS]

Consider a binary classification problem with input x being a scalar. The data generation process works as follows:

- First, a target t is sampled from $\{0, 1\}$ with equal probability.
- If $t = 0$, x is sampled from a uniform distribution over the interval $[0, 1]$.
- If $t = 1$, x is sampled from a uniform distribution over the interval $[0, 2]$.

Part (a) Write down the formula for $P(x|t = 0)$, $P(x|t = 1)$, $P(t = 0)$, and $P(t = 1)$. [4 MARKS]

Part (b) Compute the posterior probability $P(t = 0|x)$ as a function of x. [4 MARKS]

Question 9. Gaussian Discriminant Analysis [8 MARKS]

Consider a GDA model with two classes with covariance matrices Σ_1 and Σ_2.

Part (a) Write down the formula describing the decision boundary (you should be able to find this from the slides). [4 MARKS]

Part (b) If the covariance matrices are shared between two classes (i.e., $\Sigma_1 = \Sigma_2 = \Sigma$), mathematically show that the decision boundary is linear. [4 MARKS]

Question 10. Estimation Problems in Casino [15 MARKS]

You are at a casino and you decide to play a slot machine. The machine works as follows: On each round of the game, you pull an arm. The machine tells you whether you have won or lost. If you lose, it costs you \$1. If you win, you get a known value \$r, e.g., $r = 5$. You play with the machine until you win, and then it restarts, and then you play another round. In other words, each round consists of playing until a win.

The number of times that you have to pull an arm before winning (and finishing a round) is a random variable K that can take values of $0, 1, 2, \ldots$. For each round, you record this value. For example, if you play three rounds and get LLLLW (round 1), LLW (round 2), and W (round 3), you have $K_1 = 4$, $K_2 = 2$, $K_3 = 0$.

Let us model this process. The probability of winning at each arm pull is θ and the probability of losing is $1 - \theta$, for some unknown $\theta \in [0, 1]$ that depends on the machine. You can assume that the probability of winning at each arm pull is independent from each other and does not change as you play the game. With this assumption, the probability of k losses before winning follows the following distribution:

$$P(K = k|\theta) = (1 - \theta)^k \theta, \qquad k = 0, 1, 2, \ldots .$$

As a perfectly rational person, deciding whether or not to play this game should depend on the expected money that you gain.[1] You can calculate the expected gain as follows (you do not need to understand this derivation completely to answer this question): If at one round you suffer k losses before winning, you have lost $k \times \$1$ and you gained $\$r$. This happens with probability $P(K = k|\theta)$. Therefore, your expected gain is

$$\begin{aligned}\text{expected gained money} &= \sum_{k \geq 0} (-k + r) P(K = k|\theta) = r \sum_{k \geq 0} P(K = k|\theta) - \sum_{k \geq 0} k P(K = k|\theta) \\ &= r - \frac{1 - \theta}{\theta}.\end{aligned}$$

So, if $r > \frac{1-\theta}{\theta}$, playing the game is worth it as the expected gain is positive; otherwise, it is not, and you better get out of the casino as soon as possible! Likewise, you can say that if $\theta > \frac{1}{1+r}$, the game is worth playing.

Since you do not know θ, it is crucial to estimate it in order to make an informed decision. This question asks you to develop some estimators for θ.

You play for N rounds and collect a dataset of $\mathcal{D}_N = \{K_1, K_2, \ldots, K_N\}$ describing the result of each rounds. For example, $\{4, 2, 0\}$ are the values for the example above. We assume that this data has already been collected, so you do not have to worry about your gain or loss so far. You can also assume that each round is independent from the previous ones (this is not an extra assumption, as it is implied by the independence of each arm pull).

[1] In reality, we are not rational agents, but that is another story.

Part (a) Likelihood [2 MARKS]

Write the likelihood function $P(\mathcal{D}_N|\theta)$ given a dataset $\mathcal{D}_N = \{K_1, K_2, \dots, K_N\}$. It should be in the following form:

$$\theta^{\cdots\cdots} \times (1-\theta)^{\cdots\cdots},$$

where $\cdots$ should be completed by you.

Part (b) Log-likelikelihood [1 MARK]

Write the log-likelihood function $\ell(\theta) = \log P(\mathcal{D}_N|\theta)$ given a dataset $\mathcal{D}_N = \{K_1, K_2, \dots, K_N\}$.

Part (c) MLE [2 MARKS]

Find the Maximum Likelihood Estimator. You need to show your derivations in order to get any mark.

Part (d) Encoding Prior Belief [2 MARKS]

You are skeptical that a casino would setup a game such that you win money. You believe that they set their slot machine such that its θ is small enough to make you lose money in average. You can formulate this belief as a prior distribution on θ. Your skepticism can be expressed by stating that the prior probability that $\theta < \frac{1}{1+r}$ should be high.

As you are already familiar with the Beta distribution, you decide to use it as your prior. Recall that

$$\text{Beta}(\theta; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\,\theta^{a-1}(1-\theta)^{b-1},$$

where Γ is the gamma function. How do you reasonably encode your prior belief with a Beta distribution? You need to specify conditions on a and b. Note that there is no single correct answer. Specify a relation between a and b and briefly justify your answer.

Part (e) MAP [2 MARKS]

Assume that you have selected a Beta distribution with a particular choice of a and b to encode your prior belief. Find the Maximum A-Posteriori (MAP) estimate. Your answer should be a function of $\mathcal{D} = \{K_1, K_2, \ldots, K_N\}$ and a and b. You should not use the particular a and b that you found in the previous question, but write it for any a and b.

Part (f) Bayesian Posterior [2 MARKS]

Calculate the Posterior probability of θ given the prior $\text{Beta}(\theta; a, b)$ and the data $\mathcal{D}_N = \{K_1, K_2, \ldots, K_N\}$. (Hint: Notice that the Beta distribution is a conjugate prior for this likelihood, so your posterior is in the form of a Beta distribution too.)

Part (g) Bayesian Estimation of $\mathbb{E}[\theta]$ [1 MARK]

What is the expected value of the parameter θ according to the posterior distribution?

Part (h) Comparison of MLE, MAP, and Bayesian Estimation [3 MARKS]

Briefly explain what the advantages and disadvantages of each of MLE, MAP, and Bayesian Estimation are (you should be able to answer this even if you have not calculated the estimators correctly.)

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.