PRACTICE MIDTERM EXAM

STA414/2104 Winter 2021

University of Toronto

Name:

Student #:

Enrolled in course (circle one): STA414 STA2104

Enrolled in section (circle one): Monday Tuesday

Exam duration: **80 minutes**

Please check that your exam has **6 pages**, including this one. Use the back of the page if you need more space on a question. The answer to each lettered question needs no more than two sentences. The total possible number of points is 100.

Read the following instructions carefully:

1. If a question asks you to do some calculations, you must *show your work* to receive full credit.
2. You will submit your answer to each question separately on crowdmark. If you run into any technical difficulties, you can send an email to sta414-2021-tas@cs.toronto.edu attaching your solutions. Any email sent after the exam is over will not be considered.
3. Do not share the exam with anyone or in any platform!
4. Lastly, enjoy the problems!!!

1. Academic Integrity Statement. Academic integrity is a fundamental value of learning and scholarship at the UofT. Participating honestly, respectfully, responsibly, and fairly in this academic community ensures that your UofT degree is valued and respected as a true signifier of your individual academic achievement.

The University of Toronto's Code of Behaviour on Academic Matters outlines the behaviours that constitute academic misconduct, the processes for addressing academic offences, and the penalties that may be imposed. You are expected to be familiar with the contents of this document.

Potential offences include, but are not limited to:

- Working together to answer questions.
- Looking at someone else's answers.
- Letting someone else look at your answers.
- Sharing or posting the exam questions.
- Discussing answers or the exam questions with anyone else in or outside the course.
- Misrepresenting your identity or having someone else complete your exam.

Prior to beginning this exam, you must attest that you will follow the Code of Behaviour on Academic Matters and will not commit academic misconduct in the completion of this online exam. Affirm your agreement to this by rewriting the following statement:

I, [name] (type your full name here), [stnum] (type your student number here), agree to fully abide to the Code of Behaviour on Academic Matters. I will not commit academic misconduct, and am aware of the penalties that may be imposed if I commit an academic offence.

All suspected cases of academic dishonesty will be investigated following the procedures outlined in the Code of Behaviour on Academic Matters.

2. Exponential families.

2.1. *Geometric distribution.* Probability mass function of a random variable X distributed as geometric distribution with parameter γ is given as

$$\mathbb{P}(X = k) = \gamma(1-\gamma)^{k-1} \quad \text{for} \quad k = 1, 2, ...$$

(a) Show that this is a probability mass function.

(b) Write the above distribution as an exponential family, and identify its sufficient statistics, natural parameter, and normalizing function (or cumulant generating function or partition function).

(c) Assume that we observed $X_1, X_2, ..., X_n$ i.i.d. random variables from geometric distribution with an unknown parameter γ. Find the MLE for γ.

3. Logistic regression. In class, we encoded the target values for logistic regression with $t_i \in \{0, +1\}$. In this problem, you will derive an equal formulation when targets are encoded with $\tilde{t}_i \in \{-1, +1\}$.

For a dataset $\mathcal{D}_N = \{(\mathbf{x}_i, t_i)\}$ with $t_i \in \{0, +1\}$, logistic regression is defined using the following steps:

$$\begin{aligned} z &= \mathbf{w}^\top \mathbf{x} + b \\ y &= \sigma(z) \\ \mathcal{L}(y, z) &= -t \log(y) - (1 - t) \log(1 - y). \end{aligned}$$

(a) Write the equivalent error minimization problem over the training data by eliminating the intermediate variables y and z. Your cost function should only depend on variables $\mathbf{w}$ and b, and dataset $\mathcal{D}$.

(b) Show that if we encode the targets as $\tilde{t}_i \in \{-1, +1\}$, the minimization problem (your answer to the previous part) can be equivalently written in the following form.

$$\text{minimize}_{\mathbf{w},b} \sum_{i=1}^{N} \log\Big(1 + \exp\{-\tilde{t}_i(\mathbf{w}^\top \mathbf{x}_i + b)\}\Big)$$

(c) Assume that we trained a logistic regression model and our class probabilities can be found by

$$z(\mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b)$$

where $(\mathbf{w}_k, b)$ are the parameters, and we classify using the rule

$$y(\mathbf{x}) = \mathbb{1}[z(\mathbf{x}) > 0.5].$$

Show that this corresponds to a linear decision boundary in the input space.

4. Optimization.

4.1. *Minimizing training error.* Assume that you are minimizing an error function which can be written as

$$E(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(\mathbf{w}, \mathbf{x}_i, t_i),$$

where $N = 1,000,000$.

(a) Write the one-step update rules for gradient descent (GD), stochastic GD (SGD), and mini-batch SGD (mSGD) with batch size 100. You can denote the gradient of the loss with respect to $\mathbf{w}$ for each sample with $\mathbf{g}_i = \nabla\mathcal{L}(\mathbf{w}, \mathbf{x}_i, t_i)$, and your learning rate with η.

(b) Rank the computational cost of each iteration for GD, SGD, and mini-batch SGD (with batch size 100) from smallest the largest.

5. Neural networks.

5.1. *NN-1.* Draw the computation graph of the neural network defined by the logistic regression (given in question 3).

Write the backpropagation updates to derive $\bar{b}$ and $\bar{\mathbf{w}}$.

5.2. *NN-2.* The “expressibility” of a neural network, it’s ability to model different functions, is given by the number of hidden units. If we wanted to, we could simply use millions (i.e., a lot) of hidden units in order to model any kind of function we wanted. Why is this a bad idea in general? How could we avoid this problem?

6. True or False questions. Circle either True or False. Each correct answer is worth 2 points. To discourage random guessing, 2 points will be deducted for a wrong answer.

1. (True or False) Assume that you are using cross validation to choose the penalty parameter λ in L^2 regularized linear regression. As the number of training samples increases, we expect that the value of λ chosen by cross validation becomes larger.
2. (True or False) In the K-fold cross-validation procedure for selecting a model parameter λ out of m values, you fit your model $K \times m$ times.
3. (True or False) Assume that you have a dataset composed of N observations: the target $\mathbf{t}$ and features $\mathbf{X}$. You want to fit a linear regression model and find the weights $\mathbf{w}$, but you also know that more data is always helpful. Instead of fitting a model with $\mathbf{t} \in \mathbb{R}^n$ and $\mathbf{X} \in \mathbb{R}^{n \times d}$, you concatenate the data and fit a model using $\begin{bmatrix} \mathbf{t} \\ \mathbf{t} \end{bmatrix} \in \mathbb{R}^{2n}$ and $\begin{bmatrix} \mathbf{X} \\ \mathbf{X} \end{bmatrix} \in \mathbb{R}^{2n \times d}$. Running linear regression on this new dataset will give the same weights as on the original dataset.
4. (True or False) The decision boundaries resulting from linear regression with 1-of-K encoded targets are always the same those resulting from logistic regression.
5. (True or False) We use stochastic gradient descent (SGD) with very small constant step size to minimize a loss function. Assuming that we can run SGD for a very long time, eventually it will converge to a minimum of the loss function.