UNIVERSITY OF TORONTO MISSISSAUGA
DECEMBER 2018 FINAL EXAMINATION
CSC411H5F
Machine Learning and Data Mining
Anthony Bonner
Duration - 2 hours
Aids: 1 page of double-sided Letter (8-1/2 x 11) sheet with no more than 12,000 characters

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 devices, tablets, laptops, calculators, and MP3 players. 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.

Write your answers on the examination sheet in the spaces provided. You may use the backs of pages if necessary. Concise, well-written answers will receive more points than long, rambling ones.

I don’t know policy: If you do not know the answer to a question (or part), and you write “I don't know”, you will receive 20% of the marks of that question (or part). If you just leave a question blank with no such statement, you get 0 marks for that question.
<table>
<thead>
<tr>
<th>Question</th>
<th>Value</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td><strong>146</strong></td>
<td></td>
</tr>
</tbody>
</table>
1. (20 points total) True or False.

For each of the following statements, state whether it is true or false, without giving a explanation (2 points each):

(a) Gradient descent is guaranteed to find the global minimum of a loss function.

(b) Minimizing cross entropy is equivalent to maximizing likelihood.

(c) The softmax function converts a vector of real numbers into a probability distribution for a discrete random variable.

(d) The K-means algorithm is an example of supervised learning.

(e) Ensemble learning can combine weak classifiers into a stronger classifier.

(f) Linear regression can only fit a linear function to a data set.

(g) K nearest neighbours produces a linear decision boundary.

(h) ADAboost can be interpreted as minimizing an exponential loss function.

(i) In multi-class logistic regression, the output is given by a sigmoid function.

(j) It is the hidden layers that enable a neural net to have a non-linear decision boundary.

continued on page 4
2. (23 points total) *Over and Underfitting.*

(a) (3 points) Briefly describe overfitting, when it occurs and why it is a problem.

(b) (3 points) Briefly describe underfitting, when it occurs and why it is a problem.
(c) (3 points) Draw a simple example showing well-fitted, overfitted and underfitted data and label the various parts.
To prevent overfitting, we often add a regularization term $\alpha \sum_k w_k^2$ to the loss function, where $\alpha > 0$ and the sum is over all weights, $w_k$, that we are trying to learn.

(d) (3 points) Why does this reduce overfitting?

(e) (2 points) Why do we require that $\alpha > 0$.

(f) (2 points) What happens to the weights as $\alpha$ is increased?

continued on page 7
(g) (7 points) Draw two curves: training error vs $\alpha$, and test error vs $\alpha$. Draw both curves on a single pair of axes in which the horizontal axis is $\alpha$. Indentify the curves clearly. On the horizontal axis, indicate areas where underfitting occurs, and areas where overfitting occurs. Also indicate the the best value of $\alpha$. 

continued on page 8
3. (25 points total) Precision, Recall and Python

Consider a binary classification problem in which we have trained a logistic-regression classifier. The weight vector for the classifier is \( \mathbf{w} \) and the bias term is \( w_0 \). Assume \( \mathbf{w} \) is a column vector. We wish to use Python to generate a precision/recall curve for the classifier based on a set of test data stored in a Numpy matrix, \( \mathbf{X} \), in which each row is an input vector. The target values are stored in a Numpy vector, \( \mathbf{T} \). We say that input \( \mathbf{X}[n,:] \) is a true positive if \( T[n] = 1 \). Given a threshold, \( \mathbf{th} \), we say that, \( \mathbf{X}[n,:] \) is a predicted positive if \( \mathbf{X}[n,:]\mathbf{w} + w_0 > \mathbf{th} \). Precision is the fraction of predicted positives that are true positives, and recall is the fraction of true positives that are predicted positives.

(a) (6 points) Write a Python function, \( \text{pp}(\mathbf{X}, \mathbf{T}, \mathbf{w}, w_0, \mathbf{th}) \) that returns the number of predicted positives and the number of true predicted positives (i.e., predicted positives that are also true positives). Do not use any loops, and do not use any functions from \text{sklearn}.

continued on page 9
(b) (10 points) Write a Python function \( \text{PR}(X,T,w,w_0) \) that plots a precision/recall curve for the classifier. The function should compute and plot precision and recall for 1000 different thresholds equally spaced between -5 and 5. For full marks, do not use your function from part (a) here (since that is not very efficient). You may use one loop (but no nested loops). Do not use any functions from \texttt{sklearn}. 

\( \textit{continued on page 10} \)
(c) (5 points) Write a few additional lines of Python code for your function in part (b) that enable it to compute and print the area under the precision/recall curve. You may use one loop (but no nested loops). Do not use any functions from numpy or sklearn.

(d) (4 points) Draw two precision/recall curves, where one is better than the other. Indicate which curve is better, and explain why it is better.
4. (26 points total) Clustering

(a) (5 points) Briefly describe the K-means algorithm for clustering.

(b) (3 points) Recall that the probability density in a Mixture of Gaussians model is given by the following formula:

\[ p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \]

What are the parameters of this model, and what do they say about the clusters?
(c) (3 points) Describe how to generate a data point in a Mixture of Gaussians model.

(d) (4 points) Suppose we use the EM algorithm to fit a MOG model to data. How do EM and k-means differ in the way they assign data points to clusters? Draw a diagram to illustrate the difference, and explain the diagram.
(e) (6 points) Describe three other ways in which Mixtures of Gaussians (MOG) is a more flexible model of clusters than k-means (in addition to part (d)). In each case, name the parameters that the MOG model uses to achieve this flexibility.
(f) (1 point) Suppose we fit a single Gaussian, $\mathcal{N}(\mu, \Sigma)$, to a data set. Write down an equation for the maximum-likelihood estimate of $\mu$.

(g) (4 points) Suppose we use the EM algorithm to fit a Mixture of Gaussians to a data set. Write down the equation that EM uses to update $\mu_k$ at each iteration. In 40 words or less, describe this equation intuitively. (Only the first 40 words of your description will be marked)
5. (24 points total) *Ensemble Methods*

(a) (4 points) What is a bootstrap sample of a data set, and how does it differ from the data set?

(b) (5 points) Describe bagging and why it is effective.
(c) (4 points) How does bagging differ from boosting?
(d) (5 points) Give a simple example in which bagging combines several simple classifiers into a better classifier by reducing variance. Include a diagram in your answer. Clearly explain what kind of simple classifiers you are using and why they have high variance.
(e) (6 points) Recall that for a mixture of experts, the output is given by

\[ y(x) = \sum_m g_m(x) y_m(x) \]

where each \( y_m \) is an expert. The \( g_m \) are gating functions and are given by a softmax function,

\[ g_m(x) = \frac{e^{z_m(x)}}{\sum_i e^{z_i(x)}} \]

where each \( z_m \) is a learnable function. The simple error function is given by

\[ E(x) = \sum_m g_m(x)[t - y_m(x)]^2 \]

where \( t \) is the target value on input \( x \). The gradient of the error wrt \( z_m \) is given by

\[ \frac{\partial E(x)}{\partial z_m} = g_m(x)[(t - y_m(x))^2 - E(x)] \]

Interpret this gradient as an error signal for adjusting the gate, \( g_m \), during learning by gradient descent. Justify your interpretation.
6. (28 points total) Neural Nets

Consider a neural net with one hidden layer using a sigmoid activation function, and a single output also using a sigmoid activation function. The operation of the neural net can be described as follows:

\[ o = \text{sigmoid}(z) \quad z = hW + w_0 \quad h = \text{sigmoid}(u) \quad u = xV + v_0 \quad (1) \]

Here \( x \) is the input (a row vector), and \( o \) is the output (a real number). \( V \) is a matrix of weights, \( v_0 \) is a row vector of bias terms, \( W \) is a column vector of weights and \( w_0 \) is a bias term (a real number). Recall that \( \text{sigmoid}(z) = 1/(1 + e^{-z}) \).

The loss function is given by the cross entropy:

\[ C = \sum_n c(t^{(n)}, o^{(n)}) \quad \text{where} \quad c(t, o) = -t \log o - (1 - t) \log(1 - o) \]

where the sum is over training points, \( t^{(n)} \) is the target value for input \( x^{(n)} \), and \( o^{(n)} \) is the output.

(a) (4 points) Prove that \( \partial o/\partial z = o(1 - o) \).

continued on page 20
(b) (12 points) Prove that \( \frac{\partial c(t, o)}{\partial W_j} = (o - t) h_j \)
(c) (8 points) Let $T$ and $O$ be column vectors where $T_n = t^{(n)}$ and $O_n = o^{(n)}$. Let $H$ be a matrix where $H_{nj} = h_j^{(n)}$. Prove that $\partial C/\partial W = H^T(O - T)$. 

continued on page 22
(d) (4 points) Write down a formula for $\partial C/\partial w_0$ and give a justification.
This page is for answers and rough work.
This page is for answers and rough work.