hrs	/ /		EXAMS OFFICE USE ONLY

University of the Witwatersrand, Johannesburg

Course or topic No(s)	COMS3007
Course or topic name(s) Paper Number & title	Machine Learning III
Examination to be held during the months(s) of	May 2018
Year of Study	3rd
Degrees/Diplomas for which this course is prescribed	BSc
Faculty/ies presenting candidates	Science
Internal examiner(s) and telephone extension number(s)	Benjamin Rosman
External examiner(s)	Prof. S. Gruner
Special materials required	Calculator

Time allowance				
	Course nos	COMS3007	Hours	3

Instructions to candidates	See first page of question paper

* As Internal Examiner/Head of Department, I certify that this question paper is in final form, as approved by the External Examiner, and is ready for reproduction.

** As Internal Examiner, I certify that this question paper is in final form and is ready for reproduction.

* applicable to formal examinations as approved by an external examiner.

** applicable to formal tests not requiring approval by external examiner.

Delete whichever not applicable.

Name: .....

Signature: .....

(NOT FOR REPRODUCTION)

Instructions

- This is a closed book exam.
- This exam lasts for 3 hours.
- There are 4 pages excluding the cover page.
- There are 70 marks available in total.
- 70 marks = 100%
- The number of marks available for a question is given in square brackets at the end of the question.
- You may answer the questions in any order, and you may answer all questions.
- You may use a calculator.
- **Answer the question!** Writing an answer that is true but irrelevant to the question asked will waste your time and receive zero marks.
- If you are not sure about something, **put up your hand and wait for an invigilator to come to you**.

General [10 marks]

1. (a) Explain the difference between classification and regression problems. Give an example of each. [4 marks]

   (b) What is the difference between a model and a cost function? [2 marks]

   (c) Why do we need our models to generalise? [1 marks]

   (d) Explain the difference between training data, validation data, and testing data. [3 marks]

Naïve Bayes [15 Marks]

2. You are building a recommendation system for an online movie streaming service. For each user, you would like to be able to predict their rating (*good* or *bad*) on new movies, based on ones they have previously seen, using the features of $genre = \{action, drama, comedy\}$ and $budget = \{low, high\}$. You do this using a Naïve Bayes classifier. The following table shows training data for one user.

genre	budget	rating
action	high	good
action	low	bad
drama	low	bad
comedy	low	good
action	high	good
drama	high	bad
comedy	low	bad
drama	high	bad

   (a) What is the difference between a prior and a posterior distribution? [2 marks]

   (b) What is the Naïve Bayes assumption? [1 marks]

   (c) Compute $P(good)$ and $P(bad)$. [1 marks]

   (d) A new movie is released: a low-budget action. Use Naïve Bayes to compute the probability that this user gives the movie a good rating. Bayes' rule is given by: $P(y|x) = \frac{P(x|y)P(y)}{P(x)}$. [8 marks]

   (e) Another new movie is released: a high-budget drama. What is the challenge with using Naïve Bayes to classify this data point? Describe an approach to solve this. [3 marks]

Logistic Regression [15 marks]

3. (a) What is the effect of using non-linear features in logistic regression? [1 marks]

   (b) It is important to correctly set the learning rate α during gradient descent. What can happen if this is set too low, and what can happen if it is set too high? [2 marks]

   (c) Why might we use regularisation in logistic regression? [1 marks]

   (d) Give pseudocode for optimising the model parameters in the logistic regression problem using gradient descent. [3 marks]

   (e) Consider the dataset in the table below. We want to perform logistic regression, by fitting the function $y = h_\theta(x) = \sigma(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$, with error function $E(\theta) = -\log[\prod_{i=1}^{m}(h_\theta(x^i)^{y^i})(1-h_\theta(x^i))^{1-y^i}]$.

x_1	2	6	4	8
x_2	3	10	6	13
class y	0	1	0	1

   Perform one iteration of gradient descent with the *first data point*. Use initial values of $(\theta_0, \theta_1, \theta_2) = (-1, -1, 0.5)$, and $\alpha = 0.1$. You may use that updates are given by $\theta_j \leftarrow \theta_j - \alpha(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}$. [5 marks]

   (f) Plot all the data points from (e) labelled by class, as well as the decision boundary before that iteration of gradient descent. Also plot what you would expect the optimal decision boundary to look like. Label the two decision boundaries clearly. [3 marks]

Neural Networks [15 Marks]

4. Consider a neural network with one input node, one hidden layer with two nodes, and one output node. The weights between the layers (including biases) are given by:

$$\Theta^{(1)} = \begin{bmatrix} 2 & -1 \\ 1 & -2 \end{bmatrix}, \quad \Theta^{(2)} = \begin{bmatrix} 1 \\ 3 \\ -1 \end{bmatrix}.$$

   All activation functions are logistic functions $\sigma(z) = 1/(1+\exp(-z))$. We wish to train the network with a point $(x, y) = (2, 3)$.

   (a) Write down the output of the network y as a single function of the input variable x. [1 marks]

   (b) Calculate the output of the network for $x = 4$. [2 marks]

   (c) Update the weights of the network *from the input layer to hidden layer only* using the backpropagation algorithm with learning rate $\alpha = 0.2$. Recall that the derivative of cost function $J(\theta)$ wrt the weight $\theta_0 ij^{(l-1)}$ is $(\sum_m \delta_m^{(l+1)} \theta_0 mi^{(l)}) g'(z_i^{(l)}) a_j^{(l-1)}$. [6 marks]

(d) What is the effect of adding hidden layers to a network? [1 marks]

(e) Draw a neural network that takes two binary inputs x_1 and x_2, and outputs x_1 AND x_2. Indicate the weight values in your drawing, as well as any activation functions. [3 marks]

(f) Why is it a problem to initialise all weights to zero? What should be done instead? [2 marks]

Clustering [11 marks]

5. Suppose you want to apply the k-means algorithm with $k = 2$ to the dataset below consisting of points in $\mathbb{R}^2$. The initial cluster centres are $\boldsymbol{\mu}_1 = (0.1, 0.3)$ and $\boldsymbol{\mu}_2 = (0.2, 0.4)$.

x_1	x_2
0.1	0.2
0.4	0.3
0.4	0.5
0.2	0.5

(a) Calculate the squared error for each data point given the current cluster centres. [2 marks]

(b) Compute one iteration of the k-means algorithm and give the new cluster centres. [4 marks]

(c) Explain with the aid of a diagram how poor initialisation of the k-means algorithm may lead to poor clusters. [3 marks]

(d) Explain how clustering may be used to compress the colour-space of an image. [2 marks]

Practical ML [4 marks]

6. You have been asked to develop an automated system for a car rental company to determine whether or not some of their cars need to be serviced this month. They have provided you with a dataset which has monthly records for their cars. This includes information on: the model of car, distance travelled in the last month, number of months since last service, and whether or not it has been in a major accident. The records also indicate months in which the car was serviced. Describe how you could use a neural network to learn to predict if a car needs to be serviced this month. Describe and motivate how you would set up the inputs and outputs of the network, the architecture and activation functions, and any pre-processing you would do to the data. [4 marks]

End of question paper