Big data Modeling and Management System

Final Exam

Exam Date	February 13, 2023 @ 9:00AM
Time Allowed	2;00 Hrs.
Total Mark	65 pt.

FULL NAME: _____

ID: _____

SECTION: _____

Instructions

- Make sure to write your **FULL NAME**, **ID**, and **SECTION** information on each page
- Make sure this exam booklet contains 10 **multiple questions, and 5 work out questions**
- Any form of cheating will result in disqualification of the results obtained in this exam
- Make sure to put your answer **on the answer sheet provided**

PART I: Multiple Choice Questions (2.5 pt. Each)

Choose the best answer among the choices provided and place your choice letter on the space provided

1. What is the first step in the data-driven decision-making process?

A. Collecting data B. Analyzing data C. Interpreting results D. Making a decision

2. What are some of the benefits of big data applications?

A. Improved decision-making B. Better customer understanding C. Enhanced operational efficiency

D. All of the above

3. Transaction of data of the bank is a type of.

A. unstructured B. Structured data C. both a and b D. None of the above

4. What is MapReduce?

A. A programming model used for processing and analyzing large amounts of data

B. A database management system used for storing and managing big data

C. A cloud-based platform used for data storage and management

D. A machine learning algorithm used for predictive analytics

5. What is the value of the correlation coefficient when there is no relationship between the variables?

A. 0 B.1 C. -1 D. 0.6

6. What is the most commonly used evaluation metric for linear regression models?

A. Mean squared error B. Mean absolute error C. Root mean squared error D. Coefficient of determination

7. What are some practical problems with the sigmoidal activation function in neural nets?
   A. It is convex, and convex functions cannot solve nonconvex problems
   B. It does not work well with the entropy loss function
   C. It can have negative values
   D. Gradients are small for values away from 0,leading to the ”Vanishing Gradient” problem for large or recurrent neural nets
8. What is recall in classification evaluation techniques?
   A. The percentage of true positive predictions out of all positive predictions
   B. The percentage of true positive predictions out of all actual positive cases
   C. The overall accuracy of a classification model
   D. The ability of a classification model to avoid false negatives
9. What is the ROC curve used for in classification evaluation?
   A. To evaluate the accuracy of a classification model
   B. b. To compare the performance of different classification models

C. To determine the most important features for a classification model
D. To visualize the trade-off between the false positive rate and true positive rate

10. What type of problem can an ANN be used to solve?

A. Regression problems B. Classification problems C. Both A and B D None of the above

PART II: Work Out

Work efficiently. Some questions are easier, some more difficult. Be sure to give yourself time to answer all of the easy ones, and avoid getting bogged down in the more difficult ones before you have answered the easier ones.

1. Consider the problem of predicting how well a student does in her second year of college/university, given how well she did in her first year. Specifically, let x be equal to the number of "A" grades (including A-. A and A+ grades) that a student receives in their first year of college (freshmen year). We would like to predict the value of y, which we define as the number of "A" grades they get in their second year (sophomore year). (8 points)

x	y
3	2
1	2
0	1
4	3

For the training set given above answer the following questions

1. Name a suitable regression type that he could most likely select for tackling this problem. Why would he choose this regression?
2. what is the value of m?
3. What is $J(0,1)$
4. Suppose we set $\theta_0 = 0, \theta_1 = 1.5$. What is $h_\theta(2)$?

2. Describe the steps involved in data mining when viewed as a process of knowledge discovery. (8 points)
3. Mr.Abebe is a data scientist at ITSC. He has built an email filter application to classify whether emails sent to his inbox per day will be classified as spam or not spam. After training his model, he applies it to check a sample of 100 emails and classifies that 25 emails will go on to spam box.However, he found out that out of the 25 emails the model classified, only 10 of them are spam and 15 of them are not spam. In Addition, he realizes that in total, 12 of the emails in the sample of 100 actually did go on spam box, and the other 88 not spam. (8 points)

a) Complete the confusion matrix for Mr.Abebe's model.

b) Calculate the precision for the spam filter. What is the interpretation of having this value for precision i.e How would Mr.Abebe explain this to someone doesn't know how precision is calculated but still uses e-mail and gets spam e-mail.

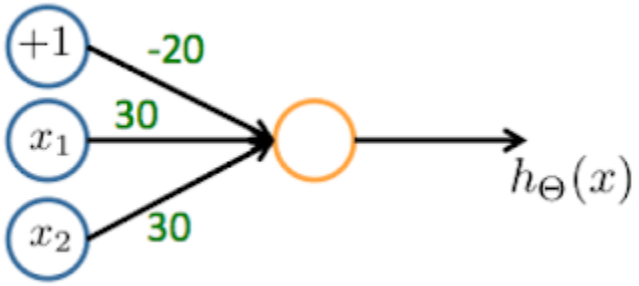
c) Calculate the recall for the spam filter. What is the interpretation of having this value for recall i.e How would Mr.Abebe explain this to someone doesn't know how recall is calculated but still uses e-mail and gets spam e-mail.

d) Mr.Abebe observe that the precision is very good but the recall is not so good.What does it mean to have high precision and low recall. What might the possible reason Mr.Abebe is seeing these results?

e) What does it mean to have high recall and low precision for a spam filter? Which of the two do you think is better i.e high precision and low recall or high recall and low precision

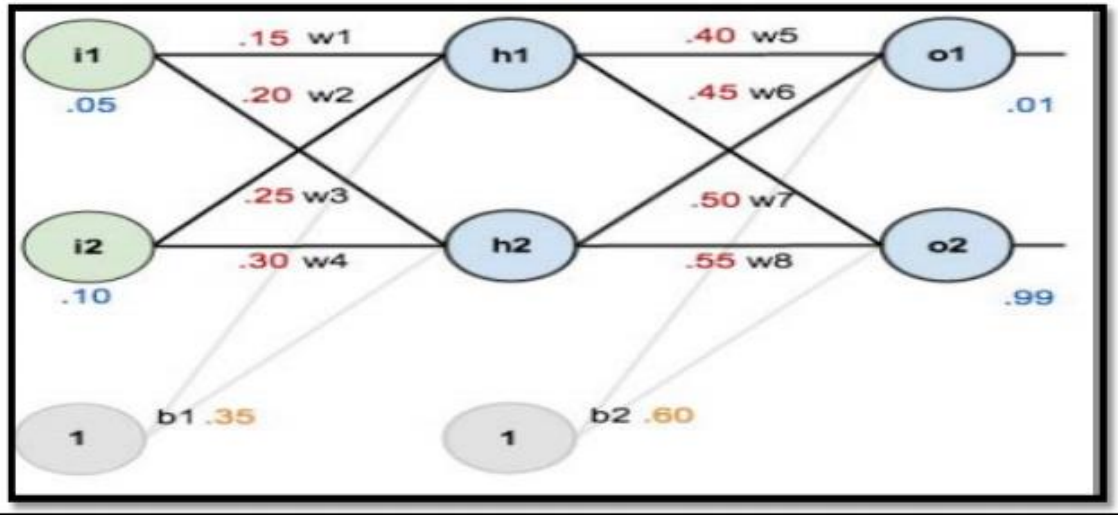
f) What is the overall accuracy of the spam filter? What does Mr.Abebe's mean when he says this spam filter has his value of accuracy?

4. Consider the following neural network which takes two binary-valued inputs $x_1, x_2 \in \{0, 1\}$ and outputs $h_\theta(x)$. Which of the following logical functions does it approximately? (8 points)



5. Consider the feedforward neural network as shown below.The initial weights, biases, and training inputs/outputs are given in the diagram.The activation function for the hidden and output neurons is the logistic sigmoid function. (8 points)

$$\frac{1}{1+e^{-x}}$$



a) What is the total number of parameters in this neural network?

b) Perform a forward pass on the network.

c) Compute the analytic form of $\frac{\partial E_{Total}}{\partial w_4}$

d) Perform a backward pass on the network.

e) Perform a further forward pass and comment on the result.

f) Test the network with (0.02,0.20).