Big Data Modeling and Management System

Final Exam

Exam Date	February 07, 2022 @ 1:30 PM
Time Allowed	2;00 Hrs.
Total Mark	50 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 **12 questions**
- Any form of cheating will result in disqualification of the results obtained in this exam
- Make sure to put your answer **on the space provided on the last page**

FULL NAME: _____ ID: _____

COLLEGE/SCHOOL/INST: _____ SECTION:___

1. Describe the steps involved in data mining when viewed as a process of knowledge discovery. (4 points)
2. Briefly discuss the major difference between database vs data warehouse, data mining vs knowledge discovery process, qualitative variable vs quantitative variables, and classification vs regression. For each of these pairs of terms, how are they similar and list one real application for each of them respectively? (4 points)
3. What is Data-Driven Decision Making and why it is so important? (4 points)
4. Define "Big Data" and what are the five V's of Big Data? (4 points)
5. Mention at least three application areas of big data and list the major benefits of applying big data application for each of them. (4 points)
6. Mr.Abebe is data scientist at ITSC, and is building a forecasting model to predict salary based on years of experience. His data is shown in the table below. (4 points)

Years of Experience	1	2	3	4	5	6	7	8	9	10
Salary	45,000	50,000	60,000	80,000	110,000	150,000	200,000	300,000	500,000	1,000,000

   a) Name a suitable regression type that he could most likely select for tackling this problem. Why would he choose this regression?
   b) Calculate the correlation coefficient r
   c) Apply the gradient descent optimization and find the equation of the regression line after two iterations.
   d) Determine RMSE
   e) Determine the coefficient of determination (R^2)
   f) Use the equation to predict the salary when years of experience is 11.

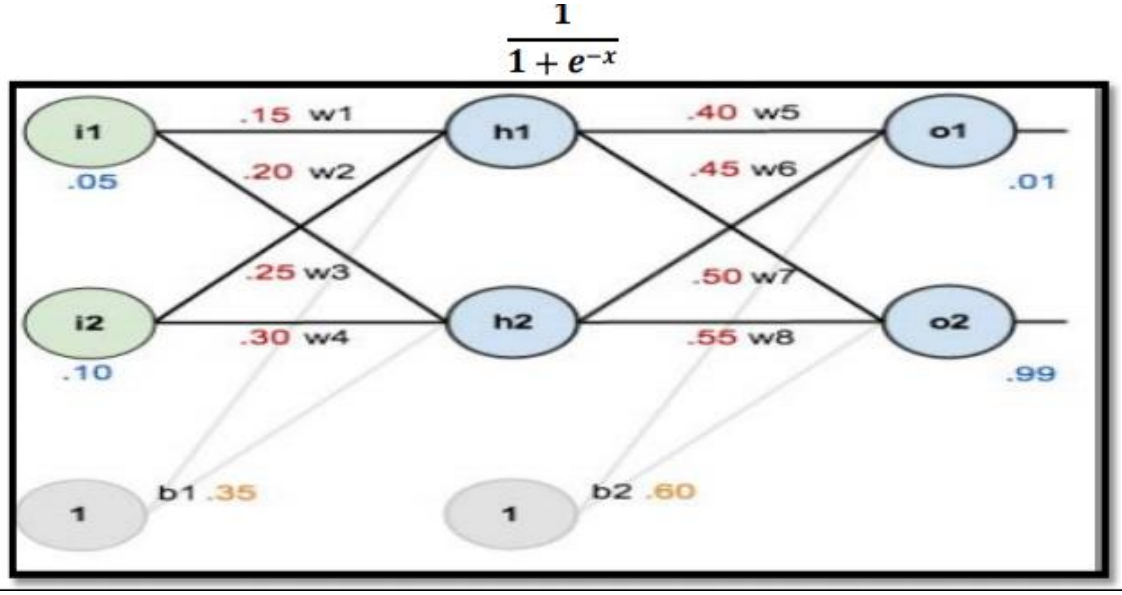
7. Suppose you are working on stock market prediction. You would like to predict whether or not a certain company will declare bankruptcy within the next 7 days (by training on data of similar companies that had previously been at risk of bankruptcy). Would you treat this as a classification or a regression problem? (4 points)
8. 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. (5 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?

9. 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 functions does it (approximately) compute? (4 points)

10. 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. (5 points)



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.

FULL NAME: _____ ID: _____

COLLEGE/SCHOOL/INST: _____ SECTION:___

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

11. Define clustering and list its applications. (4 points)

12. Use agglomerative clustering to cluster the following set of data: P1(4,4),P2(8, 4), P3(15, 8), P4(24, 4), P5(24, 12). Use Euclidean distance measure for the distance calculation. (4 points)