THE UNIVERSITY OF NEW SOUTH WALES
School of Computer Science and Engineering

Sample Final Examination– Term1, 2019
May, 2019

COMP 9321

Data Service Engineering

Reading Time: 10 Minutes

Available Mark: 50

Number of Questions : 20 + 5 + 2

Number of Parts: 3

Writing Time: 2 Hours

Permitted Materials: None

Hurdle : You have to pass the exam(25/50) to pass this course

*Answer **all** questions in this answer template(s) or booklet provided (Part1.txt,Part2.txt and an answer booklet for part3) . Answers are expected to be succinct but complete. Answers that irrelevant will be penalized. You can't access the terminal during the exam*

Part One (20 Marks)

Please put your answers in file Part1.txt,each question worth 1 Mark.

1. Which one of the following statements is true of the HTTP protocol?

   A. HTTP is stateless which means the client does not store any information about the server it is interacting with

   B. HTTP is not considered stateless because Web applications can maintain the state and store necessary data using Cookies and Sessions

   C. HTTP can be either stateless or stateful, depending on how the Web server is configured

   D. HTTP is stateless which means the server is not required to retain information about each communication partner for the duration of multiple requests

2. Which one is NOT a correct HTTP request method?

   A. HEAD B. OPTIONS C. CONNECT D. UPDATE

3. Which one of the following REST methods is neither safe nor idempotent ?

   A. HTTP HEAD B. HTTP PUT C. HTTP GET D. HTTP PATCH

4. In REST, a DELETE operation:

   A. is used to undo the last PUT operation on a resource

   B. is not supported due to security reasons

   C. must be implemented as an update operation (i.e., mark-as-delete)

   D. can be retried many times

5. Having Uniform Interfaces in RESTful Services could mean:

   A. Developers do not have to implement the operations as they are standards

   B. Developers can build more secure applications

   C. If the conventions are properly followed, understanding the interface is easy

   D. Providing standard data types for all HTTP operations

CP 1.2

6. What is Active Record?

   A. It is a style of ORM in which the domain object itself has methods like save() and delete() to manage the DB.

   B. It is a style of coding in which the objects are stored natively in an object-oriented database system

   should be domain object

   C. It is a style of ORM in which a session object is responsible for data manipulation like saving or removing objects from the database

   D. It is a style of coding in which an object-relation mapping configuration is kept separate from objects themselves

7. Which statement is NOT correct about Graphs and Charts?

   A. Both rely on a repeated pattern to show data

B. Charts are always restricted to numerical axes.

C. Graphs must have at least one numerical axe

D. They cannot be used interchangeably.

8. Assume we run PCA on a 2D-Numerical sample, what's the 3rd step of PCA?

   A. Subtract the mean

   B. Calculate the covariance matrix

   C. Ignore the components of lesser significance

   D. Calculate the eigenvectors and eigenvalues of the covariance matrix

9. Which type of graph is more suitable when you want to illustrate the correlation between two variables?

   A. Line Graph B. Scatter Plot C. Tree Diagram D. Bar Graph

10. Given d-dimensional data X, you run principle component analysis and pick P principle components. Can you always reconstruct any data point x_i for $i \in \{1...N\}$ from the P principle components with zero reconstruction error?

    A. Yes, if P < d

    B. Yes, if P = d

    C. Yes, if P > d

    D. No, you cannot

11. For which of following tasks might K-means clustering be a suitable algorithm?

    A. Given historical weather records, predict if tomorrow's weather will be sunny or rainy

    B. Given a set of news articles from many different websites, find out what topics are the main topics covered

    C. Given many emails, you want to determine if they are Spam or Non-Spam emails.

    D. Given sales data from a large number of products in a supermarket, estimate future sales for each of these products.

12. Considering the following training set of $m = 4$ training examples:

x	y
1	0.5
2	1
4	2
5	2

Table 1: Question 12

Consider the linear regression model $h_\theta(x) = \theta_0 + \theta_1 x$. What are the values of θ_0 and θ_1 that you would expect to obtain upon running gradient descent on this model? (Linear regression will NOT be able to fit this data perfectly.)

A. $\theta_0 = 1, \theta_1 = 0.5$

B. $\theta_0 = 0.5, \theta_1 = 0$

C. $\theta_0 = 0, \theta_1 = 0.5$

D. $\theta_0 = 0.5, \theta_1 = 1$

13. Suppose someone tells you that they ran PCA in such a way that "95% of the variance was retained." What is an equivalent statement to this?

A. $\dfrac{\frac{1}{m}\sum_{i=1}^{m} ||x_i - x_i^{approx}||^2}{\frac{1}{m}\sum_{i=1}^{m} ||x_i||^2} \geq 0.95$

B. $\dfrac{\frac{1}{m}\sum_{i=1}^{m} ||x_i - x_i^{approx}||^2}{\frac{1}{m}\sum_{i=1}^{m} ||x_i||^2} \geq 0.05$

C. $\dfrac{\frac{1}{m}\sum_{i=1}^{m} ||x_i - x_i^{approx}||^2}{\frac{1}{m}\sum_{i=1}^{m} ||x_i||^2} \leq 0.05$

D. $\dfrac{\frac{1}{m}\sum_{i=1}^{m} ||x_i - x_i^{approx}||^2}{\frac{1}{m}\sum_{i=1}^{m} ||x_i||^2} \leq 0.95$

14. Which of the following is NOT part of a typical data analytics architecture?

A. Data testing and generation

B. Data integration/aggregation

C. Data visualisation

D. Data management d

15. Suppose we have three cluster centroids $\mu_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, $\mu_2 = \begin{bmatrix} -3 \\ 0 \end{bmatrix}$ and $\mu_3 = \begin{bmatrix} 4 \\ 2 \end{bmatrix}$.Furthermore, we have a training example $x_i = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$. After a cluster assignment step, what will the c_i be?

A. $c_i = 3$ B. $c_i = 2$ C. $c_i = 1$ D. c_i is not assigned.

16. Which visualisation paradigm allows you to show the centre of a region or places on a map?
A. Flow map B. Point clustering C. Heat map D. Place markers

17. Suppose you have an unlabeled dataset $\{x_1 \ ... \ x_m\}$. You run K-means random initialization, and obtain 50 different clusterings of the data. What is the recommended way of choosing which one of these 50 clusterings to use (Please choose the most suitable one).

A. Always pick the final (50th) clustering found, since by that time it is more likely to have converged to a good solution,

B. The only way to do so is if we also have labels y_i for our data.

C. For each of the clusterings, compute $\frac{1}{m}\sum_{i=1}^{m} ||x_i - \mu_{c_i}||^2$ (MSE), and pick the one that minimizes this.

D. The answer is ambiguous,and there is no good way for choosing.

18. You run a movie empire, and want to build a movie recommendation system based on collaborative filtering. There were three popular review websites (which we'll call A, B and C) which users to go to rate movies, and you have just acquired all three companies that run these websites. You'd like to merge the three companies' datasets together to build a single/unified

system. On website A, users rank a movie as having 1 through 5 stars. On website B, users rank on a scale of 1 - 10, and decimal values (e.g., 7.5) are allowed. On website C, the ratings are from 1 to 100. You also have enough information to identify users/movies on one website with users/movies on a different website. Which of the following statements is true?

A. Assuming that there is at least one movie/user in one database that doesn't also appear in a second database, there is no sound way to merge the datasets, because of the missing data

B. You can combine all three training sets into one as long as your perform mean normalization and feature scaling **after** you merge the data.

C. It is not possible to combine these websites' data. You must build three separate recommendation systems.

D. You can merge the three datasets into one, but you should first normalize each dataset's ratings (say re-scale each dataset's ratings to a 0-1 range).

19. Which of the following statements is true for neural networks?

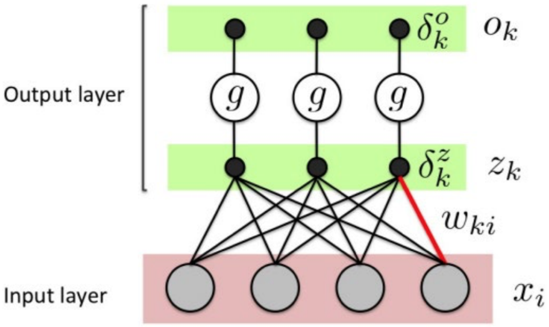
A. If we are training a neural network using gradient descent, one reasonable "debugging" step to make sure it is working is to plot $J(\theta)$(loss function) as a function of the number of iterations, and make sure it is decreasing (or at least non-increasing) after each iteration.

B. If we initialize all the parameters of a neural network to ones instead of zeros, this will suffice for the purpose of "symmetry breaking" because the parameters are no longer symmetrically equal to zero.

C. Suppose you are training a neural network using gradient descent. Depending on your random initialization, your algorithm may converge to different local optima (i.e., if you run the algorithm twice with different random initializations, gradient descent may converge to two different solutions).

D. Suppose you have a three layer network with parameters θ_1(controlling the function mapping from the inputs to the hidden units) and θ_2(controlling the mapping from the hidden units to the outputs). If we set all the elements of $\theta_1$to be 0, and all the elements of $\theta_2$to be 1, then this suffices for symmetry breaking, since the neurons are no longer all computing the same function of the input.

20. Suppose you are applying the gradient descent for a **Multi-Layer** neural network (shows below):



The error function what we used in the mean-square error(MSE). On a single training sample n, we have: $\frac{\partial E}{\partial o_k^{(n)}} = o_k^{(n)} - t_k^{(n)} := \delta_k^o$. Assume the activation function is $\frac{1}{1+e^{-x}}$, $x = z_k^{(n)}$. By given that: $\frac{\partial o_k^{(n)}}{\partial z_k^{(n)}} = o_k^{(n)}(1 - o_k^{(n)})$. What's the gradient descent update rule for this multi-layer neural network? (Hint: For a **single** layer neural network the error gradient $\frac{\partial E}{\partial w_{ki}} = \frac{\partial E}{\partial o_k}\frac{\partial o_k}{\partial z_k}\frac{\partial z_k}{\partial w_{ki}} = \delta_k^z \frac{\partial z_k}{\partial w_{ki}} = \delta_k^z \cdot x_i$)

A. $w_{ki} \leftarrow w_{ki} - \eta\frac{\partial E}{\partial w_{ki}} = w_{ki} - \eta\sum_{n=1}^{N}\delta_k^o o_k^{(n)}(1 - o_k^{(n)})x_i^{(n)}$

B. $w_{ki} \rightarrow w_{ki} - \eta\frac{\partial E}{\partial w_{ki}} = w_{ki} - \eta\sum_{n=1}^{N}\delta_k^o o_k^{(n)}(1 - o_k^{(n)})x_i^{(n)}$

C. $w_{k(i+1)} \leftarrow w_{ki} - \eta\frac{\partial E}{\partial w_{ki}} = w_{ki} - \eta\sum_{n=1}^{N}\delta_k^o o_k^{(n)}(1 - o_k^{(n)})x_i^{(n)}$

D. $w_{k(i-1)} \rightarrow w_{ki} - \eta\frac{\partial E}{\partial w_{ki}} = w_{ki} - \eta\sum_{n=1}^{N}\delta_k^o o_k^{(n)}(1 - o_k^{(n)})x_i^{(n)}$

——— *End of Part One* ———

Part Two (10 Marks)

Answer those questions in Part2.txt.

1. Describe the difference between GET request and POST request, with appropriate cases. (2 Marks)

2. Why we need data cleansing ? (2 Marks)

3. List **two** different data cleaning activities. (2 Marks)

4. How the random forest estimate of the generalization error and measure the variable importance? (2 Marks)

5. Briefly describe what's the classification and clustering, and the difference between classification and clustering. (2 Marks)

——— *End of Part Two* ———

Part Three (20 Marks)

Answer those two questions in the **provided answer booklet**.

Association Rule Mining(8 Marks)

Given the following transnational data from a restaurant. There are 9 distinct transactions and each transaction involves between 2 and 4 meal items. There are a total of 5 meal items that are involved in the transactions. For simplicity we assign the meal items short names (M1 to M5) rather than the full descriptive names (e.g.,Cherry Pie):

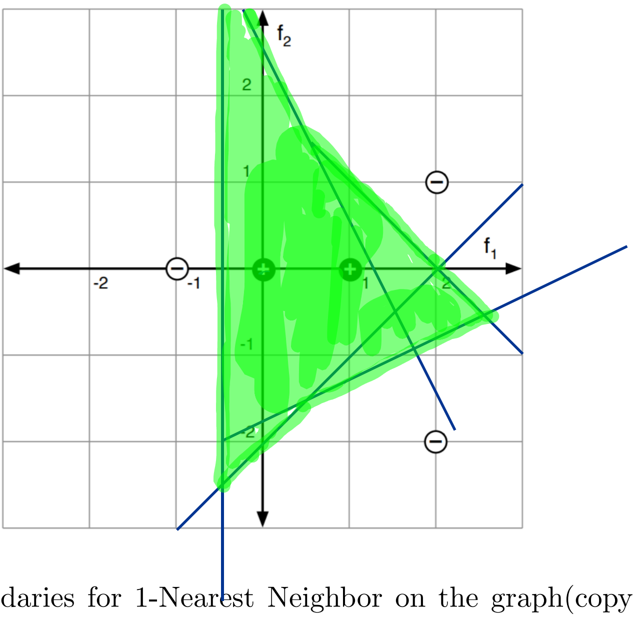
Trans-Id	Items
1	M1,M2,M5
2	M2,M4
3	M2,M3
4	M1,M2,M4
5	M1,M3
6	M2,M3
7	M1,M3
8	M1,M2,M3,M5
9	M1,M2,M3

Table 2: Data for Part Three Question one

1. We want to mine all the frequent itemsets in the data using **The Apriori Algorithm** Assume the minimum support is 20%.(You need to give the set of frequent itemsets in $L1, L2....$, candidate itemsets in $C_1, C_2, ...$).(4 Marks)

2. Find all the association rules that involves only $M1, M2, M5$(in either left of right hand side of the rule). The minimum confidence is 70%. (4 Marks)

k-Nearest Neighbors(12 Marks)



1. Draw the decision boundaries for 1-Nearest Neighbor on the graph(copy it to answer booklet). Your drawing should be accurate enough so that we can tell whether the integer-valued coordinate points in the diagram are on the boundary or, if not, which region they are in. (4 Marks)

2. What class does 1-Nearest Neighbor predict for the new point: (1, -1.01) Explain why. (4 Marks)

3. What class does 3-Nearest Neighbors predict for the new point: (1, -1.01) Explain why. (4 Marks)

——— *End of Examination* ———