CSCI585 Final Exam

June 27th, 2017

Last Name: _____

First Name: _____

Student ID: _____

Email: _____

Signature: _____

Duration: 2 hours

CLOSED book and notes. No electronic devices.

DO YOUR OWN WORK.

If you are discovered to have cheated in any manner, you will get a 0 and be reported to SJACS.

If you continue working on the exam after time is up you will get a 0.

Problem Set	Number of Points	Your Score
Q1	2+2=**4**	
Q2	1+2=**3**	
Q3	1+1+1=**3**	
Q4	1+1+1=**3**	
Q5	1+1+1+1=**4**	
Q6	4+1=**5**	
Q7	2+1=**3**	
Total	**25**	

Q1. (4 points total) SPATIAL DATABASES

a. (2 points) Name two (or more) data structures used to create spatial indices.

Answer: R trees, R+ trees, R* trees, K-d trees, K-d-b trees, Quadtrees, Octrees.

Any one data structure will get 1 point.

b. (2 points) Google Earth uses a KML format to encode spatial data. Write down how KML encodes a geological point (longitude, latitude).

Answer: <Point><coordinates>longitude, latitude</coordinates></Point>

(The order of longitude and latitude doesn't matter.)

Each tag (<point> and <coordinates>) earns 1 point.

If you didn't write down either tag but include <placemark> tag, I will give 1 point. Other somewhat reasonable explanations would get 1 point.

Q2. (3 points total) BUSINESS INTELLIGENCE (BI)

a. (1 point) What's the key factor that impacts the effectiveness of BI?

Answer: The quality of operational data / data gathered at the operational level.

I also accept the answer "operational data". Otherwise, you cannot get the credit.

b. (2 points) Which data schema is widely used in data warehouses for BI? What are the key characteristics of this schema?

Answer: Star schema. / Snowflake (1 point)

It maps multidimensional decision support data into a relational database. / Many-to-one relationship between table and each dimension table (1 point)

0-point answer: Components of star schema: Facts, Dimension, Attributes, and Attribute hierarchy. / Only mention support multidimension facts.

Q3. (3 points total) NoSQL DATABASES

a. (1 point) What are the categories (types) of NoSQL databases?

b. (1 point) Give at least one example (name of the vendor or product) for each type of NoSQL database.

Rubric:

a) and b)
Answer:
1) Key-value store: DynamoDB (Amazon), ProjectVoldemort, Redis, Tokyo Cabinet/Tyrant, memcached, Riak
2) Column-family store: Cassandra, HBase, Google BigTable, HyperTable
3) Document store: MongoDB, CouchDB, MarkLogic, ArangoDB
4) Graph store: Neo4j, HyperGraphDB, Seasame, Graphbase, FlockDB, ArangoDB, Giraph

Rubrics:
- 0.25 per one correct type of NoSQL database.
- 0.25 per one correct example of NoSQL database.
- At least one wrong example for a type of NoSQL database --> No point for that type (0.0/0.25)

c. (1 point) Why some applications prefer NoSQL databases over SQL databases?

Answer:

- Easier to distribute [big] data.

- Performance

- Other data models that fit with specific applications rather than relational model.
- Easier to partition data and store them distributedly (Scale well).
- Performance improvement (because of partition and replication)
- Prevent the expensive "join" operation.
- Schema-free, dealing with unstructure, semi-structured data.

Rubrics:
- Got full point if mentioning all of them.
- If the answer has "big data", got 0.5 point. You need to elaborate more why SQL fails to handle big data to get full point. Similar with mentioning 3 Vs.

Q4. (3 points total) MAP-REDUCE

a. (1 point) Please re-arrange steps below into the correct flow for Map-Reduce.

A. Related key/value pairs from all mappers are forwarded to a shuffler (there are multiple shufflers); each shuffler consolidates its values into a list.
B. Mapper task is run in parallel on all the segments (ie. in each cluster, therefore on each segment); each mapper produces output in the form of (key,value) pairs.
C. Shufflers forward their keys and lists, to reducer tasks; each reducer processes its list of values and emits a single value (for its key).
D. Big data is split into segments, held in a computer cluster.

Answer: D, B, A, C

b. (1 point) Could a commodity machine be used for both map and reduce tasks?

Answer: Yes.

Rubrics:
- Got full point if simply answer "Yes".
- Yes but with a wrong explanation, deduct 0.5 point.
- 0 point with "No" answer.

c. (1 point) MapReduce involves two tasks: Map and Reduce. In each task, cluster of machines are used to perform computations in parallel. Could the Reduce task start before some machines running Map task complete?

Answer: No. Map task must be completed before Reduce task starts.

An example is counting the occurrence of words in a big text file. The steps are:
- Partition the text file into segments (smaller text files).

- Map task: each segment is processed by a node that would generate <word: num_of_occurrences> key-value pairs. Note that a word "the" may be presented in multiple segments, so the output key-value pairs may have multiples ("the", 1).
- Reduce task: aggregate the values for each word. The final result should include ("the", N) where N is the number of occurences of the word "the" in the text file.
With this example (and in general), the reduce task must wait until all machines running map task to finish. If one maching running Map task hasn't finished, the reduce task may have the counter wrong for the word "the" because the segment processed by that machine may have some "the" words in it.

Q5. (4 points total) BIG DATA / DATA SCIENCE INTRO

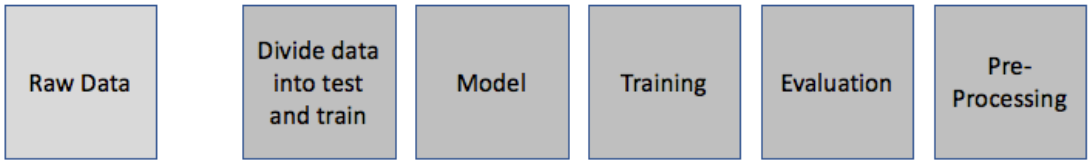
Assume you have been hired by a chocolate shop chain as a data scientist. The marketing group is suggesting that they can boost up the sales by sending the right type of sample advertisement chocolate pieces to online customers. They have already asked hundreds of customers to fill a survey which collected biographical/preference questions (raw data). Of course, the last question was "What type of chocolate do you like the most?" You are asked to come up with a machine learning approach to accomplish this task.

a. (1 point) Is this a supervised or unsupervised problem? Briefly explain.

Answer: This is supervised learning, as labels are already provided.

The term Supervised AND a correct brief explanation is required, any of these being missed means no credit, as we don't have partial credits for this question.

b. (1 point) Use the blocks below and connect them in a meaningful way to create the outline of your machine learning system.



Answer:

The diagram can be:

- Raw data > Pre processing > Divide > Training > Model > Evaluation
- Raw data > Divide > Pre processing > Training > Model > Evaluation

I also accepted this, as (somehow) theoretically it can be accepted:

- Raw data > Divide > Pre processing > Training > Evaluation > Model

Any explanation has been read and if valid has been given credit.

c. (1 point) Briefly explain the purpose of the evaluation step?

Answer:

Evaluates how system performs on test data. It is accepted if explained about a measurement of how accurate we are, even if the term “test set” was not mentioned.

d. (1 point) In another project, we are trying to classify different types of chocolate (A, B, and C). What technique could be useful for this task?

Answer:

Any of the classification techniques (or simply visualization).

They should name at least one technique
Regression (logistic/linear) is not accepted, unless it has been explained how we can deal with a multi-class classification problem using a regression model.
The term "a classification technique" is not accepted.

Q6. (5 points total) MACHINE LEARNING
a. (4 points) There are different ways to categorize the algorithms used in data mining; in class we discussed one of them with four different categories. Based on the description, write down the name of the algorithm:

[-----] involves labeling the data.

Answer: classification

- If a student answered supervised and unsupervised, instead of classification and clustering, +1 credit (instead of +2) has been given.

[-----] involves grouping data not based on pre-defined labels, but based on their similarity.

Answer: clustering

- If a student answered supervised and unsupervised, instead of classification and clustering, +1 credit (instead of +2) has been given.

[-----] in these algorithms, we try to couple the data with a continuous result values. We can use these models for predicting continuous parameters, such as amount of rain in California next week.

Answer: regression

[------] In these approaches, we try to find informative relationships within our data. For example, understanding what parameters will certainly lead to specific disease for patients.

Answer: rule extraction (association rules)

b. (1 point) What's the major problem that any of the models created by these algorithms might suffer from?

Answer: overfitting

- Any valid answer is considered correct. Low accuracy, complicated process, small number of sample inputs, etc. are not accepted as a valid answer.

Q7. (3 points total) DATA VISUALIZATION

a. (2 points) A height field is a regular array of 2D points h = f(x, y), where h is an altitude above or below a point (x, y). Height fields are often used to represent terrain data or depth underground depending on whether the 'h' are all above a point (x,y) or below it. How would you choose to visualize the height field? Describe what properties of the data it would express. HINT: Think creatively! :)

Some Possible Answers:

1. Giving different colors to heights and depths.
2. Divide the data into two maps for better visualization for ease of use.
3. The dimensions of the data, e.g. how high it is, if the height is changing over years and therefore has a time factor related to it, and whether mountains are entirely different from plains in the representation, are some of the things I expect students to write.
4. Visualization aspects of all the dimensions that the student has mentioned.

b. (1 point) Interactivity is a keyword when dealing with visualization. Mention different types of interaction that would be useful to explore the height field mentioned above.

Some Possible Answers:

1. Allow to visualize sections of the height.
2. Check and uncheck boxes for heights and depths.
3. Keep sea depths and mountain heights in different visualizations or not, tradeoffs.
4. If people can click an area and see the history of that terrain data.
5. Putting a timeline would be nice, to see how the terrain changed over time.