CSCI 585, Final Exam, Fall '20

Please read the following carefully, before starting the test.

1. The exam is open books/notes/devices - feel free to look up **whatever** you want!

2. Not counting 'Q0', there are 10 questions, each containing the 'd' word (**data**), numbered Q1..Q10, worth 5 points each. You can pick any 7, for a total of 35. ADDITIONALLY (if you want to, have time) you can answer one, two, or three more - **this means there are 15 bonus points!** We will cap your score at 35, if it exceeds that.

3. There are no 'trick' questions, or ones with long calculations or formulae, nothing that requires you to needlessly write a lot. There aren't questions whose answers are a Google search away [or available directly from the lecture notes] either! The questions do make you think, imagine, and, apply what you learned.

4. **Please do NOT cheat** - this means NOT communicating with anyone via any device/medium/channel - you will get a 0, and be reported to SJACS, if you are found to have cheated; ANY attempt to get help from others in any form is a VIOLATION, as per https://policy.usc.edu/scampus-part-b/ (https://policy.usc.edu/scampus-part-b/), sections 11.11 through 11.14 [read it, if you are not familiar with it].

5. When the time is up (90 minutes), stop your work, then spend the rest of time (30 minutes) on submission.

6. **Good luck**! Hope you do well, and enjoy coming up with the answers. Try to stay calm, take a deep breath, start!

Q0 [0 points]. You **MUST** turn this in - DO NOT omit doing so - there is a **penalty of 2 points** if you omit this.

Please write the following line, and sign it - it is your acknowledgment of having read USC's policies on academic misconduct (https://policy.usc.edu/scampus-part-b/ (https://policy.usc.edu/scampus-part-b/), 11.11-11.14) and agreement to honor them: **I have read USC's standards on academic integrity, and agree to abide by them.**

Q1. The ML revolution, as you know, is fueled by **DATA.**

Q. [1+1= 2 points] The 'supervised' in 'supervised ML' refers to the use of labeled data. Where does the labeling come from? And, how might we automate a part of the labeling task?

A. The labeling is manually done [via software] - by humans. Semi-supervised learning helps automate a part of it, by labeling only a part of the data manually, using that to train a network, then using that network to automatically label the rest.

Q. [1+1+1= 3 points] What features/columns/measurements would you need (ie. data you'd collect), in order to train an ML, on the following:

a. helping someone exercise better (eg. Kemtai (https://app.kemtai.com/setup))

A. Head posture, body posture, movements that are specific to the exercise that would be taught by the AI...

b. helping someone speak better on stage (eg. imagine a person giving a TED talk (https://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization?language=en) for 20 minutes - people have a variety of things they can be coached on, to improve themselves)

A. Gaze direction (where the speaker is looking), speech mannerisms, hand gestures, pacing the stage...

c. an AI replacement for a news reader (https://says.com/my/tech/south-korea-mbn-now-has-an-ai-news-anchor-kim-ju-ha-capable-of-working-24-7) :)

A. Pronounciation, emotional tone, facial expressions, body posture...

Q2. [1*5 = 5 points] Provide an example (using words) for each type of **data** viz indicated below - the example should not be from the viz (https://bytes.usc.edu/cs585/f20_db0DS/lectures/Viz/slides.html) lecture slides. In other words, what data would you visualize?

a. bubble plot

A. Popularity of the various social media platforms.

b. choropleth

A. Number of hours per day spent online, in every state/county in the US.

c. donut chart

A. Relatively popularity of the various types of game consoles (Switch, PlayStation, Xbox, Wii...)

d. network (ie. graph)

A. COVID-19 contact tracing.

e. histogram

A. Popularity of various coding languages in 2020

Q3. [2.5*2 = 5 points] Even though relational DBs are not being used as much anymore, compared to the past, eg. 80s and 90s [when they were the ONLY type of DB used!], their query language, SQL, still lives on, by being used for analyzing non-relational **data**. Name, and discuss, two examples of such 'living on' that we looked at.

A. Hive/HQL: a way to query large amounts of data (equivalent to a data warehouse) held in a Hadoop cluster.

CQL - Cassandra Query Language - for use with Cassandra, a column family DB.

SPARQL: SQL-like language to query RDF triples

Q4. Every major hospital is a huge repository of **data** - about illness, recovery, deaths, medications, surgical procedures, patient comfort, pain management.... For the questions below, assume you have data available from multiple (eg dozens) of hospitals across the country.

Q. [1*5 = 5 points] Pick 5 different DM algorithms, and indicate what you would use each for (ie what type of data would you feed it, what would you predict/learn).

A.

Regression tree: using a terminally ill patient's vital signs, predict how long they will live.

Clustering: based on post-discharge patient survey data, group hospitals (eg into good, average, bad)

Logistic regression: to classify a patient as at-risk/hot-at-risk, for a potential surgical procedure

SVM: to assess if a new drug would be safe/unsafe, on a patient

Neural network: use chest x-rays to train an NN, on COVID-19 detection

....

Q5. [2+1+2 = 5 points] BI is all about extracting/deriving value from massive amounts of existing (transactional) **data**, loaded into a warehouse.

Q. In such a context, what does 'rollup' mean, ie. how does it help in data analysis?

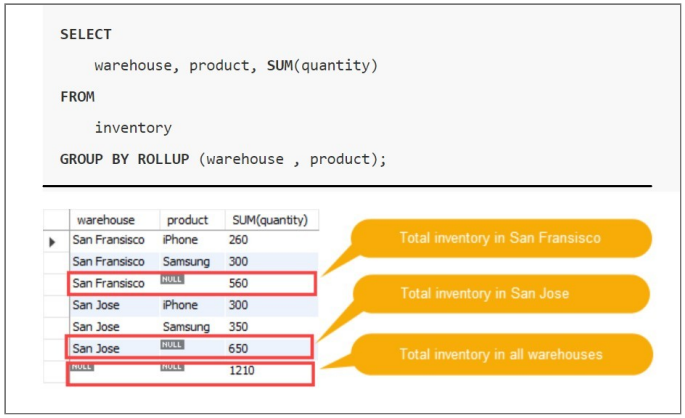
A. Rolling up involves breaking down, or its opposite - aggregating - data, along a dimension that involves a hierarchy (eg. time, space (location), product category, etc). We can get the 'big picture' by rolling up to the highest level in the hierarchy, or 'drill down' to the lowest level to see the breakdown.

Q. How is rollup analogous to viewing spatial data?

A. We can zoom in and out in a map, which is similar to rolling up or drilling down.

Q. Explain, with a simple example, how the SQL extension called ROLLUP help perform rollup.

A. Here is an example that uses a two-level hierarchy warehouses->products:



The answer doesn't need to contain SQL - a diagram/table illustrating subtotals and totals (similar to the output of the query shown above) is acceptable.

Q6. The big benefit of MapReduce is the dramatic acceleration of the processing of large volumes of **data.**

Q. [1 point] Sometimes, a local reduction is performed, at the mapping stage. Why?

A. To reduce network traffic, identical keys' values can be coalesced into a list that is then output.

Q. [1*4 = 4 points] List, and say a few words about, four alternative ways to specify data processing tasks.

A.

Java, for specifying mappers and reducers: a familiar language.

Python, for specifying mappers and reducers: also a familiar language, less verbose than Java.

Pig, for specifying data flow graphs.

Hive, for specifying data processing using the familiar SQL syntax.

Q7. [1*5 = 5 points] As ML/DM matures, we are seeing a whole suite of tools/APIs/hardware... that help with the core task, of handling **data**.

Q. What tool helps visualize a TensorFlow graph (since the graph is specified via coding)?

A. TensorBoard

Q. What is a specification for 'packaging' neural network architectures?

A. CoreML/CreateML, Turi, ONNX...

Q.What are a couple of higher-level APIs that simplify neural network creation?

A. Keras, Pytorch (and mxnet, scikit-learn...)

Q. What are a couple of tools that offer code-free data processing?

A. WEKA, KNIME, RapidMiner, Orange...

Q. What are a couple of hardware-accelerated solutions for processing (NN) data [in addition to GPUs and TPUs - do not list these!].

A. Coral, Jetson Nano, Movidius NCS (and Pixy II, etc)

Q8. [1*5 = 5 points] With regards to COVID-19, comment on governance, security, ethics, privacy, compliance, all from a **data** perspective [in other words - how to do things right, what can be a problem...]. Think broadly: disease spread, tracking, hospitalizations, drugs, virus structure, vaccines...

A. This is a 'freeform' question, with a wide variety of correct answers - but, each answer does need to address an item mentioned above.

Governance: Use of a shared/common language for communicating data (eg about hospitalization rates, testing, etc). Others examples might include ensuring lineage/provenance of data (eg virus genome sequences), ensuring reliable data collection, etc.

Security: Guarding sensitive data such as virus genome sequences, drug trial results, drug discovery details...

Privacy: Safeguarding patient medical records.

Ethics: Ensuring validity of data analyses to include minority populations, choices related to opening up data which might speed up developing vaccines...

Compliance: Legally ensuring that hospitals report accurate, timely data; legally ensuring that testing and tracing protocols are enforced...

Q9. Good use of **data** implies good practices for storing (eg table design) as well as accessing it.

Q. [1 point] What is the single biggest reason, to performance tune an RDBMS' execution?

A. Speedy processing of queries (or just, 'speed').

Q. [1*4 = 4 points] Briefly discuss 2 SQL-based query tuning ways, and 2 non-SQL-based ones.

A. From the notes... in a WHERE condition, use simple operands for column names; with multiple conditions connected via AND, list the one most likely to fail, first [with OR, last]; rule-based optimizing; cost-based optimizing; even, 'creation and use of indices'.

Q10. [2.5*2 = 5 points] We index spatial **data**, for the same reason we index non-spatial data - for speedy retrieval. Name, and briefly discuss, two different ways of indexing spatial data.

A. From the notes: R tree, k-d tree, k-d-b tree, quadtree, even recursive curve schemes such as the z curve one.