CS585 Final

Fall term, 12/11/19
Duration: 1 hour 15 min

Instructions/notes

- the exam is closed books/notes/devices/neighbors, and open mind :)
- there are 8 questions, a 'non-data-related' bonus, for a total of **35** points
- there are no 'trick' questions, or ones with long calculations or formulae
- you can write on the two blank sheets (that are at the end) if you like
- each of the eight Q1-Q8 questions contains the word 'data' :)
- **please DO NOT CHEAT; you will get a 0 if you are found to have cheated**
- **when the time is up, please STOP WRITING; you will get a 0 if you continue**
- **GOOD LUCK!**

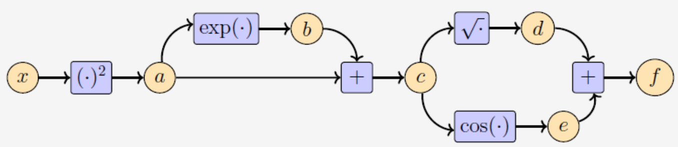
Q1 (1*2+2 = 4 points). The following diagram shows how a function f(x), possibly involving data obtained from a scientific measurement setup, could be computed [using vars to store partial results]:

Consider the function

$$f(x) = \sqrt{x^2 + \exp(x^2)} + \cos\left(x^2 + \exp(x^2)\right) \tag{5.122}$$

from (5.109). If we were to implement a function f on a computer, we would be able to save some computation by using *intermediate variables*:

$$a = x^2\,, \tag{5.123}$$
$$b = \exp(a)\,, \tag{5.124}$$
$$c = a + b\,, \tag{5.125}$$
$$d = \sqrt{c}\,, \tag{5.126}$$
$$e = \cos(c)\,, \tag{5.127}$$
$$f = d + e\,. \tag{5.128}$$



In the above, given x, we're computing f. A standard practice would be to code up a function, eg. 'fn', that encompasses the calculations [using local variables to store results, as shown in the figure], and then call it, eg. fn(4), fn(3.1415), etc.

a. What would be an alternate way to compute f? Why is it a better approach?

Answer: the alternate way is to set this up as a graph, ie. use **DATAFLOW.** It's better because computations can occur in **parallel** (sqrt and cos)**.**

b. Explain how you could further speed up the computation of f, when we have a rather large array of data (eg. 20 million long) of x values to process.

Answer: to futher speed this up, the array could be split horizontally (eg. 20 splits, each with a million values), and these could be **distributed in a cluster and processed in parallel** (using MapReduce for example)**.**

Q2 (3*1 = 3 points).

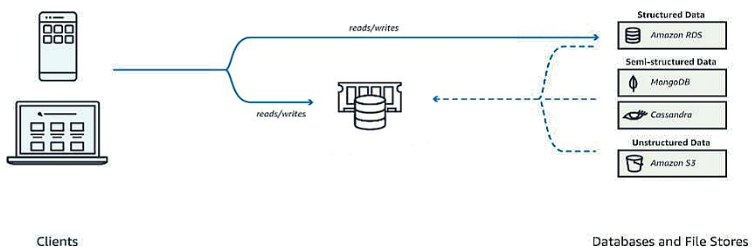
Historically, data viz has been carried out on print media (newspapers, books, magazines, journals) - these provide zero interaction, allowing for just passive consumption (and possibly leading to attendant lack of interest). Today, what else we have are interactivity, and animation (more engaging). What three other modes of data presentation can lead even more engagement and utility?

Answer: we can use **VR**, to be visually surrounded by the data (which can even be superposed over 3D scenes that led to the data creation); we can use **AR**, to have the data be visualized over existing real-world surfaces (eg. tabletops, walls), collaboratively (eg by analysts sitting around a coffee table). We can create **holograms** out of the data (eg. to show polar icecaps melting). Or we could **3D-print** the data (including in color). Or we can use **projection mapping** to project data on to surfaces.

Extra FYI: https://www.intechopen.com/books/holographic-materials-and-optical-systems/3d-capture-and-3d-contents-generation-for-holographic-imaging

Q3 (3+3=6 points).

Consider a key-value (k-v), in-memory store architecture below (shown at the center of the diagram):



a. What are 3 typical reasons why you would set up such a data store for clients?

Answer: **faster access, less load on the backend, higher throughput (more clients can be served).**

b. Typically, such a setup as shown above, would reside in your own IT infrastructure (connected to your organization's web server). **What 3 additional advantages would you get, by switching to a cloud-based service that offers a clustered version of the above (in-memory DB instances running in multiple nodes that are connected together)?**

Answer: **unlimited horizontal scaleout**, even **higher throughput** (could even use clients' location to determine nearest nodes to serve data), **no maintenance** (from our, ie data holder's, perspective).

Q4 (4 points). Association mining (eg. using the A-priori algorithm we studied), applied to ‘shopping baskets’ (large groups of transactions) produces rules of the type A->B, given ‘support’ threshold for A and a confidence for the association - in other words, it outputs items purchased together.

A refinement of the above, can produce even more specific associations - we can consider the occurrence of (A,B) pairs, in the broader context of other (unrelated to (A,B)) factors that might nevertheless influence A->B. This can help us make better use of the (A,B) associations.

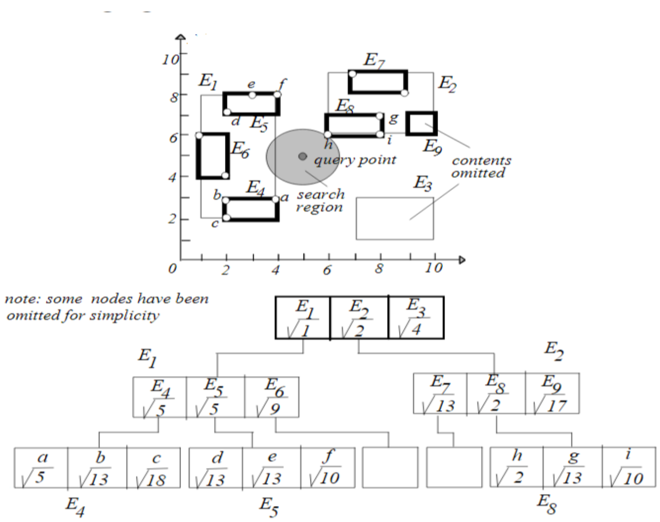
Eg. does (A,B) only occur (or NOT occur) in a certain store or groups of them? Name 4 such additional factors that can be used to analyze our data (mined associations).

Answer:

1. Time of the day
2. Time of the year (seasonal variation)
3. Ethnicity of customers (yes, such data *is* available)
4. Customer’s annual income
5 ... [eg. customer’s age]

FYI: look up differential market basket analysis [“differential MBA” :)]

Q5 (2+4 = 6 points). Consider the following diagram (from an existing source), related to nearest neighbor (NN) queries (like from your HW3 on spatial data):



The R-tree shown is in three levels, with the leaves (a..i) being items of interest, ie. what we hope to get from our NN search. There are several different algorithms for traversal, which might result in different items (any of a through i) being returned by the query.
a. Assuming we do a DEPTH-FIRST search, what (single) item will be returned by a closest-point query?

Answer: **h.**

b. What paths would you consider, to arrive at your answer above (returned item)?
Note - the numbers shown in the R-tree, indicate the closest-distances from the query point, to the bounding boxes, and to the actual leaf items [which, for simplicity, are located ON the bounding box edges and corners, but this doesn't affect the results].

Answer. We consider E1 -> E4 -> (a,b,c) - of these, we'd pick 'a' [smallest of a,b,c].
Next we'd consider E1->E5 (because E5 is also sqrt(5), like E1) -> (d,e,f), but reject all the leaves because they are bigger than sqrt(5). **We skip going down E1->E6**.

Skip E2->E7. Do E2->E8->(h,g,i), then pick 'h' as the better value than 'a' (sqrt(2) < sqrt(5)). **We'd skip going down E2->E9**.

Q6 (5+1 =6 points).

a. In the context of neural networks (which are a way to carry out supervised learning, using pre-labeled data), explain (using just two or three sentences), the following terms.

i. weights

ii. backprop

iii. loss

iv. architecture

v. pre-trained model

Answer: these are all straight from the lecture and discussion. You don't need to use wording from the slides**, your own descriptions are fine as long as they are correct.**

b. Even with a large training dataset, it seems rather easy to 'fool' a standard *convolutional neural network* (CNN) into misclassification [or in some cases, no need to explicitly attempt to fool it - it simply seems incapable of correct classification (eg. might classify an upright scooter as a parachute)]. **What is the underlying cause of such drastic failures?**

Answer: the neural net has **no additional data** beyond training data such as images, audio [knowledge of features of objects, or hierarchies (assemblies), groupings, context (eg. what objects are found where, and why), etc] - **they only learn from pixels/audio/text...** that have been previously labeled (by humans) and input to them.

Q7 (1+2 = 3 points). Machine learning ('ML'), especially deep learning (which uses massive amounts of training data that passes through deep layers of neurons), is a "revolution" that has rapidly (starting in 2012) taken over every industry and field.

But, for all its successes, there are many glaring issues, one of which is 'bias' - this has resulted in sentences unjustly imposed, medical insurance unreasonably denied, people misidentified as criminals, etc.

a. What is the source of bias, in ML?

Answer: **simply, the chief source of bias is in the dataset.** FYI - an additional source of bias could be in the NN algorithm, ie in the calculation of loss.

b. How would we fix the issue?

Answer: by analyzing ("auditing") the data for **fairness** (data cannot be incorrect/inaccurate) and **completeness/balance** (eg. cannot predominantly contain data about select labels).

Q8 (1*3 = 3 points). Here's a blue-sky ("be imaginative!) question. As voluminous as the data we process today seems, the future is sure to involve even more of it (eg. via higher resolution scientific instruments, more sensor-generated 'IoT' data, etc.). **Where do you see the following headed (in other words, what's the trend, what's coming up (even if it is in research or prototype stages)?** Think BROADLY, in terms of new technologies (including phenomena, materials, devices, designs...)! In other words, how do YOU plan to deal with these?

a. storage (to hold data)

Answer: **DNA, holographic storage, alternate materials.**

b. processing (computing)

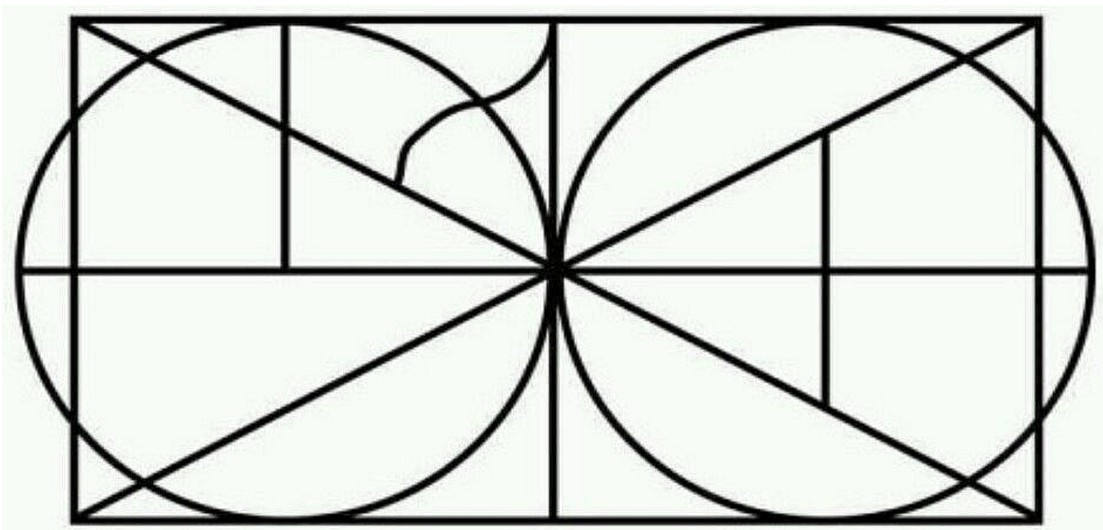
Answer: **quantum computing, DNA computing, optical computing.**

c. infrastructure (how the above two are accessed and utilized)

Answer: **edge computing** (at or near the source of the data), **custom SoCs** (eg. intelligent cameras that output labels+bounding boxes in addition to raw video), **custom 'AI' processors, newer chip architectures** (eg. NN in hardware)...

Bonus (1 point). This bonus point will count, ie. be added to your total for Q1-Q8, if the total is < 35 (if you already have a 35, it will be skipped). In other words, the max you can get for the entire exam is 35, not 36.

The figure below, represents something specific - what is it? The answer is not open-ended (eg. you can't say 'an abstract stained-glass window pattern'!). A very big hint - it's something we COMMONLY use!



Answer: **it shows A-Z (uppercase!). It also shows 0..9.**