CS585: Database Systems
Spring 2016 - Final Exam
5/5/16, 5:00-6:30 PM

Your name: _____
Your student ID: _____

Question	Your score	Max score
1		2
2		2
3		2
4		2
5		2
6		2
7		2
8		2
9		2
10		2
11		2
12		2
13		2
14		2
15		2
Bonus		1
Total		31

RULES/NOTES!

There are 15 questions, each worth 2 points, plus a bonus (non-db) question worth 1 point. That makes the total be 31, which means that each point counts towards 1% of your overall grade.

The exam might *look* long, but don't be concerned on account of that - each question has plenty of language to help clarify things, to help you think - that is what makes the test seem long! Plus, the 'CrowdMark' assessment system we're using requires us to start each question at the top of a new page..

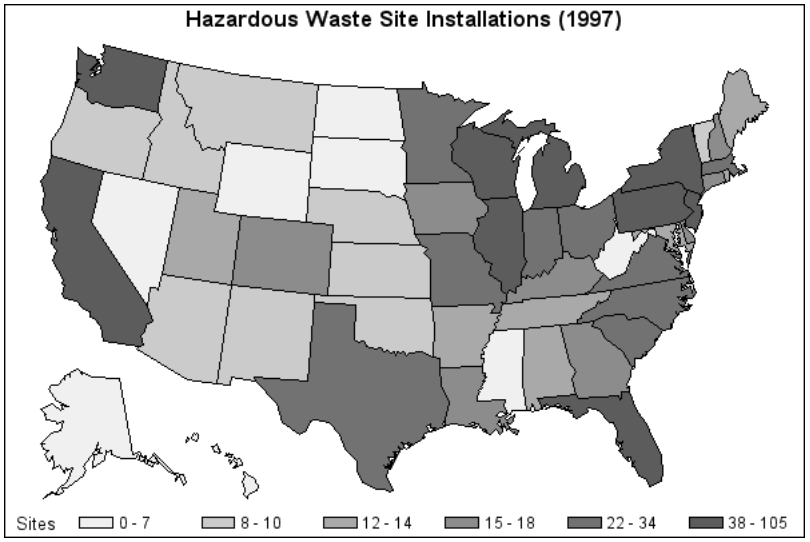
The exam is CLOSED book/notes/devices/neighbors.. If you are observed cheating, or later discovered to have cheated in any manner, you will get a 0 on the test and also be reported to SJACS.

You NEED to turn in your cheat-sheet (please write your name on it) along with your exam; if you don't, you will lose 1 point.

When we announce that the time is up, you NEED to stop writing immediately, and turn in what you have; if you don't, we will not grade your exam (ie. you will get a 0).

Have fun, and good luck! Hope you do well.

Q1 (2 points)



The above map shows spatially-varying data, for a variable being studied (number of hazardous waste sites). What is such type of depiction called?

A. A choropleth map.

Q2 (2 points)

Consider the relational table (aka "relation" or just "table"!) below:

ID	Name	Number	Type
1	John	06472643	Work
1	John	01164322	Home
2	Jane	01726443	Work
2	Jane	06243344	Mobile
3	Jack	01167343	Home

We'd like store the data in such a table, as JSON (we are going to use a NoSQL db for this). Complete the following JSON representation (where we store both phone numbers under a single key called 'phoneNumbers'):

```
{
  "id":1,
  "name":"John",
  "phoneNumbers":



  ,
  "id":2,




}
```

A.

```
[
{
  "id":1,
  "name":"John",
  "phoneNumbers": [
```

```
    "work": "06472643",
    "home": "01164322"
  ]
}
,
{
  "id":2,
 "name":"Jane",
  "phoneNumbers": [
    "work": "01726443",
    "mobile": "06243344"
  ]
}
,
{
  "id":3,
 "name":"Jack",
  "phoneNumbers": [
    "home": "01167343",
  ]
}
]
```

Notes: OK for the phone #s to be ints, and ok for the phoneNumbers keys to not have capitalizations ("home" and "Home" are both acceptable, likewise for Mobile and Work).

Q3 (2 points)

Pick four apps from your smartphone (or just think of four apps), and for each, indicate what types of data are needed to be collected/stored, in order to enable the app to function. For each app, just list field (attr) name, field type pairs (eg. sportName:string..) - no need for a table/relation /JSON/XML representation. List 4 attr:type pairs for each app.

A.

Linkedin: name:string, numConnections:int, Education:string, url:string

Facebook: name:string, numFriends:int, numNewConversations:int, numUpdates:int

iBooks: ...

Waze: ...

Q4 (2 points)

One reason why 'Big Data' is so big because people, and machines/devices/sensors are generating near-continuous data which are being stored and/or processed. Name two sources of people-originated data, and two sources from instruments, machines, devices or sensors.

A.

People: buying habits (eg. at Ralph's), clickstreams (eg. at Amazon)

Machines/sensors: FitBit device, road sensors

Q5 (2 points)

Why is 'Big Data' big *now* (why not, say, 15 years ago)? In other words, what mix of factors is responsible for (have led up to) this?

A. Because now we have the perfect storm of data storage (disk space is cheap!) and virtually unlimited computing power (via the 'cloud'), and Hadoop/MapReduce for efficient processing.

Q6 (2 points)

Imagine that scientists at an environmental agency decide to study, over the course of a year, how lead levels vary in tap water, in apartments in low-income neighborhoods, in cities all over Southern California. So they collect tap water samples each week at these locations, analyze the lead content (15 parts-per-billion is considered a safe level), and chart the data. There will be a 'lot' of data, but we can't quite call it Big Data (amount of data is not big enough).

How would you advise them to turn this into a Big Data study? In other words, how we can expand the scope of the study? List at least two ways (in addition to this one: increase the duration of the study, eg. make it 10 years instead of 1):

A. This can be answered in many ways - the study period can be extended (eg to 5 years), frequency of sample collection can be changed (eg to once a day), data can collected for a larger region (eg. Northern California too)..

Q7 (2 points)

How would you describe 'Data Science' in your own words, in a sentence or two? Don't provide a memorized regurgitation!

Data science is the systematic study of data - its collection, processing, analysis, visualization.. Again, this can be answered in many ways!

Q8 (2 points)

Studying data starts with its collection. What next, and after that...? In other words, what is the typical 'life cycle' for data? You can use words and/or a diagram to answer.

A. Collect, process (eg. clean up), analyze/mine.., visualize (optional), make changes to data generation processes, collect more data.. repeat..

Again this can have multiple ways of expressing the above.

Q9 (2 points)

What specifically does Hadoop/YARN tend to get used for? Please be precise (eg. the answer is not 'for processing Big Data' :)).

A. Hadoop/YARN allows mapping and reducing, ie. **parallel processing** of data analysis computations.

Q10 (2 points)

Disclosure: this was NOT explicitly covered in class - but with a little thinking you should be able to answer. Consider the target symbol below (from TARGET - used without permission!)

As you can see, there are two distinct sets of marks - in the middle are 'good' archers' circular marks (right on target!), and, along the periphery (closer to the circular edge), the 'bad' archers' rectangular marks.

We'd like to use a simple linear SVM to classify the data into 'good' and 'bad' [reminder: an SVM has a line that divides the data, created from two support-derived parallel lines with maximal separation]. Trouble is, there is no way to create a separation line, given our two clusters! But what we can do is draw a concentric circle to separate the data:

But we don't want a separating circle, we want a classic SVM straight line - **how would you achieve this**? In other words, what **data transformation** can you think of, to replot the archers' data, that will let us easily create an SVM line? Please try to be precise (eg. equations or formulae would be nice if possible).

PS: You'll NEVER look at a TARGET store or its logo, the same way again :)

A. TRANSFORM the data, from (x,y) for each point, to (r,theta) [polar co-ordinates]:

r = sqrt(x^2+y^2)
theta = arctan(y,x)

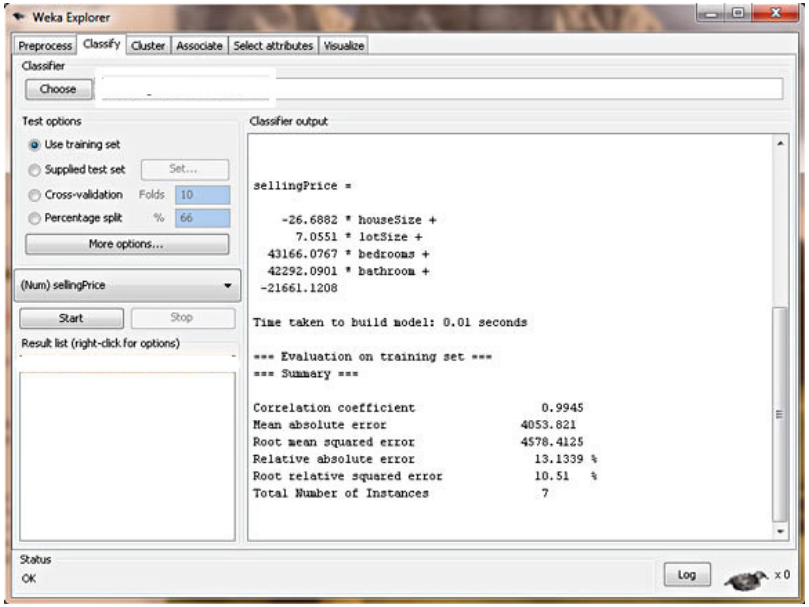
Now we can plot r vs theta :) THAT will create a nice separation in the 'r' axis, an SVM can therefore classify the (r,theta) values!

So the 'trick' here is to transform the data to a different space.

Other solutions (related to ours) are also OK, eg. plotting r vs r (!) would work too.

Q11 (2 points)

The following is an edited (parts redacted!) screenshot that shows WEKA. What data mining [classification] algorithm is being run?



A. Regression (or 'linear regression').

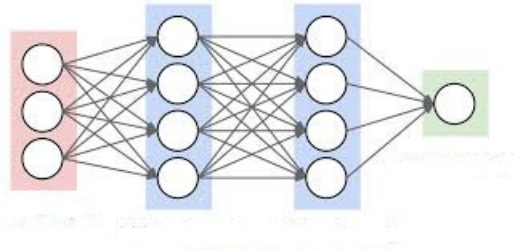
Q12 (2 points)

Scripting languages are highly popular for doing data analysis, since (among other features) they offer higher-level datatypes, eg. Python has the list and dictionary types built into it. Name two such higher-level datatypes in R:

A. Vector, matrix, array, data frame, list, factor..

Q13 (2 points)

The following shows a highly popular, and successful as of late, ML algorithm - what is it?



A. A neural network.

How does it work (ie. learn)?

A. Many ways to answer, here's one: "Neural neworks are typically organized in layers. Layers are made up of a number of interconnected 'nodes' which contain an 'activation function'. Patterns are presented to the network via the 'input layer', which communicates to one or more 'hidden layers' where the actual processing is done via a system of weighted 'connections'. The hidden layers then link to an 'output layer' where the answer is output. [from http://pages.cs.wisc.edu/~bolo/shipyard/neural/local.html]

Mentioning layers, weights, error backpropagation, activation function etc. are OK too.

Q14 (2 points)

GPUs (Graphics Processing Units), as their name implies, are chips that are meant to accelerate graphics calculations (such as image rendering from 3D data). But of late, they are being put to prominent use, for data analysis - how/why?

A. Because GPUs contain hundreds (thousands in some cases) of identical processing elements ("cores"), each of which can be used to represent a neuron - this makes learning possible in parallel (all cores in a network layer can run in parallel).

Again, OK to describe this is different ways as long as 'parallel' is mentioned.

Q15 (2 points)

In the 'old days' (eg. 1844!) it was only possible to show data in tabular form (list of numbers, example below) or using simple line/bar/pie charts. Thankfully we have come a LONG way since. What are four things that make today's data visualizations much better?

Counting-Room

ALMANAC.

1844.

	SUNDAY	MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY	SATURDAY
JANUARY.....		1	2	3	4	5	6
	7	8	9	10	11	12	13
	14	15	16	17	18	19	20

ALMANACK.

January, Firſt Month, hath xxxi Days

D M	D W	Sun rises	Sun ſets	D ſets
20	M	7 11	4 49	3 44
21	Tu	7 11	4 49	4 0
22	W	7 10	4 50	5 15
23	Th	7 9	4 51	6 15
24	Fri	7 8	4 52	ſets
25	Sat	7 7	4 53	5 32
26	Sun	7 6	4 54	6 44

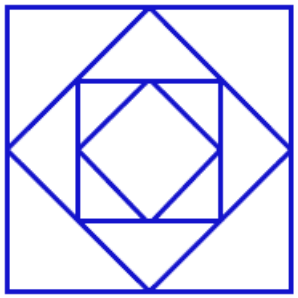
HIGH-WATER *at* New-York *this Week.*

Monday	8 min after 3	Friday	52 min after 6
Teeſday	12 min after 4	Saturday	50 min after 7
Wedneſday	0 min after 5	Sunday	46 min after 8
Thurſday	54 min after 5	Firſt Quarter on *Monday.*	

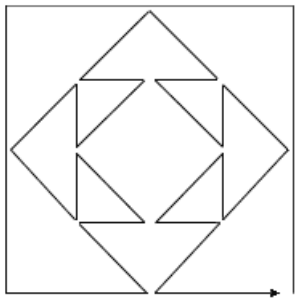
A. We can employ a variety of visualization techniques, color, interactivity, animation, even immersion VR)..

Bonus (1 point)

How would you draw the following figure using a single, continuous line: you can't lift the pen while drawing, and you can't draw over even a part of an existing line.



A.



A solution where the line crosses itself ["you can't draw over even a part of an existing line"] is not valid, but might deserve partial credit, ie. half a point.