CSE5334 Data Mining

Fall 2020, Prof. Deokgun Park

Department of Computer Science and Engineering
University of Texas at Arlington

Final Exam

Time Length: 150 minutes

Name:_____ NetID:_____

- Print your name and NetID above. Print NetID in upper right corner of every page.
- Including this cover page, this exam has 16 pages. Make sure you don't miss any.
- Look through the entire exam before getting started, and plan your time accordingly.
- No questions will be answered during the exam. If you believe something is wrong, write down your thoughts. If you feel something is confusing or ambiguous, clearly write your assumptions that you make while solving the problem.
- This is an close-book, open-notes exam. Also, calculators are permitted.
- You can assume all the logarithms in formulas are base 10.
- Any form of cheating results in a zero grade.

Problems	Part I	Part II (1)	Part II (2)
Points	60	20	20
Score			

Part I. (60 points, 3 points each) Short Questions.

(1) Explain the difference between noise/data/information/knowledge with examples

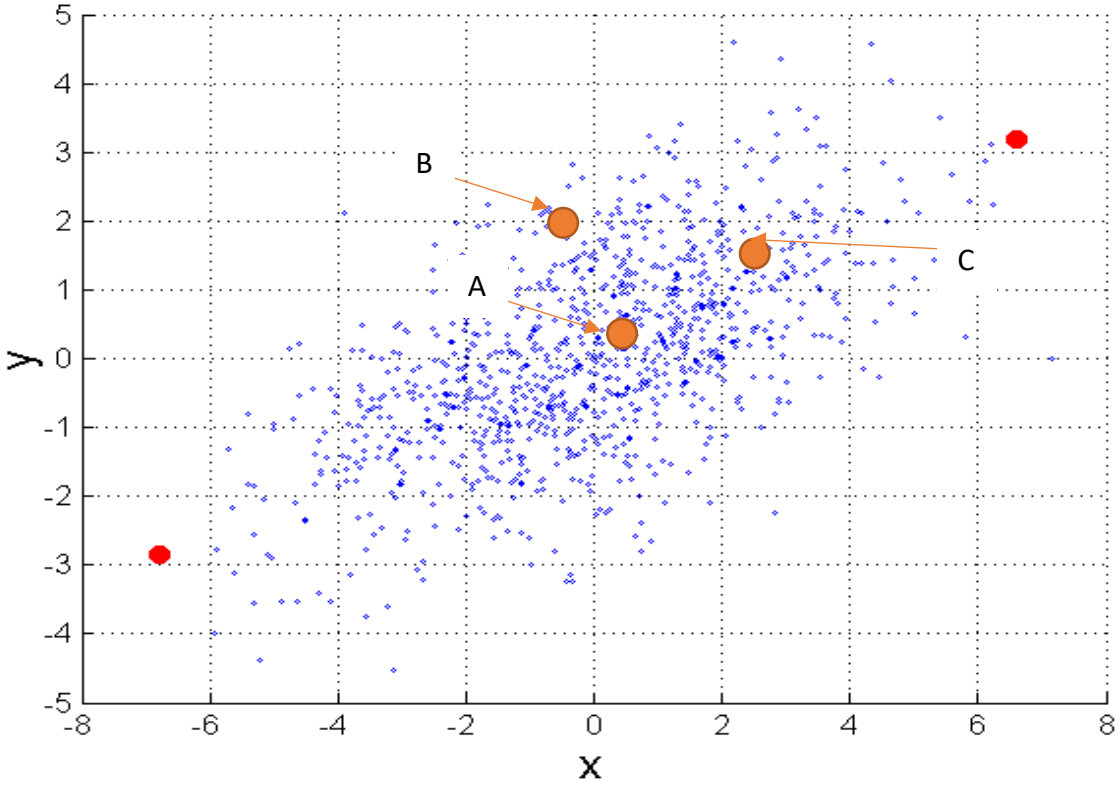
(2) Name the 4Vs of Big data and list one example for each characteristic

(3) Describe how you will choose the k in kNN algorithm.

(4) Name one example for structured/unstructured and semi-structured data

(5) According to the measurement theory, what is the four different types of attributes? Give one example and list allowed operators. ($=, \neq, <, >, \div, \times, +, -$)

(6) In the below dataset, Euclidean distance between AB and AC is same. But among B and C which one should be more similar to A? Explain why and name the appropriate distance measure for this situation.



(7) Calculate the entropy of fair coin (P[Head]=0.5) and fake coin (P[Head]=0.8).

(8) How do you know if overfitting is happening?

(9) Explain why we need to use smoothing for Naive Bayes. Give an example.

(10) Classify the doc 5 whether belong to c or j using Naïve Bayes classifier. Don't use smoothing.

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

(11) Let's say you are building Decision tree for tax fraud. You have two attributes (gender and whether the taxpayer has house for splitting data. Using attribute test for impurity using GINI index which attribute is better.

Id	Gender	Has Car	Tax Fraud
1	M	T	T
2	M	T	T
3	M	T	T
4	M	T	T
5	M	F	T
6	F	F	T
7	M	T	F
8	M	F	F
9	F	F	F
10	F	F	F
11	F	F	F
12	F	F	F

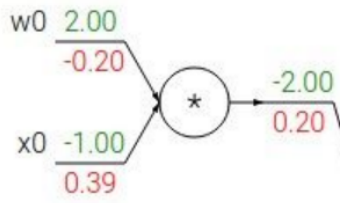
(12) Calculate the probability that anyone actually has breast cancer if anyone is diagnosed as having breast cancer using Bayes theorem

- 1% of women have breast cancer (and therefore 99% do not).
- 80% of mammograms detect breast cancer when it is there (and therefore 20% miss it). (Sensitivity = 80%)
- 9.6% of mammograms detect breast cancer when it's **not** there (and therefore 90.4% correctly return a negative result). (Specificity = 9.6%)

	Cancer (1%)	No Cancer (99%)
Test Pos	80%	9.6%
Test Neg	20%	90.4%

(13) Build a computation graph for Z = 2 (X *Y) – Max (S, 0)

(14) For the computation graph in (18) get $\frac{\partial Z}{\partial X}, \frac{\partial Z}{\partial Y}, \frac{\partial Z}{\partial S}$ when X = 3, Y= 2, S = -1 by doing forward/backward pass. Mark your answer for the forward pass in the upside of the link and backward pass in the lower side of the link like this.



(15) Consider the following set of frequent 3-itemsets:

{1,2,3}, {1,2,4}, {1,2,5},{1,3,4}, {1,3,5}, {2,3,4}, {2,3,5}, {3,4,5}

(a) List all candidate 4-itemsets obtained by the candidate generation procedure in Apriori
(b) Among those candidates, which 4-itemsets can be pruned?

Part 2. Long Questions. (40 points)

Describe everything you know.

(1) Assume you want to predict the genre of the movie based on the reviews. Describe the all steps you know. Write down your assumptions. You can use different dataset if you want.

(2) Pick one of your favorite company. Assume you are Chief Information Officer of that company. You want to create new value using the data. Describe what you will do in detail. (Hint: You will get more credit when you use the concept you learned during the course : such as cross validation, structured/unstructured, entropy, Apriori principle, and so on)