UIUC-CS412 “**An Introduction to Data Mining**” (Fall 2020)

Midterm Exam

(Tuesday, Oct. 20, 2020, 90 minutes, 100 marks)

IMPORTANT Notes

- Please provide brief explanations of your answers.
- Please sign the honor code in page 2. Your exam will not be graded unless the above agreement is signed. Please attach the signed honor code to your answer sheet.
- You can either (1) type in your answers in word or latex and submit your answer sheet in word or pdf, or (2) provide hand-written answers and submit a scanned version. For (2), Please make sure that the scanned version is clear and recognizable. Otherwise, you might loose points.

:

Name: NetID: Score:

1	2	3	4	5	6	Total

CS412 Fall 2020 Exam Honor Code

I understand that the rules of the CS412 exams in Fall 2020. That is, (1) the exams are "closed book" which means that no books or electronics are allowed, except the one for the zoom connection, and (2) I am not allowed to confer with other people about the questions or solutions to the exam (to either give or receive aid).

I have neither given nor received inappropriate aid during this exam.

I understand that my exam will not be graded unless the above agreement is signed.

NetID (print):

Name (print):

Name (signed):

Date:

1 Minkowski Distance [10 points]

Given two data points in 2-D space: $x_1 = (1, 2)'$ and $x_2 = (4, 3)'$.

(a) [2 pts] What is the L_2 distance between x_1 and x_2?

Solution: $\sqrt{10}$

(b) [2 pts] What is the L_1 distance between x_1 and x_2?

Solution: 4

(c) [2 pts] What is the L_∞ distance between x_1 and x_2?

Solution: 3

(d) [2 pts] Given another data point $x_3 = (a, b)$, where a and b are unknown real numbers. Under which condition is the L_2 distance between x_1 and x_3 less than or equal to the L_1 distance between them?

Solution: always (any a and any b)

(e) [2 pts] For x_3 given in (d), under which condition is the L_2 distance between x_1 and x_3 equal to the L_∞ distance between them?

Solution: $a = 1$ **or** $b = 2$

2 Basic Statistics and Normalization [20 points]

Table 2 provides the information of 9 randomly sampled students' midterm exam scores of an online course.

Table 1: Midterm Exam Scores of 9 Students.

Student No.	1	2	3	4	5	6	7	8	9
Final	99	82	78	100	89	92	88	60	75

(a) [3 pts] What is the median score?

Solution: 88

(b) [3 pts] If the score for student # 8 increases to 85, how would that affect the median score of this dataset?

Solution: no change

(c) [2 pts. **True or False**] If we perform normalization by decimal scaling, the normalized scores will be in the range of $[0, 1]$.

Solution: False

(d) [10 pts in total] Suppose we have N data points, where N is an even number. Here, we exclude the special case where all data points have exactly the same value. If we normalize them using *min-max* normalization to the range of [0, 2], we find out that the sample mean of the normalized data points is 1. What is the largest possible sample variance of these N normalized data points? [3 pts] When does that happen? [2 pts] What is the smallest possible sample variance of these N normalized data points? [3 pts] When does that happen? [2 pts]

Solution: $N/(N-1)$ **when all points equal to either 0 or 2;** $2/(N-1)$ **when 1 point equals to 0, 1 point equals to 2 and all other points equals to 1.**

(e) [2 pts] Suppose we have N data points, where N is an even number. We normalize them using *z-score* normalization. What is the largest possible variance of these N normalized data points?

Solution: 1 (always 1)

3 χ^2 (Chi-Square) Test [20 points]

Given the following contingency table, we want to use χ^2 test to decide if the two random variables (play chess vs. like science fiction) are correlated or not.

Table 2: Contingency Table.

	play chess	not play chess	sum (row)
like science fiction	100	400	500
not like science fiction	300	200	500
sum (column)	400	600	1000

(a) [4 pts] Under the null hypothesis (i.e., 'play chess' and 'like science fiction' are independent with each other), what is the expected number for 'play chess' and 'like science fiction'?

Solution: $4/10 * 1/2 * 1000 = 200$

(b) [4 pts] Under the null hypothesis (i.e., 'play chess' and 'like science fiction' are independent with each other), what is the expected number for 'play chess' and 'not like science fiction'?

Solution: $4/10 * 1/2 * 1000 = 200$

(c) [4 pts in total] What is the χ^2 value? [2 pts] Based on that, are these two random variables uncorrelated or not? [2 pts] [*Hint: The chi-square table is attached at the end of the exam book if needed. You can pick confidence level of 0.05 if needed.*]

Solution: 167.7. correlated

(d) [6 pts in total] Now, let us consider the following contigency table in Tabel 3, where $0 \leq x \leq 1000$ is an integer. Under which condition are the two random variables ('play chess' vs. 'like science fiction') *un-correlated* regardless the confidence level? [3 pts] Under which condition is the expected number for 'play chess' vs. 'like science fiction' larger than its observed number? [3 pts]

Table 3: Contingency Table.

	play chess	not play chess	sum (row)
like science fiction	x	$1000 - x$	1000
not like science fiction	300	200	500
sum (column)	$300 + x$	$1200 - x$	1500

Solution: $x = 600$**;** $x < 600$

(e) [2 pts.] Now, let us consider the following contingency table in Tabel 4, where a, b, c, d are positive integers. Under which condition are the two random variables ('play chess' vs. 'like science fiction') *un-correlated*, regardless the confidence level?

Table 4: Contingency Table.

	play chess	not play chess	sum (row)
like science fiction	a	b	$a + b$
not like science fiction	c	d	$c + d$
sum (column)	$a + c$	$b + d$	$a + b + c + d$

Solution: $ad = bc$

4 Principle Component Analysis (PCA) [15 points]

(a) [8 pts in total] Given the following three data points in 2-D space, $x_1 = (1, 2)$, $x_2 = (2, 4)$, and $x_3 = (-3, -6)$. What is the first principle component? [4 pts]? What is the second principle component [4 pts]?

Solution: 1st pc: $\frac{1}{\sqrt{5}}(1, 2)$**. 2nd pc:** $\frac{1}{\sqrt{5}}(-2, 1)$

(b) [3 pts] Now, let us add one more data point $x_4 = (0, 3)$. *Qualitatively* describe how the newly added data point x_4 would impact the first principle component?

Solution: 1st pc will move toward $(0, 1)$ **(somewhere between the original 1st pc and** $(0, 1)$

(c) [4 pts in total] Now, let us add an *infinite* amount of data points $x_i = (0, 3)$, $(i = 4, 5, ...)$. What is the first principle component of the entire dataset $\{x_i,\ (i = 1, 2, ...)\}$? [2 pts] If we only use the first principle component to approximate the original dataset, what is the reconstruction error? [2 pts] Note that the reconstruction error can be computed as $\frac{1}{N}\sqrt{\sum_i (x_i - \tilde{x}_i)^2}$, where $\tilde{x}_i$ is the reconstructed data point for x_i using the first principle component, and N is the total number of data points.

Solution: first pc: $(0, 1)$**, reconstruction error: 0**

5 Data Warehouse [15 points]

(a) [4 pts] Suppose we will build a data warehouse with three dimensions, including `location`, `supplier`, and `item`. If we do not consider the concept hierarchy, how many *non-base* cuboids are there in total?

Solution: 7

(b) [6 pts in total] Suppose the `location` dimension has three different values, including `Urbana`, `Chicago` and `New York City`; the *total* number of different values of `supplier` dimension and the `item` dimension 10. How many *base cells* are there *at most*? [3 pts] How many *aggregated cells* are there *at most*? [3 pts]

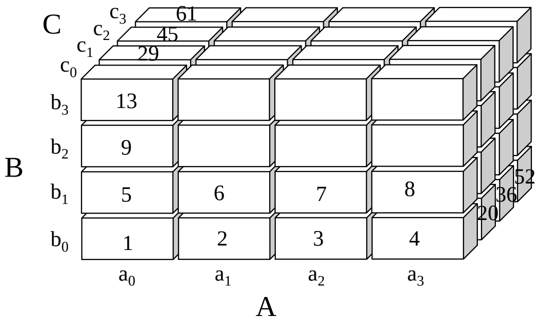
Solution: base cells: $-3x^2 + 30x$**, where** x **is the number of different values of supplier dimension. when** $x = 5$**, there are most base cells:** $3 \times 5 \times 5 = 75$ **and aggregated cells:** $-x^2 + 10x + 44$**, when** $x = 5$**, it reaches the maximum: 69**

(c) [5 pts] Now, suppose we have a new data cube with 100 dimensions. There are only two base cells in the base cuboid: $(a_1, a_2, a_3, a_4, ..., a_{100})$ and $(a_1, a_2, a_3, b_4, ..., b_{100})$, where $a_i \neq b_i (i = 4, ..., 100)$. How many *aggregated* cells are there in total?

Solution: $2^{101} - 2(\texttt{base cells themselves}) - 8(\texttt{overlapped})$

6 Data Cube Computation [20 points]

Assume our data is stored in a data cube with 3 dimensions A, B and C. We would like to do full cube computation using multi-way array aggregation. The lengths of dimension A and B are 4000 and 200, respectively. The length of dimension C is an unknown value x. We cut each dimension in quarters and get 64 chunks as follows.



(a) [4 pts] If we follow the scan order 1-2-3-4-5-6-7-8-9-10-..., what is the memory requirement to compute the whole cube?

Solution: **BC:** $50 * x/4 = 12.5x$; **AB:** $4000 * 200 = 800000$; **AC:** $4000 * x/4 = 1000x$. **In total,** $800000 + 1012.5x$.

(b) [4 pts] If we follow the scan order 1-5-9-13-2-6-10-14-3-7-..., what is the memory requirement to compute the whole cube?

Solution: **BC:** $200 * x/4 = 50x$; **AB:** $4000 * 200 = 800000$; **AC:** $1000 * x/4 = 250x$. **In total,** $800000 + 300x$.

(c) [4 pts] If we follow the scan order 1-17-33-49-5-21-37-53-9-25-..., what is the memory requirement to compute the whole cube?

Solution: BC: $200 * x = 200x$; **AB:** $1000 * 50 = 50000$; **AC:** $1000 * x = 1000x$. **In total,** $50000 + 1200x$.

(d) [8 pts in total] Under which condition is (b) the most memory-saving order **among (a), (b) and (c)**? [4 pts] Under which condition is (a) more memory-efficient than (b)? [4 pts]

Solution: $x > 833.33$. **never**

Appendix

d.f.	P=0.05	P=0.01	P=0.001
1	3.84	6.64	10.83
2	5.99	9.21	13.82
3	7.82	11.35	16.27
4	9.49	13.28	18.47
5	11.07	15.09	20.52
6	12.59	16.81	22.46
7	14.07	18.48	24.43
8	15.51	20.09	26.13
9	16.92	21.67	27.88

Table 5: Table of Chi-square Statistics

(**Opinion**).

1. I □ like □ dislike the exams in this style.
2. In general, the exam questions are □ too hard □ too easy □ just right.
3. I □ have plenty of time □ have just enough time □ do not have enough time to finish the exam questions.