National University of Computer & Emerging Sciences, Karachi
Spring 2020 CS-Department
Midterm 1
23rd February 2020, 11:00 am – 12:00 pm

Course Code: CS481	**Course Name:** Data Science
Instructor Name: Dr Muhammad Atif Tahir and Zeeshan Khan	
Student Roll No:	**Section No:**

Instructions:

- Return the question paper.
- Read each question completely before answering it. There are **3** questions and **2** pages
- Show all steps clearly.

Time: 60 minutes. **Max Marks**: 12.5 points

Question 2 [1.5 Points]: Figure 1 outlines data science progression aiming to reduce the immaturity of capabilities and capacity. Complete the table below regarding space A, B, C, D. (If u wish, you can answer directly in question paper)

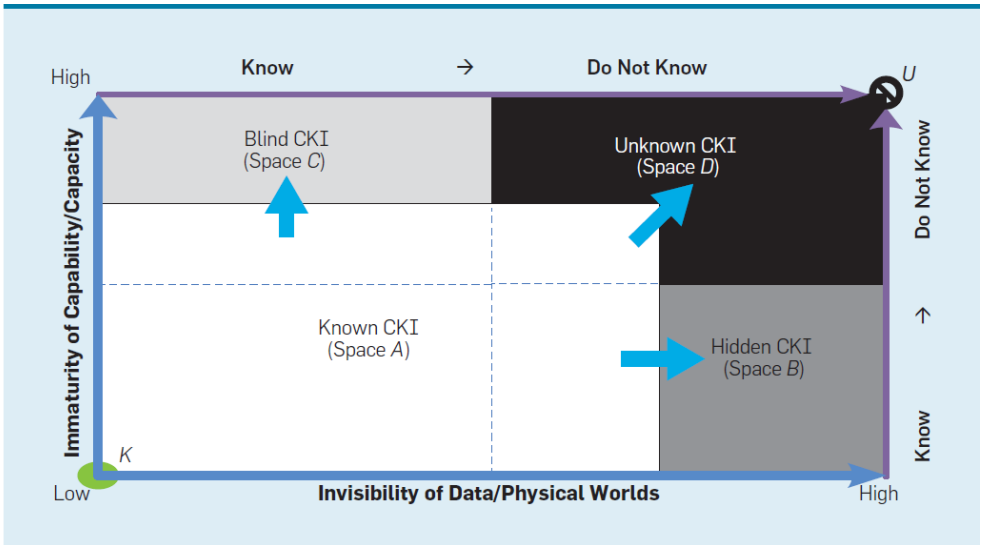

Figure 1: Data Science Space.

Space	Description	Example
A (Known space)	I know what I know	Profiling and Descriptive Analysis
B (Hidden space)		
C (Blind space)		
D (unknown space)		

Solution

Space	Description	Example
A (Known space)	I know what I know about visible world	Profiling and Descriptive Analysis
B (Hidden space)	**I know what I do not know about unseen world**	Examples include existing IID models (such as k-means and the k-nearest neighbors algorithm) that can-not handle non-IID data.
C (Blind space)	**I do not know what I know about the world**	When even established social scientists try to address a data science problem.

D (unknown space)	I do not know what I do not know	In the world of fast-evolving big data, CKI invisibility increases, resulting in an ever-larger unknown space.

Question 3 [6 Points]

Consider the algorithm below published in paper titled "Weighted k-Nearest-Neighbor Techniques and Ordinal Classification". This algorithm discusses weighted kNN classifier. You need to classify new data points from Table below using wKNN classifier. Use k = 4, city block distance: $d(p,q) = \sum_i |p_i - q_i|$, and compute kernel using inversion kernel (1/|d|). Note that I(x) is the indicator function which evaluates to 1 when the argument x is true and 0 otherwise

Instance #	Att1	Att2	Actual Class
1	2	3	0
2	1	5	1
3	4	2	1
4	2	5	0
5	6	8	0

Instance #	Att1	Att2	Predicted Class
1	3	1	?
2	2	2	?

Training and Test Data for Question 3.

Weighted k-Nearest-Neighbor classification (wkNN)

1. Let $L = \{(y_i, x_i), i = 1, \ldots, n_L\}$ be a learning set of observations x_i with given class membership y_i and let x be a new observation, whose class label y has to be predicted.
2. Find the $k + 1$ nearest neighbors to x according to a distance function $d(x, x_i)$.
3. The $(k + 1)$th neighbor is used for standardization of the k smallest distances via
$$D_{(i)} = D(x, x_{(i)}) = \frac{d(x, x_{(i)})}{d(x, x_{(k+1)})} \quad .$$
4. Transform the normalized distances $D_{(i)}$ with any kernel function $K(.)$ into weights $w_{(i)} = K(D_{(i)})$.
5. As prediction for the class membership y of observation x choose the class, which shows a weighted majority of the k nearest neighbors
$$\hat{y} = \max_r \left(\sum_{i=1}^{k} w_{(i)} I(y_{(i)} = r) \right) \quad .$$

Figure 2: wkNN classifier.

For test sample 1

D1 = |3-2| + |1-3| = 1 + 2 = 3; W1 = 3 / 10 = 0.3, K(x,x(1)) = 1 / 0.3 = 3.34
D2 = |3-1| + |1-5| = 2 + 4 = 6, W2 = 6/10 = 0.6, K(x,x(2)) = 1/0.6 = 1.67
D3 = |3-4| + |1-2| = 1 + 1 = 2, W3 = 0.2, K(x,x(3)) = 1/0.2 = 5
D4 = |3-2| + |1-5| = 1 + 4 = 5, W4 = 0.5, K(x,x(4)) = 1/0.5 = 2
D5 = |3-6| + |1-8| = 3 + 7 = 10, W5 = 1,

For 0; 3.34 + 2 = 5.34 and For 1; 1.67+5 = 6.67 thus belongs to 1

For test sample 2

D1 = |2-2| + |2-3| = 0 + 1 = 1; W1 = 1 / 10 = 0.1, K(x,x(1)) = 1 / 0.1 = 10
D2 = |2-1| + |2-5| = 1 + 3 = 4, W2 = 4/10 = 0.6, K(x,x(2)) = 1/0.4 = 2.5
D3 = |2-4| + |2-2| = 2 + 0 = 2, W3 = 0.2, K(x,x(3)) = 1/0.2 = 5
D4 = |2-2| + |2-5| = 0 + 3 = 3, W4 = 0.3, K(x,x(4)) = 1/0.3 = 3.33
D5 = |2-6| + |2-8| = 3 + 7 = 10, W5 = 1,

For 0; 10 + 3.33 = 13.33 and For 1; 2.5+5 = 7.5 thus belongs to 0

BEST OF LUCK!