**IS709 Introduction to Data Science**

**Final Exam**

**18January 2020**

Note that you can use Python, Excel or R environment if you need to make any calculations. You can copy and paste the figures from these environments. But, essentially, you do not necessarily run codes or use an environment to be able to answer the questions. Read the questions carefully.

You are not allowed to work together to answer the questions. You can use your slides and notes.

You can use this document to write down your answers. You can also prefer to print these pages first and then write manually. But ensure to scan the pages carefully. The final document should be readable. I suggest you to use a professional application such as Office Lens for scanning.

**Question 1 (25 pts):**

The dataset includes the following attributes:

Feature A: Categorical attribute which takes either a1,a2,or a3 values

Feature B: Ordinal attribute which takes either small, medium or large values

Feature C: Continuous attribute which comes from a Gaussian distribution

Feature D: A sequential unique key between 1 and 300

The descriptive statistics are as follows:

|  |  |  |
| --- | --- | --- |
| **Feature Name** |  | **Number of points** |
| **Feature A** | **a1** | 100 |
|  | **a2** | 100 |
|  | **a3** | 100 |
| **Feature B** | **Small** | 50 |
|  | **Medium** | 100 |
|  | **Large** | 150 |

|  |  |  |
| --- | --- | --- |
| **Feature Name** | **Mean and Std** | **Number of points** |
| **Feature C** | 20±3 | 300 |

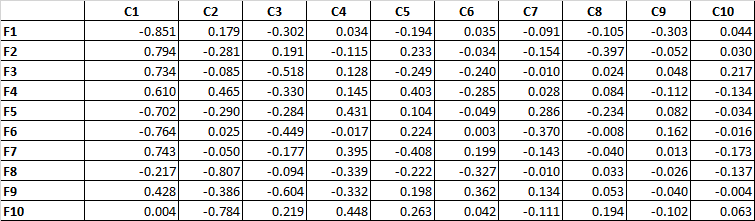
Assume that Feature C values are rewritten as 10 explicitly when A=a1 and B=medium (these values were replaced with 10 for some reasons and you downloaded the values with these replaced values. But you do not know this fact).

1. Which of the data quality problem refers to this situation?
2. How can you identify this problem in your database? Explain in detail and give justifications.
3. Provide a solution as a remedy of this quality problem which will result in losing no data at the end.

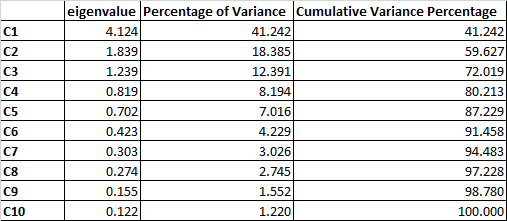
Hint: For instance, A feature may refer to a person’s identifier. Feature D might be the code of a garment. Feature C might be the price and B could be the size of the garment. Feature D is an automatically generated ID for the garment.

**Question 2 (25 pts):**

**Table 1:** It shows the loadings of the ten features in the dataset to the principal components.



**Table 2:** It shows the eigenvalues and percentage of variance values.



Answer the following questions according to Table1 and Table2.

1. Select two features which are very similar to each other as examples. For instance, are F1 and F2 similar or F3 and F5? Give your reasons.
2. How many principal components would you choose? Give your reasons.
3. Two experiments were carried out as follows: (i) 10 input features (F1, …, F10) that are used for classification with kNN where k=3. (ii) 2 input features (F1, F6) that are used for classification with kNN where k=3.

Discuss how each model is expected to perform in each case while focusing on the features and preprocessing issues.

**Question 3 (25 pts):**

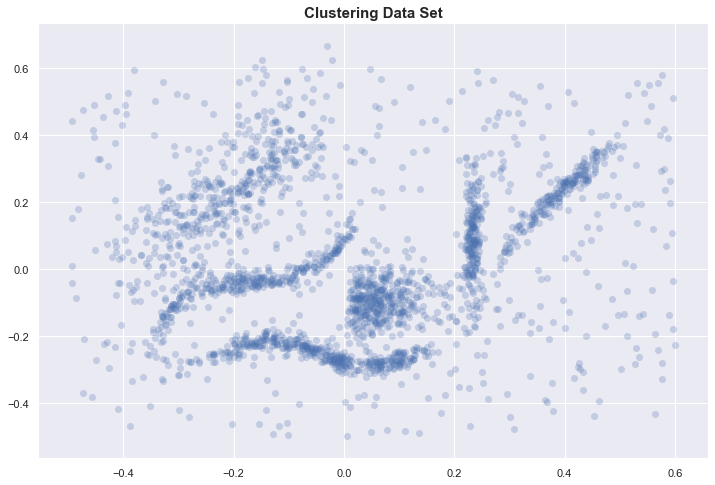
Consider the following excerpt taken from a study:

*“The difference between being raised in a bookless home compared to being raised in a home with a 500-book library has as great an effect on the level of education a child will attain as having parents who are barely literate (three years of education) compared to having parents who have a university education (15 or 16 years of education). . . . Both factors, having a 500-book library or having university-educated parents, propel a child 3.2 years further in education, on average.” From “Books in home as important as parents’ education in determining children’s education level”, Science Daily, May 21, 2010.”*

Assume that the features used in this study, which are the number of books (numBooks) and the number of years in education (numYears) features do not come from a normal distribution.

1. How will you pose a research question and which statistical test/s would you use? Give your reasons. Explain and show them in detail.
2. You are asked to group the numBooks variable and numYears into three categories. How would you do that? Then, how would you pose research questions based on these categories? In this case which statistical tests will you use? Give your reasons. Explain them in detail.

**Question 4 (10 pts):**



A dataset comprising two features is shown on the figure. Consider this figure while answering the questions:

You are going to apply a clustering algorithm on the dataset. Can you use “Silhouette score” for cluster number selection? Give your reasons. How well do you expect this measure to work on this dataset?

**Question 5 (15 pts):**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | **Ground Truth** | | | | |
|  |  | **A** | **B** | **C** | **D** | **E** |
| **Predicted** | **A** | 8144 | 3967 | 2600 | 1827 | 754 |
| **B** | 334 | 1056 | 1118 | 1539 | 4314 |
| **C** | 200 | 100 | 500 | 10 | 20 |
| **D** | 100 | 200 | 30 | 7000 | 30 |
| **E** | 10 | 3 | 20 | 10 | 50 |

The above table shows the confusion matrix of a classifier. Which error metrics would you prefer to use and why? Give one example metric which is not ideal for this case. Give your reasons based on the results of these metrics.