**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.

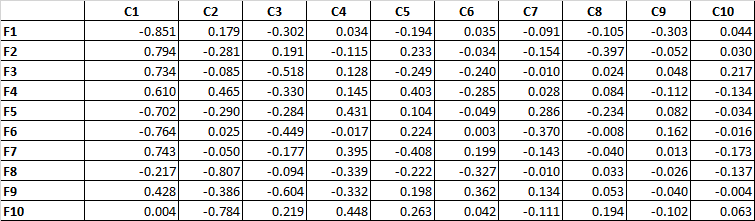
Hint: For instance, A feature may refer to a person’s identifier. Feature D might be the code of a garment. Feature C might be the price and B could be the size of the garment. Feature D is an automatically generated ID for the garment.

**Answer 1**

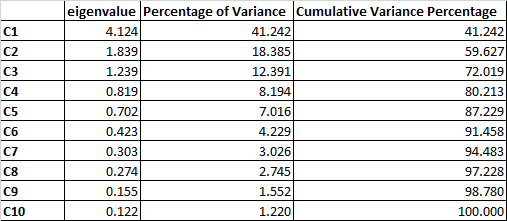
1. Disguise missing data since it was replaced with 10 with any medium sized garment by the same person. (5 pts)
2. We can identify the problem using QQplots. For instance, we plot feature C vs theoretical quantiles (Gaussian) then we can see consistent values on the plot (I mentioned that C comes from Gaussian distribution) (10 pts).
3. We treat the problem as missing data problem. We can look for relationships between Feature C with the rest of the attributes and try to predict them with regression techniques or regression trees etc. You can also use maximum likelihood, mean-median imputation methods as well. But we can’t use complete case analysis since we aim to lose no data at the end (10 pts)

**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.

**Answer 2 (25 pts):**

Based on the Table 2 results, we can select three components since the eigenvalues of the components drop significantly after the third component. Besides, since there are 10 components in total, each component is expected to be greater than 1/10 so C4 has a percentage of variance having 8.194 which is less than this number. According to Kaiser criterion, the eigenvalue greater than 1 is suggested. Hence, we select three components. **(10 pts)**

Based on this selection, we can answer the remaining questions.

Based on C1, C2 and C3, we can look for the similar loadings among the features. For instance, F1 and F6 are very similar to each other. However, F1 and F2 are not the same. F3 and F8 are totally unrelated to each other. We look at the direction and the magnitude of the values of the loadings. **(5 pts)**

Knn makes use of Euclidean distance hence we need to select features which are not correlated with each other. Otherwise, you give more importance (weight) to those variables inherently. Hence, it is important to include significant but not correlated variables.

F1 and F6 are very similar to each other. Hence, we give more emphasis to this variable and ignore the other remaining important features. Consequently, this will not be a good model. For instance, F1 and F6 were loaded to C1 and C3 significantly but not to C2. To C2, F8 and F10 were loaded into C2 significantly. **(5 pts)**

All the input features can be used. However, there are correlated features among them. Again, we give more weigh on these particular variables. But at the end we have all the relevant features for the analysis. **(5 pts)**

You should not focus on k=3 part and computational cost issues as they are out of scope in this question. I asked specifically pre-processing issues. I expected you to mention the consequences of correlated variables in modeling.

The dataset and analysis results belong to: decathlon dataset

See <http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/118-principal-component-analysis-in-r-prcomp-vs-princomp/> (section Quantitative variables) You can run the code and see the results.

**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. (13 pts)
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. (12 pts)

**Answer 3 (25 pts):**

1. We can simply the problem as follows:

H0= There is no difference between the children’s level of education between raised in a bookless home compared to being raised in a home with a 500-book library.

Ha= There is a significant difference between the two groups.

Since they do not come from a normal distribution, we have to use nonparametric test. In addition, it should be unpaired. We can use Mann Whitney-U test since there are two groups. It tests whether samples originate from the same distribution.

We can use the same test for investigating the differences between children who have parents with university education and not.

Some of you suggested Kruskas Wallis between these four groups. It is ok if you assume the points are independent in all the groups.

Another alternative research question could be where the null hypothesis is the mean of children’s level of education is same in both groups, and in the alternative hypothesis, their means are different.

You could also check µ1-µ2≥3.2 as null hypothesis. These were posed by you and the tests were chosen accordingly so I gave full marks.

If the distributions came from normal distribution, we could test whether the mean difference is significant (3.2 years). However, I specifically mentioned that they do not come from a normal distribution in the question.

1. Assume that we clustered both numBooks and numYears and obtained three clusters for each. Each child will belong to one of the clusters. For instance, one solution could be:

Group A: L, L, M, M, M, H

Group B: M, M, M, H, H, H

L: Low, M: Medium, H: High education

Group A: Children raised in a bookless home

Group B: Children raised in a home with a 500-book library.

We could calculate the observed and expected frequencies for each category (L, M, H) in each group and then check their significance with chi-square analysis.

Some of you discretized the values in an ordinal way such as: <350, 350<= to <650 and 650=<. Then based on these groupings, Kruskal Wallis was suggested. I gave full marks for these answers .(However, you are required to show the groupings as ordinal clearly (– 6 pts if you failed to show evidences and did not explain how you could discretize/cluster)).

(-3 pts): If you didn’t pose your research question correctly. Some of you posed a null hypothesis which is ambiguous or includes statements such as “There is an effect…”, “There is a dependency..”.

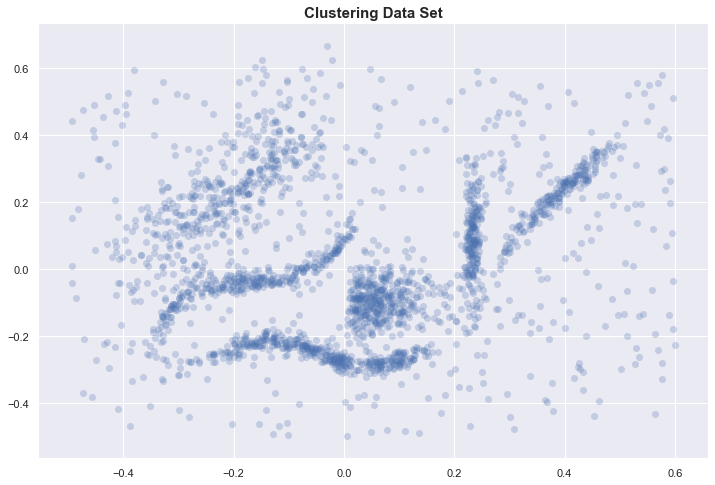
To be more specific, “The null hypothesis is generally denoted as H0. It states the exact opposite of what an investigator or an experimenter predicts or expects. It basically defines the statement which states that there is no exact or actual relationship between the variables.

The alternative hypothesis is generally denoted as H1. It makes a statement that suggests or advises a potential result or an outcome that an investigator or the researcher may expect. It has been categorized into two categories: directional alternative hypothesis and non directional alternative hypothesis”

Ref: <https://www.statisticssolutions.com/null-hypothesis-and-alternative-hypothesis/>

* You cannot use ANOVA since it assumes that the data is normally distributed.

**Question 4 (10 pts):**

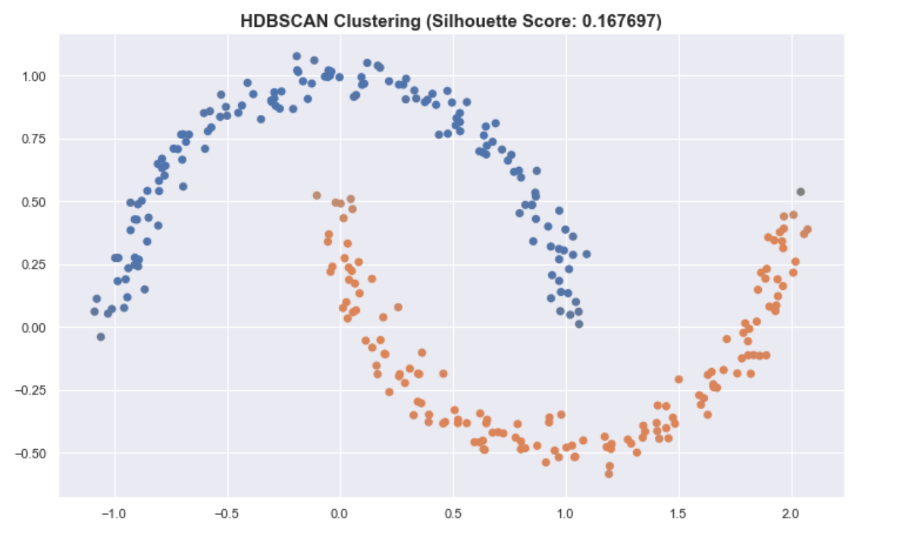


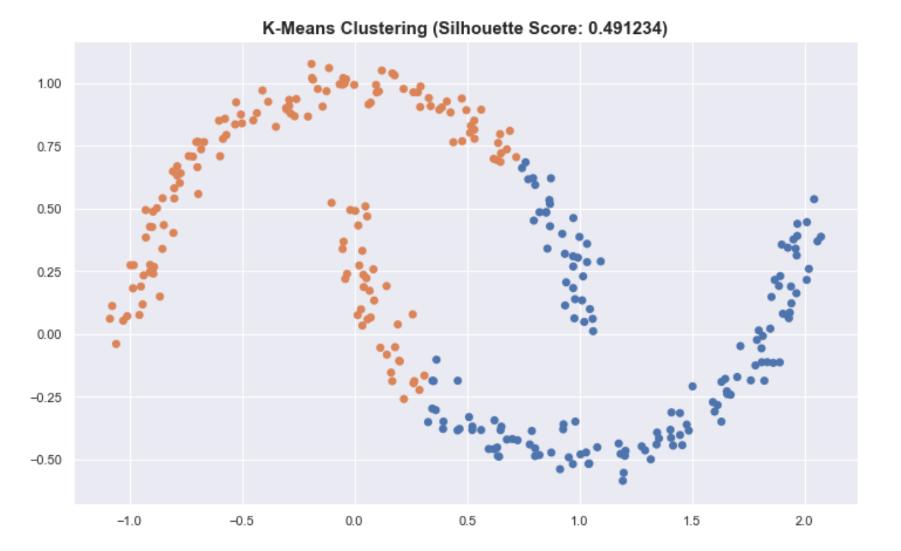
A dataset comprising two features is shown on the figure. Consider this figure while answering the questions:

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?

**Answer 4**

If the clusters are in arbitrary shape, silhouette will result in misleading values. Some metrics, such as the *silhouette score,*work best when the clusters are round. For instance, K-means has a better silhouette score than the result of HDBSCAN even though we see that the clusters in HDBSCAN are better.





This is because of the distance calculations of each point in the clusters.

I expect you to mention the relationship between k-means and silhouette analysis in the exam while elaborating more on the shape (non-convex shape’s problems) on the clusters.

Ref: https://www.kdnuggets.com/2020/02/understanding-density-based-clustering.html

**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.

**Answer 5:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **Ground Truth** | | | | |  |  |  |  |  |  |  |  |  |
|  |  | **A** | **B** | **C** | **D** | **E** | **Total** | **TP** | **TN** | **FN** | **FP** | **Recall** | **Precision** | **F1** |  |
| **Predicted** | **A** | 8144 | 3967 | 2600 | 1827 | 754 | 17292 | 8144 | 16000 | 644 | 9148 | 0.93 | 0.47 | 0.62 |  |
| **B** | 334 | 1056 | 1118 | 1539 | 4314 | 8361 | 1056 | 21305 | 4270 | 7305 | 0.20 | 0.13 | 0.15 |  |
| **C** | 200 | 100 | 500 | 10 | 20 | 830 | 500 | 29338 | 3768 | 330 | 0.12 | 0.60 | 0.20 |  |
| **D** | 100 | 200 | 30 | 7000 | 30 | 7360 | 7000 | 23190 | 3386 | 360 | 0.67 | 0.95 | 0.79 |  |
| **E** | 10 | 3 | 20 | 10 | 50 | 93 | 50 | 28725 | 5118 | 43 | 0.01 | 0.54 | 0.02 |  |
|  | **Total** | 8788 | 5326 | 4268 | 10386 | 5168 | 33936 | 16750 | 118558 | 17186 | 17186 |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  | **Weighted-F1:** | | 0.45 |  |  |  |
|  |  |  |  |  |  |  |  |  |  | **Macro-F1:** | | 0.36 |  |  |  |
|  |  |  |  |  |  |  |  |  |  | **Micro-F1:** |  | 0.49 |  |  |  |
|  |  |  |  |  |  |  |  |  |  | **Macro-precision:** | | 0.54 |  |  |  |
|  |  |  |  |  |  |  |  |  |  | **Macro-recall:** | | 0.39 |  |  |  |
|  |  |  |  |  |  |  |  |  |  | **Micro-precision:** | | 0.49 |  |  |  |
|  |  |  |  |  |  |  |  |  |  | **Micro-recall:** | | 0.49 |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

We need to use F1-macro or F1-weighted metrics (-4 if you didn’t mention any of multiclass comparison metrics). Since it is unbalanced, it is not recommended to use F1-micro since it is corresponding to accuracy and affected negatively from unbalanced classes. I expect you to mention about unbalanced datasets in your response (8 pts). Also see above that F1-micro provided quite good values compared to the other metrics.

Some of you calculated precision and recall for each class. This is ok however, when you compare different classifiers, it might be hard to go between different precision and recall values for each class.

Some of you suggested MCC however it is designed for binary classification. You can calculate it for each class. However, later, you need to combine them similar to F1-weighed score.

I would expect you to state one of the combined metrics in your response. If you stated accuracy, you were expected to discuss its limitations properly.

If you calculate one of the suggested metrics, you get 7 pts.

Recall from our lecture notes: “Macro-averaging gives equal weight to each class, whereas micro-averaging gives equal weight to each per-document classification decision. Because the F1 measure ignores true negatives and its magnitude is mostly determined by the number of true positives, large classes dominate small classes in micro-averaging.

Micro-averaged results are therefore really a measure of effectiveness on the large classes in a test collection. To get a sense of effectiveness on small classes, you should compute macro-averaged results

micro-F1 = micro-precision = micro-recall=accuracy.”