**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.

**Answers:**

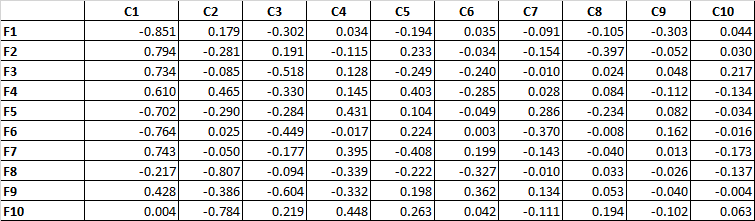
1. There is an accuracy problem because the new data contains error and this error may not be easily identified.
2. When C=10 with some conditions explained above, normal distribution of C can be violated after changes since number of values when A=a1 is 100 and number of vales when B=medium is 100. When A=a1 and B=medium, number of their combinations will be high, and it distorts normal distribution of C. Also, C’s mean is 20 and standard deviation is 3. That means when the number of 10 increase in dataset, it influences C’s normality distribution badly. When we apply normality check of C, we can identify the problem.

Moreover, Plotting C's values with respect to A, the value of C will always show 10 when A = a1. The same situation will be seen when the B graph is plotted with respect to C, when B = medium. Thanks to graphical representation, we can identify this problem.

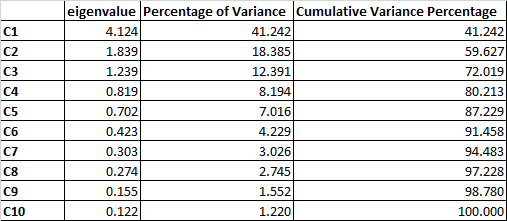
1. After realizing the problem, we see disguised data and treat them as missing values. Then we fill the missing data with the one of the appropriate simple imputation methods like mean imputation because C is distributed normally. Thus, we get rid of erroneous data without losing data.

**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.

**Answers:**

1. We have determined the two variables according to the similar values they received while loading the components, so F1 and F6 can be selected as very similar features since their values related to loading components are similar such that the two most important components are C1 and C2. F2 and F3 is loading these components similarly regarding loading signs (negative and positive) Also, remaining components are loaded by F2 and F3 similarly.
2. We can consider the proportion of variance accounted for. In the original space representation, each of the 10 features contains 10% (1/10) of the total variance explained by all the features. The principal component retained should then explain at least 10% of the total variance. As a result, the application of this rule allows the selection of only the first 3 components that are C1, C2 and C3 according to percentage of variance in Table 2. Also, we can consider the number of components with Kaiser criterion (The eigenvalue-one criterion): You retain and interpret any component with an eigenvalue greater than 1.00. As a result, As a result, the application of this rule allows the selection of only the first 3 components that are C1, C2 and C3 according to eigen value in Table 2.
3. Since PCA can be applied to numerical variables, we can say that variables are numeric. Since kNN is distance-based algorithm, we first normalize values of attributes in both case (i) and (ii). When we consider case (i), all features will exist kNN. The performance of the model will be positively affected since important variables are included in the variable set in case(i). However, some of variables are not significant regarding values related to eigen values that leads to decrease in kNN performance.

When we consider case(ii), F1 and F6 have chosen as input features. These variables are similar to each other since the signs and values they received while loading the components are similar. Therefore, we can say that these variables contain similar information about dataset. When we construct classification model with F1 and F6, there might be multicollinearity problem that harms good classification and some of the information contained in the data set is not used in classification.

**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.

**Answers:**

1. Our assumption is follows: numBooks: number of book in library at home, numYears: number of education years of parents. target variable: number of education years of children. We have two different groups that are 500-book library group and university education group since we have investigated is there any difference between books in home and well-educated parents regarding education level of children.

Research question: Comparison of the education years(level) of children with different conditions, taking into account the number of books in the library at home and the education levels of the parents.

Since we have two different groups (two samples-unpaired) with continuous endpoint and we have investigated education years of children in two different groups. In addition to that, non-parametric statistical test should be used due to normality assumption. Therefore, Mann Whitney U Test can be used for hypothesis testing. Also, we will check whether there is a difference between education years of children in two different groups.

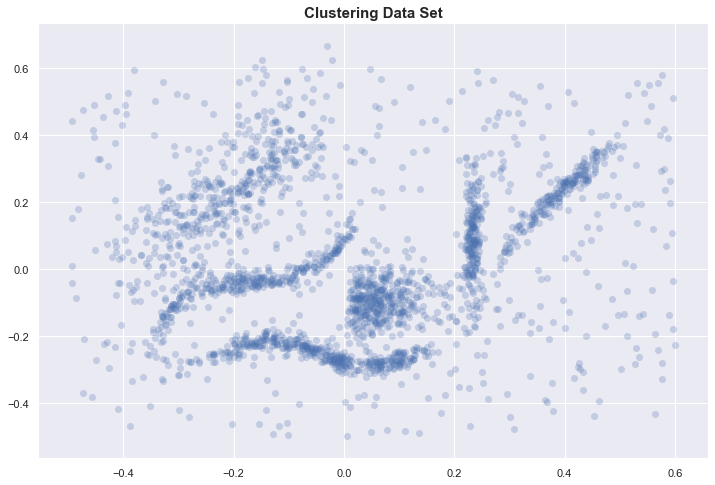
H0: 𝜇1=𝜇2 (the means of education years(level) of children are equal regarding two different groups)

H1: 𝜇1≠𝜇2 (the means of education years(level) of children are not equal regarding different groups)

1. One of the discretized methods (binning, Chi-merge etc) can be used to categorization. Categorized education years comparison can be made for each of the categories regarding research questions such that new research question might be like this: Considering two different conditions regarding newly categorized features, do the education years of the children fall into different categories?

Now, we should do unpaired group comparison of >2 categorical outcomes, so Kruskal-Wallis statistical test can be applied.

**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:**

Points on the graph have different densities and clear density does not appear and points are distributed in wide range of areas. Silhouette score can use for cluster number selection since it takes into account how close each point in one cluster is to points in the neighboring clusters. That means it considers both within cluster and between clusters. Also, it evaluates the goodness of a clustering by considering how well the clusters are separated and how compact the clusters are. The silhouette score depends on which algorithm(k-means, DBSCAN etc) will be used with which parameters(number of k, eps and minpoints). Since the points does not separated well and most of the points does not compact in graph, we expect that this Silhouette score does not quite well in this dataset due to algorithm performance. Most probably, Silhouette score will be less than 0.50 silhouette average.

**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.

**Answers:**

The dataset can be consider as unbalanced since percent of ground truth is a bit far away with each other. That means these are away from uniform distribution. These are 0.25, 0.15,0.12,0.30 and 0.15. Total number of values are 33936. There are huge differences between number and percent of ground truth and predicted ones. For example, while number of ground truth A and percent of ground truth A is 8788 and 0.26 respectively, number of predicted A and percent of predicted A is 17292 and 0.50 respectively. There are huge differences between ground truth A and predicted A. High number of false positive leads to that problem. To detect this issue, precision can be used since it includes false positive (FP) in its calculations. Since we have more than two categories and unbalance data, we can use macro-average precision and it is 0.54.

Micro-averaged results are 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. Therefore, micro average precision is not ideal for this case. Its value is 0.49.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **Ground Truth** | | | | |  |  |
| **Predicted** |  | **A** | **B** | **C** | **D** | **E** | **Total** | **Percent** |
| **A** | 8144 | 3967 | 2600 | 1827 | 754 | 17292 | 0,509547 |
| **B** | 334 | 1056 | 1118 | 1539 | 4314 | 8361 | 0,246376 |
| **C** | 200 | 100 | 500 | 10 | 20 | 830 | 0,024458 |
| **D** | 100 | 200 | 30 | 7000 | 30 | 7360 | 0,216879 |
|  | **E** | 10 | 3 | 20 | 10 | 50 | 93 | 0,00274 |
|  | **Total** | 8788 | 5326 | 4268 | 10386 | 5168 |  |  |
|  | **Percent** | 0,258958 | 0,156942 | 0,125766 | 0,306047 | 0,152287 |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP** | **TN** | **FN** | **FP** | **Precision** | **Macro-average precision** | **Micro-average precision** |
| 8144 | 16000 | 644 | 9148 | 0,470969 | 0,537680184 | 0,493576143 |
| 1056 | 21305 | 4270 | 7305 | 0,126301 |  |  |
| 500 | 29338 | 3768 | 330 | 0,60241 |  |  |
| 7000 | 23190 | 3386 | 360 | 0,951087 |  |  |
| 50 | 28725 | 5118 | 43 | 0,537634 |  |  |
| 16750 | 118558 | 17186 | 17186 |  |  |  |
|  |  |  |  |  |  |  |