**Implementing the k-means & k-medians clustering algorithms**

**Data Mining & Visualisation CA2**

Panashe Mundondo

STUDENT NUMBER: 201542826

# **INTRODUCTION**

The purpose of this assignment was to implement the k-means and k-medians clustering algorithms in python. This would subsequently allow us to compare the results we obtained from both algorithms on both normalised and unnormalised data. The purpose of both algorithms is to group the input data into different clusters based on their levels of similarity allowing further inferences to be made.

**Question 1**

Refer to the python file.

**Question 2**

Refer to the python file

**Question 3**

Chart

Description automatically generated

FIG 1. K-Means graph of precision, recall, f-score when the data is unnormalised

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **K=1** | **K=2** | **K=3** | **K=4** | **K=5** | **K=6** | **K=7** | **K=8** | **K=9** |
| **Precision** | 32.8 | 66.23 | 81.39 | 86.38 | 91.97 | 92.49 | 92.69 | 92.55 | 92.55 |
| **Recall** | 100.00 | 100.00 | 98.81 | 87.42 | 67.86 | 62.38 | 54.10 | 47.53 | 42.27 |
| **F-Score** | 47.23 | 74.68 | 87.36 | 86.38 | 75.67 | 70.11 | 62.25 | 60.76 | 54.98 |

Table 1. B-Cubed measures for unnormalized K-Means

**Question 4**

Chart

Description automatically generated

FIG 2. K-means graph of precision, recall, f-score when the data is normalised

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **K=1** | **K=2** | **K=3** | **K=4** | **K=5** | **K=6** | **K=7** | **K=8** | **K=9** |
| **Precision** | 32.87 | 66.23 | 81.39 | 86.36 | 91.98 | 92.49 | 94.00 | 92.55 | 92.55 |
| **Recall** | 100.00 | 100.00 | 98.81 | 87.43 | 67.86 | 62.38 | 55.87 | 47.53 | 42.27 |
| **F-Score** | 47.23 | 74.68 | 87.36 | 86.37 | 75.67 | 70.11 | 66.59 | 60.76 | 54.98 |

Table 2. BCubed measures for normalized K-Means

**Question 5**

Chart

Description automatically generated

FIG 3. K-Medians graph of precision, recall, f-score when the data is unnormalised

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **K=1** | **K=2** | **K=3** | **K=4** | **K=5** | **K=6** | **K=7** | **K=8** | **K=9** |
| **Precision** | 32.87 | 66.23 | 79.32 | 92.43 | 87.69 | 87.17 | 92.21 | 91.63 | 92.15 |
| **Recall** | 100.00 | 100.00 | 95.99 | 92.51 | 64.45 | 60.72 | 56.18 | 49.30 | 43.54 |
| **F-Score** | 47.23 | 74.68 | 85.26 | 92.42 | 71.40 | 67.30 | 67.21 | 62.39 | 56.44 |

Table 3. BCubed measures for unnormalized K-Medians

**Question 6**

Chart, line chart

Description automatically generated

FIG 4. K-Medians graph of precision, recall, f-score when the data is normalised

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **K=1** | **K=2** | **K=3** | **K=4** | **K=5** | **K=6** | **K=7** | **K=8** | **K=9** |
| **Precision** | 32.87 | 66.23 | 79.32 | 86.26 | 86.80 | 87.17 | 92.21 | 91.63 | 92.15 |
| **Recall** | 100.00 | 100.00 | 95.99 | 87.57 | 63.95 | 60.72 | 56.17 | 49.30 | 43.47 |
| **F-Score** | 47.23 | 74.68 | 85.26 | 86.37 | 70.63 | 67.30 | 67.21 | 62.39 | 56.37 |

Table 4. BCubed measures for normalized K-Medians

**Question 7**

To assess which of the algorithms produced the best result, we first need to know what precision, recall, and F-score are. Precision is simply the number of correctly classified classes in comparison to the total classifications of that case. Precision is more important for example when deciding who to show a targeted advert to, as one would want to be sure that the person the advert is shown to would be interested in it, and false positives would be a wasted cost. To clarify, this is:

Precision: True Positives (TP)/ TP + False Positives (FP).

Recall, on the other hand, is the amount of total true positives that were correctly classified, regardless of the number of false negatives. Recall is more important when the consequences of a false negative are graver than that of a false negative. An easy example is cancer diagnosis, it is better to detect all true positives and mistakenly misdiagnose a few patients as having cancers, rather than never misdiagnosing a patient as having cancer, but then not detecting a few patients who actually had cancer. The formula is:

Recall: TP/(TP+FN)

F-Score considers both measures and finds their harmonic means and as a result, uses a balance of both precision and recall. The formula for the F-Score is:

F-Score = (2\*Precision\*Recall)/ (Precision + Recall)

When running my algorithms, each iteration resulted in differing graphs. So for the sake of reproducibility, the algorithms were initiated with a seen of ten (10) resulting in the graphs in fig 1., fig 2., fig 3, and fig 4. Generally, all four graphs followed a similar trend. Recall would always begin at 1 at k = 1, as the cluster would be assigned the truth label of countries as it had the highest count. All 161 countries would be in the cluster resulting in a recall of 100. The precision would also always be low as all data points being in the same cluster means there will naturally be a lot of false positives.

For all four algorithms, the B-Cubed measures tended to converge at k = 4 which makes sense as there were four true categories within our data. As a result, there would be the optimal intersection of precision (as there are enough clusters for each true category to have its own) and recall falls from 1 as there begins to be a few miscalculations, however, it is still fairly high. Table 5 (below) shows the optimal measures were found at K = 4 using the K medians algorithm on unnormalized data with the measures of precision, recall, and F-Score all being above 90%. The reason k-medians may have provided a better score is the K-means algorithms are sensitive to outliers, so a single outlier can completely skew the mean and the centroid as a result leading to worse clustering. Using the median is more robust against outliers and as a result may provide better centroids throughout the iterations.

|  |  |  |  |
| --- | --- | --- | --- |
| **At K = 4** | **Precision** | **Recall** | **F-Score** |
| **K-means Unnormalised** | 86.38 | 87.42 | 86.38 |
| **K-means Normalised** | 86.36 | 87.43 | 86.37 |
| **K-medians Unnormalised** | 92.43 | 92.51 | 92.42 |
| **k-medians Normalised** | 86.26 | 87.57 | 85.37 |

Table 5. B-Cubed measures of the four algorithms at K = 4.

At most other K values, the differences in the B-cubes measures seemed more arbitrary with the K-means algorithms scoring slightly higher than the K-medians algorithms at most intervals for K except k = 4.