Introduction to Python Programming

DSCI-D 590 (Topics in Data Science)

Final Project

Arin Esterbrook

Project Statement: This project involves using the k-means clustering algorithm to analyze the Wisconsin Breast Cancer data set to classify patients as either class 2 (benign) or class 4 (malignant).

Phase 1: This phase involves importing the dataset into the program as a data frame. This phase also handles values in the dataset listed as ‘?’ and replacing the values with the mean value of the entire column. The mean, median, standard deviation, and variance are calculated for each of the attributes. The value frequency for each attribute was modeled using histograms. The final output of the program is a table listing the summary statistics for each of the attributes. A total of nine histograms were generated to describe the value frequency of each attribute.

Phase 2: This phase includes three processes for implementing the k-means algorithm. The first step initializes the centroids randomly. The next calculates the distance from the centroids for each of the 699 data points and assigns each data point to either the ‘2’ or ‘4’ cluster in the newly generated ‘Predicted\_Class’ column depending on the shortest distance. Once the data points have been assigned to either cluster, the means for each cluster are calculated. The centroids are then updated depending on the cluster means. This process is iterated until the centroids not longer change or there were 50 iterations performed. The final output of phase 2 is comprised of the randomly selected initial centroids, the final centroids after recompilation, and final cluster alignment.

Phase 3: This phase analyzes the quality of the k-means clustering algorithm as performed in Phase 2. The program initially calculates the total error by analyzing the discrepancies between the ‘Class’ and ‘Predicted\_Class’ as computed in Phase 2. If the error is greater than 50% the program will then swap the values in the ‘Predicted\_Class’ column. The program then summarizes the number of data points, provides an output for the ‘Predicted\_Class’ error data points, and calculates the error rate for both class 2 and class 4.

Phase 1 Results:

Attribute A2 -----------

Mean: 4.4

Median: 4.0

Variance: 7.9

Standard Deviation: 2.8

Attribute A3 -----------

Mean: 3.1

Median: 1.0

Variance: 9.3

Standard Deviation: 3.0

Attribute A4 -----------

Mean: 3.2

Median: 1.0

Variance: 8.8

Standard Deviation: 3.0

Attribute A5 -----------

Mean: 2.8

Median: 1.0

Variance: 8.2

Standard Deviation: 2.9

Attribute A6 -----------

Mean: 3.2

Median: 2.0

Variance: 4.9

Standard Deviation: 2.2

Attribute A7 -----------

Mean: 3.5

Median: 1.0

Variance: 13.0

Standard Deviation: 3.6

Attribute A8 -----------

Mean: 3.4

Median: 3.0

Variance: 5.9

Standard Deviation: 2.4

Attribute A9 -----------

Mean: 2.9

Median: 1.0

Variance: 9.3

Standard Deviation: 3.0

Attribute A10 -----------

Mean: 1.6

Median: 1.0

Variance: 2.9

Standard Deviation: 1.7

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Phase 2 Results:

Randomly selected row 362 for centroid mu\_2.

A2 3.0

A3 2.0

A4 2.0

A5 1.0

A6 4.0

A7 3.0

A8 2.0

A9 1.0

A10 1.0

Name: 362, dtype: float64

Randomly selected row 402 for centroid mu\_4.

A2 5.0

A3 3.0

A4 3.0

A5 1.0

A6 2.0

A7 1.0

A8 2.0

A9 1.0

A10 1.0

Name: 402, dtype: float64

Program ended after 8 iterations.

Final centroid for mu\_2:

A2 3.047210

A3 1.302575

A4 1.446352

A5 1.343348

A6 2.087983

A7 1.378755

A8 2.105150

A9 1.261803

A10 1.109442

Name: 0, dtype: float64

Final centroid for mu\_4:

A2 7.158798

A3 6.798283

A4 6.729614

A5 5.733906

A6 5.472103

A7 7.873391

A8 6.103004

A9 6.077253

A10 2.549356

Name: 1, dtype: float64

Final cluster assignment:

Scn Class Predicted\_Class

0 1000025 2 2

1 1002945 2 4

2 1015425 2 2

3 1016277 2 4

4 1017023 2 2

5 1017122 4 4

6 1018099 2 2

7 1018561 2 2

8 1033078 2 2

9 1033078 2 2

10 1035283 2 2

11 1036172 2 2

12 1041801 4 2

13 1043999 2 2

14 1044572 4 4

15 1047630 4 2

16 1048672 2 2

17 1049815 2 2

18 1050670 4 4

19 1050718 2 2

20 1054590 4 4

Phase 3 Results:

Output when Clusters are swapped:

Total errors 95.9 %

Clusters are swapped

Swapping Predicted\_Class

Number of Data points in Predicted Class 2: 465

Number of Data points in Predicted Class 4: 234

Error data points, Predicted\_Class 2:

Scn Class Predicted\_Class

12 1041801 4 2

15 1047630 4 2

23 1057013 4 2

50 1108370 4 2

51 1108449 4 2

57 1113038 4 2

59 1113906 4 2

63 1116132 4 2

65 1116998 4 2

101 1167439 4 2

103 1168359 4 2

105 1169049 4 2

222 1226012 4 2

273 428903 4 2

348 832226 4 2

356 859164 4 2

455 1246562 4 2

489 1084139 4 2

Error data points, Predicted\_Class 4:

Scn Class Predicted\_Class

1 1002945 2 4

3 1016277 2 4

40 1096800 2 4

196 1213375 2 4

252 1017023 2 4

259 242970 2 4

296 616240 2 4

315 704168 2 4

319 721482 2 4

352 846832 2 4

434 1293439 2 4

Number of all data points: 699

Number of error data points: 29

Error rate for class 2: 3.9 %

Error rate for class 4: 4.7 %

Total error rate: 4.1 %

Output when Clusters are not swapped:

Total errors 4.3 %

Number of Data points in Predicted Class 2: 466

Number of Data points in Predicted Class 4: 233

Error data points, Predicted\_Class 2:

Scn Class Predicted\_Class

12 1041801 4 2

15 1047630 4 2

23 1057013 4 2

25 1065726 4 2

50 1108370 4 2

51 1108449 4 2

57 1113038 4 2

59 1113906 4 2

63 1116132 4 2

65 1116998 4 2

101 1167439 4 2

103 1168359 4 2

105 1169049 4 2

222 1226012 4 2

273 428903 4 2

348 832226 4 2

356 859164 4 2

455 1246562 4 2

489 1084139 4 2

Error data points, Predicted\_Class 4:

Scn Class Predicted\_Class

1 1002945 2 4

3 1016277 2 4

40 1096800 2 4

196 1213375 2 4

252 1017023 2 4

259 242970 2 4

296 616240 2 4

315 704168 2 4

319 721482 2 4

352 846832 2 4

434 1293439 2 4

Number of all data points: 699

Number of error data points: 30

Error rate for class 2: 4.1 %

Error rate for class 4: 4.7 %

Total error rate: 4.3 %

Conclusion: The initial phase provides summary statistics and histograms for each attribute in the dataset. The mean range for the attributes was between 1.6 – 4.4, the median range was between 1.0 -4.0, the variance range was 4.9 – 13.0 and the standard deviation range was from 1.7 – 3.6. During Phase 2, the final centroids for class 2 and class 4 were calculated after eight iterations and produced a final cluster alignment with a predicted class attribute. The error rate of ‘Class’ compared to ‘Predicted\_Class’ was calculated during phase 3. Overall, the k-means algorithm produces an error rate of approximately 4.1% - 4.3% when comparing the correct ‘Class’ interpretation versus the ‘Predicted\_Class’ interpretation. The calculated error rate describes how well the final ‘Predicted\_Class’ centroids represent the correct ‘Class’ clusters.