Assignment 7: Clustering Seeds

Sophia J Quinton

Grand Canyon University

DSC-540: Machine Learning

Dr. Aiman Darwiche

15 December 2021

Assignment 7: Clustering

Questions

1. What metrics help predict the wheat type?
2. What accuracy does the algorithm have and is there any way to increase the model accuracy?
3. What are some applications of this information?

The data was derived from the uci database of datasets. The data given contains metrics derived from x-ray images. In an agricultural setting what would be the benefit of quickly sorting grains. Question 1: *What metrics help predict the wheat type?* It is important to ask this question because the data gives several continuous variables that can help describe the kernels. The variables that are given include: area A, perimeter P, compactness $C = \frac{4A\pi}{P^2}$, length of kernel, width of kernel, asymmetry coefficient, and length of kernel groove.

The question is used to help determine the success of the model and compare false positives. Question 2: *What accuracy does the algorithm have and is there a way to increase the model accuracy?* In the world of data science, the question of accuracy is important because it helps determine if the model should be used in production. In can also help a data scientist fine tune their algorithm.

The last question is important in terms of the greater applications of this data. Often times one needs to ask: I have these results, so what? The so what question is extremely important when presenting use cases. Question 3: *What are some applications of this information?* This question is beneficial because it is important to apply the data to a greater audience and to the real world.

K-means clustering

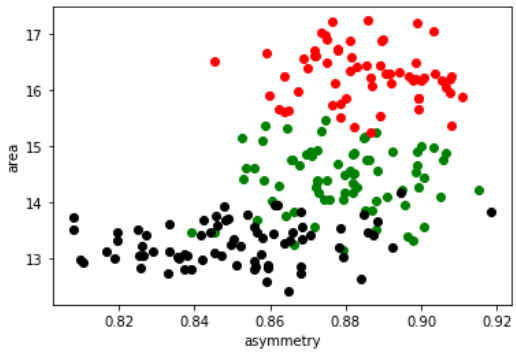
The data seven variables and a target variable for the type of kernel. It contains 210 samples in the dataset. Three clusters were used because there are a total of three different kernels in the dataset. The data was obtained from the UCI data repository (Charytanowicz et. al., 2010) (Dua & Graff, 2019). To determine the most successful two variable pairing, a nested four loop was used to run multiple k-means algorithms and calculate the accuracy scores. Because k-means is random when choosing where to start the algorithm, I conducted the analysis three times. From there I analyzed the data table to chose the combination of variables that produced the best result at least once. The asymmetrical coefficient with the area had an 89% rate. The data was plotted with matplotlib to show the different labels (AskPython, 2020) (Hunter, 2007). A fuzzy k-means was used to help see if that produced a better result. A fuzzy k-mean is useful because it allows points to exist is multiple groupings (Gopal, 2019). The fuzzy k-mean produced a similar result.

	first	second	third
0	0.152381	0.252381	0.252381
1	0.842857	0.342857	0.842857
2	0.342857	0.247619	0.004762
3	0.342857	0.247619	0.342857
4	0.347619	0.290476	0.347619
5	0.342857	0.247619	0.842857
6	0.000000	0.252381	0.252381
7	0.247619	0.366667	0.366667
8	0.371429	0.009524	0.852381
9	0.361905	0.252381	0.361905
10	0.423810	0.085714	0.847619
11	0.380952	0.238095	0.842857
12	0.004762	0.342857	0.842857
13	0.366667	0.366667	0.000000
14	0.414286	0.390476	0.414286
15	0.142857	0.395238	0.819048
16	0.552381	0.423810	0.390476
17	0.647619	0.647619	0.642857
18	0.342857	0.409524	0.342857
19	0.852381	0.852381	0.852381
20	0.414286	0.771429	0.414286
21	0.247619	0.014286	0.014286
22	0.195238	0.409524	0.252381
23	0.538095	0.742857	0.538095
24	0.247619	0.004762	0.247619
25	0.361905	0.252381	0.000000
26	0.142857	0.261905	0.819048
27	0.347619	0.247619	0.347619
28	0.252381	0.180952	0.409524
29	0.423810	0.033333	0.423810
30	0.023810	0.895238	0.347619
31	0.847619	0.423810	0.847619
32	0.385714	0.195238	0.261905
33	0.390476	0.395238	0.566667
34	0.252381	0.409524	0.409524

```
# A_Coef - A (AskPython, 2020) (Hunter, 2007)
Group = data[["A_Coef", "A"]]
Kmeans_model = KMeans(n_clusters=3)
label = Kmeans_model.fit_predict(Group)
accuracy = accuracy_score(data["target"], label)
print("The accuracy of LK-P: ", accuracy)

# (AskPython, 2020) (Hunter, 2007)
tlabel_0 = data[label == 0]
tlabel_1 = data[label == 1]
tlabel_2 = data[label == 2]
plt.scatter(tlabel_0["C"], tlabel_0["P"], color="green")
plt.scatter(tlabel_1["C"], tlabel_1["P"], color="red")
plt.scatter(tlabel_2["C"], tlabel_2["P"], color="black")
plt.xlabel("asymmetry")
plt.ylabel("area")
plt.show()
```

```
The accuracy of LK-P:  0.8952380952380953
```



```
: from sklearn.metrics import classification_report
  print(classification_report(np.array(data["target"]), label))
```

	precision	recall	f1-score	support
0	0.82	0.87	0.85	70
1	0.98	0.87	0.92	70
2	0.89	0.94	0.92	70
accuracy			0.90	210
macro avg	0.90	0.90	0.90	210
weighted avg	0.90	0.90	0.90	210

Analysis

From this analysis, the asymmetrical coefficient and the area produced the best result with an accuracy of 89%. Looking at the classification results, the precision for the three target kernels is high showing a majority of the classifications is true, and the recall is high showing that the results are relevant to this study (Larose & Larose, 2019). Going forward, I would use this to help classify in coming data bay making these center consistent. I would extract the centers and use those for clustering new information.

Question 1: *What metrics help predict the wheat type?* The metrics that most accurately determine wheat type based on the algorithm above is asymmetrical coefficient and area.

Question 2: *What accuracy does the algorithm have and is there a way to increase the model accuracy?* The alogirthm has an accuracy of 89%. Out of the multiple runs and combinations, this algorithm had the most accurate score. The precision and recall values also support the success of this algorithm. The fuzzy k-means algorithm produced a similar result but not a better result than the normal k-means algorithm.

Question 3: *What are some applications of this information?* This information would be used to classify different wheat seeds. The most straightforward application of this data would be to use this algorithm to build an automated wheat sorter after the harvest. This way, the farmers could easily distribute the wheat and sell it according to the species. Different wheat species are used for different food products (Wheat, 2016). It would automate a human-intensive process. This would help the farm increase productivity and sales because the focus could be moved

elsewhere. Another application is to help identify the wheat when these species cross-pollinate. Plants can easily be cross-pollinated, so being able to sort the seeds would be useful. This data could be applied to a greater set of data. There are many more examples of the application of this project, but these are a few that come to mind.

Ethics

The ethics of data management are important for privacy and for giving authors proper credit. Data repositories are a blessing for data scientists to explore techniques and learn new skills, but they should always be properly cited. The UCI Machine Learning Data Repository is a university-based data repository where the seed data was accessed from. It is a reputable source as the data is an archive of data managed by the university (Dua & Graff, 2019). In terms of privacy, it is important to hide data that reveals personal information. This is a common issue when working with HIPPA or FERPA related information. It is best practice to remove personal data such as names, addresses, and dates of birth.

If the data HIPPA related, then the storage of the data is important. There needs to be proper backups and clear data retention times (Barrangton, 2019). In general, it is best to have consistent data that is properly stored.

Overall, it is important to give authors credit when credit is due and protect private data.

References

AskPython. (2020, October 27). *How to plot K-means clusters with python?* AskPython. Retrieved December 12, 2021, from https://www.askpython.com/python/examples/plot-k-means-clusters-python.

Barrangton, J. (2019, October 1). *5 reasons why data management is important to any organization?* Medium. Retrieved December 16, 2021, from https://medium.com/technicalwords/https-medium-com-jessie-barrangton-5-reasons-why-data-management-is-important-to-any-organization-99b069f720b0

Charytanowicz, M., Niewczas, J., Kulczycki, P., Kowalski, P. A., Łukasik, S., & Żak, S. (2010). Complete gradient clustering algorithm for features analysis of X-ray images. *Advances in Intelligent and Soft Computing*, 15–24. https://doi.org/10.1007/978-3-642-13105-9_2

Dias, M. L. D. (2019). fuzzy-c-means: An implementation of Fuzzy C-means clustering algorithm. doi:10.5281/zenodo.3066222

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Gopal, M. (2019). *Applied machine learning*. McGraw-Hill Education.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. doi:10.1109/MCSE.2007.55

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Wheat. (2016, October 19). *New World Encyclopedia*, . Retrieved 15:04, December 15, 2021 from https://www.newworldencyclopedia.org/p/index.php?title=Wheat&oldid=1000768.

Github: https://github.com/squinton-gcu/Data-Science/tree/main/DSC-540/Assignment7