Sreeti Ravi

Final Project Phase 3

I590 Python Programming

1 December 2020

Analysis of K-means Clustering on Breast Cancer Data

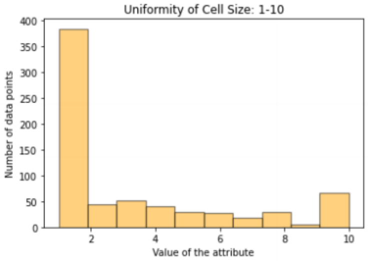
Introduction

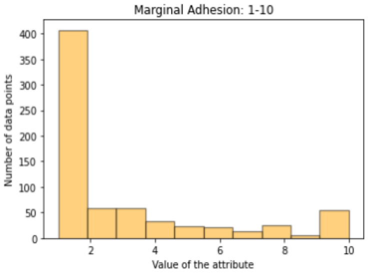
Breast cancer is the second most common cancer among women in the United States and affects around one in every eight women. When breast cancer is found early, there are more treatment options and a higher chance for survival because the cancer is more likely to be contained and not have spread elsewhere. According to the National Breast Cancer Foundation Inc., when it is detected early the five-year relative survival rate is 100%. In this project, we used the k-means algorithm on the Wisconsin Breast Cancer dataset to separate benign and malign cells into two different clusters and then calculated how accurate the algorithm was. The k-means algorithm uses centroids, which are averages and assigns them to a cluster based on a specified feature. This algorithm uses a random set of points as the initial points for the cluster and then performs the same calculation repetitively until the cluster assignments do not change. The results of this experiment showed that the algorithm has a good success rate.

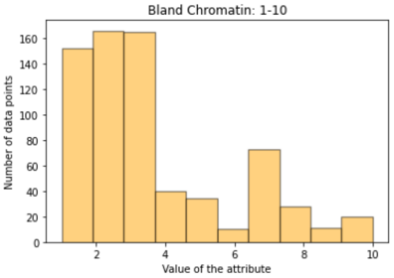
Data

This data and the work behind this data occurred because of Dr. William H. Wolberg, who wanted to be able to diagnose breast masses using Fine Needle Aspiration. Dr. Wolberg identified nine characteristics using a Fine Needle Aspiration from the breast mass. With the help of two of his students, Rudy Setiono and Kristin Bennett, he created a classifier based on the nine characteristics that successfully diagnosed 97% of cases. The dataset that was created from that process is the dataset that was used in this project and is more commonly known as the "Wisconsin Breast Cancer Data." This research has resulted in the creation of a software system called Xcyt, which Dr. Wolberg uses today for the diagnosis process. The process to collect data has changed since Dr. Wolberg first started his research. Today, the process begins with a fine needle aspiration from the breast mass, which is then viewed under a microscope slide and stained to highlight the cellular nuclei. Boundaries of each nucleus are drawn and Xcyt computes values for each of the ten characteristics for each nuclei. The mean, standard error and extreme values are also calculated and results in thirty nuclear features for each sample.

The data used for the k-means algorithm consisted of 699 datapoints and eleven columns: scn, a2, a3, a4, a5, a6, a7, a8, a8, a9, a10 and class. a2 through a10 represented the following features respectively: clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses, and class and the first column was for an ID number. All the features except class are identified from one to ten. Class is identified by a two for benign and four for malignant.







Visualizing the data by creating histograms showed that the data for most of the characteristics were right skewed, which tells us that the mean was greater than the median. This can be seen in the histograms above. Most of the data had a similar standard deviation of between two and three with exception of two characteristics that had a standard deviation under two. While not pictured above, the class histogram shows that there were twice as many benign cells as malign.

Methods (or procedure)

To construct the k-means algorithm, the Wisconsin Breast Cancer Data was loaded into a data frame using pandas. Column 7 contained missing values, which were indicated with a “?”. The missing values were replaced with the mean because outliers were not a large concern with this dataset because all the data contained values between one and ten. The first and last columns were removed while constructing the algorithm because they were not necessary for calculations. The first step to implement the k-means algorithm was ‘initialization’ and required initializing two means, μ_2 and μ_4. One random row was chosen for each initial mean and both were nine-dimensional vectors that contained data for columns a2 through a10. The next step was ‘assignment’. To do this, the Euclidian distance of every datapoint was calculated from each initial mean. The datapoint was assigned to cluster 2 if the distance was closer from μ_2 rather than μ_4 and cluster 4 otherwise. At the end of this step, all the datapoints were assigned to either cluster 2 or cluster 4. The next step ‘recalculation’ updated the initial means with the mean of each cluster. For example, the mean of all the cluster 2 datapoints was calculated and μ_2 was updated with that mean. The same occurred with cluster 4. This process was iterated until the datapoints did not move cluster assignments. After the k-means algorithm was finished, the result was two clusters and two lists of final means. To test the quality of the algorithm and its clustering, three error rates were calculated. One was the error rate of how often the algorithm predicted a datapoint as malign when it was benign. Another was the vice-versa; the datapoint was predicted to be benign but actually malign. The last error rate was total error rate.

Results

Using this k-means algorithm, there were 464 datapoints predicted to be benign and 235 predicted to be malign, while there were actually 458 benign and 241 malign. There were 17 samples that were malignant but classified as benign and 11 benign samples classified as malignant. The error rate for benign cells was 0.023 and for malign cells the error rate was 0.072. The total error rate was 0.04.

Discussion

The error rates mentioned above, 0.023 for benign cells and 0.072 for malign cells. These error rates are small enough for the algorithm to be considered for use in classifying cells as benign or malign. The error rate for benign cells is lower, so the algorithm seems to be better at identifying benign cells over malign cells. Because poor initialization was a concern, the initial means μ_2 and μ_4 were chosen randomly every time the algorithm was run and the algorithm was run numerous times. Most of the runs resulted in the same results. Some of the algorithm runs resulted in much different results of 224 actual malignant samples being classified as benign and 447 actual benign samples being classified as malignant. The error rate for benign cells was 1.90 and the error rate for malign cells was 0.48. This could be due to poor initialization meaning that the algorithm works better with certain initial means than others.

For this project only a smaller version of the dataset was used, so that could have contributed to the results because the data that was not included could have altered the results. This project also only required a limited knowledge of the original dataset and there may have been some errors and inconsistencies that we were unaware about. To improve this model, I would continue testing the algorithm with an even larger dataset and look at whether the error rates change. If the new results are smaller or equal to the existing error rates, then the algorithm can be used for identifying samples as malign or benign.

Conclusion

The objective of this project was to apply the k-means algorithm to the Wisconsin Breast Cancer Dataset to assign cells to two different clusters, benign and malign, and evaluate how well the algorithm performed. After applying the algorithm, the results had a small error rate and showed that the algorithm performed well. While this is in relation to the dataset used for this project, it could easily be applied to a larger dataset to confirm the results. The results indicate that the k-means algorithm has the potential to classify benign and malign cells in breast cancer diagnosis. This could open a larger conversation to other uses for the k-means algorithm in other diagnoses and the medical field in general.

References

"Breast Cancer Early Detection." *National Breast Cancer Foundation,* www.nationalbreastcancer.org/early-detection-of-breast-cancer/.

Final Project PDF

Machine Learning for Cancer Diagnosis and Prognosis, pages.cs.wisc.edu/~olvi/uwmp/cancer.html.