# CS7641 Assignment 3

# Unsupervised Learning and Dimensionality Reduction

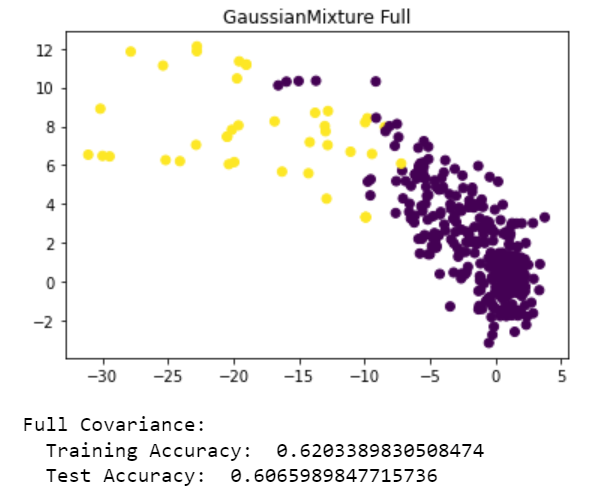
Datasets and hypotheses

The Data sets explored in this assignment come from Assignment 1. The first Data set is Credit card Fraud Analysis. Credit card fraud detection is a critical task for financial institutions to safeguard their customers' assets and maintain trust in the banking system. With the increasing prevalence of online transactions, detecting fraudulent activities has become more challenging yet crucial. This Dataset has a lot of data a lot of dimensions. It has various features and a calcification column to determine whether the transaction was fraud or not. In this assignment, I will explore this dataset to prove my hypothesis that dimensionality reduction will improve the performance of the clustering algorithms. I think the clustering algorithms will fit well with the data. In HW 1 it was found that the data performs well on neural networks, but I think it will perform better with calcification, and therefore my explorations in this assignment will be centered around that.

The second dataset I explored was Data Science salary prediction. Determining job salaries in data science involves predicting the salary range for a given set of job attributes such as location, experience, education, skills, and industry. It's a crucial task for both employers and job seekers to ensure fair compensation and competitiveness in the job market. This data set has 7 features which include position, employment, location, etc. For this dataset, I hypothesize that location will be the most important feature in determining the salary. I think this data set will not benefit as much from dimensionality reduction as Its features all seem to hold equal value.

Clustering

The Clustering Algorithms I ran were Gaussian Mixture and Agglomerative Clustering. For both of these, I explored various factors. From all covariance types Diagonal fit the data set 1 the best. And for dataset2 full coverage performs the best. Which was well expected. Since the data sets are very different, one of them has interdependent features while the other does not, so they will not have the same covariance type fit. [**¶**](http://localhost:8888/notebooks/Downloads/Dimensionality%20reduction%20%2B%20clustering.ipynb#Agglomerative-Clustering)

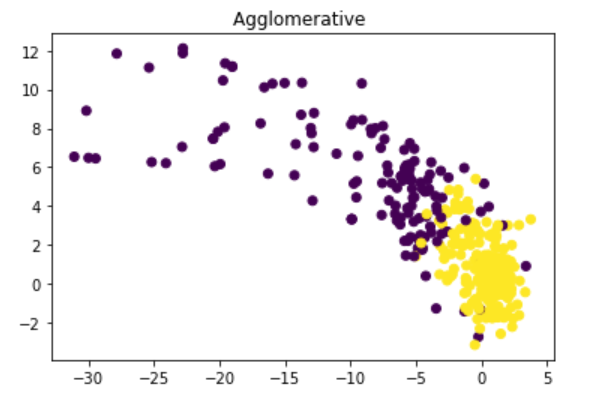
A chart with yellow dots and numbers

Description automatically generatedA graph with yellow and purple dots

Description automatically generatedA chart with yellow dots and numbers

Description automatically generatedA graph with numbers and dots

Description automatically generated A graph of a number

Description automatically generated with medium confidence

For Agglomerative clustering, I found that the model did not fit the data well however amongst all linages ward linkage performed the best. The results are shown below. A number of numbers and symbols

Description automatically generated with medium confidenceA number of numbers and symbols

Description automatically generated with medium confidence

I think both of these problems performed well on both data sets. However, the First dataset, although was of higher dimension yelled better performance. This was probably because this data was already designed for clustering problems. Which a is not the case for the Salaries Data set. It was not specifically designed for these kinds of problems and needs a lot of altering. To improve performance further the data set can be further simplified.

Dimensionality Reduction

For the dimensionality reduction I used all the required algorithms to reduce the dataset down to two dimensions. The distribution of eigenvalues and kurtotic are shown below.

A graph with blue rectangular bars

Description automatically generated A graph with blue rectangular bars

Description automatically generated A graph of a graph

Description automatically generated with medium confidence A graph of a bar

Description automatically generated with medium confidence

A graph with a bar

Description automatically generated A graph showing the difference between a number of objects

Description automatically generated with medium confidence

Their was little to no variation when I re-ran the random projections, that shows that noise does not affect each algorithm much. The rank of the data is 2 after I reduced it down to 2 components. The first data frame is more correlated then the second. As the ica plot ablve shows there are a few independent features in the data.m

Dimensionality Reduction and Clustering

For dimensionality reduction and clustering experments, I found that I got about the same or better clusters after reduction for both data set. Below you can see the data plotted.

A chart with yellow and purple dots

Description automatically generated

In order to save space on this report I will further compare and evaluate the function performance based on accuracy scores instead of visual plots. For PCA + Full Covariance Accuracy was 0.889 and Ward linkage Accuracy was 0.569. Which is approximately 10% more than the performance without dimensionality reduction. Similarly for data set 2 PCA + Full Covariance Accuracy: 0.590 and Ward linkage Accuracy was 0.47.

For ICA on the other had the results were mixed. ICA + Full Covariance Accuracy was 0.911 and Ward linkage Accuracy was 0.151. For dataset2 Full Covariance Accuracy: 0.80 and Agglomerative Clustering Accuracy was 0.473.

For Randomized projection Full Covariance Accuracy: 0.682 and Agglomerative Clustering Accuracy: 0.627. For data set 2 Full Covariance Accuracy: 0.256 and Agglomerative Clustering accuracy: 0.463

For MLA + Full Covariance Accuracy: 0.346 Agglomerative Accuracy: 0.293. For dataset 2 Full Covariance Accuracy: 0.398 Agglomerative Clustering: 0.409

Overall I found that the accuracies after MLA were the lowest. Which was well expected as the data was correlated.

Dimensionality Reduction, CNN and Clustering

When re-running the neural network algorithms with dimensionally reduced data the performance was found to be much better. Additionally, the K NN also worked much faster than before. I think this difference originates from the fact that now the data is much more condensed. KNN accuracy after tsne was 0.934 after Randomized projection was 0.929, after ICA was 0.831, and after PCA was 0.949 for data set 1. The accuracies were in the high 80 and 90s for the second dataset as well. This can be because of redundant features. High-dimensional datasets often contain redundant or correlated features. Additionally, Dimensionality Reduction rescued and removed overfitting from the model which helped accuracies a lot.

For CNN after clustering I performed clustering obtained the labels then used the labels for the clusters and original features as input for KNN. For which the accuracies were exceptionally high. The Accuracy for KNN on Data set 1 after Expectation Maximization was 0.98 and 1.0 after Agglomerative clustering. The Accuracy for KNN on Data set 1 after Expectation Maximization was 0.99 and the same after Agglomerative clustering. I think this exceptionally high accuracy might be because the dataset was clustered and the features were already extracted and used for nearest neighbors. This allowed the undesirable information to be dropped off and reduced overfitting.

Overall, I failed to reject all my hypotheses and it was proven that the models performed much better after dimensionality reduction.

Resources

<https://www.nagwa.com/en/explainers/402106373582/#:~:text=3%208%20%EF%81%88%20.-,Answer,be%20between%200%20and%202>.

<https://realpython.com/python-linear-algebra/>