Exploring Clustering and Dim Redn Algorithms

# Datasets

For the project two datasets were chosen *Dry Beans* and *Wine Quality* sourced from UCI ML data repo. The datasets have all numeric features to circumvent requirement of any special treatment for categorical features. The datasets also have a target variable defined to help verify the unsupervised clustering against ground truth.

In pair plots of *Dry Beans* exhibit clear clustering and strongly correlated features. There are no outliers. *Wine Quality* data set however does not demonstrate clear clustering across features, no strong correlation and has outliers. This should provide for a good contrast on the algos performance.

# Dimension Reduction

StandardScaler was applied to all datasets before all algorithms of clustering /dimension reduction.

***- WHY? Does this help with dealing with outliers?***

*Dim Redn – how many components were required –*

*Give the skree plots 4 plots*

*Give the distrb plot after dim redn 4 plots*

*Explain if clusters formed – if not why might they not have formed*

## Dry Beans

A diagram of different colored dots

Description automatically generatedA diagram of a number of different colored dots

Description automatically generated with medium confidenceA diagram of different colored dots

Description automatically generatedA map of different colored dots

Description automatically generated

A graph with a line

Description automatically generatedA graph with a line

Description automatically generated A graph with a line graph

Description automatically generated

A graph with a line

Description automatically generatedA graph of different colored lines

Description automatically generated

## Wine Quality

A diagram of a variety of wine quality

Description automatically generated with medium confidenceA diagram of a number of purple dots

Description automatically generatedA diagram of a variety of wine quality

Description automatically generated with medium confidence A diagram of a wine quality scatterplot

Description automatically generated

A graph with a line

Description automatically generatedA graph with a line

Description automatically generated A graph with a line

Description automatically generated

A graph of a line

Description automatically generatedA graph of different colored lines

Description automatically generated

# Clustering

We chose KNN and Gaussan – KNN sticks with circular blobs so has limitations unless there is clear groups.

Gaussian should be more adaptable to the

Clustering raw data – show the elbow + sil score, and the BIC 3 plots

Data with clustering2 plots

## Dry Beans

### KMeans

A graph with a line

Description automatically generatedA graph with a line

Description automatically generated A chart of different types of beans

Description automatically generated

### Gaussian Mixture

A graph with blue lines

Description automatically generated A diagram of a line graph

Description automatically generated with medium confidence

## Wine Quality

### KMeans

A graph of a line

Description automatically generatedA graph with a line

Description automatically generatedA diagram of a wine quality

Description automatically generated

### Gaussian Mixture

A graph with blue lines

Description automatically generatedA diagram of different colored dots

Description automatically generated

# Clustering with Dim reduction

KNN Clustering with dim reduction –

Based on dim reduction analysis – chose the 2 primary factors.

*• Explanations of methods. This is your opportunity to demonstrate nuances needed to support your*

*hypotheses.*

KNN – what is the best cluster – we don’t see an elbow

Gaussian Mix – BIC is all over the place

High dimensionality leads to issue in clustering. No adv

Dimension Reduction

Compare - how did you choose the number of components to use? – did you see improved clustering/grouping in plot ?

PCA

ICA

GaussianProj

T-SNE

## Clustering

**Single Linkage Clustering**

Hierarchical agglomerative

Deterministic. - O(n^3)

Edge Length – minimum spanning Tree

The silhouette score is specialized for measuring cluster quality when the clusters are convex-shaped

**K-Means Clustering**

* Optimisation is more like hill climbing -> select the next neighbour to add based on lower the distance
* An like hill climbing, K-means can get stuck on local optimum -> try selecting centroid based on the sample analytis, try random inits.

**Soft-Clustering**

Allows for the possibility for a point to belong to two clusters.

1. Max-Likelihood Gaussian
2. Expectation Maximisation

**Clustering properties: ( No algo can achieve all these at once)**

1. Richness
2. Scale-invariance
3. Consistency

## Feature Selection

* Need this form interpretability and insigh
* Avoid the curse of dimensionality (more dimension -> exp growth of data req.)

**Filtering**  - faster but w/o context of the learning

* Information gain
* Variance, entropy
* Indepednent

**Wrapping** – slower as it depends on the learning

* Hill climbing : subset of the fearture determined by the leaning algo output
* Randomised Opt

• Grounded descriptions of resulting clusters. Support descriptions with data-driven evidence.

• Analyses of your results. Why did you get the clusters you did? Do they make ”sense”? If you used

data that already had labels (for example data from a classification problem from assignment #1) did the

clusters line up with the labels? Do they otherwise line up naturally? Why or why not? Compare and

contrast the different algorithms. What sort of changes might you make to each of those algorithms to

improve performance? How much performance was due to the problems you chose? Be creative and think

of as many questions you can, and as many answers as you can. Take care to justify your analysis with

data explicitly.

• Can you describe how the data looks in the new spaces you created with the various dimensionality

reduction algorithms? For PCA, what is the distribution of eigenvalues? For ICA, how kurtotic are the

distributions? Do the projection axes for ICA seem to capture anything ”meaningful”? Assuming you

only generate k projections (i.e., preforming dimensionality reduction), how well is the data reconstructed

by the randomized projections? How much variation did you get when you re-ran your random projections

several times? How does noise affect each algorithm? What is the rank of your data? Can you describe

how colinear your data is both qualitatively and quantitatively? How might specific properties of your

data influence outputs of various algorithms?

• When you reproduced your clustering experiments on the datasets projected onto the new spaces created

by ICA, PCA, and RP, did you get the same clusters as before? Different clusters? Why or why not?

Remember to justify why one output might be more interesting when choosing your demonstrations.

• When you re-ran your neural network algorithms were there any differences in performance? Speed?

Consider how you might judge differences in performances and include these notes in your discussion

# References

<https://towardsdatascience.com/kmeans-hyper-parameters-explained-with-examples-c93505820cd3>

<https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/>

<https://towardsdatascience.com/gaussian-mixture-model-clusterization-how-to-select-the-number-of-components-clusters-553bef45f6e4>

<https://towardsdatascience.com/gaussian-mixture-model-clearly-explained-115010f7d4cf>

<https://jakevdp.github.io/PythonDataScienceHandbook/05.12-gaussian-mixtures.html>

<https://scentellegher.github.io/machine-learning/2020/01/27/pca-loadings-sklearn.html>

<https://towardsdatascience.com/introduction-to-ica-independent-component-analysis-b2c3c4720cd9>

<https://medium.com/@violante.andre/an-introduction-to-t-sne-with-python-example-47e6ae7dc58f>

<https://distill.pub/2016/misread-tsne/>