# Introduction

Brief description of your datasets, and hypotheses you want to highlight in your report.

# Methods

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

hypotheses.

## Clustering

**Single Linkage Clustering**

Hierarchical agglomerative

Deterministic. - O(n^3)

Edge Length – minimum spanning Tree

**K-Means Clustering**

[**https://towardsdatascience.com/kmeans-hyper-parameters-explained-with-examples-c93505820cd3**](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://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-clusterization-how-to-select-the-number-of-components-clusters-553bef45f6e4)

* 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

**https://scikit-learn.org/stable/modules/mixture.html#gmm**

**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