



K Q-flats Clustering

Machine Learning - Project
Arnaud Ruymaekers - S5131820

Motivations

- Interest for clustering techniques
- Curiosity about clustering more complex data

Method Description

Same steps as Lloyd's Algorithm for K-means:

1. Assignment to cluster
2. Update of the cluster

Instead of clusters with points as centers:

Hyperplanes of q dims (1 = line, 2 = plane, >2 = hyperplane)

When $q=0$, the algorithm is simply k-means

Method Description - Cluster Assignment

Each point “i” is assigned to the closest cluster by using:

$$\min_{\ell=1,\dots,k} |A_i w_\ell^j - \gamma_\ell^j|$$

Where:

- Each plane “l” is defined by: $P_\ell := \{x \mid x \in R^n, x' w_\ell = \gamma_\ell\}$
- And “A” is the matrix of all the data points

Method Description - Cluster Update

Updating the hyperplane coefficients w and γ :

- w becomes the eigenvector corresponding to the smallest eigenvalue of:

$$B(\ell) := [A(\ell)]' \left(I - \frac{ee'}{m(\ell)} \right) A(\ell)$$

- and γ becomes:
$$\frac{e' A(\ell) w_{\ell}^{j+1}}{m(\ell)}$$

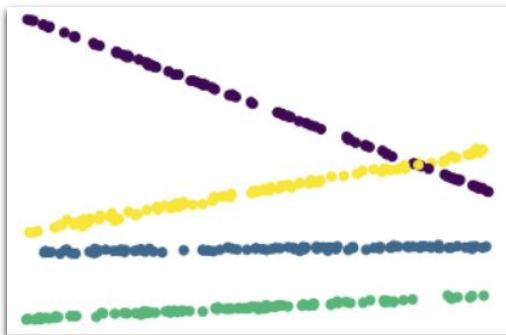
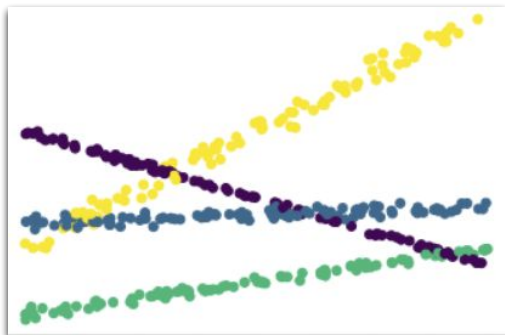
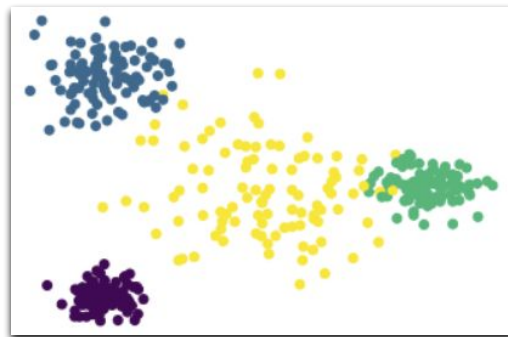
Where:

- $A(l)$ are the points assigned to cluster “l”
- $m(l)$ is the amount of point assigned to cluster “l”

Discussion of assessment

Different synthetic data sets:

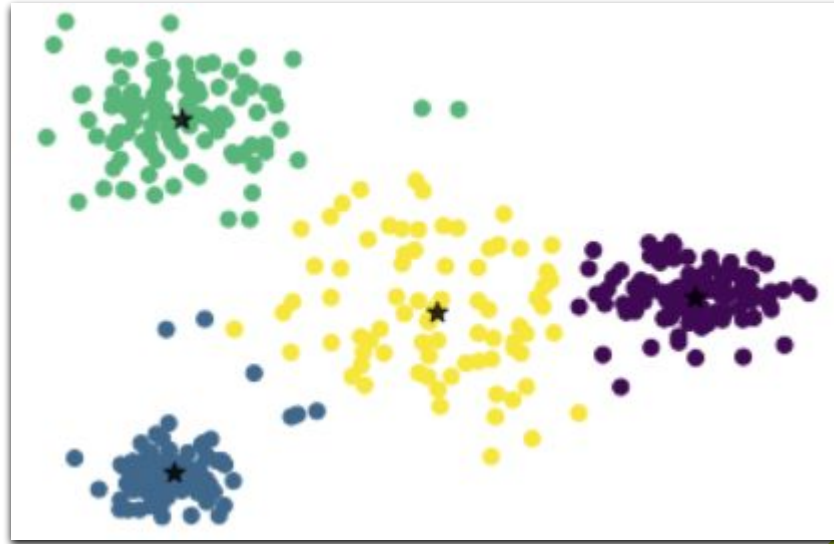
- Blobs (same generation as in k-means lab)
- Linear (with added Gaussian noise)



Discussion of assessment - Blobs

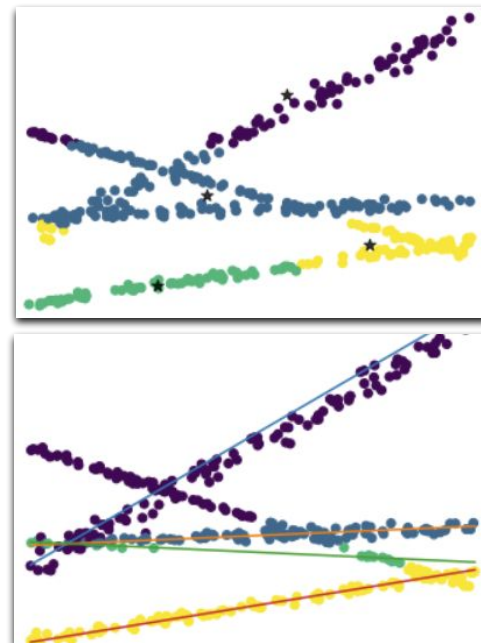
As shown in lab, k-means (k q-flats with $q=0$) performs well to find the clusters when the clusters don't overlap too much

(accuracy 94.5%)



Discussion of assessment - Lines

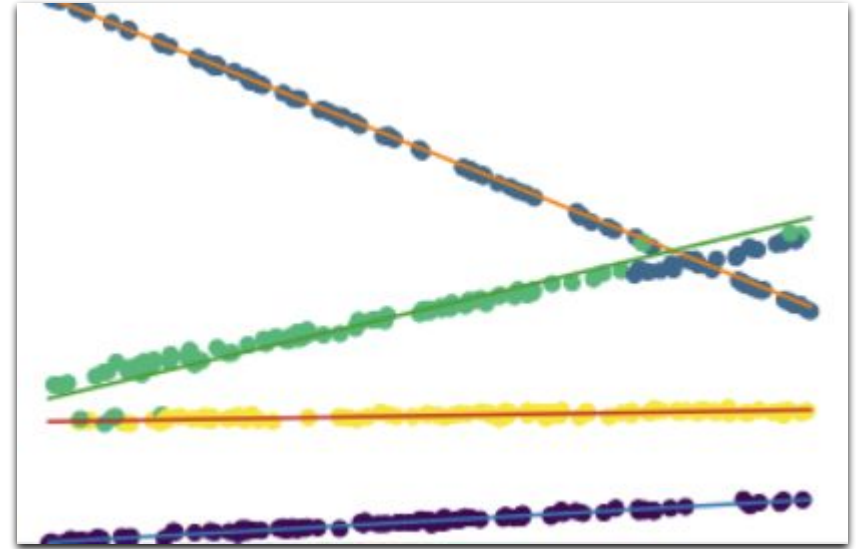
- With $q=0$, very bad results
- With $q=1$, better
 - However it didn't manage to pick up the cluster with a downward slope intersecting the other lines
 - Had to run the algorithm a significant amount of times before getting satisfactory results
 - Doesn't perform well when clusters overlap, like k-means



Discussion of assessment - Lines (spread out)

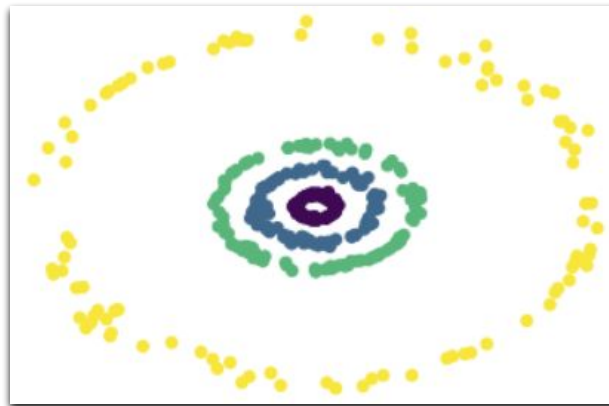
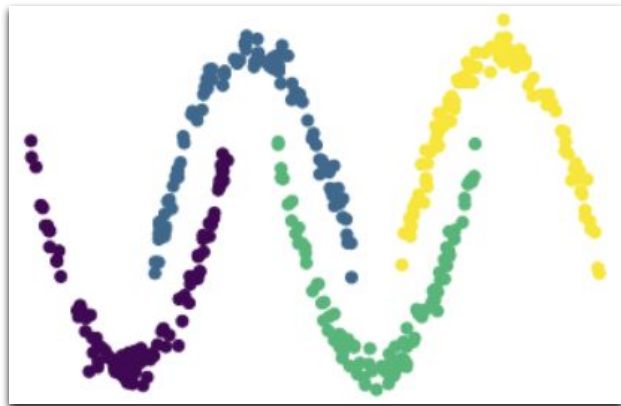
With more spread out data, the algorithm reacts much better

Minor misclassification around intersection



Future work

- Can be combined with feature maps to work on non-linear data



References

- <https://minds.wisconsin.edu/bitstream/handle/1793/66118/98-08.pdf?sequence=1&isAllowed=y>
- https://en.wikipedia.org/wiki/K_q-flats
-