Clustering with a Density-Based Similarity Measure

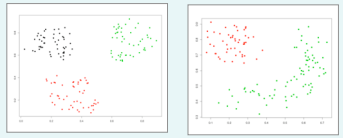
Student: Xiaoyi Fei, Advisor: Rebecca Nugent

Carnegie Mellon University, Pittsburgh, Pennsylvania

Contact Information:
Xiaoyi Fei
Email: xiaoyifei@gmail.com
Phone: (703) 945-8075

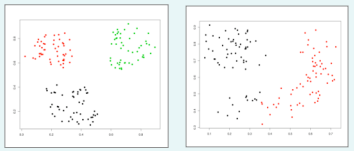
Introduction

What is Clustering?
The goal of clustering is to identify distinct groups in a data set and assign a group label to each observation.



A Desirable Clustering Partitions Observations Such That:
•Observations in the same cluster are very similar to each other
•Observations in different clusters are very different from each other.

One Commonly used Algorithm: *K-means*
user asks for k clusters, the algorithm finds k cluster centers that minimize the within-cluster sum of squared Euclidean distances from each cluster's points to its center



K-means only works well with spherical data

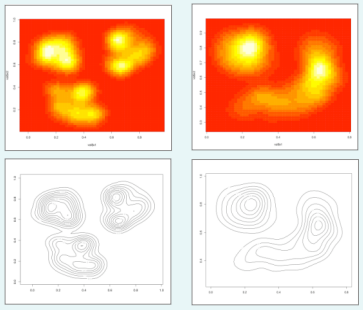
Many popular clustering algorithms depend on *Euclidean distance*.

Clustering Based on Density

The ideal clustering algorithm would be able to identify clusters of different sizes and shapes.

High Density Area: Area where observations occur frequently
Low Density Area: Area where observations occur rarely

Clusters are then identified as connected high *density* areas separated by low *density* areas.

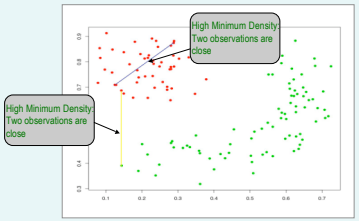


Minimum-Density Clustering "Distance"

Define the distance between two observations to be minimum of the density along the line segment connecting them.

$$d(i,j) = \min_{t\in[0,1]} f(t\cdot \underline{x}_i + (1-t)\cdot \underline{x}_j)$$

Observations in same cluster have high d(i,j)
Observations in different cluster have low d(i, j)



How does it work?

Calculate Distance Matrix:
- Use *Gaussian Kernel Density Estimate* to estimate the true density of the data.
- The bandwidth of the density estimate is chosen by least squares cross validation.
- Use grid search to estimate the minimum density between pairs of observations.

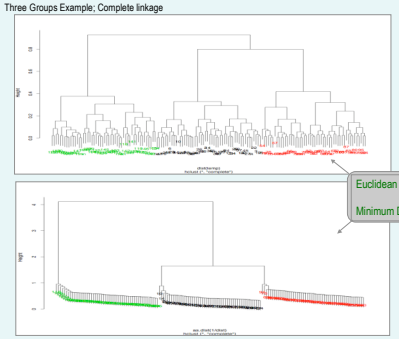
Note: in high dimensions, the data are sphered prior to distance matrix calculation.

Partition Data into Clusters
Use hierarchical clustering/linkage method on distance matrix.
Start with all observations, eventually combine into one cluster.

Steps:
1. Start with each observation as its own group.
2. The two closest groups are merged into a new group. (linked)
3. Update the distance among all groups.
4. Repeat 2) and 3) until left with one group.

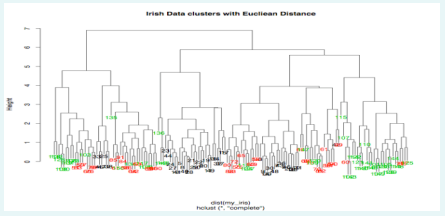
Pruning the Tree: extract k clusters from the tree by cutting the tree at the height where there are k branches.

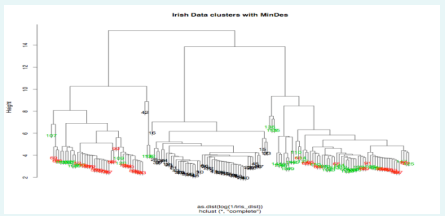
Three Groups Example; Complete linkage



Algorithm In Practice

Irish Data: Irish data; 50 specimens each of 3 types of iris. Sepal.length and width, petal length and width.





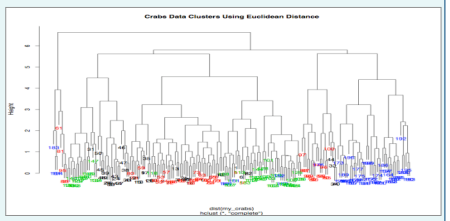
Euclidean Distance

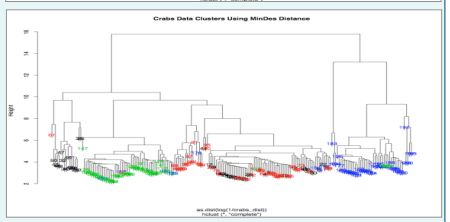
	C1	C2	C3
S1	26	18	6
S2	0	28	22
S3	3	31	16

Minimum Density

	C1	C2	C3
S1	49	1	0
S2	0	28	22
S3	0	13	37

Crabs Data: 50 crabs each of two species (orange/blue) and both genders. 200 total. 5 morphological measurements.





Euclidean Distance

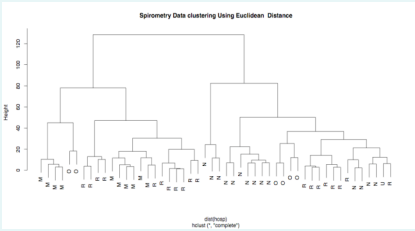
	C1	C2	C3	C4	C5
BM	37	13	0	0	0
BF	4	45	0	1	0
OM	1	0	49	0	0
OF	0	1	6	40	3

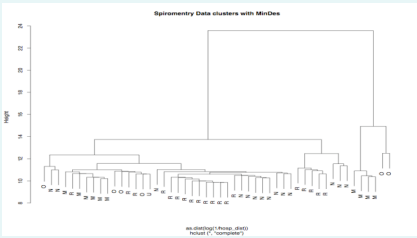
Minimum Density

	C1	C2	C3	C4	C5
BM	27	19	3	0	0
BF	30	0	15	5	0
OM	34	16	0	0	0
OF	5	0	0	3	41

In Practice (Continued)

Spirometry Data: Sets of results from three pulmonary function tests run on 50 internal medicine patients; the type of pulmonary disease is classified by physicians as Normal, Obstructive, Restrictive and Mixed.
(Data are standardized for age, gender ethnicity and BMI.)





Performance Comment

- The algorithm handles different size of groups
- The shape of the data set has minimal effect on the performance of the algorithm.
- The bottom of the tree are the modes of the data
- The top points tend to be outliers
- May need to prune more groups in order to extract substantial clusters.

New Ideas:
Prune the tree from the bottom instead of from the top.
Will find modes, outliers will need to be assigned.

For Further Information

Please contact Xiaoyi Fei at xiaoyifei@gmail.com. For suggestions, questions and comments on this research project.

Acknowledgement

Special thanks for Dr. Rebecca Nugent.