Hierarchical Clustering and Anomaly Detection via DBSCAN

Ryan Miller

Grinnell College
Statistics

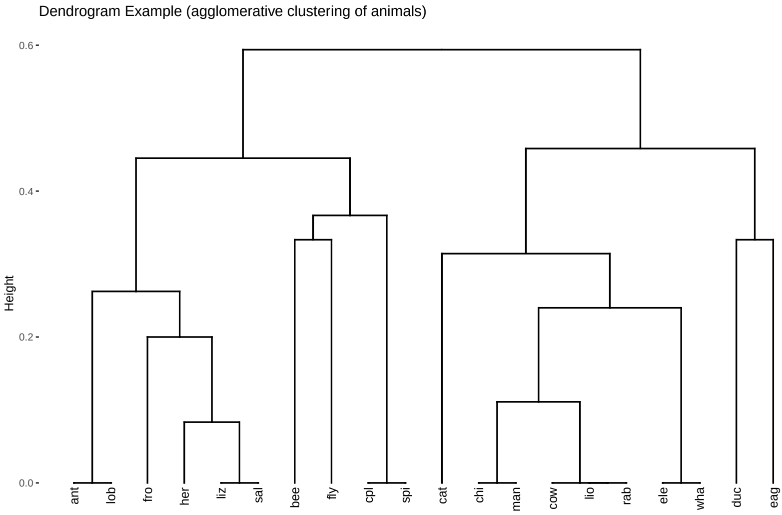
Introduction

- Unsupervised methods identify patterns in the data without specifying an outcome measure
  - *k*-means clustering finds groupings of similar data-points using *prototypes*
  - *k*-means is an example of **partitional clustering**, as it produces clusters without any overlap

Hierarchical Clustering (overview)

- **Hierarchical clustering** organizes data-points into a tree-like structure known as **dendogram** that stores a series of nested (overlapping) groupings
- There are two major types of hierarchical clustering algorithms:
  1. **Agglomerative** - each data-point begins as its own cluster and pairs of clustered are merged using a *linkage criterion*
  2. **Divisive** - all data-points begin in a single cluster that is recursively subdivided until each data-point is its own cluster
- We'll focus on agglomerative clustering since divisive clustering algorithms aren't currently offered in `sklearn`

Hierarchical Clustering (dendrogram example)



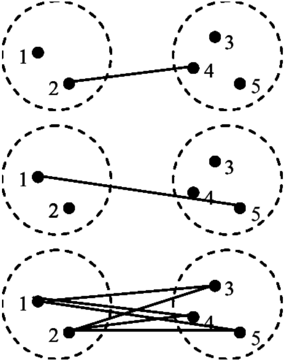
Agglomerative clustering - linkage

Three of the most popular ways to merge clusters are very straightforward:

- *Single linkage* - find the minimum pairwise distance between data-points in different clusters and merge their clusters
- *Complete linkage* - find the maximum pairwise distance between data-points in each pairing of clusters and merge the two clusters with the smallest maximum
- *Average linkage* - find the average pairwise distances between points in each pairing of clusters and merge the two clusters with the smallest average distance

Agglomerative clustering - linkage

Diagram illustrating single, complete and average linkage:



$$d_{24}$$

$$d_{15}$$

$$\frac{d_{13}+d_{14}+d_{15}+d_{23}+d_{24}+d_{25}}{6}$$

Agglomerative clustering (Ward linkage)

The most common linkage criterion used in agglomerative clustering is *ward's linkage*, which minimizes Δ_{c_a,c_b}, the increase in sum of squared error accumulated by merging clusters:

- $\Delta_{c_a,c_b} = SSC_{c_a,c_b} - SSS_{c_a,c_b}$
  - $SSC_{c_a,c_b} = \sum_{i \in c_a \cup c_b} ||\mathbf{x}_i - \boldsymbol{\Psi}_{c_a \cup c_b}||^2$
  - $SSS_{c_a,c_b} = \sum_{i \in c_a} ||\mathbf{x}_i - \boldsymbol{\Psi}_{c_a}||^2 + \sum_{i \in c_b} ||\mathbf{x}_i - \boldsymbol{\Psi}_{c_b}||^2$

Note that c_a and c_b index the data-points in belonging to two different clusters and $\boldsymbol{\Psi}_{c_a}$ denotes the center of cluster c_a.

Choosing a linkage criterion

- *Single linkage* - tends to create large chains of one-at-a-time additions but can be good at identifying irregular patterns
- *Complete linkage* - robust to outliers and tends to favor similarly sized clusters at each "level" of the dendrogram
- *Average linkage* - lower variability than complete linkage but more impacted by outliers
- *Ward's linkage* - clusters tend to be the most compact (desirable) but the method is also the most computationally expensive

Agglomerative clustering vs. *k*-means

- *k*-means forms *spherical clusters* (thus it excels with the "blobs" data) but struggles to identify irregularly shaped clusters (ie: "moons" data)
  - Agglomerative clustering tends to be more flexible when applied to unusual data sets
- Agglomerative clustering can also better handle outliers and do not involve random initialization
- The main downside of agglomerative clustering is its computational burden when n is large

DBSCAN

- DBSCAN finds clusters using two parameters: a radius, `eps`, and a minimum number of data-points, `min_samples`
  - The algorithm surrounds each data-point with a hypersphere (or a circle in 2 dimensions)
- These hyperspheres are used to label each data-point as one of three types: *core points*, *border points*, and *noise*
  - Core points contain at least `min_samples` neighbors within their hypersphere
  - Border points contain at least 1 neighbor within their hypersphere
  - Noise points are at least `eps` away from any other data-point
- Cluster membership can then be determined by connected density regions

The diagram below illustrates DBSCAN:

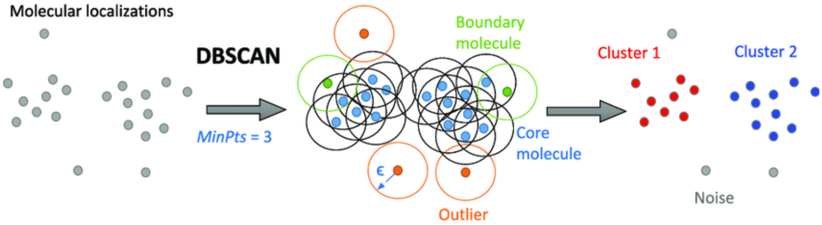


Image Source: A Review of Super-Resolution Single-Molecule Localization Microscopy Cluster Analysis and Quantification_Methods

Outlier (anomaly) detection

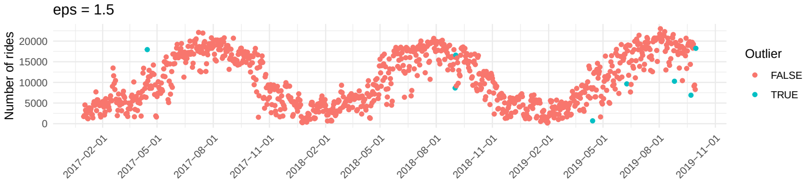
- In statistics, an outlier is a data-point that is significantly far from other observations
  - A simple example is the "3 sigma rule", which will classify anything more than 3 standard deviations from the mean as an outlier
    - This amounts most extreme 0.3% of data-points under a normal model

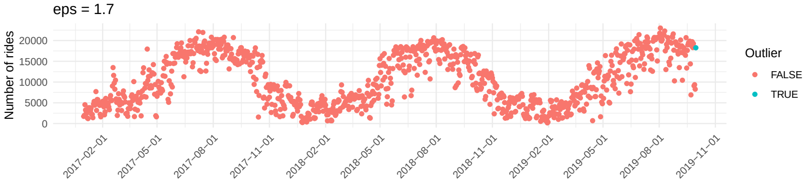
Outlier (anomaly) detection

- In statistics, an outlier is a data-point that is significantly far from other observations
  - A simple example is the "3 sigma rule", which will classify anything more than 3 standard deviations from the mean as an outlier
    - This amounts most extreme 0.3% of data-points under a normal model
- DBSCAN provides a flexible method of outlier detection governed by the eps hyperparameter
  - This can be set using domain-specific knowledge, or tuned so that a certain percentage of the data is classified as outliers

Example (Chicago Divvy bikeshare data)

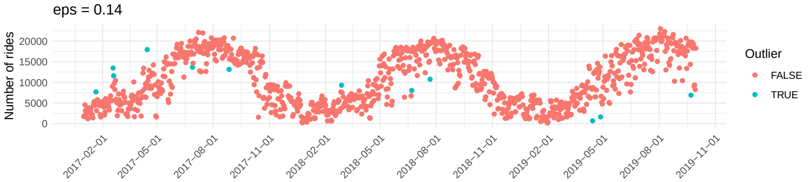
Standardizing the number of rides only:

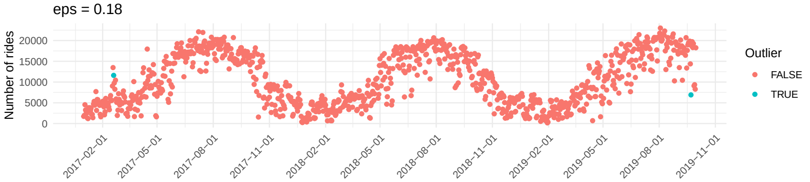




Example (Chicago Divvy bikeshare data)

Standardizing the number of rides and day:





Grinnell College
Statistics