Clustering: Deterministic/Algorithms

Rebecca Nugent
Department of Statistics, Carnegie Mellon University
http://www.stat.cmu.edu/∼rnugent/PCMI2016

PCMI Undergraduate Summer School 2016

July 14, 2016

What did we think about last time?

- Linear Discriminant Analysis
- Quadratic Discriminant Analysis
- Visualizing (Dis)Similar High-Dim Observations
- Icons/Glyphs
- What it's like after being a Math Major

Now we'll try

- looking for and visualizing high-dim structure
- clustering observations with specific structure goals

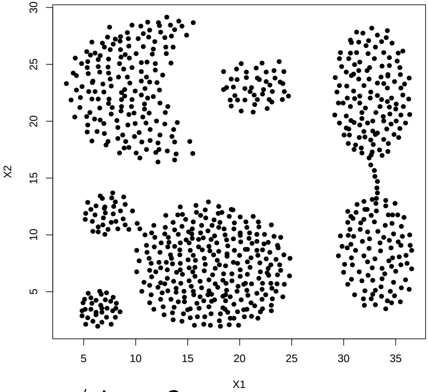
Clustering

Often we're interested in determining the presence (or absence) of group structure in our data

- sets of genes with similar expression patterns
- food samples with similar infrared spectra
- voting preferences in the United States
- marketing segments (who buys what?)
- learning trajectories over time
- online chatter changing topics

Goal: to identify distinct groups in a data set and assign a group label to each observation; observations are partitioned such that observations in one subset are more similar to each other than to observations in different subsets

Clustering



- What is a group/cluster?
- How many are there?
- How sure are we?
- What do they look like?
- What properties do they have?
- What happens when we get new observations?

Clustering Approaches

Most approaches can be very loosely binned into two categories:
Deterministic/Algorithm

- Clusters often defined by distance (or dissimilarity) measure
- Largely data-driven
- Structure determined by algorithm, user-chosen parameters
- Often used for very large data sets
- Research often concentrates on approximations
- Change the data, change the clusters

Statistical:

- Assume data have an underlying population distribution
- Groups are features of the unknown population density
- Estimate the density; estimate the clusters
- Can assign probabilistic labels; clusters have well-established statistical properties
- Suffers from problems associated with density estimates

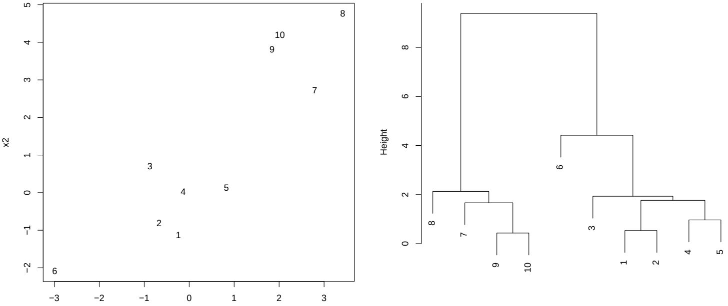
Newer methods borrow strength from both sides

Hierarchical Clustering

Algorithm that links observations in order of closeness in a hierarchically linked structure (*dendrogram*); deterministic

Most common version is *agglomerative*

- Every observation starts as its own group
- Compute all intergroup distances*
- Merge the two closest groups; update distances
- Repeat the previous step until have one group



Hierarchical Clustering

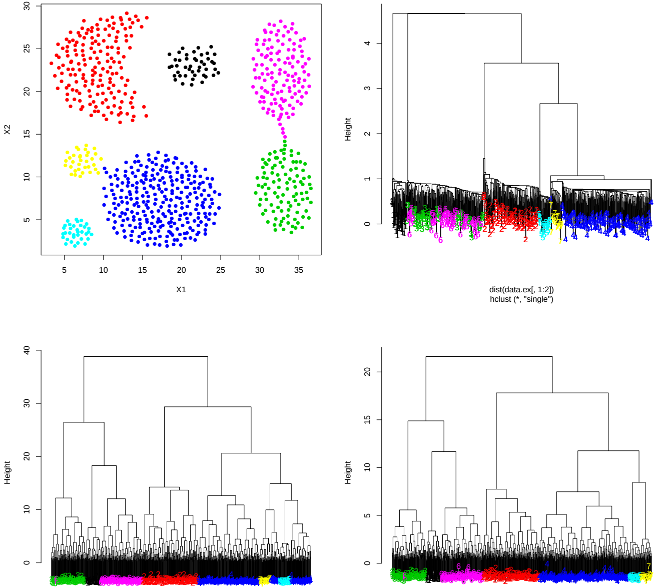
What is the intergroup distance? User needs to choose

- Single Linkage: $d(G_1, G_2) = \min_{x_i \in G_1, x_j \in G_2} d(x_i, x_j)$
  Chaining effects; walking through nearest neighbors; cool theoretical properties
- Complete Linkage: $d(G_1, G_2) = \max_{x_i \in G_1, x_j \in G_2} d(x_i, x_j)$
  Tends to chunk data into compact spheres; popular in practice
- Average Linkage: $d(G_1, G_2) = average_{x_i \in G_1, x_j \in G_2} d(x_i, x_j)$
- Other linkage types include: Ward's method, median, centroid, prototype

So how many clusters do we have? User chooses cut threshold. Reasons can be

- theoretical
- application-driven
- subjective

Back to our odd example



K-means

Algorithm to partition observations into spherical clusters

Measure "quality" of clusters: *within-cluster squared-error criterion*

$$\sum_{k=1}^{K} \sum_{x_i \in C_k} (x_i - \bar{x}_k)^2$$

Required: Set the number of clusters, K, in advance.

Given a set of K initial cluster centers, alternate between:

- Assign each observation to the closest center
- Recompute the centers given the current assignments

Stop when the cluster assignments/centers no longer change.

Each step decreases the within-cluster criterion.

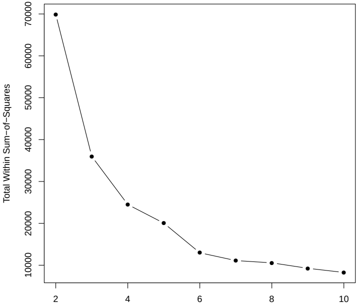
Theoretical results tell us we'll converge to the global optimum.
Real life laughs in the face of theory.

K-means:

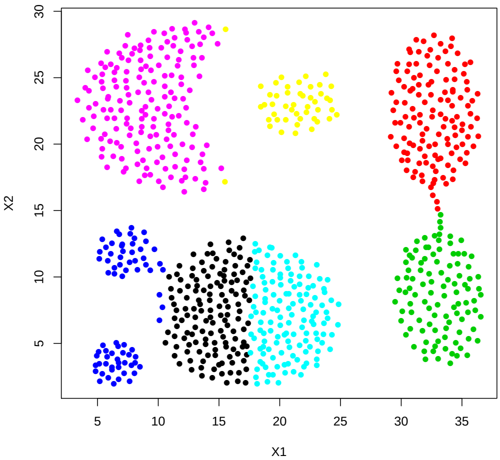
In practice:

- First few steps correspond to large drops in the criterion; later steps correspond to negligible drops.
- Use K randomly chosen observations as the starting centers (but don't have to; can choose specific centers)
- Have an idea of what K should be in advance

If we increase K, what happens to the within-cluster criterion?
We use an *elbow graph* to determine a "useful" K.



Back to odd data



K-means is also dependent on the set of starting centers you choose; solutions can vary widely. Often people simulate lots of K-Means solutions and search for the most stable one.