
CLUSTERING
Paul Speaker

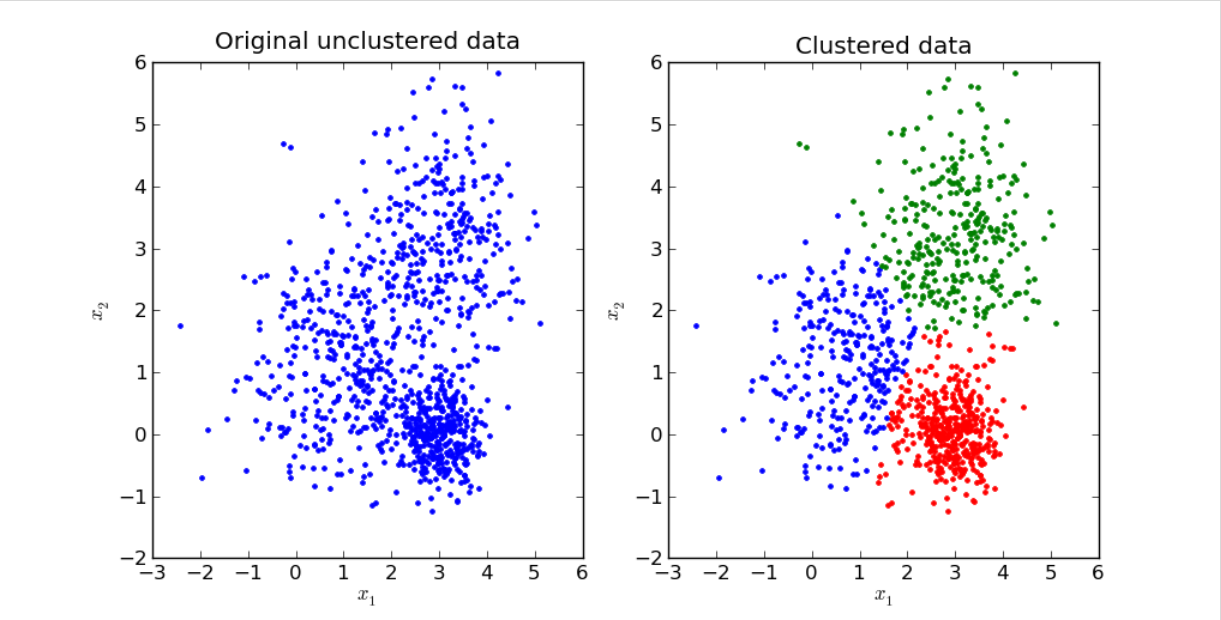
Clustering

- For many reasons, we can analyze data without a specific target in mind
  - "Unsupervised learning"
- Why?
  - For predictive models, different clusters of data can behave qualitatively different
  - Creates convenient labels for data
  - Association analysis (recommender enginers)

Types of Clustering

- There are several different approaches to find clusters. They can be determined by
  - Distance
    - K-means clustering
  - Density
    - DBSCAN
  - Distributional
    - Gaussian Mixture (GMM)
  - Hierarchical (Clusters within clusters)
    - Agglomerative/Divisive
- We will look in more detail at K-Means and Gaussian Mixture

Distance Based Clustering



- The most common approach to clustering is k-means clustering
- Pre-determined number of clusters
- Iterative approach to find cluster "location"
  - Location for a cluster determined by center of mass
- Since distance measures treat different fields the same way, they need to be normalized if they have different scales

K-Means Algorithm

- Scale data
- Cluster initialization
  - Randomly determine center points for n clusters
  - Usually come from sampling k values
- The distance between each data point and center point is calculated
- Each point is assigned to clusters by which center point is closest
- Once all points are assigned, the new center points are calculated for the points in each cluster
- Recalculate the distances, assignments, and center points until a maximum number of iterations is reach, or the points stop moving

K-Means Algorithm

- Since the number of clusters is pre-determined, some trial and error might be needed to arrive at the "right" number
- This can happen with a visual inspection of the data for low-dimensional problems
- A trickier issue that that the final results is often dependent on the initial locations of the centers
  - If you do not in fact get different results after several runs, then you are likely to havehave a good number of clusters
- Again, if the scales of the fields are different sizes, it is important to normalizes so that the distances are comparable

Gaussian Mixture Models

- A Gaussian Mixture model is a statistical approach to clustering by fitting a linear combination of multivariate Gaussian models to data.
- As a statistical process, fitting is a parameter estimation process
- The linear combination parameters represent the fraction of data points in the given Gaussian cluster