CMSC 170 Lecture 10 Notes April 24, 2018

Clustering

CLUSTERING

- *Unsupervised Learning*: detecting patterns in unlabeled data
- Key tasks in unsupervised learning: clustering, dimensionality reduction
- *Clustering:* grouping together unlabeled data such that similar objects belong to one cluster
- Uses distance metrics (e.g. Euclidean, Manhattan) to compute similarity
- Has applications in computer vision (segmentation), recommender systems, social network analysis, market research, and other fields

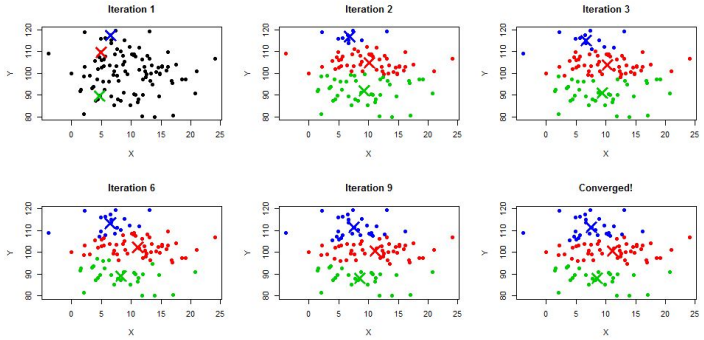
K-MEANS CLUSTERING

- *Assumption*: dataset has K clusters
- K is a parameter → user will set this
- Randomly assign K centers
- Group data points around the centers → data point will find the center nearest to it and join that group
- After forming the clusters, re-compute the center of each cluster by finding the average of the cluster points
- Repeat the process, this time using the newly computed cluster center points
- Stop when it has converged → no changes anymore
- Quality of clustering depends on initial centers (that's why we randomize)

HIERARCHICAL CLUSTERING

- Aka agglomerative clustering
- Initially, each data point is its own cluster
- We pick two clusters that are closest to each other and merge them
- Repeat until there's only one cluster left
- We can form a **dendrogram:** which shows the connections (which clusters got merged) and the corresponding distance (how close the two clusters are)
- To find distances of clusters with multiple elements, we can use any of the ff: (1) closest pair, (2) farthest pair, (3) average of all pairs

K-Means



Hierarchical Clustering