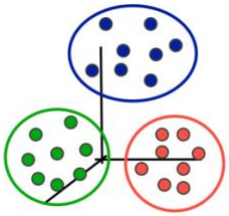
Notes for Sept20, 22:

Clustering: a grouping of data objects such that the objects within a group are similar (or near) to one another and dissimilar (or far) from the objects in other.



We capture this by minimize intra-cluster distances, maximize inter-cluster distances.

We have, stand-alone tool to gain insight into the data for visualization, preprocessing step for other algorithms like indexing or compression often relies on clustering.

Applications include but not limited to the following:
Image processing: cluster images based on their visual content
Web mining, cluster groups of users based on their access patterns on webpages, and cluster webpages based on their content
Bioinformatics: cluster similar proteins together (similarity wrt chemical structure and/or functionality etc)

Three types of clustering:
Partitional: each object belongs in exactly one cluster
Hierarchical: a set of nested clusters organized in a tree
Density Based: clustering is defined based on the local density of the points

Partitional algorithms:
partition the n objects into k clusters
• each object belongs to exactly one cluster
• the number of clusters k is given in advance
The k-means problem
• consider set X={x1,...,xn} of n points in Rd
• assume that the number k is given
• problem: find k points c1,...,ck (named centers or means)
so that the cost:

$$\sum_{i=1}^{n} \min_{j} \left\{ L_2^2(x_i, c_j) \right\} = \sum_{i=1}^{n} \min_{j} ||x_i - c_j||_2^2$$

is minimized.

The k-means (Lloyd's) algorithm
1.randomly (or with any other method) pick k cluster centers {c1,...,ck}
2.for each j, set the cluster Xj to be the set of points in X that are the closest to center cj
3.for each j let cj be the center of mass of cluster Xj (mean of the vectors in Xj)
4.repeat (go to step 2) until convergence

Properties of the k-means algorithm:
• finds a local optimum
• often converges quickly but not always
• the choice of initial points can have large influence in the result

The initialization is:
• random initialization
• random, but repeat many times and take the best solution
• helps, but solution can still be bad
• pick points that are distant to each other
• k-means++
• provable guarantees

Advantages of applying K-means:
• easy to implement
• provable guarantee
• works well in practice