**kNN** •Classify a new instance by taking a vote of the $k1$ nearest exemplars. •vote among all neighbors within a fixed radius r •combine the two, stopping when $count > k$ or $dist. > r$ •distance weighting, the closer an exemplar is to the instance, the more its vote counts. **ties** Preference to the 1NN, or Random choice. **Pros** train $O(n)$, simple, intuitive **Cons** test $O(n)$ require a good deal of storage, and canâĂŹt easily represent a specific boundary geometry, rely on a useful distance metric.
**Clustering vs. classification** In a classifier, possible class labels are provided. In a clustering problem, possible labels are the cluster labels learned from the training set. **Assigning** of labels or clusters to data points is classification.
**Clustering** find clusters (groupings) that are compact with respect to the distance metric.
**Scatter Matrix** The scatter of X is defined as the trace of the scatter matrix.(The trace is the sum of the diagonal elements of a square matrix).
**Kmeans** NP-complete, no efficient solution to find the optimal clustering (data partition). **K-means algorithm** heuristic algorithm, not optimal, converge to local optimal, works quite well in most cases. run several times (with a random starting point) and then the best solution is selected(the solution with the smallest within-cluster scatter).
**Kmedoids** medoid of a set of points is the point with the minimal average dissimilarity (distance) to all other points in the set.