K-means Clustering

Lecture 8

Marina Santini

Acknowledgements
Slides borrowed and adapted from:
Data Mining by I. H. Witten, E. Frank and M. A. Hall

Lecture 8: Required Reading

Daume' III (2015: 32-33)

Witten et al. (2011: 138-141)

Clustering

- Clustering techniques apply when there is no class to be predicted
- Aim: divide instances into “natural” groups
- Clusters can be:
  - disjoint vs. overlapping
  - deterministic vs. probabilistic
  - flat vs. hierarchical
- We'll look at a classic clustering algorithm called *k-means*
  - *k-means* clusters are disjoint, deterministic, and flat

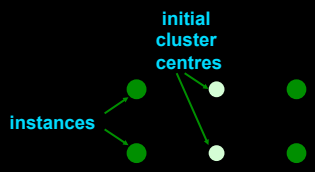
The *k*-means algorithm

To cluster data into *k* groups:
(*k* is predefined)

1. Choose *k* cluster centers
   - e.g. at random
2. Assign instances to clusters
   - based on distance to cluster centers
3. Compute *centroids* of clusters
4. Go to step 1
   - until convergence

Discussion

- Algorithm minimizes squared distance to cluster centers
- Result can vary significantly
  - based on initial choice of seeds
- Can get trapped in local minimum
  - Example:



- To increase chance of finding global optimum: restart with different random seeds
- Can we applied recursively with $k = 2$

Clustering: how many clusters?

- How to choose k in k-means? Possibilities:
  - Choose k that minimizes cross-validated squared distance to cluster centers
  - Use penalized squared distance on the training data (eg. using an MDL criterion)
  - Apply k-means recursively with $k = 2$ and use stopping criterion (eg. based on MDL)
    - Seeds for subclusters can be chosen by seeding along direction of greatest variance in cluster (one standard deviation away in each direction from cluster center of parent cluster)
    - Implemented in algorithm called X-means (using Bayesian Information Criterion instead of MDL)

Summary

- Initial seeds
- All data points are assigned to the nearest centroid
- Iteration continues until thre are no changes in the clusters.

The end