K-Means Clustering

3/3/17
Unsupervised Learning

• We have a collection of *unlabeled* data points.
• We want to find underlying structure in the data.

Examples:
• Identify groups of similar data points.
  • Clustering
• Find a better basis to represent the data.
  • Principal component analysis
• Compress the data to a shorter representation.
  • Auto-encoders
Unsupervised Learning

• We have a collection of *unlabeled* data points.
• We want to find underlying structure in the data.

Applications:

• Generating the input representation for another AI or ML algorithm.
  • Clusters could lead to states in a state space search or MDP model.
  • A new basis could be the input to a classification or regression algorithm.
• Making data easier to understand, by identifying what’s important and/or discarding what isn’t.
The Goal of Clustering

Given a bunch of data, we want to come up with a representation that will simplify future reasoning.

Key idea: group similar points into clusters.

Examples:

• Identifying objects in sensor data
• Detecting communities in social networks
• Constructing phylogenetic trees of species
• Making recommendations from similar users
EM Algorithm

E step: “expectation” … terrible name
• Classify the data using the current model.

M step: “maximization” … slightly less terrible name
• Generate the best model using the current classification of the data.

Initialize the model, then alternate E and M steps until convergence.

Note: The EM algorithm has many variations, including some that have nothing to do with clustering.
K-Means Algorithm

Model: k clusters each represented by a centroid.

E step:
• Assign each point to the closest centroid.

M step:
• Move each centroid to the mean of the points assigned to it.

Convergence: we ran an E step where no points had their assignment changed.
K-Means Example
Initializing K-Means

Reasonable options:

1. Start with a random E step.
   • Randomly assign each point to a cluster in \{1, 2, ..., k\}.

2. Start with a random M step.
   a) Pick random centroids within the maximum range of the data.
   b) Pick random data points to use as initial centroids.
K-Means in Action

https://www.youtube.com/watch?v=BVFG7fd1H30
Another EM Example: GMMs

GMM: Gaussian mixture model

- A Gaussian distribution is a multivariate generalization of a normal distribution (the classic bell curve).
- A Gaussian mixture is a distribution comprised of several independent Gaussians.
- If we model our data as a Gaussian mixture, we’re saying that each data point was a random draw from one of several Gaussian distributions (but we may not know which).
EM for Gaussian Mixture Models

Model: data drawn from a mixture of k Gaussians

E step:
• Compute the (log) likelihood of the data
  • Each point’s probability of being drawn from each Gaussian.

M step:
• Update the mean and covariance of each Gaussian.
  • Weighted by how responsible that Gaussian was for each data point.
How do we pick K?

There’s no hard rule.

• Sometimes the application for which the clusters will be used dictates k.
• If k can be flexible, then we need to consider the tradeoffs:
  • Higher k will always decrease the error (increase the likelihood).
  • Lower k will always produce a simpler model.
Hierarchical Clustering

• Organizes data points into a hierarchy.
• Every level of the binary tree splits the points into two subsets.
• Points in a subset should be more similar than points in different subsets.
• The resulting clustering can be represented by a dendrogram.
Direction of Clustering

Agglomerative (bottom-up)
• Each point starts in its own cluster.
• Repeatedly merge the two most-similar clusters until only one remains.

Divisive (top-down)
• All points start in a single cluster.
• Repeatedly split the data into the two most self-similar subsets.

Either version can stop early if a specific number of clusters is desired.