Clustering

• **Unsupervised** technique.
• Task is to identify groups of similar entities in the available data.
  – And sometimes remember how these groups are defined for later use.
People are outstanding at this
People are outstanding at this
People are outstanding at this.
Dimensions of analysis
Dimensions of analysis
Dimensions of analysis
Dimensions of analysis
How would you do this computationally or algorithmically?
Machine Learning Approaches

• Objective functions to optimize for other tasks
  – Maximum Likelihood/Maximum A Posteriori
  – Empirical Risk Minimization
  – Loss function Minimization

• What makes a good cluster?
Cluster Evaluation

• **Intrinsic** Evaluation
  – Measure the compactness of the clusters.
  – or similarity of data points that are assigned to the same cluster

• **Extrinsic** Evaluation
  – Compare the results to some **gold standard** or labeled data.
  – Not covered today.
Intrinsic Evaluation
class cluster variability

• Intercluster Variability (IV)
  – How different are the data points within the same cluster?

• Extracluster Variability (EV)
  – How different are the data points in distinct clusters?

• Goal: Minimize IV while maximizing EV

\[
\text{Minimize } \frac{IV}{EV} \\
IV = \sum_{C} \sum_{x \in C} d(x, c) \\
EV = \frac{1}{N} \sum_{i} \sum_{j} \delta(C(x_i) \neq C(x_j)) d(x_i, x_j)
\]
Similarity and Dissimilarity Between Objects

- Distances are normally used measures

- **Minkowski distance**: a generalization

  \[ d(i,j) = \sqrt[q]{|x_{i_1} - x_{j_1}|^q + |x_{i_2} - x_{j_2}|^q + \ldots + |x_{i_p} - x_{j_p}|^q} \quad (q > 0) \]

- If \( q = 2 \), \( d \) is **Euclidean distance**

- If \( q = 1 \), \( d \) is **Manhattan distance**

- **Weighted distance**

  \[ d(i,j) = \sqrt[q]{w_1 |x_{i_1} - x_{j_1}|^q + w_2 |x_{i_2} - x_{j_2}|^q + \ldots + w_p |x_{i_p} - x_{j_p}|^q} \quad (q > 0) \]
Euclidean Distance

\[ d(i, j) = \sqrt{|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \ldots + |x_{i_p} - x_{j_p}|^2} \]

Now to find the distance between two points, say the origin and the point (3,4):

\[ d_E(O, A) = \sqrt{3^2 + 4^2} = 5 \]

Simple and Fast! Remember this when we consider the complexity!
Degenerate Clustering Solutions
Degenerate Clustering Solutions
Two approaches to clustering

• Hierarchical
  – A set of nested clusters that are organized as a tree
  – Either merge or split clusters.

• Partitional
  – A simply a division of the set of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset
Partitional Clustering
Partitional Clustering
Partitional Clustering
Partitional Clustering
Partitional Clustering
Partitional Clustering
Partitional Clustering
Partitional Clustering
Partitional Clustering
Agglomerative Clustering
Agglomerative Clustering
Agglomerative Clustering
Agglomerative Clustering
Agglomerative Clustering
Agglomerative Clustering
Agglomerative Clustering
Agglomerative Clustering
Agglomerative Clustering
Agglomerative Clustering
K-Means

• K-Means clustering is a **partitional** clustering algorithm.
  – Identify different partitions of the space for a fixed number of clusters
  – Input: a value for K – the number of clusters
  – Output: K cluster centroids.
K-Means Algorithm

- Given an integer K specifying the number of clusters
- Initialize K cluster centroids
  - Select K points from the data set at random
  - Select K points from the space at random
- For each point in the data set, assign it to the cluster center it is closest to \( \arg\min_{C_i} d(\vec{x}, C_i) \)
- Update each centroid based on the points that are assigned to it
  \[ C_i = \frac{1}{|C_i|} \sum_{\vec{x} \in C_i} \vec{x} \]
- If any data point has changed clusters, repeat
How does K-Means work?

• When an assignment is changed, the distance of the data point to its assigned cluster is reduced.
  – IV is lower

• When a cluster centroid is moved, the mean distance from the centroid to its data points is reduced
  – IV is lower.

• At convergence, we have found a local minimum of IV
K-means Clustering
K-means Clustering
K-means Clustering
K-means Clustering
K-means Clustering
K-means Clustering
K-means Clustering
K-means Clustering
Potential Problems with K-Means

• Optimal?
  – Is the K-means solution optimal?
  – K-means approaches a local minimum, but this is not guaranteed to be globally optimal.

• Consistent?
  – Different seed centroids can lead to different assignments.
Suboptimal K-means Clustering
Inconsistent K-means Clustering
Inconsistent K-means Clustering
Inconsistent K-means Clustering
Inconsistent K-means Clustering
Soft K-means

• In K-means, each data point is forced to be a member of exactly one cluster.
• What if we relax this constraint?

\[ p(x, C_i) = \frac{\exp\{-d(x, c_i)\}}{\sum_j \exp\{-d(x, c_j)\}} \]
Soft K-means

- Still define a cluster by a centroid, but now we calculate a centroid as a **weighted** center of all data points.

\[
C_i = \frac{\sum_x x \cdot p(x, C_i)}{\sum_x p(x, C_i)}
\]

- Now convergence is based on a stopping threshold rather than changing assignments.
Another problem for K-Means
Next Time

• Evaluation for Classification and Clustering
  – How do you know if you’re doing well