Machine Learning and Data Mining

Clustering

(adapted from) Prof. Alexander Ihler
Overview

• What is clustering and its applications?
• Distance between two clusters.
• Hierarchical Agglomerative clustering.
• K-Means clustering.
What is Clustering?

- You can say this “unsupervised learning”
- There is not any label for each instance of data.
- Clustering is alternatively called as “grouping”.
- Clustering algorithms rely on a distance metric between data points.
  - Data points close to each other are in a same cluster.
Unsupervised learning

• Supervised learning
  – Predict target value ("y") given features ("x")

• Unsupervised learning
  – Understand patterns of data (just "x")
  – Useful for many reasons
    • Data mining ("explain")
    • Missing data values ("impute")
    • Representation (feature generation or selection)

• One example: clustering
Clustering Example

Clustering has wide applications in:

- Economic Science
  - especially market research
  - Grouping customers with same preferences for advertising or predicting sells.

- Text mining
  - Search engines cluster documents.
Overview

- What is clustering and its applications?
- **Distance between two clusters.**
- Hierarchical Agglomerative clustering.
- K-Means clustering.
Cluster Distances

There are four different measures to compute the distance between each pair of clusters:

1- Min distance
2- Max distance
3- Average distance
4- Mean distance
Cluster Distances (Min)

\[ D_{\text{min}}(C_i, C_j) = \min_{x \in C_i, y \in C_j} \| x - y \|^2 \]
Cluster Distances (Max)

\[ D_{\text{max}}(C_i, C_j) = \max_{x \in C_i, \ y \in C_j} \|x - y\|^2 \]
Cluster Distances (Average)

\[ D_{\text{avg}}(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i, y \in C_j} \|x - y\|^2 \]

Average of distance between all nodes in cluster \( C_i \) and \( C_j \)
Cluster Distances (Mean)

\[ D_{\text{means}}(C_i, C_j) = \| \mu_i - \mu_j \|^2 \]

\[ \mu_i = \text{Mean of all data points in cluster } i \text{ (over all features)} \]
Overview

- What is clustering and its applications?
- Distance between two clusters.
- Hierarchical Agglomerative clustering.
- K-Means clustering.
Hierarchical Agglomerative Clustering

Initially, every datum is a cluster

- Define a distance between clusters
- Initialize: every example is a cluster.
- Iterate:
  - Compute distances between all clusters
  - Merge two closest clusters
- Save both clustering and sequence of cluster operations as a Dendrogram.
Iteration 1

Merge two closest clusters
Iteration 2

Cluster 2  Cluster 1
Iteration 3

- Builds up a sequence of clusters ("hierarchical")

- Algorithm complexity $O(N^2)$ (Why?)

In matlab: “linkage” function (stats toolbox)
Dendrogram

Stop the process whenever there are enough number of clusters.
Overview

• What is clustering and its applications?
• Distance between two clusters.
• Hierarchical Agglomerative clustering.
• K-Means clustering.
K-Means Clustering

- Always there are K cluster.
  - In contrast, in agglomerative clustering, the number of clusters reduces.
- Iterate: (until clusters remains almost fix).
  - For each cluster, compute cluster means: (\( X_i \) is feature i)
    \[
    m_k = \frac{\sum_{i:C(i)=k} x_i}{N_k}, \quad k = 1, \ldots, K.
    \]
  - For each data point, find the closest cluster.
    \[
    C(i) = \arg \min_{1 \leq k \leq K} \| x_i - m_k \|^2, \quad i = 1, \ldots, N
    \]
Choosing the number of clusters

- Cost function for clustering is:
  \[ C(z, \mu) = \sum_i \| x_i - \mu_{z_i} \|^2 \]

  what is the optimal value of \( k \)?
  (can increasing \( k \) ever increase the cost?)

- This is a model complexity issue
  - Much like choosing lots of features – they only (seem to) help
  - Too many clusters leads to over fitting.

- To reduce the number of clusters, one solution is to penalize for complexity:
  - Add \((\# \text{ parameters}) \times \log(N)\) to the cost
  - Now more clusters can increase cost, if they don’t help “enough”
Choosing the number of clusters (2)

- The Cattell scree test:

Scree is a loose accumulation of broken rock at the base of a cliff or mountain.

The best K is 3.
2-means Clustering

• Select two points randomly.
2-means Clustering

- Find distance of all points to these two points.
2-means Clustering

- Find two clusters based on distances.
2-means Clustering

- Find new means.
- Green points are mean values for two clusters.
2-means Clustering

- Find new clusters.
2-means Clustering

- Find new means.
- Find distance of all points to these two mean values.
- Find new clusters.
- Nothing is changed.
Summary

- Definition of clustering
  - Difference between supervised and unsupervised learning.
  - Finding labels for each datum.
- Clustering algorithms
  - Agglomerative clustering
    - Each data is a cluster.
    - Combine closest clusters
    - Stop when there is a sufficient number of cluster.
  - K-means
    - Always K clusters exist.
    - Find new mean value.
    - Find new clusters.
    - Stop when nothing changes in clusters (or changes are less than very small value).