Clustering

Sravani Vadlamani

Schedule

Date	Topic
April 12	Principal Component Analysis, Examples and Applications
April 14	Clustering Methods, K-means algorithm
April 19	Final Exam Review
April 21	Introduction to Deep Learning
April 26	Final Exam in Class
May 5 (8 – 10 AM IST 1017)	Final Project Presentations

Final Project - Deadlines

Date	Deliverable
March 22	Project Proposal Due
April 3	Exploratory Data Analysis
April 10	Analysis/Results
April 17	Draft Final Report
April 27	Final Report Due
April 27	Final Presentation Due

Unsupervised Learning

- Agenda
  - What is Unsupervised Learning?
  - Principal Component Analysis
  - Clustering

Supervised vs Unsupervised Methods

- Supervised Learning
  - Goal is to predict the response variable Y using a set of features X_1, X_2, ….X_p measured on n observations
  - Examples include regression and classification methods
- Unsupervised Learning
  - There is no response variable, and we are not interested in prediction
  - The goal is to discover interesting subgroups or patterns among the variables or among the observations. Another way of visualizing the data

Unsupervised Methods

- Two Methods
  - Principal Component Analysis
    - A useful tool for data visualization and data pre-processing before applying supervised techniques
  - Clustering
    - A range of methods to discover unknown subgroups in the data

Challenges of Unsupervised Methods

- Supervised methods include well developed set of tools to estimate trained models and assess their quality using test sets and cross validation techniques
- Unsupervised methods are more subjective as there is no specific goal like prediction of a response variable. They are often performed as part of exploratory data analysis and the results are hard to assess.
- Unsupervised methods are used in various fields
  - Subgroup breast cancer patients by their gene expression measurements
  - Group shoppers by their browsing and purchase histories
  - Group movies by ratings assigned by viewers

Clustering

- The task of grouping a set of objects in a way that objects in the same group (called a cluster) are more similar (in some sense or other) to each other than to those in other groups (clusters)
- We seek to partition observations into distinct groups such that observations within each group are quite similar while observations in different groups are quite different from each other
- How do we define observations to be similar or different?

Clustering Algorithms

- Many exist but two most popular
  - Ward's Hierarchical Clustering
    - Do not know in advance how many clusters we want
    - Bottom-up approach (grouping similar observations together)
    - Results in a tree-like visual representation of the observations called a dendogram
  - K-means Clustering
    - Seek to partition the observations into a pre-specified number of clusters
    - Top-down approach

Hierarchical Clustering

- Very simple algorithm
- Start by defining some sort of dissimilarity measure between each pair of observations like Euclidean distance
- Iterative algorithm
    - Each of the n observations is treated as its own cluster
    - The two clusters that are most similar are then fused resulting in n-1 clusters
    - Repeat above step until all the observations belong to one single cluster

Hierarchical Clustering

- The concept of dissimilarity between a pair of observations needs to be extended to a pair of group of observations
- This is achieved by developing the notion of linkage, which defines the dissimilarity between two groups of observations.
- Four most common types of linkage – complete, average, single, centroid

Linkages

- Complete
  - Maximal inter cluster dissimilarity
  - Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B and record the largest of these dissimilarities
- Single
  - Minimal inter cluster dissimilarity.
  - Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B and record the smallest of these dissimilarities
  - This can result in extended, trailing clusters in which single observations are fused one-at-a-time

Linkages

- Average
  - Mean inter cluster dissimilarity
  - Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B and record the average of these dissimilarities
- Centroid
  - Dissimilarity between the centroid for cluster A and the centroid for cluster B
  - Can result in undesirable inversions

Euclidean Distance

- Square root of sum of squares of difference on x plus sum of squares of differences on y
- Biased by variables with a larger scale and hence standardization of data is important

Dissimilarity Measures

- Choice of dissimilarity measure is very important as it has a strong effect on the resulting cluster
- Correlation-Based Distance
  - Alternative dissimilarity measure
  - Considers two observations to be similar if their features are highly correlated, even though the observed values may be far apart in terms of the Euclidean distance

Dendrogram

- Each leaf of the dendrogram represents one of the n observations
- Moving up the tree, leaves begin to fuse into branches corresponding to observations that are similar to each other
- Moving higher up, branches fuse with leaves or other branches
- The earlier (lower in the tree) fusions occur, the more similar the groups of observations are to each other
- Observations that fuse later (near the top of the tree) can be quite different

Dendrogram

- The height of the fusion, as measured on the vertical axis indicates how different the two observations are
  - Observations that fuse at the very bottom of the tree are quite similar to each other
  - Observations that fuse close to the top of the tree will tend to be quite different
- Draw conclusions about the similarity of two observations based on the location on the vertical axis where branches containing those two observations first are fused

Identifying Cluster in Dendrogram

- Make a horizontal cut across the dendrogram
- The distinct sets of observations beneath the cut can be interpreted as clusters
- Hence a dendrogram can be used to obtain any number of clusters
- Researchers often look at the dendrogram and select by eye a sensible number of clusters based on the heights of the fusion and the number of clusters desired
- The choice of where to cut the dendrogram is not always clear

K-means Clustering

- The desired number of clusters K is first specified
- The K-means algorithm will assign each observation to exactly one of the K clusters
- C_1, …..C_k denote sets containing the indices of the observations in each cluster satisfying two properties
  - Each observation belongs to at least one of the K clusters i.e.,
    - $C_1 \cup C_2 \cup \ldots\ldots C_k = \{1, 2, \ldots.n\}$
  - Clusters are non overlapping i.e., no observation belongs to more than one cluster
    - $C_k \cap C_{k'} \neq 0$ for all $k \neq k'$

Goal of K-means Clustering

- A good cluster will have smallest within-cluster variance
- The within-cluster variation for a cluster C_k is a measure $W(C_k)$of the amount by which the observations within a cluster differ from each other.
- Goal is to partition the observations into K clusters such that the total within-cluster variation summed over all K clusters is as small as possible

Definition of Within-Cluster Variation

- Most commonly used metric is the square of the Euclidean distance

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i\prime \in c_k} \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2$$

- $|C_k|$ denotes the number of observations in cluster k

K –means Clustering Algorithm

- Randomly assign a number from 1 to K to each of the observations. These serve as initial cluster assignments for the observations
- Iterate until the cluster assignments stop changing
  - For each of the K clusters, compute the cluster centroid. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster
  - Assign each observation to the cluster whose centroid is closest (closest is defined using Euclidean distance)

Local Optimum

- K-means algorithm will find a local optimum rather than a global optimum
- Results obtained will depend on the initial (random) cluster assignment of each observation
- It is important to run the algorithm multiple times from different random starting values and then select the best solution

Issues with Clustering

- Determining the number of clusters to retain
- Cross validation of clusters and cluster sizes
- All or none decision process (either in or out of a cluster)
- What happens to observations that don’t belong to any cluster
- Consequences of choices among linkage, dissimilarity measure, cutting dendrogram
- Sensitivity to data perturbations

Recommendations

- Perform clustering with different parameter choices and look at the full set of results to identify patterns
- Cluster subsets of the data in order to get a sense of the robustness of the clusters obtained
- Careful about how the results are reported
  - Results should not be taken as the absolute truth
  - Constitute a starting point for the development of a scientific hypothesis and further study, preferably on an independent data set