Notes

Association

quality metrics

Support is the fractio nof transaction containing the items #{itemsList}/n.of transaction

Confidence is the frequency of x in transaction containing x sup(x,..)/sup(x)

Given a set of transaction T association rule mining is the exctration of rules that satisfy the constraints:

1. support >= minsup threshold
2. confidece >= minconf threshold

The result is complete when all the rule satisfy the constraint, it is correct only some rules satisfy both

Requent itemset generation

It is computational expensive if brute force is used

Apriori principle

The support of an itemset can never exceed the support of any of is subsets, that means that if a subset is unfrequent all the superset of it will be

Algorithm

Level-based approach: at each iteration extracts itemsets of a given length k Two main steps for each level:

1. Candidate generation
   a. Join Step generate candidates of length k+1 by joining frequent itemsets of length k
   b. Prune Step apply Apriori principle: prune length k+1 candidate itemsets that contain at least one k-itemset that is not frequent
2. Frequent itemset generation
   a. scan DB to count support for k+1 candidates
   b. prune candidates below minsup

The issue is that candidate sets may be huge

Clustering

Clustering is finding group of objects such that the objects in a group will be similar or relato to one another and different or unrelated from the objects in other groups.

Partitional clustering A division data objects in not overlapping subsset such that each object in exactly in one subset **Hierarichical clustering** A set of nested cluster organize as a hierarchical tree

exclusive vs non-exclusive Cannot/can belong to multiple cluster **fuzzy vs non-fuzzy** a points belong to every cluster with some weight between 0 and 1 **partial vs complet** cluster a part/all of the data **Heterogeneous vs. homogeneus** Cluster with different/similar size shape density

Type of cluster

1. Well-Separated Clusters: A cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.
2. Center-based A cluster is a set of objects such that an object in a cluster is closer (more similar) to the "center" of a cluster, than to the center of any other cluster
3. Contiguous Cluster (Nearest neighbor or Transitive) A cluster is a set of points such that a point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster.
4. Density-based A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density. Used when the clusters are irregular or intertwined, and when noise and outliers are present.

K-means clustering

Partitional clustering approach 1. Each cluster is associated with a centroid (center point) 2. Each point is assigned to the cluster with the closest centroid 3. Number of clusters, K, must be specified

```
Select K points as th einitial centroids
repeat
  Form K cluster by assigling all point to the closest centroid
  Recompute the centroid for all the cluster
until the centroids do not change
```

The initial centroid are often chosen randomly and the closeness is measured by euclidean distance cosine similaritycorrelation etc.. Most of the convergense happens in few iteraations. The choice of the initial centroid it is quite importatnt to obtain a meaningful result. The most commont measur is Sum of Squared Error(SSE) where we sum the quared distance of each point from the nearest cluster

To set K we can use the elbow or knee approack where we plot the K vs SSE to identify when the gain of adding a centroid is negligible

Pre and post processing 1. Pre-processing: Normalize the data and eliminate outliers 2. Post-processing: Eliminate small cluster (outliers), split loose cluster (high SSE), Merge close cluster (low SSE)

Hierarchical clustering

Starts with a cluster for each point and merge themm according to the **inter-cluster* similarity that can be evaluated as : 1. MIN 2. MAX 3. Group avereage 4. Distance between centroids 5. Other…

The problems are that that once that two cluster are merge they can not be unmerged, no objective function ids directlu miniomized, some apporach are sensityve to noise and outliers other cannot handle differet sized cluster and convex shapes or they break large clusters

DBScan

It is a density based algorithm where the density is the number of points within a specified radius eps. A point is a **core point** if it has more than a specified number of points (MinPts) within Eps. A **border point** has fewere than MinPts within Eps but is in the neighbourood of a core point whiel a **noise point** is a point that is neither core nor border The algorithm for DBScan aims to eliminate noise point an perform clustering on the remaining ones. It is resistant to noide and can handle cluster of different shapes and sizes. It does not work well when we have high dimensional data and varying densities

Cluster validity

Internal measures: 1. Cluster cohesion: measure how closely related are the objects within a cluster 2. Cluster separation: measure how distinct or well separated a cluster is from othe clusters

Clssification

The objective of classification is thep prediction of a class label by defining an **intepretable** model of a fiven phenomenon. To do so there are different approachees :

1. decision trees
2. bayesian classification
3. classification rules
4. random forest
5. neural networks
6. k-nearest neighbours