Clustering With K Means

Objective:

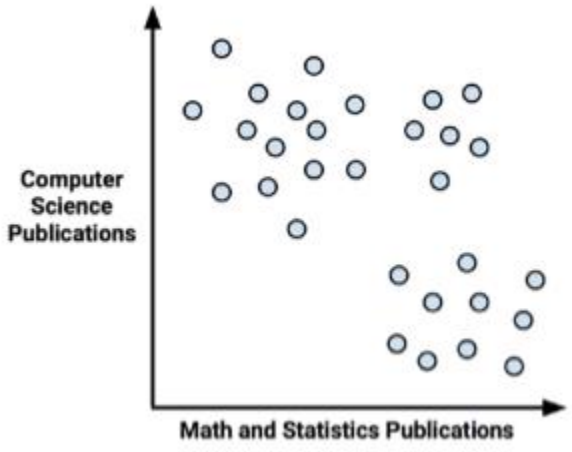
- ❖ Comprehending Clustering as a unsupervised Learning Algorithm.
- ❖ Data is segmented by a similarity criterion.

Clustering

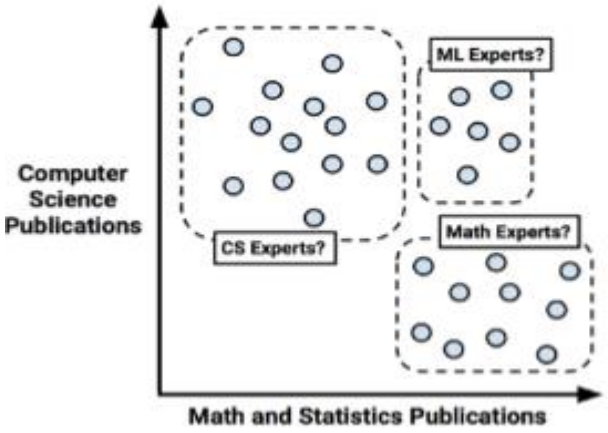
- This is an unsupervised classification technique creates groups with similarities.
- This technique is not for prediction but for exploratory knowledge discovery and envisioning inherent pattern
- The data segmentation creates homogeneous sub groups which results in alleviating the complexity of the data.
- Clustering can intuitively be considered as unsupervised classification
- The classification generated has to be interpreted and deciphered as actionable data patterns

Visualization

- If hypothetically a researcher wishes to classify a group of scientist as per the diverse domains.



- Patterns can be observed in the data and these have to be unraveled.
- Domain specific knowledge is required to distinguish the one pattern from the other.
- Ideally it is preferred if the boundaries were created by ML algorithm on the basis of similarity measures rather than by the subjective perspective of humans.



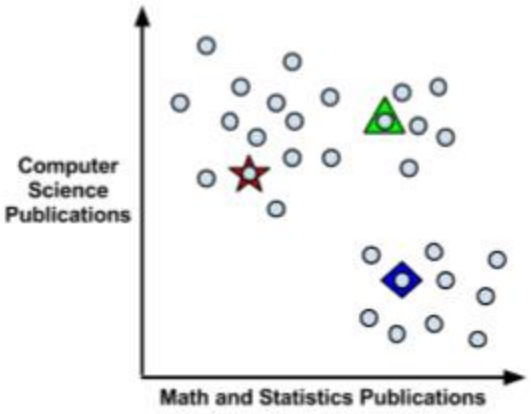
- Sometimes a two fold process is used where in the first phase the class labels are created by clustering and then a supervised learner like decision trees is used to find the most important predictors from these predictors. This will be under the umbrella of semi supervised classification.

K Means Clustering:

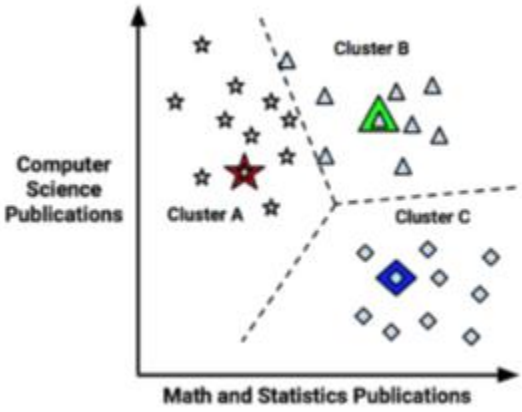
- K-means is a very popular clustering technique which is prevalently used in the contemporary context
- K-means does not use very complex statistical methods and is more simplistic as compared to other complicated clustering techniques.
- It is flexible and gives good outcomes by adjustments.
- Due to inherent randomness it is not assured that optimal clusters will be obtained.
- To determine the number of clusters a reasonable guess has to be made.
- This clustering technique ie k-means cannot be used for non-spherical clusters.
- K-means algorithm take all the n examples and divides them into k clusters.
- For a good classification both the number of examples n and k clusters should be large in order to obtain optimal clusters.
- The kmeans starts with an initial guess then heuristically ascertains the local optimal clusters which is iteratively improved to finally obtain homogenous clusters.

Distance Similarity Metric Algorithm

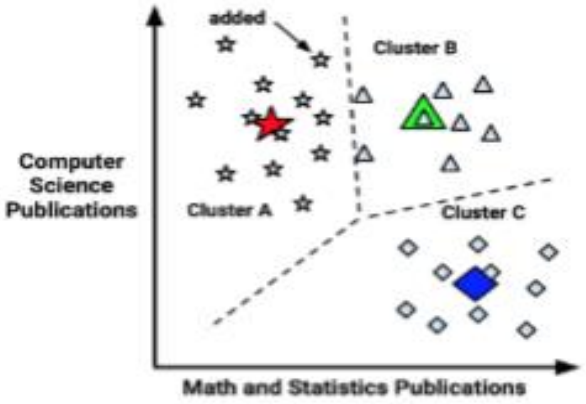
- k-means represent each feature as a dimension therefore usually 10 features implies that the space is a 10 dimensional space.
- For instance for the research paper illustration the data was represented by a two dimensional data.
- If k=3 then initially the algorithm randomly assigns three centers for the clusters.



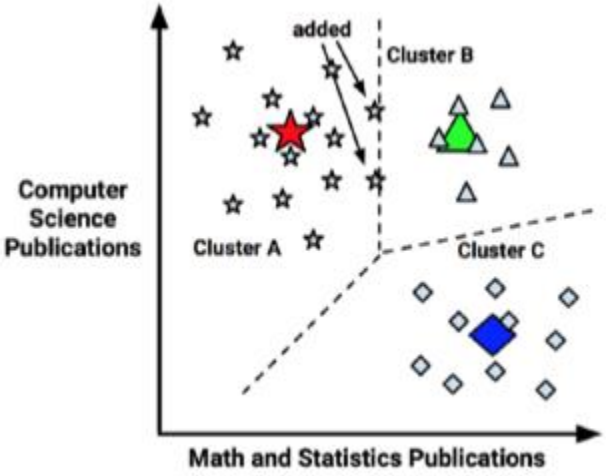
- K means is very sensitive to the starting center therefore sometimes optimal centers are not obtained.
- To offset this issue many solutions have come up like initial centers need not be a value from the data examples themselves. K-means++ improves the performance by a mathematical algorithm that allows for achieving optimality.
- Once the cluster center is determined the examples are assigned to the appropriate cluster depending on the their distance from each of the centers.
- $dis(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
  This Euclidean distance formula is used to find the distance between each example and each center. Each example is assigned to the center which has the smallest distance from it.



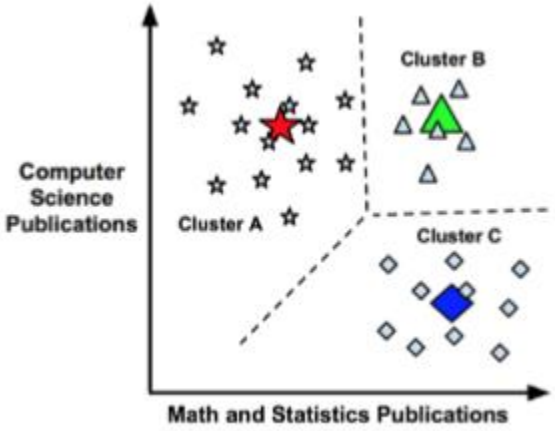
- The three clusters formed are separated into Cluster A and Cluster B and Cluster C. The cluster boundaries are called Voronoi Diagram represents areas closed to a center. The vertex of the three clusters is mathematically farthest from the three centers.
- Subsequent to the initial phase the new center is updated by centroid calculated from the current examples assigned to each cluster.
- The distance measurement is now recalculated to reassign all the examples to the appropriate cluster.



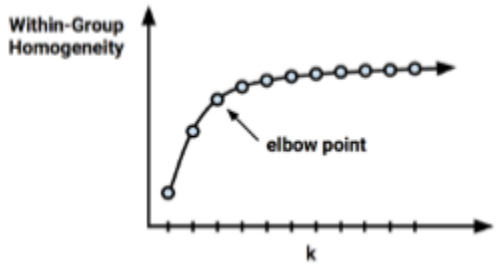
- The process works recursively and reassignment of centers and examples takes place.

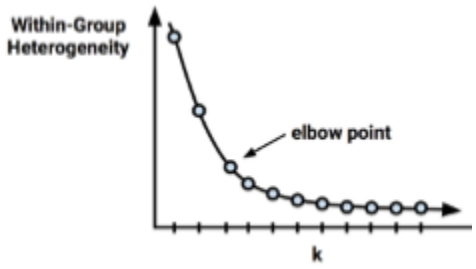


- Once the reassignment becomes stable and does not change then the cluster assignments are finals.



- Finally cluster centers are finally determined and the examples are fully segmented.
- K-means is sensitive to the number of clusters. If k is too large then the clusters are more homogeneous but the data can get overfitted.
- Mostly a priori knowledge can facilitate the decision making for the number of clusters. For examples we would decide on the number of clusters on the basis of movie genres for a movies dataset.
- The decision of k could be determined by business decisions.
- A well implemented technique called Elbow method helps to determine the size of k based upon the criterion of increasing homogeneity and decreasing heterogeneity.
- The objective is to determine the balance point beyond which the homogeneity or heterogeneity does not change.





- Numerous Statistical techniques are used to create the elbow graph to provide insights into the data.