Weekly report of lessons

Name: Suryam Arnav Kalra
Roll No: 19CS30050
The week: 27th September 2021 to 1st October 2021

The topics covered:

- Determining the value of K
- Cluster Validity Indices
- Stability check based clustering
- Generalizing K-means
- Mixture of Gaussians
- Expectation Maximization Algorithm
- Hierarchical clustering
- Hierarchical clustering algorithm
- Clique graphs
- Transforming graph to a clique graph
- Corrupted clique problem
- Distance graphs
- Parallel classification with cores (PCC algorithm)
- CAST
- DBSCAN

Summary topic wise:

- Determining the value of K: We can use cluster validity index and stable clustering results with random initialization to get a good value of K.
- Cluster Validity Indices: We can use external indices using a reference partition information as well as internal indices by looking at variance distribution.
    - Silhouette Index: $s(x) = (a(x) - b(x))/\max(a(x), b(x))$ where a(x) is the average distance of points within the cluster from x and b(x) is the min average of points of other clusters from x.
    - Calinski-Harabasz (CH) index: $CH(K) = \frac{(J(1) - J(K))/(K-1)}{J(K)/(n-K)}$ where J(i) is the SSE with K = i.
    - Normalized Mutual Information (NMI): NMI = 2I(Y;C)/(H(Y)+H(C)) where I(Y;C) = H(Y) – H(Y|C)
    - Fraction of same pairs in same clusters (FM index)
- Stability Check based clustering: We can use repeated clustering having similar partition and **Wang's method** of cross-validation.
- Generalizing K-means: $P(x) = \sum_{i=1}^{K} P(x|G_i)P(G_i)$ where K is the number of components (a hyper-parameter), G_i defines the i^{th} group.
- Mixture of Gaussians: Each cluster center is augmented by a covariance matrix, whose values are re-estimated from corresponding samples using the **Mahalanobis** distance function.
- Expectation Maximization Algorithm: Start with an initial set : $\{\pi_k, \mu_k, \sum_k\}$
    - E-Step (Expectation stage): Compute probability of x belonging to k^{th} cluster and assign x to the m^{th} cluster whose probability is maximum.

    - M-Step (Maximization stage): Re-estimate parameters $\{\pi_k, \mu_k, \sum_k\}$ from class distribution and iterate these two steps till it converges.
- Hierarchical clustering: Build hierarchy of groups (bottom – up approach) and uses a distance matrix among the samples.
- Hierarchical clustering algorithm: The algorithm takes a nXn distance matrix d of pair-wise distances between points as an input and forms n clusters each with one element, then it finds two closest clusters C1 and C2 and merges them together.
- Clique graphs: A **clique** is a graph with every vertex connected to every other vertex. A **clique graph** is a graph where each connected component is a clique.
- Transforming a graph to a clique graph: By addition or removal of edges a graph can be made into a clique graph.
- Corrupted cliques problem: The smallest number of additions and removals of edges that will transform G into a clique graph. It is a NP-hard problem.
- Distance Graphs: Feature vectors represented as vertices in the graph. Choose a threshold distance θ and if the distance is less than θ draw an edge between them.
- PCC algorithm: Let S be a set of n elements and G be a distance graph and k be the number of clusters.
    - Randomly select S' a subset of S and S'' a subset of S - S'.
    - For all k partitions in S' , obtain extended partition in S through two stages of extension and choose one which has the maximum score.
- CAST: It is a practical and fast algorithm based on the notion of features close to cluster C or distant from cluster C. It finds the nearest close feature not in C and adds it to C.
- DBSCAN: It requires no explicit computation of distance graph and grows regions of connected core points from a seed.

Concepts challenging to comprehend:

Hierarchical clustering and PCC Algorithm are a little bit challenging to comprehend.

Interesting and exciting concepts:

Clique graphs are quite interesting and exciting to learn.

Concepts not understood:

After going through the book and the video lectures the concepts are clearly understood.

Any novel idea of yours out of the lessons:

Clique graph can be used as a technique in Naïve Bayes estimator as well. We can represent the relationship between variables and their interdependency using clique graphs. The connected components of clique can represent the nodes which are strongly dependent on each other the interconnection between different clique graph connected components can show the dependency of these highly dependent features.