Machine Learning (CS60050) – Weekly Report

Kaushal Banthia (19CS10039)

Week 7: 22th – 24th September, 2021

Topics Covered:

- Supervised Learning vs Unsupervised Learning
- Clustering
- K-means clustering
- Strengths and Weaknesses
- Various Initialization Approaches
- K – Means++

Summary (Topic Wise):

- Supervised Learning vs Unsupervised Learning

  - Supervised Learning entails learning with labeled data, to learn a mapping from the input to an output (labels provided by a supervisor).
  - Unsupervised Learning is learning from only input data. No labels of instances are available. No supervisor present to provide mapping between input and output. (The aim is to find the regularities / structures / patterns in the input).

- Clustering

  - Clustering is the task of organizing objects into groups (clusters) whose members are similar in some way or the other. Clusters are a collection of objects similar to each other, but dissimilar to the objects belonging to other clusters. They might be regions of homogeneity in an image or they might be similar kind of people with similar behaviors. Basically, clustering involves grouping of similar components.
  - There is a difference between a class and a cluster. A class is a well-studied group of objects identified by their common properties or characteristics, whereas a cluster is a group with a loosely defined similarity among the objects.
  - A cluster has the potential to form a class.
  - We perform clustering due to the following reasons:
    - Finding representatives for homogeneous groups (To reduce data).
    - Discovering natural groups or categories (To describe by their unknown properties) or finding relevant groups.
    - Detecting unusual data objects (These unusual data objects are called as outliers.

- K – means Clustering

  - Given N d – dimensional data points, K – means clustering involves the computation of K partitions (clusters) in them, so that it minimizes the sum of square of distances between a data point and the center of its respective partition (cluster).
  - $E = \sum_k \sum_{\forall x \in c_k} ||x - c_k||^2, where\ c_k = \frac{1}{|C_k|} \sum_{\forall x \in C_k} x$

- The optimization problem at hand is the minimization of E (Sum of squared errors, abbreviated as SSE).
- We can try out K – means exhaustively, but there is an issue with that.
  - The number of ways a set of N objects partitioned into K non-empty groups turns out to be the Stirling Number of the second kind.
  - $S(N,K) = \frac{1}{K!}\sum_{i=0}^{K}(-1)^{K-i} * \binom{K}{i} * i^N \approx \frac{K^N}{K!}$
  - Checking all the possible combinations is prohibitive, due to exponential order of the input size. Thus, for K > 1, it is an NP – hard problem.
- To avoid this, we try out the Lloyd's Algorithm (also called as Batch K – means).

  - `Step 1: Given K initial centers, assign a point to the cluster represented by its center, if it is the closest among them.`
  - `Step 2: Update the centers.`
  - `Step 3: Iterate steps 1 and 2, till the centers do not change their positions.`

- We can also try a more conservative approach, since the Lloyd algorithm maybe fast, but not necessarily giving better convergences. A more conservative approach would be to move one data point at a time, provided the overall cost reduces.
- This can be done greedily by transferring a data point from one class to another, which causes the maximal cost reduction at that step.

- <u>Strengths and Weaknesses</u>

  - Strengths
    - Convergence guaranteed at a quadratic rate.
    - Linear time complexity in N, d and K.
    - Versatile, simple and invariant to the data ordering.

  - Weaknesses
    - Only detects well separated, compact, hyper spherical clusters.
    - Value of K needs to be provided
    - Sensitive to noise and outlier points (due to squared Euclidean distance).
    - May get stuck at local minima, as it is highly sensitive to the selection of the initial centers. Also, improper initialization may lead to empty clusters, slower convergence, and a higher chance of getting stuck in bad local minima.

- <u>Various Initialization Approaches</u>

  - Assign each point randomly to one of the clusters
  - First K points are selected as the centers (This is sensitive to data ordering. Thus, it is better to choose them randomly. However, outliers may still get selected).
  - Repeated K-means is a K-means on J random subsets. It involves merging all centers and running K-means repeatedly on them. We then have to choose the best set of centers, minimizing the error, and use them for iterative convergence.

- <u>K – Means++</u>

  - In K – Means++, the first center c_1 is chosen randomly.
  - The i^{th} center c_i $(for\ i = 2, 3, \dots, K)$ is chosen as x' with a probability proportional to the square of the minimum distance from the selected $i - 1$ centers.
  - $p(x') = \frac{min_{j=1,2,\dots,i-1}||x' - c_j||^2}{\sum_x min_{j=1,2,\dots,i-1}||x - c_j||^2}$

Concepts Challenging to Comprehend: None yet.

Interesting and Exciting Concepts: K – means Clustering

Concepts not understood: None yet.

A novel idea: We can set the initialization of the K cluster centers by either having a manual look at the dataset (if the data can be easily visualized), so that the error margin is reduced and the convergence is quicker. We could also do the initialization by setting the K cluster centers to K data points from within the dataset (without replacement). That would make sense, as the centers would be closer to the accurate centers, than if they were randomly assigned any value.