INTERPRET RESULT & ADJUST CLUSTERING

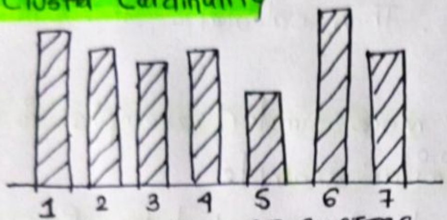
→ Because clustering is unsupervised, no "truth" is available to verify results. The absence of truth complicates assessing quality.

So to interpret the results, do following steps —

Step 1 — Quality of clustering

- Checking the quality of clustering is not a rigorous process because clustering lacks "truth".
- Normally 3 metries we check — 1) Cluster cardinality 2) Cluster magnitude 3) Performance of downstream system.

1) Cluster Cardinality



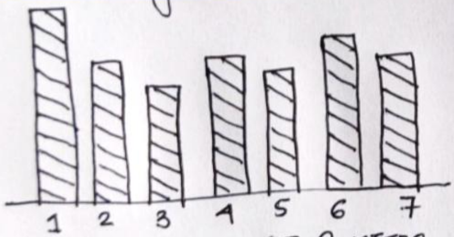
CARDINALITY OF CLUSTERS

→ Cluster cardinality is the number of example per cluster. Plot the cluster cardinality for all clusters and investigate clusters that are major outlier.

Investigate cluster number 6.

→ Cardinality means number of elements in a set. High Cardinality means lots of unique value. A column with hundred of zip codes is an example of high cardinality.

2) Cluster magnitude

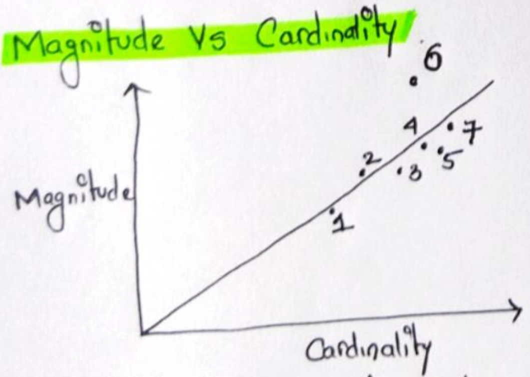


MAGNITUDE OF CLUSTERS

→ Cluster magnitude is the sum of distances from all the examples to the centroid of Cluster. similar to the ~~magnitude~~ cardinality, check magnitude varies across the cluster and investigate anamolies.

Investigate cluster number 1.

Magnitude Vs Cardinality



Cardinality vs Magnitude of clusters

→ Normally higher cluster cardinality tends to result in a higher cluster magnitude, which intuitavely make sense.

→ Clusters are anamalous when cardinality doesn't correlate with magnitude relative to the other clusters.

→ Find anamolous clusters by plotting magnitude against cardinality.

→ Fitting a line to cluster metrics shows that cluster number 6 is anomalous.

Anamoly refers to identification of items or event that do not conform to an expected pattern or to do items present in the dataset.

3) Performance of Downstream System

Since clustering output is often used in downstream ML systems, check if downstream system's performance improves when your clustering process changes. Impact will be quality of clustering.

Question to Investigate if problem are found - 1) Data Scaled? 11) Similarity measure correct? 3) Algorithim assumption match the data?

Step 2 — Performance of the Similarity measure

→ Simplest check is to identify pairs of examples that are known to be more or less similar than other pairs. Then calculate the similarity measure for each pair.

→ Ensure that the similarity measure for more similar examples is higher than the similarity measure of less similar example.

→ Measurement of similarity between 2 objects is computed by distance function/metric. For example Mahhatten distance / Euclidean distance. etc.