upGrad

#LifeKoKaroLift

Data Science Certification Program

02/06/19

upGrad

Course : Machine Learning

Lecture On : Cluster Analysis

Instructor : Sandeep Muttha

Poll 1

Clustering is a form of supervised learning model

1. True
2. False

Poll 1(Answer)

Clustering is a form of supervised learning model

1. True

2. **False**

There is no target variable in the data used for training a cluster model, which makes it an unsupervised learning model.

upGrad

Agenda

1. Cluster Analysis
2. K-means, K-means ++
3. Hierarchical clustering
4. More clustering algorithms; Pros & Cons
5. Demos in python

What is cluster analysis?

upGrad

- **Grouping ‘similar’ objects into sets.**
  - In Data Science, optimally classifying data into ‘k’ clusters
  - Similarity determined by one or more parameters
  - It is an unsupervised learning algorithm, for verifying hypothesis on data distribution, and sometimes for data pre-processing on large datasets

Types of clustering and approaches..

upGrad

There two types of clustering, in general

- **Hard clustering:** Exclusive (each item belongs to only one cluster)
- **Soft clustering:** Not exclusive (items may belong to multiple clusters)

Two approaches to clustering

- **Agglomerative (bottoms-up):** Begins with single observation in a cluster and merges clusters till a stopping criteria is reached
- **Divisive (tops-down):** Begins with all observations in a single cluster and performs splitting until a stopping criteria is reached

Different clustering algorithms..

upGrad

- **Connectivity based:**
  - Data-points connected based on distance to each other. Different clusters are formed at different distance thresholds. Represented by Dendograms which look like a hierarchy (called hierarchical clustering)

- **Centroid based:**
  - This is an iterative algorithm. Clusters are formed based on how close the data-point is to the centroid of the cluster. K-means is a centroid based algorithm.

Different clustering algorithms..

upGrad

- **Distribution based:**
  - Clusters are formed based on the probability that the data-point can fit into a certain (probability) distribution. Data-points in a cluster belong to same distribution. Eg. Gaussian Mixture Models, based on expectation maximization.

- **Density based:**
  - This approach scans/searches the n-dimensional data space to measure the density of data-points. Clusters are identified as data-points belonging to spaces with high-density. The spaces with low density are identified as borders for the cluster. DBSCAN is an example of density based clustering algorithm.

Pre-processing..

upGrad

- **Scaling**
  - Observations are clustered based on the distance between each other and the distance to clusters. Scaling is advised to bring all features on a comparable scale, as distance between observations is highly susceptible to the scale.
  - Eg:
    - Jack: 6ft, 70kg; Jill: 5ft, 69kg; Johny: 5.5ft, 65kg
    - Use standard scaler, min_max scaler, etc.
- **Missing value imputation**
  - Impute missing, null, inf, values
  - Remove/impute

Poll 2

Which of the following are clustering problems:

1. Document classification into sets based on the text content, tags, titles

2. Identifying potential defaulters in a pool of loan applicants

3. Customer segmentation for telecom company based on usage patterns

4. Predicting the total ticket sales of the Avengers movie

Poll 2

Which of the following are clustering problems:

1. Document classification into sets based on the text content, tags, titles
2. Identifying potential defaulters in a pool of loan applicants
3. Customer segmentation for telecom company based on usage patterns
4. Predicting the total ticket sales of the Avengers movie

upGrad

#LifeKoKaroLift

K-means clustering

The K-means clustering algorithm

upGrad

1. K centroids are initialized at random for K clusters
2. Assign the data-points to the nearest cluster based on a distance metric
3. Recalculate the centroids of the clusters
4. Repeat step 2 for all data-points with the new centroids calculated in step 3
5. Repeat the process until a stopping criteria is reached or there is no change in assignment of data-points to clusters in successive iterations

K-means++

upGrad

K-means allots cluster centers randomly

K-means++ only allots only one cluster center randomly and searches other centers given the cluster center one.

- Converges faster due to better initialization of the cluster centers
- Performs better as it is less likely to converge at local optimum

Evaluating the clusters..

upGrad

Silhouette score:

- Silhouette value is a measure of how similar a data-point is to own cluster than to other clusters
- Can be calculated based on any distance metric (euclidean, manhattan, etc.)

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

a(i) avg distance to points in own cluster
b(i) avg minimum distance to points on other clusters

- Mean Silhouette coefficient is calculated over all samples after clustering
- Ranges from -1 to 1
- 1 indicates good clustering and 0 indicates overlapping clusters.

from sklearn.metrics import silhouette_score

Cluster tendency..

upGrad

Hopkins statistic:

To measure the cluster tendency of a data-set

$$H = \frac{\sum_{i=1}^{m} u_i^d}{\sum_{i=1}^{m} u_i^d + \sum_{i=1}^{m} w_i^d}$$

u_i, the distance of $y_i \in Y$ from its nearest neighbour in X, and
w_i, the distance of $x_i \in X$ from its nearest neighbour in X.

- Ranges from 0 to 1
- Value of 1 indicates highly clustered data, 0.5 for randomly generated data, and 0 for uniformly distributed data

Optimum number of clusters..

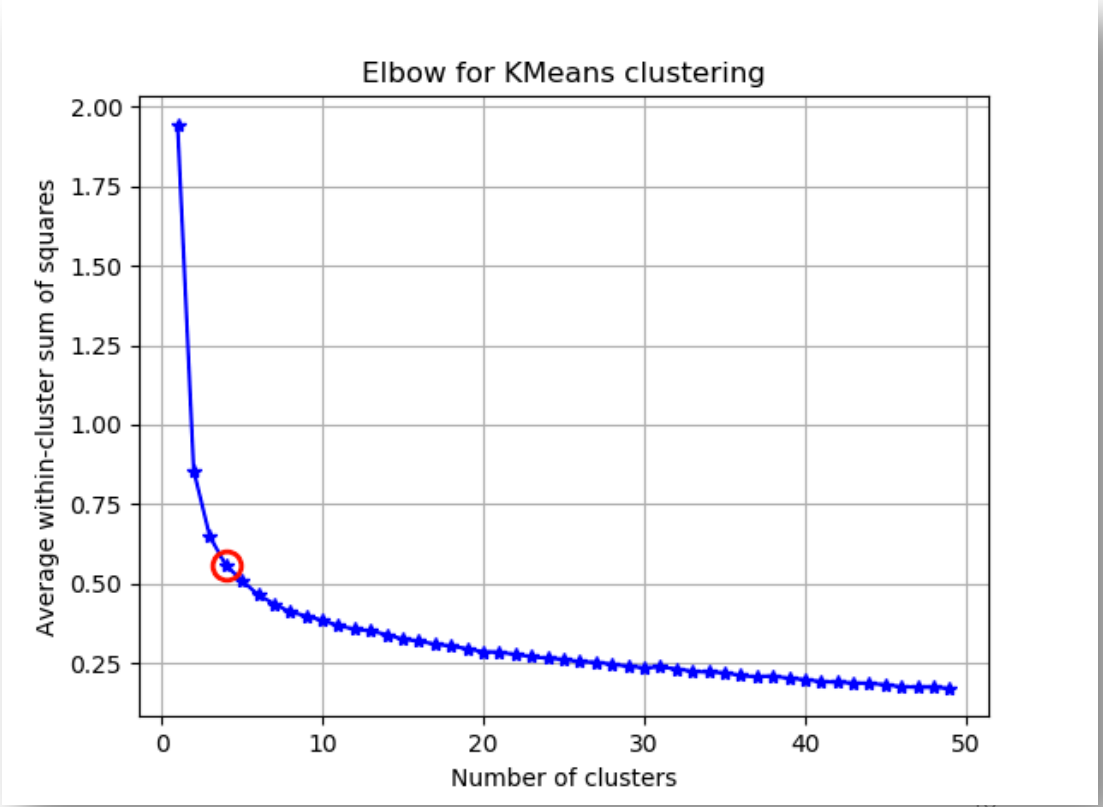
upGrad

Elbow curve:

Plot SSE (sum of squared distances of points from the cluster centers) for different cluster sizes

Identify the elbow in the curve to determine the optimal number of clusters

*KMeans().intertia_*



upGrad

#LifeKoKaroLift

Hierarchical clustering

Hierarchical clustering

upGrad

It is an agglomerative approach, where we combine smaller clusters starting from each observation till the optimal clusters are formed.

Key questions:

- How do you represent the clusters?
- How to measure the nearness of the clusters?
- What do we use as a stopping criteria?

The hierarchical clustering algorithm

upGrad

1. Each point starts in a cluster of its own (k clusters)
2. Create a distance matrix, with distance of each cluster from the rest of the clusters
3. Merge the closest pair of clusters to create a new cluster and compute the center of the new cluster
4. Re-create the distance matrix in step 2 (for k-1 clusters)
5. Repeat step 3, and continue until there is only one cluster

Distance metric & Linkage methods

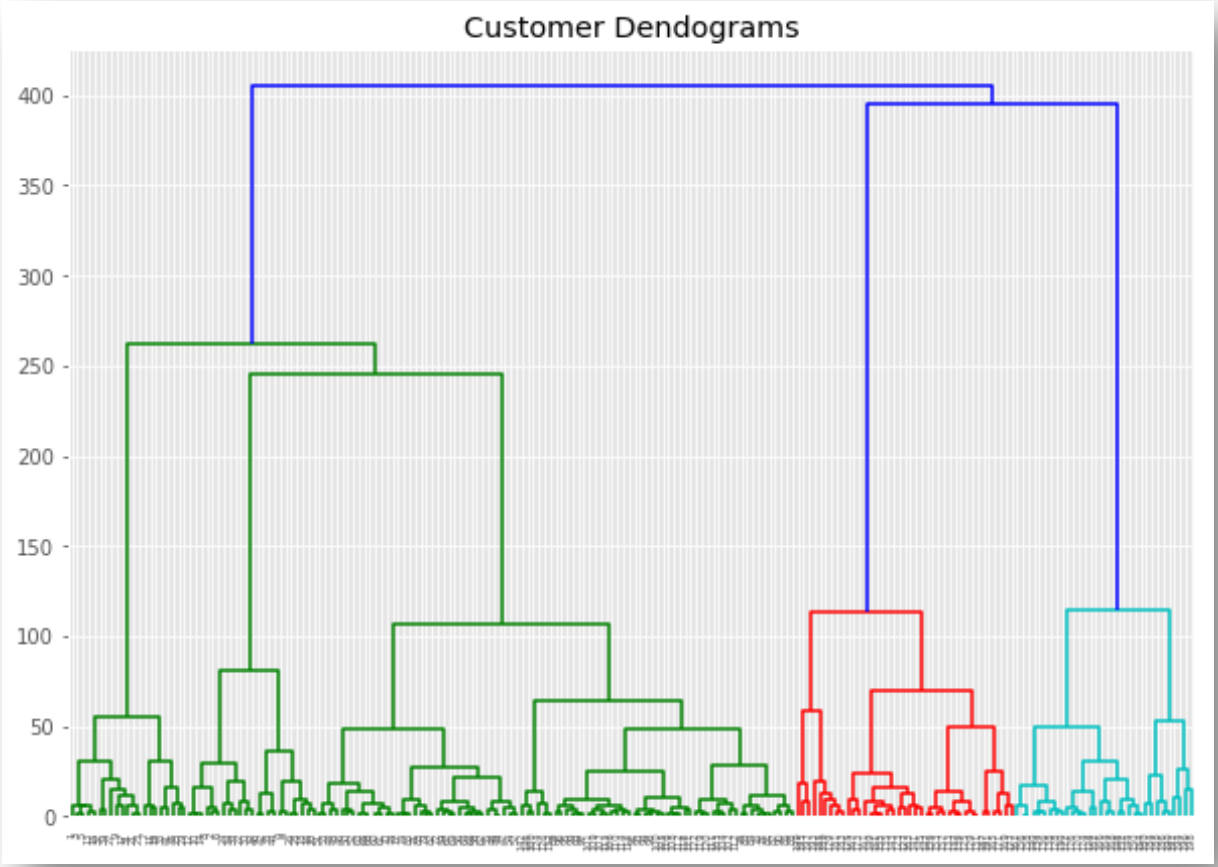
upGrad

- There are multiple distance measures that can be used
  1. Euclidean, Manhattan, Minkowski, Jaccard (for categorical data – after one hot encoding), Cosine, etc.

- There are also several ways to measure the distance between clusters. These are called Linkage Methods (to link two clusters to form a new cluster)
  1. Complete linkage: Maximum distance
  2. Single linkage: Minimum distance
  3. Average linkage: Average of distances between clusters
  4. Centroid linkage: Distance between centroids

Dendogram, interpretation..

upGrad



Cutting the Dendogram..

upGrad

- Distance threshold
  - Specify a minimum distance between the clusters and draw a line at that threshold.
  - No pairs of clusters thus formed are at a distance less than the threshold.

- Inconsistency coefficient
  - Height of merge point is the distance between clusters
  - Compare height of merge to average merge heights below the cut.
  - If the top merge is substantially higher than the average height of merges below the cut. Choose an inconsistency threshold and make the cut.

No correct method. Use Intuition, Heuristics, Application specific knowledge.

upGrad

#LifeKoKaroLift

More algorithms, Pros & Cons

More algorithms..

upGrad

- K-medians clustering
- K-prototype clustering
- Spectral & Graph clustering

Categorical clustering algorithms

- K-modes clustering
- Squeezer, LIMBO, Cobweb, STIRR, ROCK, CLICK, CACTUS, COOLCAT, CLOPE

Pros and Cons of different models..

upGrad

K-means clustering

- Needs us to pre-specify number of clusters, which can be hard in some cases
- Sensitive to outliers as mean is used
- Sensitive to the cluster centroid initialization point and can sometimes converge on a local minima. May get different results when we change the ordering of the data
- It is a lazy learner, and only modeled when triggered. Each query creates a model from scratch
- May require large memory

Pros and Cons of different models..

upGrad

Hierarchical clustering

- Not necessary to specify the number of clusters
- Hierarchical clustering is sensitive to outliers, creating exclusive clusters for them, and sometimes forcing real clusters to merge
- Number of clusters need to be determined subjectively after the clustering is done

Pros and Cons of different models..

upGrad

Gaussian Mixture Models

- Have strong theoretical foundation but are quite susceptible to over-fitting.

DBSCAN

- Density based model need inputs to identify cluster borders.

K-medians

- Works well even with outliers

K-modes

- Works with even categorical data where distance cannot be computed.

K-prototype

- Works with mixed attribute data-points (containing both numerical & categorical data)

Performance measures..

upGrad

- **Adjusted rand index**

  sklearn.metrics.adjusted_rand_score

  The Rand Index computes a similarity measure between two clusterings by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clusterings.

- **Mutual information based scoring**

  sklearn.metrics.mutual_info_score

  The Mutual Information is a measure of the similarity between two labels of the same data

- **Homogenity, completeness and v-measure**

  sklearn.metrics.homogeneity_score

  sklearn.metrics.completeness_score

  sklearn.metrics.v_measure_score

  A clustering result satisfies homogeneity if all of its clusters contain only data points which are members of a single class.

  A clustering result satisfies completeness if all the data points that are members of a given class are elements of the same cluster.

upGrad

#LifeKoKaroLift

Thank You!