2022-05-19

Familiarize unsupervised learning and k-means Clustering

UNSUPERVISED LEARNING

- uses machine learning algorithm to analyze and cluster unlabeled datasets
- With out human interaction, the algorithms can discaver hidden patterns or data groupings
- Can ideally be used for exploratory data analysis, cross-selling strategies, customer segmentation, and image recognition because it can discover similarities and differences

COMMON UNSUPERVISED LEARNING APPROACHES

CLUSTERING

- groups unlabeled data based on similarities and differences
- clustering algorithms used process raw, unclassified data objects

EXCLUSIVE AND OVERLAPPING CLUSTERING

"Hard" k means clustering

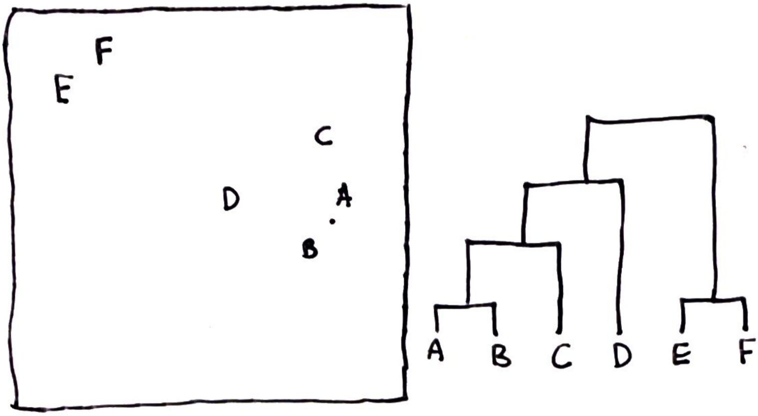
- Form of grouping that stipulates a data point can exist only in one cluster
- k-means clustering:
  - data points are assigned into K groups
  - K represents the number of clusters based on the distance from each group's centroid
  - Data points closest to a given centroid will be clustered under the same category
  - ↑ K value, smaller groupings with more granularity
  - ↓ K value, larger " " less "
  - market segmentation, document clustering, image segmentation, and image segmentation

soft k means cluster

- overlapping cluster allows data points to belong to multiple clusters

HIERARCHAL CLUSTERING

- Agglomerative or divisive.
- agglomerative clustering is considered a "bottoms up approach"
- datapoints are isolated as seperate groupings initially
- Then they are merged together iteratively on the basis of similarity until one clustering has been achieved.



PROBABILISTIC CLUSTERING

- Used to solve density estimation or "soft" clustering problems
- Clustered based on the liklihood that they belong to a particular distribution
- Gaussian Mixture Model(GMM) is the most commonly used probabilistic clustering methods

OTHER Examples of UNSUPERVISED LEARNING

- Association Rules
  - Apiori algorithms, dimensionality reduction, principal component analysis, singular value decomposition, auto encoders

UNSUPERVISED LEARNING APPLICATIONS

- News Section
- computer vision
- Medical imaging
- Anomaly detections
- customer personas
- Reccomendation Engines

UNSUPERVISED VS SUPERVISED LEARNING

- Supervised learning algorithms use labeled data
- Using the data it either predicts future outcomes or assigns data to specific categories based on the regression or classification problem at hand
- Higher accuracy with supervised learnings because it requires human interaction
- Supervised learning avoids computational complexity

CHALLENGES WITH UNSUPERVISED LEARNING

- longer training times
- computational complexity due to a high volume of training data
- higher risk of innacurate results
- human intervention to validate output variables
- lack of transparency into the basis on which data was clustered

K-means Clustoring Algorithm with Python Tutorial

1. define the number (k) of clustering to the split into

2. select k random points with the data.

3. Calculate distance between centroid and other points
4. Assign the points to the closest centroid
5. calculate the centre of each clustre
6. Repeat steps 3-5