Reference:Introduction to Machine Learning with Python

April 21, 2021

1 What is Unsupervised learning algorithm

Unsupervised learning subsumes all kinds of machine learning where there is no known output, no teacher to instruct the learning algorithm.
In unsupervised learning, the learning algorithm is just shown the input data and asked to extract knowledge from this data.

2 What is Clustering

Clustering is the task of partitioning the dataset into groups, called clusters.

3 Goal of Clustering

Split up the data in such a way that points within a single cluster are very similar and points in different clusters are different.

Clustering algorithms assign (or predict) a number to each data point, indicating which cluster a particular point belongs to.

4 k-Means Clustering

One of the simplest and most commonly used clustering algorithms.

It tries to find cluster centers that are representative of certain regions of the data.

The algorithm alternates between two steps: assigning each data point to the closest cluster center, and then setting each cluster center as the mean of the data points that are assigned to it.

The algorithm is finished when the assignment of instances to clusters no longer changes.

5 Failure cases of k-means

Even the "right" number of clusters for a given dataset are known , k-means might not always be able to recover them.

Each cluster is defined solely by its center,so each cluster is a convex shape. As a result of this, k-means can only capture relatively simple shapes.

k-means also assumes that all clusters have the same "diameter" in some sense; it always draws the boundary between clusters to be exactly in the middle between the cluster centers.

6 Pro / Con of k-means

6.1 Pro

Relatively easy to understand and implement.

Runs relatively quickly, k means scales easily to large datasets.

6.2 Con

It relies on a random initialization, which means the outcome of the algorithm depends on a random seed.
By default, scikitlearn runs the algorithm 10 times with 10 different random

initializations, and returns the best result.

Relatively restrictive assumptions made on the shape of clusters, and the requirement to specify the number of clusters you are looking for.