Unsupervised Machine Learning Algorithms

How do you find the underlying structure of a dataset? How do you summarize it and group it most usefully? How do you effectively represent data in a compressed format? These are the goals of unsupervised learning, which is called "unsupervised" because you start with unlabeled data.

Machine Learning Algorithms

Supervised machine learning

***Unsupervised machine learning**

Reinforced machine learnings

*Clustering

The goal of clustering is to create groups of data points such that points in different clusters are dissimilar while points within a cluster are similar.

*k-means clustering

With k-means clustering, we want to cluster our data points into k groups. A larger k creates smaller groups with more granularity, a lower k means larger groups and less granularity.

The output of the algorithm would be a set of "labels" assigning each data point to one of the k groups. In k-means clustering, the way these groups are defined is by creating a centroid for each group. The centroids are like the heart of the cluster, they "capture" the points closest to them and add them to the cluster.

Application: one real-life application of k-means clustering is classifying handwritten digits. Suppose we have images of the digits as a long vector of pixel brightness's. Let's say the images are black and white and are 64x64 pixels. Each pixel represents a dimension. So the world these images live in has 64x64=4,096 dimensions. In this 4,096-dimensional world, k-means clustering allows us to group the images that are close together and assume they represent the same digit, which can achieve pretty good results for digit recognition.

*Hierarchical clustering

Hierarchical clustering is similar to regular clustering, except that you're aiming to build a hierarchy of clusters. This can be useful when you want flexibility in how many clusters you ultimately want. In terms of outputs from the algorithm, in addition to cluster assignments you also build a nice tree that tells you about the hierarchies between the clusters. You can then pick the number of clusters you want from this tree.

Application: For example, imagine grouping items on an online marketplace like Etsy or Amazon. On the homepage you'd want a few broad categories of items for simple navigation, but as you go into more specific shopping categories you'd want increasing levels of granularity, i.e. more distinct clusters of items..

*Dimensionality Reduction

Looks a lot like compression. This is about trying to reduce the complexity of the data while keeping as much of the relevant structure as possible.

*Principal Component Analysis

Principal component analysis (PCA) is a mathematical algorithm that reduces the dimensionality of the data while retaining most of the variation in the data set. It accomplishes this reduction by identifying directions, called principal components, along which the variation in the data is maximal. By using a few components, each sample can be represented by relatively few numbers instead of by values for thousands of variables. Samples can then be plotted, making it possible to visually assess similarities and differences between samples and determine whether samples can be grouped..

Application: An interesting example of clustering in the real world is marketing data provider Acxiom's life stage clustering system, Personicx. This service segments U.S. households into 70 distinct clusters within 21 life stage groups that are used by advertisers when targeting Facebook ads, display ads, direct mail campaigns, etc. Their white paper reveals that they used centroid clustering and principal component analysis, both of which are techniques covered in this section..

*Singular value decomposition (SVD)

Matrix decomposition, also known as matrix factorization, involves describing a given matrix using its constituent elements, and the most known and widely used matrix decomposition method is the Singular-Value Decomposition, or SVD. All matrices have an SVD, which makes it more stable than other methods, such as the eigendecomposition.

Application: SVD has been used in various applications such as: (a) *Automatic Background removal* - e.g. photo background removal service for E-commerce websites to increase there sales volume. (b) *Topic Modeling* - e.g. Sentiment Analysis on Social Media for Stock Market Prediction or providing insight into how significant topics such as the Great East Japan Earthquake, the Arab Spring, and the Boston Bombing affect individuals.