Clustering

Monday, September 2, 2024 7:24 PM

Clustering Introduction

Clustering: an unsupervised machine learning technique designed to group unlabeled datapoints based on their similarity to one another. Clustering is commonly used for:

- Market segmentation / targeted marketing
- Personalized medicine
- Locating facilities
- Image analysis
- Initial data investigation

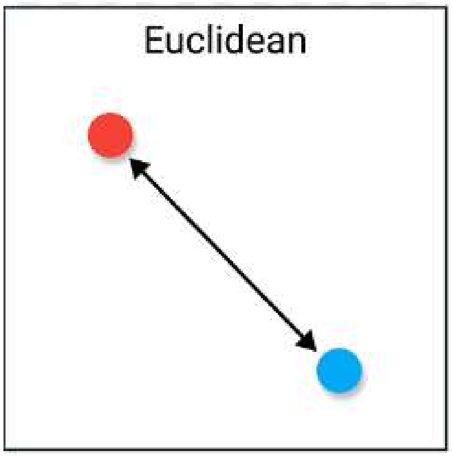
Distance Norms

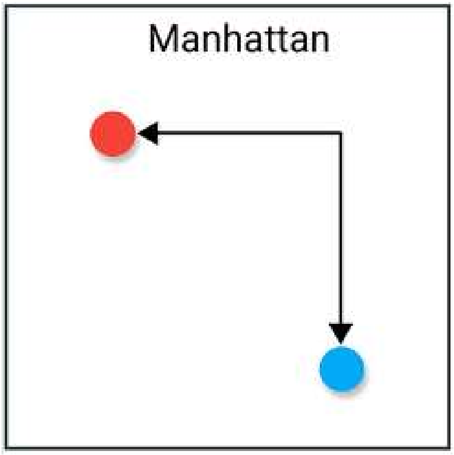
Since clustering is grouping datapoints based on their similarity, mathematically this translates to grouping datapoints that are nearest to each other on a graph. There are multiple ways to calculate distance.

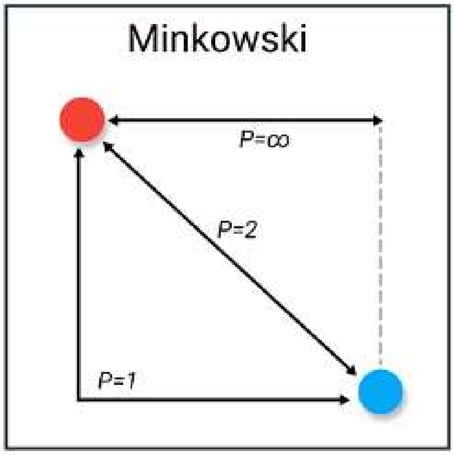
Euclidean (straight-line) distance: $distance = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$

Rectilinear/Manhattan/L1 distance: $distance = |x_1 - y_1| + |x_2 - y_2|$

Both the Euclidean and Manhattan distance formulas can be generalized in what is known as the p-norm distance (Minkowski distance): $distance = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p}$. One of the most common p values used in this formula is ∞, which means that the largest of the n terms comes to dominate the others. Thus the sum of the terms raised to infinity is almost equal to the largest term raised to infinity. This is known as the infinity-norm: $\sqrt[\infty]{\sum_{i=1}^{n} |x_i - y_i|^\infty} = max_i|x_i - y_i|$





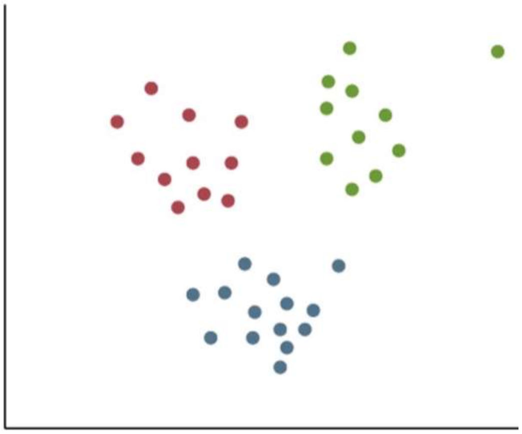


k-Means Clustering

The traditional clustering algorithm is time consuming and computationally hefty. A simpler solution is k-Means Clustering, which generally operates as follows:

1. Pick *k* cluster centers (centroids) within the range of data.
2. Temporarily assign each datapoint to its nearest cluster center.
3. Recalculate centroids.
4. Repeat steps 2-3 until there are no changes; until no datapoints change which clusters they are assigned to.

k-Means Clustering is heuristic, which means it is fast and good at coming close to the ideal solution, but not guaranteed to find the absolute best solution. The algorithm is also an Expectation-Maximization, where the expectation step (finding centroids) is alternated with the maximization step (assigning points to clusters).

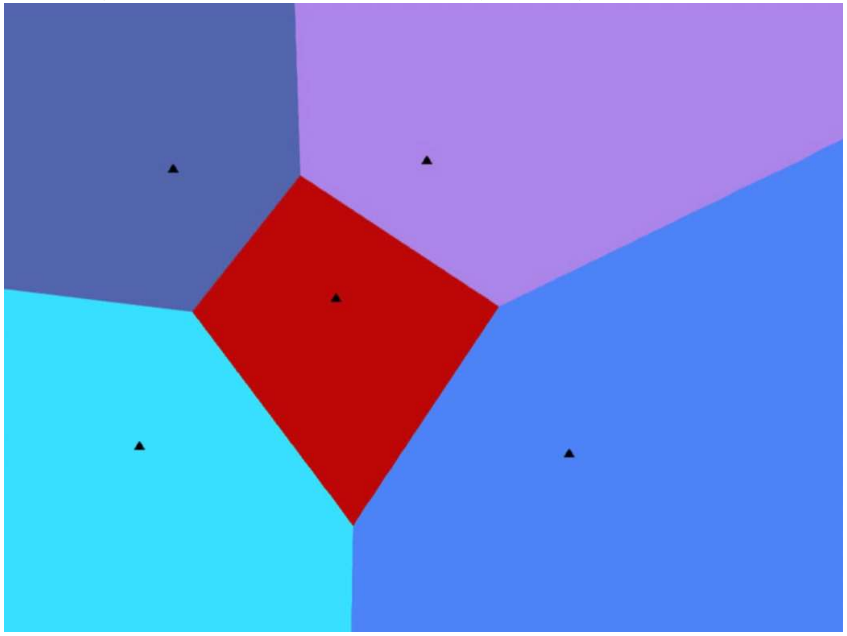


In some cases, it may be appropriate to remove outliers from your dataset so it doesn't artificially pull the centroid of its cluster in one direction. But in most cases, the best approach is to learn more about what sets the outlier apart; what is it? What makes its attributes so different? What are the implications of including it in its nearest cluster?

Because k-Means Clustering is heuristic, it can be quickly run several times to test different aspects such as:

- Different initial centroids to potentially find more optimal final clusters
- Different *k* values to find the most effective number of clusters (useful tools like an elbow graph may be used to do this)

When predicting for previously unseen datapoints, we can pre-determine which cluster new points would be assigned to by observing which centroid is nearest. The type of diagram illustrated below is known as a Voronoi diagram.



Clustering vs Classification

There are two primary approaches to machine learning.

Supervised Learning: uses labeled training data so that the models have a baseline understanding of what the correct output should be. The algorithm "learns" from the training dataset by iteratively making predictions on the data and adjusting to get the correct answer. These algorithms tend to be more accurate than unsupervised ones, but also require more work to get the data into an appropriate state to use. Common types of supervised algorithms include:

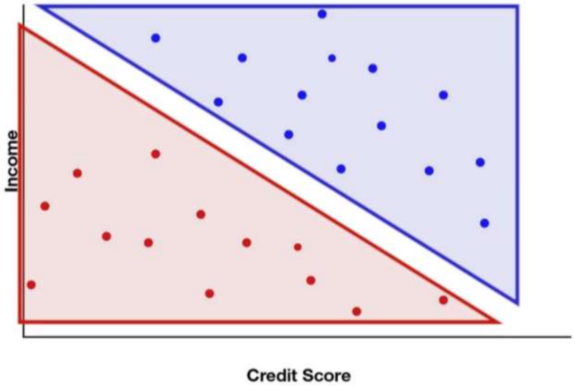
- Regression
- Classification

Unsupervised Learning: uses unlabeled training data so the models work on their own to discover the inherent structure of the data.

- Clustering
- Anomaly detection
- Principle Component Analysis

Classification

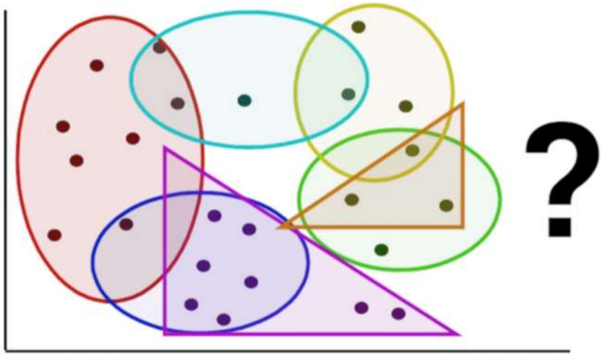
Grouping data points



Correct classification of data points is already known

Clustering

Grouping data points



Correct classification of data points is **not** known