Module 2

Distance Metrics: Euclidean and Manhattan Distance

Distance Metrics Overview

- Different measures of distance, such as Euclidean distance, Manhattan distance, Cosine similarity, and Jaccard distance, are essential for clustering algorithms.
- The choice of distance metric significantly impacts the performance of clustering algorithms.

Euclidean and Manhattan Distances

- Euclidean distance (L2 distance) is calculated using the square root of the sum of squared differences between points, applicable in higher dimensions.
- Manhattan distance (L1 distance) sums the absolute differences and is more effective in high-dimensional spaces, often used in business cases.

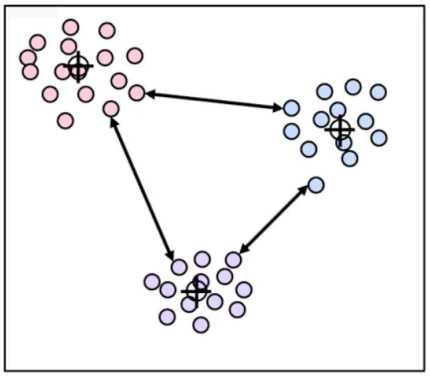
Empirical Evaluation

- Selecting the appropriate distance metric may require empirical evaluation to determine which metric best achieves clustering goals.
- Understanding the strengths and appropriate use-cases of each distance metric is crucial for effective clustering.

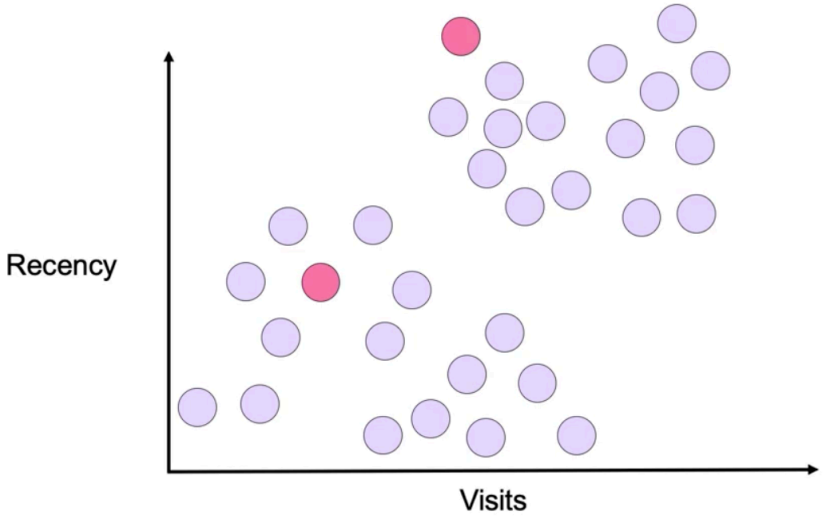
Distance Metric Choice

Choice of distance metric is extremely important to clustering success.

Each metric has strengths and most appropriate use-cases…



Euclidean Distance



Distance Metrics: Cosine and Jaccard Distance

Cosine Distance

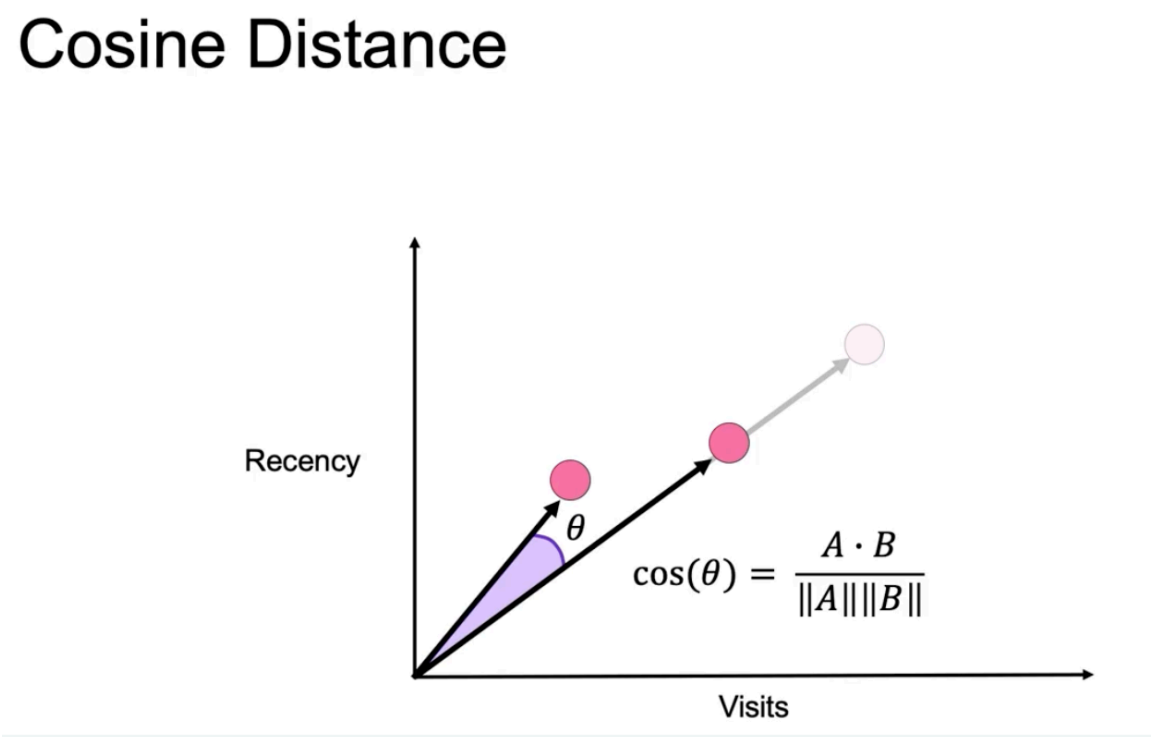
- Measures the angle between two points in a vector space, focusing on the direction rather than the magnitude.
- Remains insensitive to scaling, meaning points along the same ray from the origin have a cosine distance of 0, indicating similarity.

Jaccard Distance

- Used for comparing sets, calculated as 1 minus the ratio of the intersection of two sets to their union.
- Useful for text data, allowing for the grouping of similar topics based on unique word occurrences.

Summary of Distance Metrics

- Discussed Euclidean distance, Manhattan distance, cosine similarity, and Jaccard distance.
- Emphasized the importance of selecting the appropriate distance metric based on the specific application and data characteristics.



Curse of Dimensionality Notebook - Part 1

Curse of Dimensionality

- As dimensions increase, data points become more sparse and further apart, complicating clustering and classification tasks.
- The demo aims to illustrate how higher dimensions lead to increased empty space in data representation.

Visualization Techniques

- The demonstration includes creating a circle within a square and a sphere within a cube to visualize the concept of space coverage.
- It emphasizes the importance of maintaining equal scales on axes to accurately represent shapes in plots.

Modeling Implications

- The content discusses how points outside a defined area (like a circle) can be harder to classify, indicating they are further from the mean.
- It concludes with a mathematical explanation of the area outside the circle in relation to the square, highlighting the percentage of space that remains uncovered as dimensions increase.

Curse of Dimensionality Notebook - Part 2

Understanding Three-Dimensional Data

- The transition from a square and circle to a cube and sphere is discussed, emphasizing how values are represented in three dimensions.
- Each dimension corresponds to a different feature (covariate), normalized between -1 and 1, with a mean of 0 and standard deviation of 1.

Plotting in Three Dimensions

- The Axes3D library is introduced for creating 3D plots, allowing visualization of data points along the x, y, and z axes.
- The process of plotting a cube and a sphere is explained, including how to create combinations of points and visualize them in three-dimensional space.

Outlier Analysis in 3D

- The concept of outliers is revisited, with 48% of values identified as outliers when moving to three-dimensional space, compared to 21% in two dimensions.
- The volume of the sphere and cube is calculated to understand the distribution of data points and their relation to standard deviations.

Curse of Dimensionality Notebook - Part 3

Understanding Dimensions

- The move from two to three dimensions shows an increase in the percentage of values that are more than one unit away from the mean, from 21% to 48%.
- The concept of being "within the ball" or "outside the ball" is introduced to determine if points are within one standard deviation from the mean.

Generalizing to Higher Dimensions

- A function is created to determine the percentage of points within an n-dimensional ball compared to an n-dimensional cube.
- The function generates random points and calculates their Euclidean distance to assess whether they fall within the defined range.

Analyzing Results

- The lecture discusses how the percentage of points within the ball decreases as the number of dimensions increases, indicating a rise in outliers.
- A plot is created to visualize the relationship between dimensions and the percentage of points within the ball, demonstrating a steep decline as dimensions increase.

Curse of Dimensionality Notebook - Part 4

Understanding High Dimensionality

- High dimensionality can negatively affect model performance, leading to overfitting.
- The curse of dimensionality makes clustering difficult as the number of features increases.

Methods to Combat Dimensionality Issues

- Feature selection involves using domain knowledge to reduce the number of features to only those that are informative.
- Feature extraction employs techniques like PCA (Principal Component Analysis) to transform raw data into a lower-dimensional space while preserving variability.

Practical Application and Results

- A classification dataset is created using the `make_classification` function, demonstrating how the number of features affects model accuracy.
- Increasing the number of informative features can lead to reduced model performance, emphasizing the importance of balancing the number of features with the amount of data available.