Netflix Movies & TV Shows Clustering

Sayan Bandopadhyay
Data science trainee,
AlmaBetter, Bangalore

Abstract

Netflix is an American subscription streaming service and production company which offers a library of films and television series through distribution deals. It is one of the most popular as well as most preferred streaming services out there.

Provided with a Netflix dataset where no such target variables can be identified makes it clear that this particular problem statement is an unsupervised machine learning problem. Performed clustering analysis specifically for text-based clusters using one of the most popular unsupervised machine learning algorithm 'K-means clustering' and furthermore creating a recommendation system by using 'cosine similarity'.

Keywords: *Netflix, unsupervised machine learning, cluster, K-means clustering, recommendation, cosine similarity*

Problem Statement

This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flixable which is a third-party Netflix search engine. In 2018, they released an interesting report which shows that the number of TV shows on Netflix has nearly tripled since 2010. The streaming service's number of movies has decreased by more than 2,000 titles since 2010, while its number of TV shows has nearly tripled. Main purpose here is to explore what all other insights can be obtained from this particular dataset and how text-based clusters helps in creating recommendations.

Steps Involved

Following are the different steps been taken in the overall analysis process right from the beginning to the end:

<u>Framing Questions</u>: This is the very first step of the analysis process. Going through the problem statement deeply and to extricate some hidden questions that problem statement is throwing to us is the main objective here as based on which the analysis will choose its path.

<u>Data Inspection</u>: Once the objective of the problem statement becomes clear, now it's the time to go through the dataset thoroughly in order to understand different features that the dataset is comprised of and also to check some statistical scores and a bit of more information regarding the dataset based on which next steps of the process can be designed.

<u>Data Scrubbing & Pre-processing</u>: Raw datasets usually have missing values or null values. To make sure that the models later does not malfunctions or provides some biased insights, data cleaning is essential and this is what is being done in this particular step. This includes null or missing values treatment, outlier treatment and duplicate value check.

EDA: Exploratory data analysis is the step where we just visualize data based on different features of the cleaned dataset so that we can come across some meaningful insights.

Data Preparation: This step basically includes arrangement of data, text cleaning and feature engineering so that machine learning algorithms can be implemented without any hassle.

Model Implementation: Sixth step is all about implementing the best suited model in order to perform machine learning operations.

Model Evaluation: In this particular case cluster evaluation is been done via elbow method and silhouette method in order to determine the optimal number of cluster.

Conclusions: Last step is all about summing up the insights and model performances observed from analysis process.

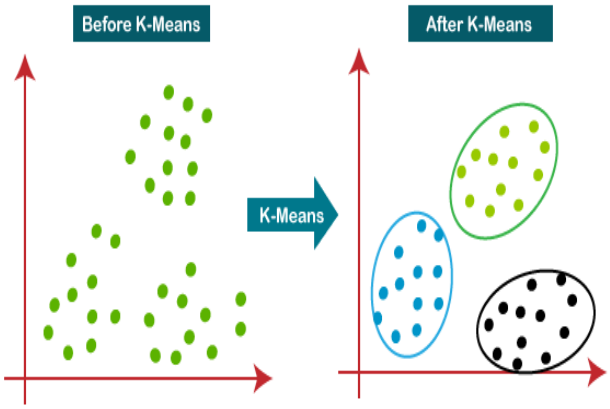
Algorithms

K-MEANS CLUSTERING

The model used in this particular case is 'K-Means' clustering. K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties. K-means clustering mainly performs two tasks:

- Determines the best value for K center points or centroids by an iterative process.
- Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

Hence, each cluster has datapoints with some commonalities, and it is away from other clusters. The below diagram explains the working of the K-means Clustering Algorithm:



Hyperparameter Tuning

Hyperparameters are model configuration properties that define the model and remain constant during the training of the model. The design of the model can be trained by tuning the hyperparameter. For K-Means clustering there are 3 main hyperparameters to set-up to define the best configuration of the model:

- Initial values of clusters
- Distance measures
- Number of clusters

Initial values of clusters greatly impact the clustering model and there are various algorithms to initialize the values. Distance measures are used to find points in clusters to the cluster center and different distance measures yield different clusters. The number of clusters (**k**) is the most important hyperparameter in K-Means clustering. If we already know beforehand, the number of clusters to group the data into, then there is no use to tune the value of k. Two such methods used to find the optimal number of clusters in this case are ‘Elbow method’ & ‘Silhouette method’.

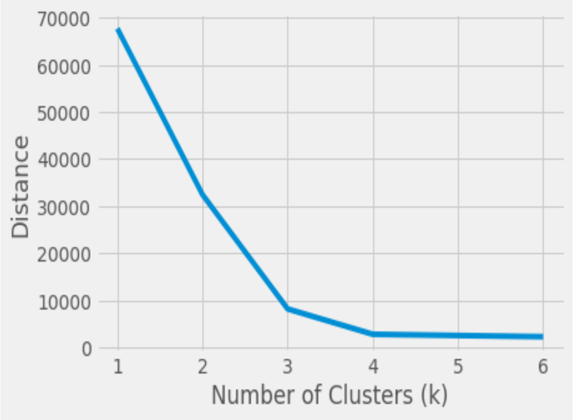
SILHOUETTE METHOD

Silhouette is a measure of how a clustering algorithm has performed. After computing the silhouette coefficient of each point in the dataset, plot it to get a visual representation of how well the dataset is clustered into k clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like the number of clusters visually. This measure has a range of [-1, 1]. Important points:

- ✓ The Silhouette coefficient of ‘+1’ indicates that the sample is far away from the neighbouring clusters.
- ✓ The Silhouette coefficient of ‘0’ indicates that the sample is on or very close to the decision boundary between two neighbouring clusters.
- ✓ The Silhouette coefficient of ‘-1’ indicates that those samples might have been assigned to the wrong clusters or are outliers.

ELBOW METHOD

Elbow Method is an empirical method to find the optimal number of clusters for a dataset.



In this method a range of candidates value of ‘K’ is picked and the average distance of each point in a cluster to its centroid in found. The value of ‘K’ is picked where the average distance falls suddenly. For example, in the above diagram 3 will be considered as the optimal number of ‘K’.

Conclusion

At the end, after all the way through right from data inspection to model implementation it was a rigorous process to be followed which finally led a way to create a recommendation system that suggests TV shows and movies based on individual search taste.

Reference:

1. GeeksforGeeks

2. My Great Learning

3. Analytics Vidhya

4. JavaTpoint