greatlearning
Learning for Life

DATA MINING (Week 2)

greatlearning
Learning for Life

DSBA CURRICULUM DESIGN

FOUNDATIONS

- Data Science Using Python
- Statistical Methods for Decision Making
- Advanced Statistics

CORE COURSES

- Data Mining (Week-2/2)
- Predictive Modelling
- Machine Learning
- Time Series Forecasting
- Data Visualization

DOMAIN APPLICATIONS

- Financial Risk Analytics
- Web &Social Media Analytics
- Marketing Retail Analytics

greatlearning
Learning for Life

LEARNING OBJECTIVE OF THIS MODULE

- PCA
- Clustering Techniques

greatlearning
Learning for Life

LEARNING OBJECTIVES OF THIS SESSION -

- Hierarchical Clustering
- K Means Clustering

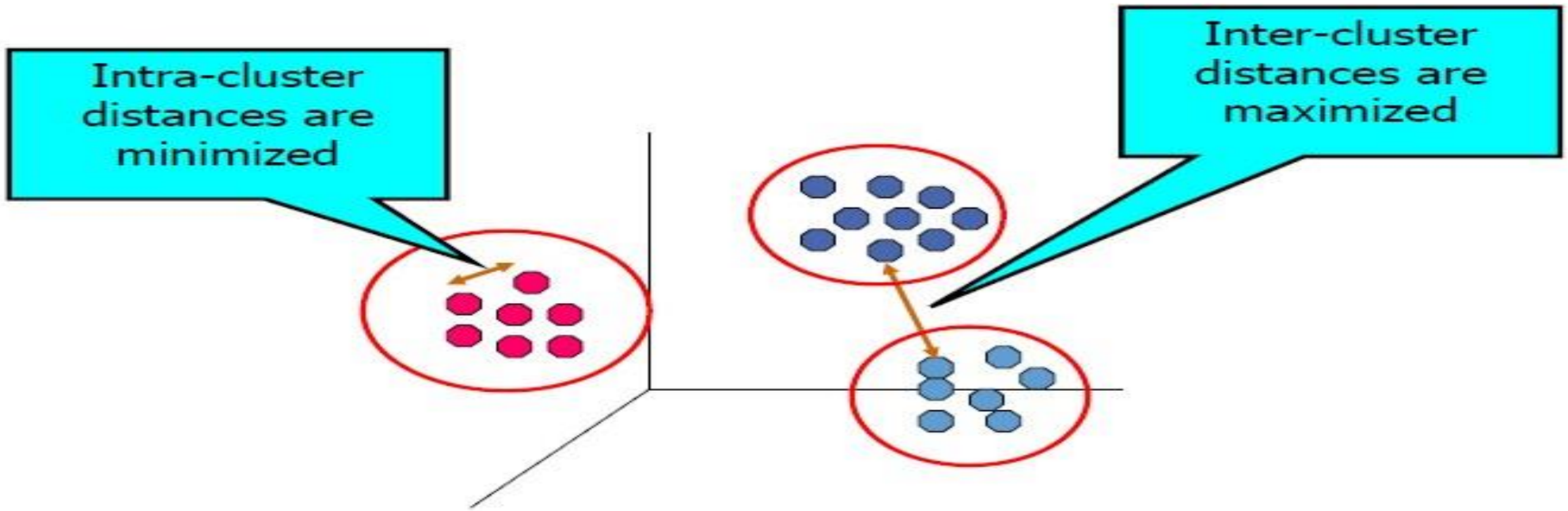
greatlearning
Learning for Life

TRY ANSWERING THE FOLLOWING

- What is the formula to calculate Euclidean Distance?
- Is Clustering a part of Supervised learning?
- Which clustering technique is preferred for dealing with large data- Hierarchical or K-Mean Clustering?

greatlearning
Learning for Life

BROAD OVERVIEW



This is classic case of Unsupervised learning

greatlearning
Learning for Life

Few Application of Clustering

Image processing	Web	Market Segmentation	Market structure analysis	Finance
• cluster images based on their visual content	• Cluster groups of users based on their access patterns on webpages • Cluster webpages based on their content	• customers are segmented based on demographic and transaction history information, and a marketing strategy is tailored for each segment	• identifying groups of similar products according to competitive measures of similarity	• cluster analysis can be used for creating *balanced portfolios*

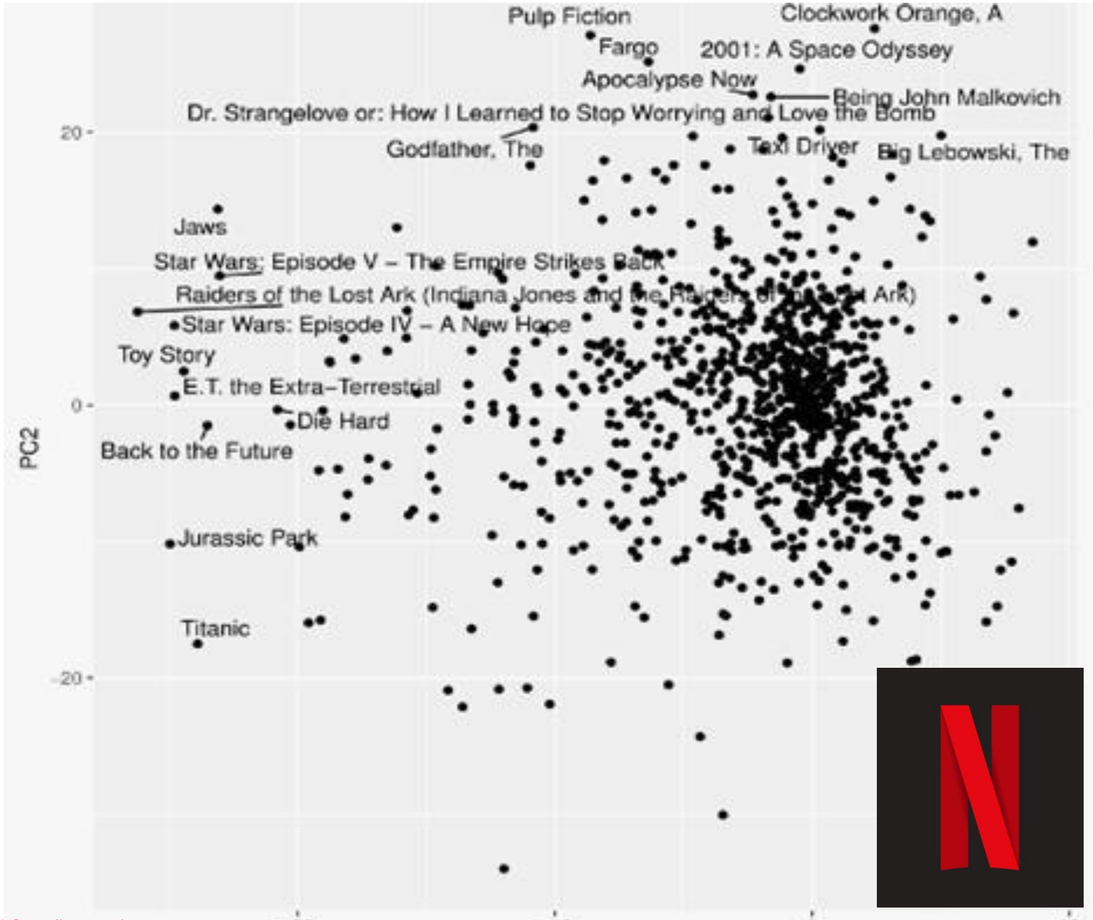
greatlearning
Learning for Life

Industry Application - Netflix' Movie Recommendation System

In 2006 Netflix rolled out a contest with a prize money of 1 million $ to develop a movie recommendation system which could deliver 10% reduction in terms of root mean square error compared to their own in-house developed algorithm called Cinematch.

They had a huge database of over 100 million ratings of 17,770 movies from 4,80,189 customers. Now anything that requires sharper customer targeting can never be achieved without forming clusters. But basis what features or attributes do we form cluster here?

Genre is the answer. It is quite evident that every movie lover has a liking for certain type of movies e.g. Action-Adventure-Science Fiction or Crime-Thrillers etc. Around 10 clusters were formed using 19 different genres. More meaningful recommendations could be made knowing the cluster membership of subscribers, and finally the desired accuracy was achieved after 3 years of rigorous competition.



Reference: https://www.netflixprize.com/
https://www.thrillist.com/entertainment/nation/the-netflix-prize

greatlearning
Learning for Life

Industry Application - Clustering the stocks by monthly returns

It is a common practice to apply machine learning algorithms in order to predict stock price movement i.e. Upwards or Downwards i.e. essentially a classification problem. But is it wise to club all the listed stocks together and try to fit them to a model? Well, not really!

Let's say we take a dataset where we have stock returns from January to November, and basis this we want to predict if a given stock's price will move up or go down in December. It is observed that the prediction accuracy on such datasets without cluster analysis is around 50-55% whereas the same significantly improves to ~70% when models were developed specific to each cluster. These clusters can be obtained using K-means clustering, typically with the value of k between 3 to 5.
Cluster analysis helps group the stocks of similar nature together and hence improves prediction accuracy.

Similar Reference https://econwpa.ub.uni-muenchen.de/econ-wp/fin/papers/0509/0509022.pdf

greatlearning
Learning for Life

CASE STUDY – Strategy to group Engineering Colleges

You are an independent trainer who would like to pitch your Data Science training program to a set of Engineering colleges. You have data of 26 collages after survey using questionnaires. Each college has been given a score for 5 performance criteria-Teaching, Fees, Placement, Internship & Infrastructure. Ratings are in the standardized scale of 1 to 5 where 5 has a higher weightage than 1. Segment the collages into groups and come up with your pitch recommendations for each segment.

CASE STUDY - Clustering of Banks

We have a transaction details of 515 banks which include number of DD taken, Withdrawals, Deposits, Area of the branch and Average Walk-Ins. Profile the banks into segments and come up with recommendations for each segment.

Data Science @ Work

greatlearning
Power Ahead

Apply **Data Science at your workplace** to gain some instant benefits:

- Get noticed by your management with your outstanding analysis backed by data science.
- Create an impact in your organization by taking up small projects/initiatives to solve critical issues using data science.
- Network with members from the data science vertical of your organization and seek opportunities to contribute in small projects.
- Share your success stories with us and the world to position yourself as a subject matter expert in data science.

greatlearning
Learning for Life

ANY QUESTIONS

greatlearning
Learning for Life

HAPPY LEARNING