أنماط البيانات
DataPatterns
Unleash the power of data

CL203: Unsupervised Learning : Clustering - NO-CODE TRACK

BY: SRINIVAS RAO

ASST:

CODELESS DATA SCIENCE – TARGET AUDIENCE

- Program / Project Managers
- Data Analysts
- Data Scientists
- Data Engineers
- Business Analysts
- Researchers in various domains
- Subject Matter Experts
- Freshers from any background

CODELESS TOOLS: EXAMPLES

alteryx

rapidminer

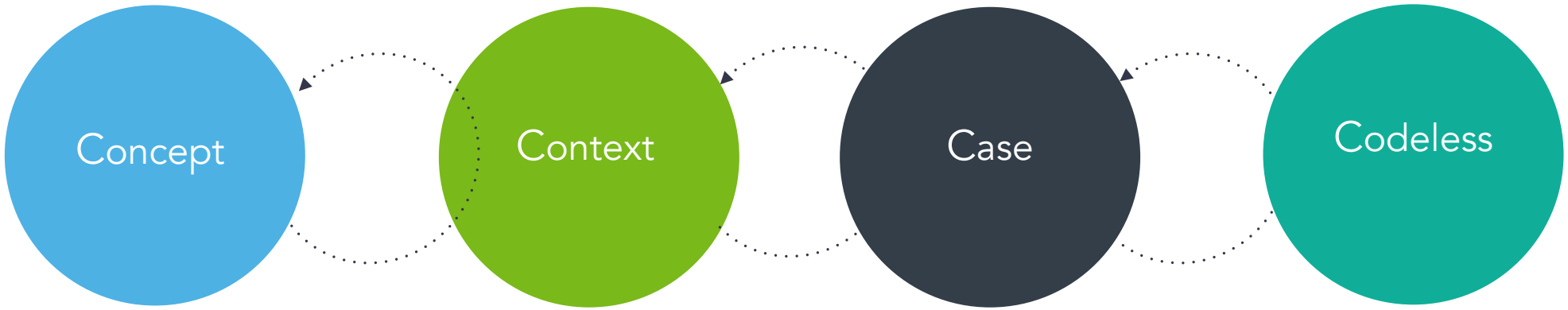
IBM Watson

IBM SPSS software

intellicus

OPEN SOURCE

THE 4-C METHODOLOGY



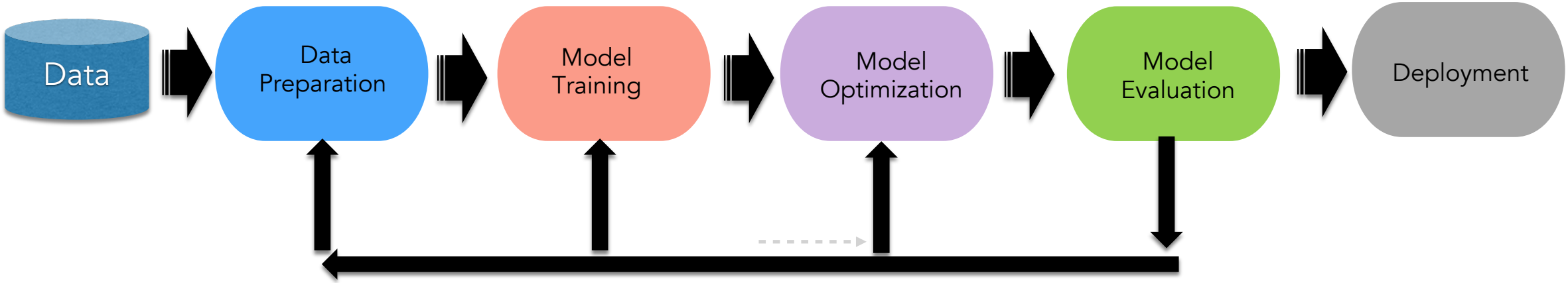
The Concept of each technique is explained in detail

The Context in which each technique is applied is explained

Cases / Business Problems relevant to each Concept is taken up for detailed analysis

Cases / Business Problems are solved using Codeless methodologies

DATA --> DEPLOYMENT CYCLE



Exploratory Data Analysis

a. Measures of Central Tendency
b. Measures of Dispersion
c. Distributions
d. Confidence Intervals
e. Hypothesis Testing
f. Visualization
g. Univariate
h. Bivariate

Data Cleaning / Data Preparation

a. Outlier analysis
b. Handling Missing data
c. Standardization / Normalization
d. Discretization / Binning / Grouping
e. Data Transformation
f. One Hot Encoding

Feature Engineering / Feature Extraction
Feature Selection

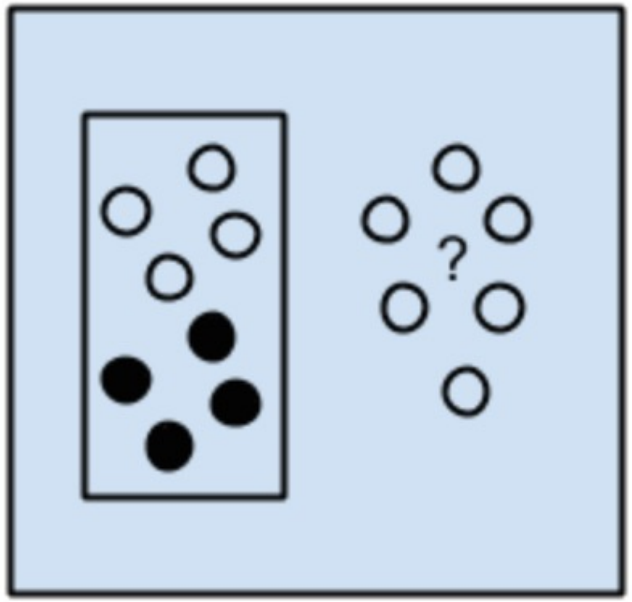
Model Training
Bag of Models
Model Selection
Ensemble Models

Parameter Tuning
Parameter Optimization
Regularization

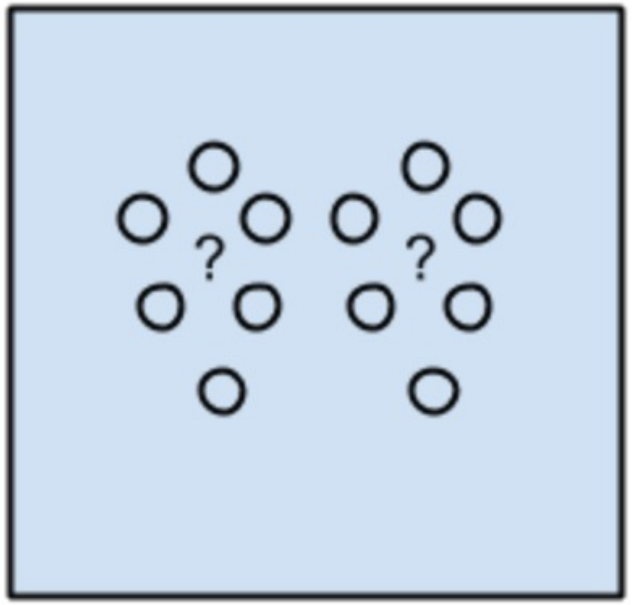
Performance Measures
Accuracy
ROC
Cross Validation

Files & DBs
Dashboards
REST API
Reporting

SUPERVISED & UNSUPERVISED LEARNING



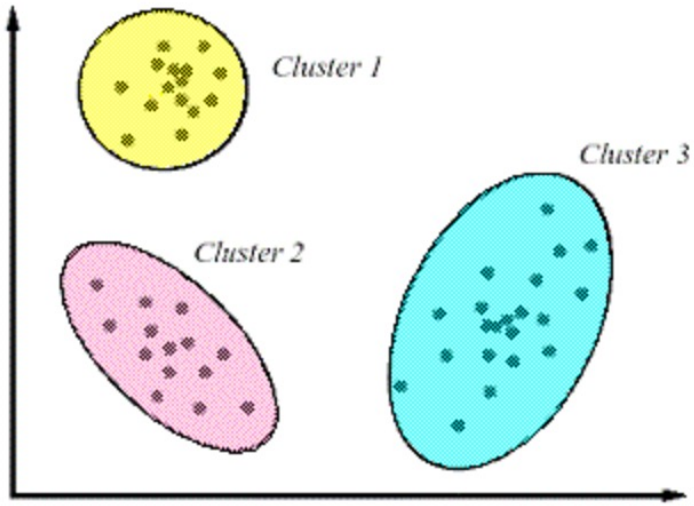
Supervised Learning
Algorithms

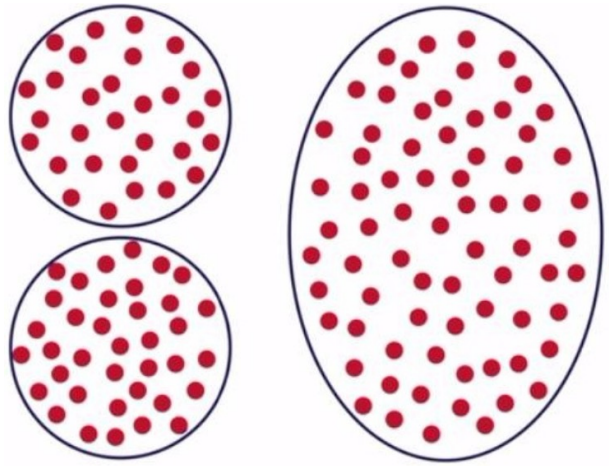


Unsupervised Learning
Algorithms

CLUSTERING

- Clustering" is the process of grouping similar entities together.
- It is an unsupervised machine learning technique to find similarities in the data points and group them together.
- While carrying out clustering, the basic objective is to group the input points in such a way as to **maximise the inter-cluster variance and minimise the intra-cluster variance**.





CLUSTERING

Centroid Based

The Centroid based is one of iterative clustering algorithm in which the clusters are formed by the closeness of data points to the *centroid* of clusters. The cluster centre, i.e. *centroid* is constructed such that the distance of data points is minimum with the centre. e.g k-Means

Density-Based

In this clustering model, there will be searching for data space for areas of the varied density of data points in the data space. It isolates various density regions based on different densities present in the data space. e.g DBScan

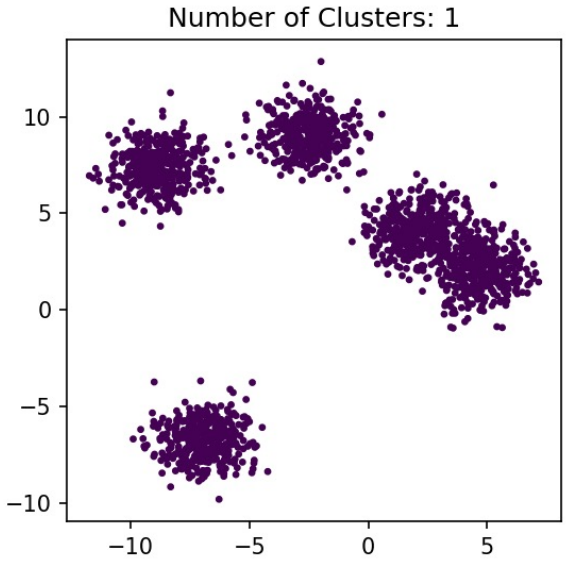
Hierarchical Based

In this method, clusters are constructed as a tree-type structure based on the hierarchy. They have two categories, namely, Agglomerative (Bottom-up approach) and Divisive (Top-down approach). e.g Hierarchical Clustering

k-MEANS CLUSTERING

- k-Means is a centroid-based algorithm, or a distance-based algorithm, where we calculate the distances to assign a point to a cluster.
- In k-Means, each cluster is associated with a centroid.
- The main objective of the K-Means algorithm is to minimize the sum of the squares of the distances between the points and their respective cluster centroid.

k-MEANS – Clusters with different k-Values

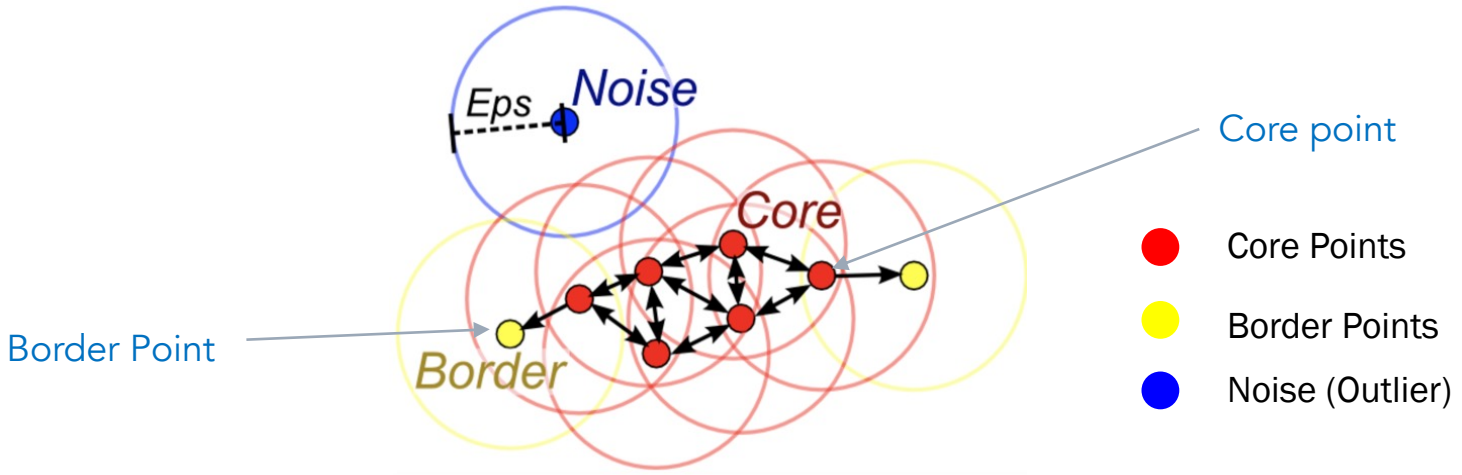


DENSITY BASED CLUSTERING

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

- It is a method that identifies distinctive clusters in the data, based on the key idea that a cluster is a group of high data point density, separated from other such clusters by regions of low data point density.
- The main idea is to find highly dense regions and consider them as one cluster.
- It can easily discover clusters of different shapes and sizes from a large amount of data, which is containing noise and outliers.

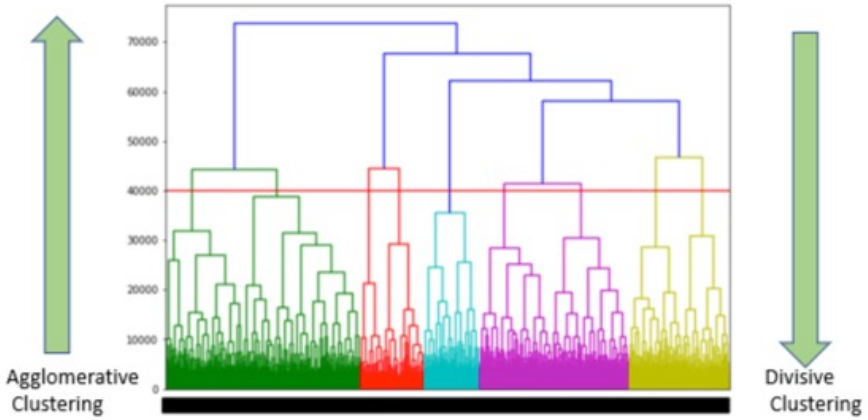
DBSCAN

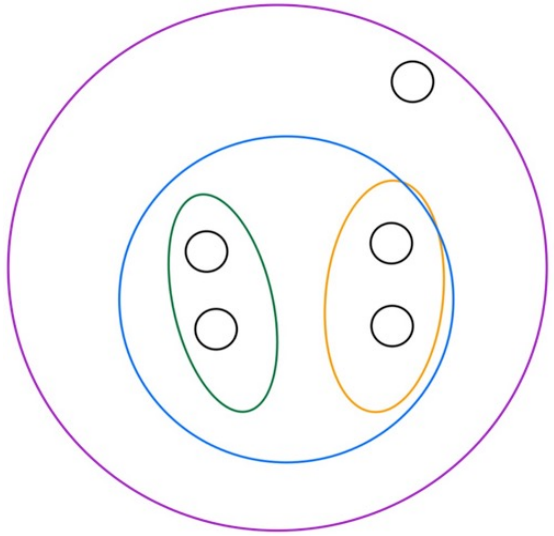


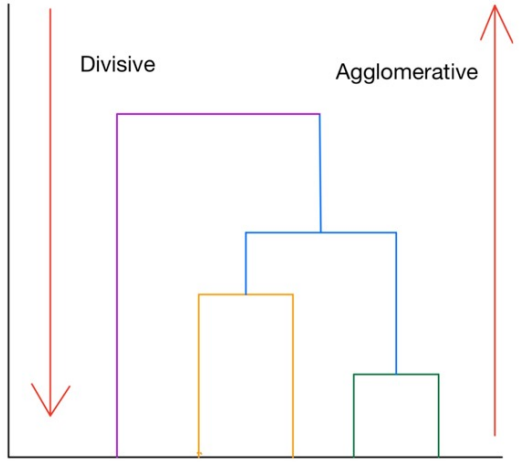
- **Core point:** A point is a core point if there are at least minPts number of points (including the point itself) in its surrounding area with radius eps.
- **Border point:** A point is a border point if it is reachable from a core point and there are less than minPts number of points within its surrounding area.
- **Outlier:** A point is an outlier if it is not a core point and not reachable from any core points.

HIERARCHICAL CLUSTERING

1. Agglomerative hierarchical clustering
2. Divisive Hierarchical clustering







KEY TAKEAWAYS FROM THE COURSE

- CONCEPTUAL CLARITY ON CLUSTERING
- HANDS ON EXPERIENCE WITH THE CODELESS METHODOLOGY
- HANDS ON PRACTICE SESSIONS WITH 1 OR 2 CASES
- HOME WORK EXERCISES TO PRACTICE THE CODELESS METHODOLOGY
- 2 DATASETS TO PRACTICE CLUSTERING
- PRESENTATION
- EXCEL WORK BOOK CONTAINING THE EXERCISES
- CSV / EXCEL FILES CONTAINING DATASETS

THANK YOU