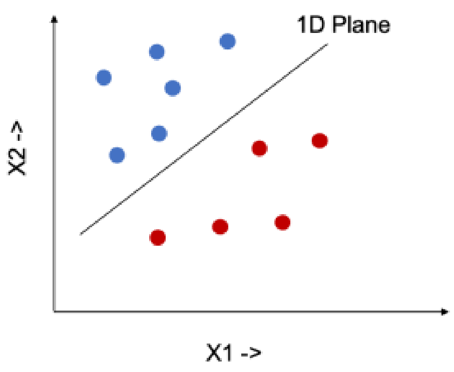
Support Vector Machine: Dimensionality Reduction and Clustering

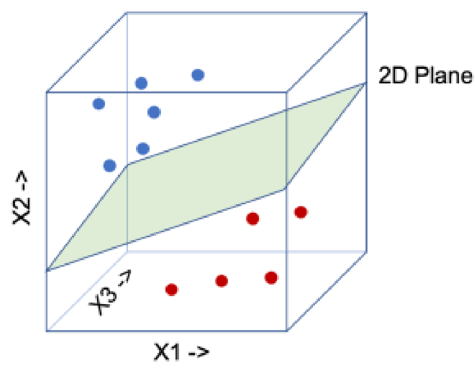
SVM ~

Support Vector Machines (SVMs) are a type of **supervised learning algorithm** used for **classification** or **regression tasks**. They deal in:

1. **Binary Classification**:
   - SVM aims to find an **optimal hyperplane** that maximally separates two classes in the training data.
   - The **hyperplane** is the decision boundary.
   - It maximizes the **margin** (distance) between the hyperplane and the closest data points from each class.
   - New data is classified based on which side of the hyperplane it falls.
2. **Multiclass Classification**:
   - SVM inherently performs binary classification.
   - For multiclass problems, we create **binary classifiers** for each class.
   - Each classifier decides whether a data point belongs to that class or not.
   - The classifier with the highest score is chosen as the output of the SVM.
3. **Kernelized SVM**:
   - For **non-linearly separable data**, we use kernelized SVM.
   - It maps data to a higher dimension, making it linearly separable.
   - Common kernels include **polynomial**, **radial basis function (RBF)**, and **sigmoid**.

**SVMs are powerful for handling high-dimensional data and nonlinear relationships.





MultiClass Classification:

Multiclass classification is a type of classification task where we aim to **assign data points** to **more than two classes**. Unlike binary classification, where we have only two possible outcomes (e.g., spam or not spam), multiclass classification deals with **three or more distinct classes**.

Here are some key points about multiclass classification:

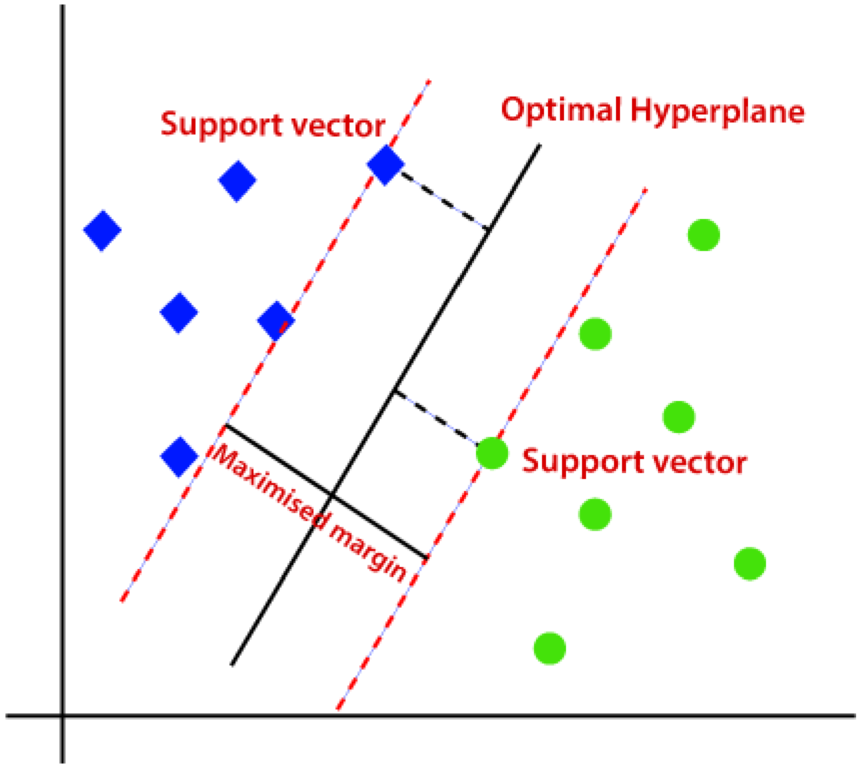
1. **Example**:
    - Imagine we have a dataset of **fruit images**. Each image can be an **orange**, an **apple**, or a **banana**.
    - Our goal is to classify each image into one of these **three classes**.

2. **Algorithms for Multiclass Classification**:
    - Several algorithms can handle multiclass problems:
        - **Naive Bayes**
        - **Decision Trees**
        - **Support Vector Machines (SVM)**
        - **Random Forest**
        - **K-Nearest Neighbours (KNN)**
        - **Logistic Regression**
3. **One-vs-One (OVO) and One-vs-All (OVA)**:
    - **OVO**: Builds a binary classifier for each pair of classes.
    - **OVA**: Trains one classifier per class against the rest.

Multiclass classification is about **handling more than two classes, making it essential for tasks like **image recognition**, **natural language processing**, and **species classification**.

How SVM Works?

- SVM aims to find an **optimal boundary (hyperplane)** between different classes.
- It transforms the data based on a selected **kernel function** to maximize separation boundaries.
- The **support vectors** (closest data points to the hyperplane) play a crucial role.



2. **Multiclass SVM**:

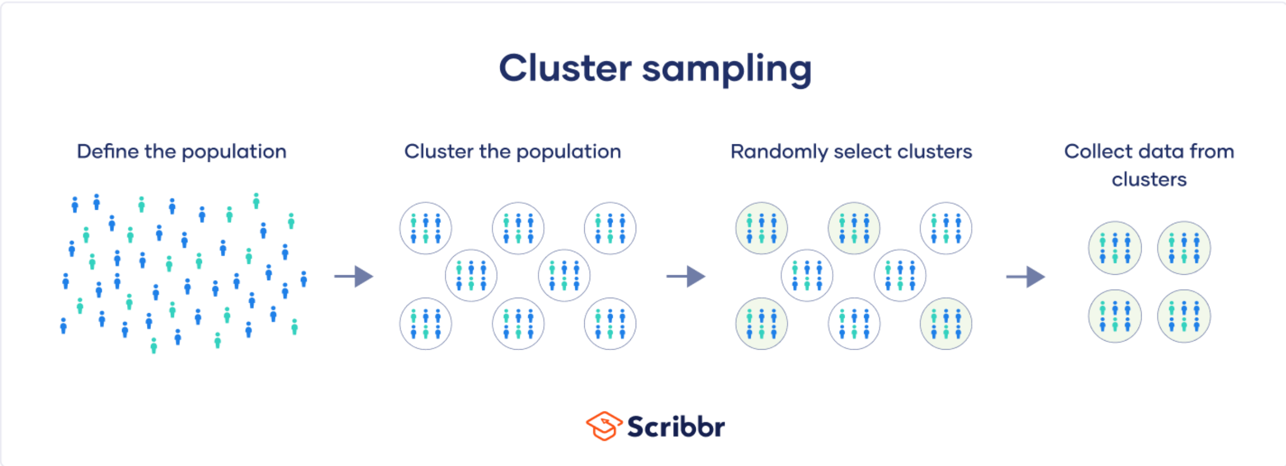
- SVM **doesn’t** natively support multiclass problems.
- We break down multiclassification into **multiple binary classification problems**.
- Common approaches:
  - **One vs One (OVO)**
  - **One vs All (OVA)**
  - **Directed Acyclic Graph (DAG)**: Hierarchical approach for multiclass SVM.

Cluster Analysis:

A technique used in **unsupervised machine learning** to **group similar data points** together based on their intrinsic characteristics.

Unlike supervised learning, where we have labelled data, clustering works with **unlabelled data**. The primary goal of clustering is to identify **patterns**, **similarities**, and **structures** within the data ***without*** any predefined classes or target variables.

- Clustering aims to organize data points into **clusters** based on their similarity.
- A **cluster** is a group of data points that share common characteristics or exhibit similar behaviour.
- Clustering algorithms evaluate the similarity between data points using metrics like **distance**, **density**, or **probability**.



1. **Applications of Clustering in Data Analytics**:
    - **Market Segmentation**: Businesses use clustering to group similar customers for targeted marketing campaigns. For example, identifying high-spending households or specific consumer segments.
    - **Streaming Services**: Clustering helps streaming platforms understand user behaviour and tailor content recommendations based on usage patterns.
    - **Sports Science**: Sports teams use clustering to group players with similar performance metrics for customized training and practice sessions.
    - **Email Marketing**: Clustering assists businesses in segmenting customers based on email engagement patterns, optimizing email campaigns.
    - **Health Insurance**: Actuaries use clustering to identify clusters of consumers with specific health insurance usage patterns.

2. **Common Clustering Algorithms**:
   - **K-Means**: A centroid-based algorithm that partitions data into K clusters. It iteratively adjusts cluster centroids to minimize the sum of squared distances from data points to their assigned centroids.
   - **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**: Identifies dense regions in data by connecting neighbouring points. It can find clusters of arbitrary shapes and handles noise effectively.
   - **Agglomerative Hierarchical Clustering**: Builds a hierarchy of clusters by iteratively merging or agglomerating data points. It forms a dendrogram that allows choosing the desired number of clusters.
   - **Affinity Propagation**: Uses message-passing between data points to find exemplars (representative points) that define clusters.
   - **Mean Shift**: Iteratively shifts the centroid of a kernel density estimate to find dense regions in the data.
3. **Example**:
   - Consider a retail company collecting information on households (income, household size, occupation, etc.).
   - Using K-Means, they identify clusters like “small family, high spenders” or “large family, low spenders” for targeted marketing.

Reference:

- https://www.baeldung.com/cs/svm-multiclass-classification
- https://www.analyticsvidhya.com/blog/2021/05/multiclass-classification-using-svm/
- https://www.geeksforgeeks.org/support-vector-machine-algorithm/
- https://www.techopedia.com/definition/30364/support-vector-machine-svm
- https://www.mygreatlearning.com/blog/multiclass-classification-explained/
- https://scikit-learn.org/stable/modules/multiclass.html
- https://builtin.com/machine-learning/multiclass-classification
- https://www.statology.org/cluster-analysis-real-life-examples/
- https://careerfoundry.com/en/blog/data-analytics/what-is-cluster-analysis/
- https://www.csee.umbc.edu/courses/graduate/CMSC691/ds_fall20/Clustering.pdf
- https://www.datacamp.com/blog/clustering-in-machine-learning-5-essential-clustering-algorithms
- https://www.surveymonkey.com/market-research/resources/how-cluster-analysis-identifies-market-and-customer-segments/
- https://www.analyticssteps.com/blogs/cluster-analysis-types-and-applications
- https://www.freecodecamp.org/news/8-clustering-algorithms-in-machine-learning-that-all-data-scientists-should-know/
- https://developers.google.com/machine-learning/clustering/clustering-algorithms
- https://machinelearningmastery.com/clustering-algorithms-with-python/
- https://www.analytixlabs.co.in/blog/types-of-clustering-algorithms/

Lokesh Patra

FET-BDS-2022-26-020

Baccalaureus Technologiae 2nd Year 4th Semester, Introduction to Data Analytics

Faculty Of Engineering & Technology

Sri Sri University, Cuttack