Machine Learning Research

# Supervised Learning

Supervised Learning Involves training a model on labelled data where a desired outcome is known. The model learns to map inputs to outputs based on the provided examples

## Logistic Regression

Models Probability of a binary outcome, outputs probability and classifies instances by setting a threshold

* Simple and easy
* Linear relationship between input and odds of outcome

Area of Applications – Email spam, disease diagnosis, credit scoring

## Support Vector Machines(SVM)

Finds the most eperate classes by maximizing the margin between them

* Effective in high dimension
* Good for both linear and non linear classifications via kernel trick
* Senstivie to kernel choice

Area of Applications – Image classification, text categorization , bioinformatics

## k-Nearest Neighbours

Classifies via majorit class among the k-nearest neighbours in the space

* Simple
* No explicit training phase
* Lazy learner
* Sensitive to choice of k and distance metric

Area if Application – pattern recognition, anomaly detection

## Naïve Bayes

Uses Bayes theorem with assumption of feature indepence to classify instances

* Fast
* Well Performing
* Assumption of Independence might not hold in application

Area of Application – Text classification, spam filtering

## Decision Trees

Decision tree split data into subsets based on value input

* Interpreability
* Both numerical and categorical data
* Prone to overbranching

Area of Application – Risk assessment, fraud dectection , customer segmentation

## Random Forest

A collection of decision trees that improves accuracy and controls overbranching by averaging multiple trees trained on different subsets of data

* Reduces over branching
* Handles large Datasets
* Requires large computational power

Area of application – image classification, healthcare diagnostics

## Gradient Boosting

Builds models sequentially to correct errors via prior models optimising accuracy

* High accuracy
* Multiple data types

Area of application – Web search ranking, risk prediction

# Regression

## Linear Regression

Models Relationship between dependent and independent variables linearly

* Simple
* -assumes linear

Area of application – risk management, sales forecasting

## Lasso Regression

Adds L1 Regularization to linear regression to perform feature selection by shrinking coeffections to zero

* Requires tuning
* Performs feature selection

Area of application – Model selection, finances

## Support Vector Regression

Uses support vector machines for regression tasks by finding a function that deviates form actualised target values by values no greater than a specified margin

* Effective in high dimensions
* Can Multi Data

Area of application – Medical diagnosis , engineering

# Unsupervised Learning

Unsupervised learning works with unlabelled data and aims to find hidden patterns or intrinsic structures within the input data

# Clustering

## K-Means

Partitions data into k clusters based on similarity, minimizing distances from each point to the centroid of its assigned cluster

* Simple
* Sensitive to initial placement of centroids
* Assumes clusters are spherical

Area of Application – Customer segmenting, market Research, image compression

## Hierarchical Clustering

Builds a hierarchy of clusters using either a bottom up or top down approach

* Does not require a predefined number of clusters
* Produces a dendrogram for visualization
* Computationally intensive (shit big o notation)

Area of Application – social network analysis, document clustering

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

Groups points that are close to each other based on distance and densityand indetifies outliers as points that lie alone in low density regions

* Can find arbitrary shaped clusters
* Robust to noise and outliers
* Requires tuning

Area of application – Geographical data analysis, fraud detection, biology

## Gaussian Mixture Models(GMM)

GMM assumes data is generated from a mixture of Gaussian distribtuitions each representing a cluster

* Can model clusters with different shapes and sizes
* Users probabilistic soft assignments

Area of application – anomaly detection, finance

# Dimensionality Reduction

## Principal Component Analysis(PCA)

PCA Reduces the dimensionality of data by transforming it to a new set (principal components) that capture the maximum variability

* Reduces complexity of data
* Helps visualization
* Assumes linear relationships

Area of applications – Data compression, noise reduction , feature extraction

## t-Distributed Stochastic Neighbour Embedding(T-SNE)

Reduces Dimension for visualization by preserving local data structure, making points stay close together

* Effective for visualization of high Dimension data
* Computationally Intensive
* Does not preserve Global Structure

Area of Application – Cluster Visualization, anomaly detection

## Linear Discriminant Analysis(LDA)

Reduces Dimension by maximiing class seperability

* Assumes normal class distribution

Area of application – Pattern Recognition, face Recognition

## Independent Component Analysis(ICA)

ICA separates a signal into additive, independent components

* Useful for blind source separation
* Sensitive to noise

Area of Application – Brain imaging, finance

## UMAP(Uniform Manifold Approximation and Projection)

UMAP reduces dimensions wile preserving the GLOBAL structure of the data using a manifold learning technique

* Preservers LOCAL and GLOBAL Structure
* Computationally efficient
* Requires toning

Area of Application – Data visualization, clustering , pattern recognition

# Association

## APriori Algorithm

Identifies frequent itemsets in transactional data and generates associative rules

* Simple
* Large dataset easy
* Computationally expensive

Area of application – Market analysis, text mining

## Eclat Algorithm

Uses Depth-first Search to find frequent itemsets, improving efficiency by reducing number of database scans

* More efficient than Apriori
* Uses vertical data format
* Requires sufficient memory

Area of Application – Market Analysis, text mining

# Reinforcement Learning

Involves Training Agents to make a sequence of decisions via reward and penalization

## Q-Learning

Learns value of actions in states to maximize reward, updates Q values based on Bellman Equation

* Off-policy learning method
* Can handle problems with stochastic transitions
* Convergence can be slow

Area of application – Robotics, game playing

## Deep Q-Network

Uses deep learning to approximate Q-Values, Enabling reinforcement learning in high dimension spasces

* Combines q learning with deep neural networks
* Handles large state spaces
* Requires extensive training

Area of application – Video games, robotics , control systems

## SARSA(State-Action-Reward-State-Action)

Learns value of policy being followed by updating Q-values based on state-Action pairs encontered

- On policy Learning Method

Takes into account policy behaviour

Sensitive to choice policy

Area of Application – Path Planning, robotics