100 Days of Machine Learning (Day : 1 – 50)

Types of machine learning:

1. Supervised Learning
2. Unsupervised Learning
3. Semi supervised learning
4. Reinforcement Learning

Types of supervised learning:

Classification
Regression
Ordinal Regression
Multi-output Regression

Available algorithm in supervised learning:

1. Linear Regression
2. Logistic Regression
3. Decision Trees
4. Random Forest
5. Support Vector Machines (SVM)
6. Naive Bayes
7. K-Nearest Neighbors (KNN)
8. Gradient Boosting Machines (GBM)
9. Neural Networks

Types of unsupervised learning:

1. Clustering
2. Dimensionality Reduction
3. Anomaly Detection
4. Association Rule Learning
5. Generative Modeling

Available algorithms unsupervised learning:

1.K-Means Clustering
2. Hierarchical Clustering
3. Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
4. Principal Component Analysis (PCA)
5. t-Distributed Stochastic Neighbor Embedding (t-SNE)
6. Autoencoders
7. Anomaly Detection Algorithms
8. Association Rule Learning
9. Generative Adversarial Networks (GANs)

Types of semi- supervised learning:

1. Self-training
2. Co-training
3. Semi-supervised Support Vector Machines (S3VM)
4. Graph-based Methods
5. Generative Models

Available algorithms of semi-supervised learning:

1. Self-training

2. Co-training
3. Semi-supervised Support Vector Machines (S3VM)
4. Graph-based Methods
5. Generative Models
6. Entropy Regularization

Reinforcement learning types:

1. Model-Based RL
2. Model-Free RL
3. Exploration Strategies
4. Multi-Agent RL

Available algorithms reinforcement learning:

1. Q-Learning
2. Deep Q-Networks (DQN)
3. Policy Gradient Methods
4. Actor-Critic Methods
5. Multi-Agent RL Algorithms
6.Evolutionary Strategies

Types of ML based on training:

- Online learning
- Offline learning

Online :

1. Online Gradient Descent
2. Perceptron
3. Stochastic Gradient Descent (SGD)
4. Online Support Vector Machines (SVM)
5. Adaptive Learning Rate Methods
6. Online Random Forests
7. Memory-Based Methods
8. Reinforcement Learning

Offline:

1. Supervised Learning
2. Unsupervised Learning
3. Semi supervised learning
4. Reinforcement Learning
5. Feature Selection and Engineering

Types of ML based on learning:

- Instance based learning
- Model based learning

Instance based learning:

1. K-Nearest Neighbors (KNN)
2. Locally Weighted Learning (LWL)
3. Case-Based Reasoning (CBR)
4. Learning Vector Quantization (LVQ)
5. Adaptive Resonance Theory (ART)

Model based learning:

1. Linear Regression
2. Logistic Regression
3. Decision Trees
4. Random Forests
5. Gradient Boosting Machines (GBM)
6. Neural Networks

What is tensors:

Is nothing but data structure

Types:

1. Scalar (0D Tensor)
2. Vector (1D Tensor)
3. Matrix (2D Tensor)
4. 3D Tensor and Higher-Dimensional Tensors
5. Sparse Tensor

Data gathering techniques:

- CSV file
- Fetch API
- JSON/SQL
-Web Scraping
-Public Datasets
-APIs (Application Programming Interfaces)

-Data Augmentation

-Data Labeling Services

-Data Purchase or Licensing

Methods of understanding data of ML:

1. Data Exploration and Visualization
2. Summary Statistics
3. Data Cleaning and Preprocessing
4. Correlation Analysis
5. Dimensionality Reduction
6. Feature Importance and Selection
7. Time Series Analysis
8. Cluster Analysis
9. Association Rule Mining

What are the basic questions should we ask while getting a data set:

1. How big is the data
2. How does the data look like
3. What is the data type of cols
4. Are there any missing values
5. How does the data look mathematically
6. Are there duplicate values
7. How is the correlation between cols

Exploratory Data Analysis (EDA) : Pandas Profiler

Some key aspects of EDA in machine learning:

1. Data Inspection
2. Summary Statistics
3. Data Visualization
4. Feature Distribution Analysis
5. Target Variable Analysis
6. Correlation Analysis
7. Dimensionality Reduction
8. Data Preprocessing Insights

Types of EDA:

1. Univariate Analysis
- Histograms
- Bar plots
- Box plots
- Descriptive statistics

2. Bivariate Analysis
- Scatter plots
- Pair plots
- Correlation analysis

3. Multivariate Analysis
- Heatmaps
- Dimensionality reduction techniques

4. Distribution Analysis
- Kernel density estimation

- QQ plots

5. Temporal Analysis
- Time series plots
- Decomposition

6. Spatial Analysis
- Choropleth maps
- Spatial autocorrelation

Pandas Profiler:

*various types of information in its report, including:
1. Summary Statistics
2. Distribution Statistics
3. Correlation Analysis
4. Missing Values Analysis
5. Categorical Variables Analysis
6. Warnings and Recommendations

Feature Engineering:

Mainly 4 types of feature engineering
-Feature Transformation
-Feature Construction
-Feature selection
-Feature Extraction

*Types of feature engineering:

1. Feature Selection
2. Feature Creation
3. Feature Encoding
4. Handling Missing Values
5. Feature Scaling
6. Feature Transformation

*Steps of feature engineering:

1. Data Understanding
2. Exploratory Data Analysis (EDA)
3. Feature Selection
4. Feature Creation
5. Feature Encoding
6. Handling Missing Values
7. Feature Scaling
8. Feature Transformation
9. Validation and Iteration
10. Documentation and Reproducibility

Feature Transformation

-Missing value imputation

-Handling categorical features

-Outliers detection

-Feature scaling

* Feature Scaling - Standardization algorithms:

-Z-score Standardization
-MinMax Scaling
-Robust Scaling
-MaxAbs Scaling.
-Quantile Transformation

*Feature Scaling - Normalization algorithms:

- MinMax Scaling (Normalization)
- Standardization
- Robust Scaling
- MaxAbs Scaling
- Unit Vector Scaling (also known as L2 normalization)

Encoding categorical data and it's types:

1.One-Hot Encoding
2.Label Encoding
3.Ordinal Encoding
4.Frequency Encoding
5.Target Encoding

One-Hot Encoding : binary representation

*Types of One-Hot Encoding:

-Basic One-Hot Encoding

-Dummy Encoding
-One-Hot Encoding with Drop
-One-Hot Encoding with Pandas

Column Transformation

Key Features of ColumnTransformer:
-Selective Transformation
-Pipeline Integration
-Handling Missing Values
-Parallelization

Machine learning pipelines

Steps:
1. Data Preprocessing
2. Feature Engineering
3. Model Training
4. Model Evaluation
5. Model Deployment

Machine learning transformers:

1. Function transformer
2. Power transformer
3. Quartile transformer

* 1. Function Transformer / Variable transformation
-Log Transform
-Reciprocal Transform

-Square Root Transform

* 2. Power transformer

-Box - Cox Transform

- Yeo -Johnson Transform

Binning , Binarization and Discretization:

-Quantile Binning

-K-Means Binning

Feature Engineering:

* Handling mixed variables:

1. Separate Numerical and Categorical Variables
2. Apply Appropriate Preprocessing Techniques
- For Numerical Variables
- For Categorical Variables
3. Use Feature Engineering Techniques
4. Combine Numerical and Categorical Variables
5. Use Pipeline for Automated Preprocessing
6. Consider Model-specific Requirements

*Handling date and time variables

1. Extract Components
2. Create Time-based Features
3. Encode Cyclical Features

4. Handle Periodicity
5. Consider Time Zones and Daylight Saving Time
6. Deal with Missing Values
7. Use Domain Knowledge
8. Visualize Temporal Patterns

Handling missing data
- Numerical Data
- Categorical Data

*Numerical Data:
1. Imputation
2. Interpolation
3. **Drop Missing Values

* Categorical Data:
1. Imputation
2. Label Encoding
3. One-Hot Encoding
4. Drop Missing Values
5. Use a Separate Category
6. Consider Multiple Imputation

* More topic on feature engineering (Missing values):
-Missing Indicator
-Random Sample Imputation
-KNN Imputer
-Multivariate Imputation

-Multivariate Imputation by Chained Equations for Missing Value

-MICE Algorithm

-Iterative Imputer

Outliers

*Handling Outliers steps:

1. Data Understanding
2. Visualization
3. Statistical Methods
4. Transformations
5. Winsorization
6. Trimming
7. Robust Algorithms
8. Domain Knowledge

* Algorithms to handle outliers:

1. Z-Score Method
2. Interquartile Range (IQR) Method
3. Modified Z-Score Method
4. Winsorization
5. Trimming
6. Robust Statistical Methods
7. Data Transformation
8. Clustering-Based Methods
9. Isolation Forest
10. Elliptic Envelope
11. Local Outlier Factor (LOF)

12. One-Class SVM

Feature Construction

-Feature Splitting

*Feature Construction

1. Polynomial Features
2. Interaction Terms
3. Derived Variables
4. Time-Based Features
5. Text-Based Features

*Feature Splitting

1. Categorical Feature Splitting
2. Numerical Feature Binning
3. Date Feature Decomposition
4. Text Feature Tokenization
5. Geospatial Feature Decomposition

Feature selection and extraction for dimensionality reduction

The Curse of Dimensionality

1. Sparsity of Data
2. Increased Computational Complexity
3. Difficulty in Visualization

4. Increased Risk of Overfitting

*Techniques to Mitigate the Curse of Dimensionality:

1. Feature Selection
2. Feature Extraction
3. Regularization
4. Dimensionality Reduction
5. Clustering and Data Compression
6. Domain Knowledge

Principle Component Analysis -PCA (Feature extraction technique):

*Key Concepts of PCA:

1. Variance Maximization
2. Orthogonality
3. Dimensionality Reduction
4. Eigenvalue Decomposition

*Steps of PCA:

1. Standardization
2. Covariance Matrix Computation
3. Eigenvalue Decomposition
4. Principal Component Selection
5. Projection

*Applications of PCA:

1. Dimensionality Reduction
2. Data Visualization.
3. Feature Extraction

*Available algorithms for PCA:

1. Eigenvalue Decomposition (EVD)
2. Singular Value Decomposition (SVD)
3. Randomized PCA
4. Incremental PCA (IPCA)
5. Kernel PCA