Data Science & Machine Learning Basics

Rafiq Islam

2024-09-20

Table of contents

This page is my personal repository of most common and useful machine learning algorithms using Python and other data science tricks and tips.

Data Science

Data science involves extracting knowledge from structured and unstructured data. It combines principle from statistics, machine learning, data analysis, and domain knoledge to understand and interpret the data

Data Collection & Accuisition

- **Web srcaping:** Data collection through Webscraping
- API integration
- Data Lakes, Data Warehouses

Data Cleaning & Preprocessing

- Handling Missing Values
- Data Transformation
- Feature Engineering and Selection
- Encoding Categorical Variables
- Handling Outliers

Exploratory Data Analysis (EDA)

- Descriptive Statistics
- Data Visualization
- Identifying Patterns, Trends, Correlations

Statistical Methods

- **ANOVA - Categorical Features':** How do we treat the categorical features for our data science project?
- Hypothesis Testing
- Probability Distributions
- Inferential Statistics
- Sampling Methods

Big Data Techniques

- Hadoop, Spark

- Distributed Data Storage (e.g., HDFS, NoSQL)
- Data PipeLines, ETL (Extract, Transform, Load)

Machine Learning Algorithms

Supervised Learning

(Training with labeled data: input-output pairs)

Regression

Parametric

- Simple Linear Regression
- Multiple Linear Regression
- Polynomial Regression

Non-Parametric

- K-Nearest Neighbor (KNN) Regression
- Decesion Trees Regression
- Random Forest Regression
- Support Vector Machine (SVM) Regression

Classification

Parametric

- Logistic Regression
- Naive Bayes
- Linear Discriminant Analysis (LDA)

- Quadratic Discriminant Analysis (QDA)

Non-Parametric

- KNN Classification
- Decision Tree Classification
- Random Forest Classification
- Support Vector Machine (SVM) Classification

Multi-Class Classification

- Multi-class Classification

Bayesian or Probabilistic Classification

- What is Bayesian or Probabilistic Classification?
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)
- Naive Bayes
- Bayesian Network Classifier (Tree Augmented Naive Bayes (TAN))

Non-probabilistic Classification

- Support Vector Machine (SVM) Classification
- Decision Tree Classification
- Random Forest Classification
- KNN Classification
- Perceptron

Unsupervised Learning

(Training with unlabeled data)

Clustering

- k-Means Clustering
- Hierarchical Clustering
- DBSCAN (Density-Based Spatial Clustering)
- Gaussian Mixture Models (GMM)

Dimensionality Reduction

- Principal Component Analysis
- Latent Dirichlet Allocation (LDA)
- t-SNE (t-distributed Stochastic Neihbor Embedding)
- Factor Analysis
- Autoencoders

Anomaly Detection

- Isolation Forests
- One-Class SVM

Semi-Supervised Learning

(Combination of labeled and unlabeled data)

- Self-training
- Co-training
- Label Propagation

Reinforcement Learning

(Learning via rewards and penalties)

- Markov Decision Process (MDP)
- Q-Learning
- Deep Q-Networks (DQN)
- Policy Gradient Method

Deep Learnings

- PyTorch
- Artificial Neural Networks (ANN)
- Convolutional Neural Networks (CNN)
- Recurrent Neural Networks (RNN)
- Long Short-Term Memory (LSTM)
- Generative Adversarial Networks (GAN)

Model Evaluation and Fine Tuning

Model Evaluation Metrics

- **For Regression:** Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R^2 score
- **For Classification:** Accuracy, Precision, Recall, F1 Score, ROC-AUC
- **Cross-validation:** kFold, Stratified k-fold, leave-one-out

Model Optimization

- **Bias-Variance:** Bias Variance Trade off

- **Hyperparameter Tuning:** Grid Search, Random Search, Bayesian Optimization
- **Features Selection Techniques:** Recursive Feature Elimination (RFE), L1 or Rasso Regurlarization, L2 or Ridge Regularization
- **Model Interpretability:** SHAP (Shapley values), LIME (Local Interpretable Model-agnostic Explanations)

Ensemble Methods

- **Bagging:** Random Forest, Bootstrap Aggregating
- **Boosting:** Gradient Boosting, AdaBoost, XGBoost, CatBoost
- **Stacking:** Stacked Generalization

Learning Type	Parametric	Non-Parametric
Supervised	• Simple Linear Regression • Multiple Linear Regression • Polynomial Regression • Logistic Regression • Naive Bayes	• KNN Regression and Classification • Decision Trees • Random Forest • Support Vector Machine (SVM)
Unsupervised	Principle Component Analysis (PCA) Gaussian Mixture Model (GMM) Latent Dirichilet Allocation (LDA)	K-Means Hierarchial Clustering Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
Semi-Supervised	Self-training	
Reinforcement Learning	Q-Learning DQN Policy Gradient	
Dimensionality Reduction	Principle Component Analysis (PCA) Linear Discriminant Analysis (LDA)	t-SNE Autoencoders
Ensemble Methods	Bagging Gradient Boosting	Stacking

Learning Type	Parametric	Non-Parametric
Deep Learning	Artificial Neural Networks (ANN) Convolutional Neural Networks (CNN) Recurrent Neural Networks (RNN) Long Short-Term Memory (LSTM) Generative Adversarial Networks (GAN)	

Techniques	Description
Categorical Features	How do we treat the categorical features for our data science project?
Webscraping	Data collection through Webscraping
Bias-Variance	Model Fine Tuning: Bias-Variance Trade Off
Regularization	Model Fine Tuning: Regularization

You may also like