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 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**](../../dsandml/datacollection/index.qmd)**:** 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’**](../../dsandml/dataengineering/index.qmd)**:** 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)

(Training with labeled data: input-output pairs)

#### **Regression**

##### Parametric

* [Simple Linear Regression](../../dsandml/simplelinreg/index.qmd)
* [Multiple Linear Regression](../../dsandml/multiplelinreg/index.qmd)
* [Polynomial Regression](../../dsandml/polyreg/index.qmd)

##### Non-Parametric

* [K-Nearest Neighbor (KNN) Regression](../../dsandml/knn/index.qmd)
* [Decesion Trees Regression](../../dsandml/decisiontree/index.qmd)
* [Random Forest Regression](../../dsandml/randomforest/index.qmd)
* [Support Vector Machine (SVM) Regression](../../dsandml/svm/index.qmd)

#### **Classification**

##### Parametric

* [Logistic Regression](../../dsandml/logreg/index.qmd)
* [Naive Bayes](../../dsandml/naivebayes/index.qmd)
* [Linear Discriminant Analysis (LDA)](../../dsandml/lda/index.qmd)
* Quadratic Discriminant Analysis (QDA)

##### Non-Parametric

* [KNN Classification](../../dsandml/knn/index.qmd)
* [Decision Tree Classification](../../dsandml/decisiontree/index.qmd)
* [Random Forest Classification](../../dsandml/randomforest/index.qmd)
* [Support Vector Machine (SVM) Classification](../../dsandml/svm/index.qmd)

##### Multi-Class Classification

* [Multi-class Classification](../../dsandml/multiclass/index.qmd)

##### Bayesian or Probabilistic Classification

* [What is Bayesian or Probabilistic Classification?](../../dsandml/bayesianclassification/index.qmd)
* [Linear Discriminant Analysis (LDA)](../../dsandml/lda/index.qmd)
* Quadratic Discriminant Analysis (QDA)
* Naive Bayes
* Bayesian Network Classifier (Tree Augmented Naive Bayes (TAN))

##### Non-probabilistic Classification

* [Support Vector Machine (SVM) Classification](../../dsandml/svm/index.qmd)
* [Decision Tree Classification](../../dsandml/decisiontree/index.qmd)
* [Random Forest Classification](../../dsandml/randomforest/index.qmd)
* [KNN Classification](../../dsandml/knn/index.qmd)
* Perceptron

(Training with unlabeled data)

##### Clustering

* [k-Means Clustering](../../dsandml/kmeans/index.qmd)
* Hierarchical Clustering
* DBSCAN (Density-Based Spatial Clustering)
* Gaussian Mixture Models (GMM)

##### Dimensionality Reduction

* [Principal Component Analysis](../../dsandml/pca/index.qmd)
* Latent Dirichlet Allocation (LDA)
* t-SNE (t-distributed Stochastic Neihbor Embedding)
* Factor Analysis
* Autoencoders

##### Anomaly Detection

* Isolation Forests
* One-Class SVM

(Combination of labeled and unlabeled data)

* Self-training
* Co-training
* Label Propagation

(Learning via rewards and penalties)

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

* [PyTorch](../../dsandml/pytorch/index.qmd)
* Artificial Neural Networks (ANN)
* Convolutional Neural Networks (CNN)
* Recurrent Neural Networks (RNN)
* Long Short-Term Memory (LSTM)
* Generative Adversarial Networks (GAN)

#### Model Evaluation Metrics

* **For Regression:** Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), score
* **For Classification:** [Accuracy, Precision, Recall, F1 Score, ROC-AUC](../../dsandml/classificationmetrics/index.qmd)
* **Cross-validation:** kFold, Stratified k-fold, leave-one-out

#### Model Optimization

* **Bias-Variance:** [Bias Variance Trade off](../../dsandml/biasvariance/index.qmd)
* **Hyperparameter Tuning:** Grid Search, Random Search, Bayesian Optimization
* **Features Selection Techniques:** Recursive Feature Elimination (RFE), [L1 or Rasso Regurlarization](../../dsandml/regularization/index.qmd), [L2 or Ridge Regularization](../../dsandml/regularization/index.qmd)
* **Model Interpretability:** SHAP (Shapley values), LIME (Local Interpretable Model-agnostic Explanations)

#### Ensemble Methods

* **Bagging:** [Random Forest](../../dsandml/randomforest/index.qmd), Bootstrap Aggregating
* **Boosting:** [Gradient Boosting](../../dsandml/gradientboosting/index.qmd), AdaBoost, XGBoost, CatBoost
* **Stacking:** Stacked Generalization

| Learning Type | Parametric | Non-Parametric |
| --- | --- | --- |
| Supervised | * [Simple Linear Regression](../../dsandml/simplelinreg/index.qmd) * [Multiple Linear Regression](../../dsandml/multiplelinreg/index.qmd) * [Polynomial Regression](../../dsandml/polyreg/index.qmd) * [Logistic Regression](../../dsandml/logreg/index.qmd) * [Naive Bayes](../../dsandml/naivebayes/index.qmd) | * [KNN Regression and Classification](../../dsandml/knn/index.qmd) * [Decision Trees](../../dsandml/decisiontree/index.qmd) * [Random Forest](../../dsandml/randomforest/index.qmd) * Support Vector Machine (SVM) |
| Unsupervised | [Principle Component Analysis (PCA)](../../dsandml/pca/index.qmd)  Gaussian Mixture Model (GMM)  Latent Dirichilet Allocation (LDA) | [K-Means](../../dsandml/kmeans/index.qmd)  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)](../../dsandml/pca/index.qmd)  Linear Discriminant Analysis (LDA) | t-SNE  Autoencoders |
| Ensemble Methods | Bagging  [Gradient Boosting](../../dsandml/gradientboosting/index.qmd) | Stacking |
| 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](../../dsandml/dataengineering/index.qmd) | How do we treat the categorical features for our data science project? |
| [Webscraping](../../dsandml/datacollection/index.qmd) | Data collection through Webscraping |
| [Bias-Variance](../../dsandml/biasvariance/index.qmd) | Model Fine Tuning: Bias-Variance Trade Off |
| [Regularization](../../dsandml/regularization/index.qmd) | Model Fine Tuning: Regularization |

**You may also like**