Data Science with Python (Coding Repo) Outline

This Github repository is the part of resource development activity for EHSAS Center. This complete git codebook targets students and other audience from data science and machine learning background. This actually targets students specifically from course Data Science, Machine Learning offered as an elective in Department of Computer Science, Habib University and individuals who are interested in learning basics of machine learning and data science.

Also, this codebook contains Data folder which contains variety of datasets which individuals can use for different purpose of data analysis tasks and practice as well.

The codebook is available online at below link:

Topics Covered in Data Science with Python Repo →

1. Introduction to Numpy
    a. Initialization of 1D array in Python
    b. Initialization of 2D array in Python
    c. Slicing and Indexing in Numpy
    d. Understanding - Pass By Reference
    e. Broadcast Operation in Numpy
    f. Matrix Operations
    g. Statistical Functions
    h. Miscellaneous Functions

2. Pandas for Machine Learning
    a. Introduction to Pandas
    b. Understanding Series & DataFrames
    c. Loading CSV,JSON
    d. Connecting databases
    e. Descriptive Statistics
    f. Accessing subsets of data - Rows, Columns, Filters
    g. Handling Missing Data
    h. Dropping rows & columns
    i. Handling Duplicates
    j. Function Application - map, apply, groupby, rolling, str
    k. Merge, Join & Concatenate
    l. Pivot-tables
    m. Normalizing JSON

3. Plotting
    a. Introduction to matplotlib

4. Linear Models for Regression & Classification
    a. Simple Linear Regression using Ordinary Least Squares
    b. Gradient Descent Algorithm
    c. Regularized Regression Methods - Ridge, Lasso, ElasticNet
    d. Logistic Regression for Classification
    e. OnLine Learning Methods - Stochastic Gradient Descent & Passive Aggrasive
    f. Robust Regression - Dealing with outliers & Model errors
    g. Polynomial Regression
    h. Bias-Variance Tradeoff

5. Data Pre-processing using scikit-learn
    a. Introduction to Preprocessing
    b. StandardScaler
    c. MinMaxScaler
    d. RobustScaler
    e. Normalization
    f. Binarization
    g. Encoding Categorical (Ordinal & Nominal) Features
    h. Imputation
    i. Polynomial Features
    j. Custom Transformer
    k. Text Processing
    l. CountVectorizer
    m. Tfldf
    n. HashingVectorizer
    o. Image using skimage

6. Decision Trees
    a. Introduction to Decision Trees
    b. The Decision Tree Algorithms
    c. Decision Tree for Classification
    d. Decision Tree for Regression
    e. Advantages & Limitations of Decision Trees

7. Naive Bayes
    a. Generative vs Discriminative Models
    b. Introduction Bayes' Theorem
    c. Maximum Likelihood Estimate
    d. Naive Bayes Classifier
    e. Gaussian Naive Bayes
    f. Multinomial Naive Bayes
    g. Burnolis' Naive Bayes
    h. Naive Bayes for out-of-core

8. Composite Estimators using Pipeline & Feature Unios
    a. Introduction to Composite Estimators
    b. Pipelines
    c. Transformed Target Regressor
    d. FeatureUnions
    e. ColumnTransformer
    f. GridSearch on pipeline

9. Model Selection & Evaluation
    a. Cross Validation
    b. Hyperparameter Tuning
    c. Model Evaluation
    d. Model Persistence
    e. Validation Curves
    f. Learning Curves

10. Feature Selection Techniques
    a. Introduction to Feature Selection
    b. VarianceThreshold
    c. Chi-squared stats
    d. ANOVA using f_classif
    e. Univariate Linear Regression Tests using f_regression
    f. F-score vs Mutual Information
    g. Mutual Information for discrete value
    h. Mutual Information for continues value
    i. SelectKBest
    j. SelectPercentile
    k. SelectFromModel
    l. Recursive Feature Elimination

11. Nearest Neighbours
    a. Fundamentals of Nearest Neighbor
    b. Unsupervised Nearest Neighbors
    c. Nearest Neighbors for Classification
    d. Nearest Neighbors for Regression
    e. Nearest Centroid Classifier

12. Clustering Techniques
    a. Introduction to Unsupervised Learning
    b. Clustering
    c. Similarity or Distance Calculation
    d. Clustering as an Optimization Function
    e. Types of Clustering Methods
    f. Partitioning Clustering - KMeans & Meanshift
    g. Hierarchical Clustering - Agglomerative
    h. Density Based Clustering - DBSCAN
    i. Measuring Performance of Clusters
    j. Comparing all clustering methods

13. Anomaly Detection
    a. What are Outliers ?
    b. Statistical Methods for Univariate Data
    c. Using Gaussian Mixture Models
    d. Fitting an elliptic envelope
    e. Isolation Forest
    f. Local Outlier Factor
    g. Using clustering method like DBSCAN

14. Support Vector Machines
    a. Introduction to Support Vector Machines
    b. Maximal Margin Classifier
    c. Soft Margin Classifier
    d. SVM Algorithm for Classification
    e. SVM
    f. SVM for Regression
    g. Hyper-parameters in SVM

15. Dealing with imbalanced classes
    a. What are imbalanced classes & their impact ?
    b. OverSampling
    c. UnderSampling
    d. Connecting Sampler to pipelines
    e. Making classification algorithm aware of Imbalance
    f. Anomaly Detection

16. Ensemble Methods
    a. Introduction to Ensemble Methods
    b. RandomForest
    c. AdaBoost
    d. GradientBoostingTree
    e. VotingClassifier