Advanced Supervised and Unsupervised Learning Algorithms

By Dr. Vishwanath Rao

Python for Data Science

- Python Data Science-Centric Libraries
- SciPy
- NumPy
- pandas
- Scikit-learn
- Matplotlib
- Seaborn
- Python Dev Tools and REPLs
- IPython
- Jupyter Notebooks
- Anaconda
- Summary

Data Visualization in Python

- Why Do I Need Data Visualization?
- Data Visualization in Python
- Getting Started with matplotlib
- A Basic Plot
- Scatter Plots
- Figures
- Saving Figures to a File
- Seaborn
- Getting Started with seaborn
- Histograms and KDE
- Plotting Bivariate Distributions
- Scatter Plots in seaborn
- Pair plots in seaborn
- Heatmaps
- A Seaborn Scatterplot with Varying Point Sizes and Hues
- Summary

Introduction to NumPy

- What is NumPy?
- The First Take on NumPy Arrays
- The ndarray Data Structure
- Understanding Axes
- Indexing Elements in a NumPy Array
- Re-Shaping
- Commonly Used Array Metrics
- Commonly Used Aggregate Functions
- Sorting Arrays

Vectorization
Vectorization Visually
Broadcasting
Broadcasting Visually
Filtering
Array Arithmetic Operations
Reductions: Finding the Sum of Elements by Axis
Array Slicing
2-D Array Slicing
The Linear Algebra Functions
Summary

Introduction to pandas

What is pandas?
The DataFrame Object
The DataFrame's Value Proposition
Creating a pandas DataFrame
Getting DataFrame Metrics
Accessing DataFrame Columns
Accessing DataFrame Rows
Accessing DataFrame Cells
Deleting Rows and Columns
Adding a New Column to a DataFrame
Getting Descriptive Statistics of DataFrame Columns
Getting Descriptive Statistics of DataFrames
Sorting DataFrames
Reading From CSV Files
Writing to a CSV File
Summary

Repairing and Normalizing Data

Repairing and Normalizing Data
Dealing with the Missing Data
Sample Data Set
Getting Info on Null Data
Dropping a Column
Interpolating Missing Data in pandas
Replacing the Missing Values with the Mean Value
Scaling (Normalizing) the Data
Data Preprocessing with scikit-learn
Scaling with the scale() Function
The MinMaxScaler Object
Summary

Defining Data Science

What is Data Science?
Data Science, Machine Learning, AI?
The Data Science Ecosystem
Tools of the Trade

The Data-Related Roles
Data Scientists at Work
Examples of Data Science Projects
The Concept of a Data Product
Applied Data Science at Google
Data Science and ML Terminology: Features and Observations
Terminology: Labels and Ground Truth
Label Examples
Terminology: Continuous and Categorical Features
Encoding Categorical Features using One-Hot Encoding Scheme
Example of 'One-Hot' Encoding Scheme
Gartner's Magic Quadrant for Data Science and Machine Learning Platforms (a Labeling Example)
Machine Learning in a Nutshell
Common Distance Metrics
The Euclidean Distance
Decision Boundary Examples (Object Classification)
What is a Model?
Training a Model to Make Predictions
Types of Machine Learning
Supervised vs Unsupervised Machine Learning
Supervised Machine Learning Algorithms
Unsupervised Machine Learning Algorithms
Which ML Algorithm to Choose?
Bias-Variance (Underfitting vs Overfitting) Trade-off
Underfitting vs Overfitting (a Regression Model Example) Visually
ML Model Evaluation
Mean Squared Error (MSE) and Mean Absolute Error (MAE)
Coefficient of Determination
Confusion Matrix
The Binary Classification Confusion Matrix
The Typical Machine Learning Process
A Better Algorithm or More Data?
The Typical Data Processing Pipeline in Data Science
Data Discovery Phase
Data Harvesting Phase
Data Cleaning/Priming/Enhancing Phase
Exploratory Data Analysis and Feature Selection
Exploratory Data Analysis and Feature Selection Cont'd
ML Model Planning Phase
Feature Engineering
ML Model Building Phase
Capacity Planning and Resource Provisioning
Communicating the Results
Production Roll-out
Data Science Gotchas
Summary

Overview of the scikit-learn Library

The scikit-learn Library
The Navigational Map of ML Algorithms Supported by scikit-learn

- Developer Support
- scikit-learn Estimators, Models, and Predictors
- Annotated Example of the LinearRegression Estimator
- Annotated Example of the Support Vector Classification Estimator
- Data Splitting into Training and Test Datasets
- Data Splitting in scikit-learn
- Cross-Validation Technique
- Summary

Classification Algorithms (Supervised Machine Learning)

- Classification (Supervised ML) Use Cases
- Classifying with k-Nearest Neighbors
- k-Nearest Neighbors Algorithm Visually
- Decision Trees
- Decision Tree Terminology
- Decision Tree Classification in the Context of Information Theory
- Using Decision Trees
- Properties of the Decision Tree Algorithm
- The Simplified Decision Tree Algorithm
- Random Forest
- Properties of the Random Forest Algorithm
- Support Vector Machines (SVMs)
- SVM Classification Visually
- Properties of SVMs
- Dealing with Non-Linear Class Boundaries
- Logistic Regression (Logit)
- The Sigmoid Function
- Logistic Regression Classification Example
- Logistic Regression's Problem Domain
- Naive Bayes Classifier (SL)
- Naive Bayesian Probabilistic Model in a Nutshell
- Bayes Formula
- Document Classification with Naive Bayes
- Summary

Deep Study on Supervised learning

- Linear Models
- Linear and Quadratic Discriminant Analysis
- Kernel ridge regression
- Support Vector Machines
- Stochastic Gradient Descent
- Nearest Neighbors
- Gaussian Processes
- Cross decomposition
- Naive Bayes
- Decision Trees
- Ensemble methods
- Multiclass and multioutput algorithms
- Feature selection

Semi-supervised learning
Isotonic regression
Probability calibration
Neural network models (supervised)

Unsupervised Machine Learning Algorithms

PCA
PCA and Data Variance
PCA Properties
Importance of Feature Scaling Visually
Unsupervised Learning Type: Clustering
Clustering vs Classification
Clustering Examples
k-means Clustering
k-means Clustering in a Nutshell
k-means Characteristics
Global vs Local Minimum Explained
Summary

Deep study on Unsupervised learning

Gaussian mixture models
Manifold learning
Clustering
Biclustering
Decomposing signals in components (matrix factorization problems)
Covariance estimation
Novelty and Outlier Detection
Density Estimation
Neural network models (unsupervised)