Table of Contents

01.09: Deep Dive into Pandas and Data Manipulation

Introduction & Getting Started with Pandas

- Overview of pandas.
- Setting up the environment.
- Intro to DataFrames: Creation, basic operations.

Pandas Operations and Configurations

- Top functions: `head()`, `tail()`, `describe()`, etc.
- Configuring pandas options for a tailored environment.

Advanced Data Handling

- Data type conversions.
- String operations.
- Date-time operations.
- Handling missing data: Impute, drop, etc.

Data Transformation

- `apply()`, `map()`, and `applymap()`: When and how to use.
- Transforming DataFrames.
- Groupby, aggregation, and summary statistics.
- Reshaping DataFrames: Pivot, stack, and their utilities.

Merging and Concatenation Techniques

- Merging: Using `merge()` and `join()`.
- Concatenation: Using `concat()`.
- Mapping variables to groups: Real-world use cases.

04.09.: Exploratory Data Analysis & Visualization Mastery

Introduction to Data Visualization

- Basics of plotting with pandas.
- Correlation and statistical functions overview.
- Intro to Matplotlib: Purpose and importance.

Matplotlib Essentials

- Setting marker type, colors.
- MATLAB vs. Object-oriented syntax.
- Setting titles, labels, and customizing axes.
- Incorporating grids, legends, and aesthetics.

Advanced Plotting Techniques

- Deep dive into Histograms: Bin sizing, stacking, etc.
- Working with Subplots: Creating complex visual narratives.
- Matplotlib wrappers: Integrating pandas and Seaborn for easier plotting.
- Heatmaps: Importance and generation using Seaborn.

Hands-on Workshop

Students get a dataset to perform:

- Data manipulation using pandas.
- EDA and visualization.
- Insights generation and presentation.

Discussion, Presentations, & Wrap-up

Students present their insights from the workshop.

04.10: Fundamentals of Machine Learning and Scikit-Learn

Introduction and Setting the Stage (30 mins)

The importance of machine learning in modern research.
Overview of the four-day workshop.

Understanding Machine Learning (1.5 hrs)

Recap: What is machine learning? A theoretical and historical perspective.
Recap: Supervised vs. Unsupervised learning.

Getting Started with Scikit-Learn (2 hrs)

A deep dive into Scikit-Learn: Origin, features, and benefits.
Live coding session: Input, load data, and basic operations.

Activity

Review an approach that utilizes Scikit-Learn for its machine learning applications.

05.10: Diving into Supervised Learning Techniques

Linear Models in Scikit-Learn (1.5 hrs)

Recap: The theory behind linear regression and logistic regression.
Practical application using Scikit-Learn.

Tree-based Models and Ensemble Techniques (2.5 hrs)

Dive into decision trees: What is it all about?
Introduction to ensemble methods: Bagged trees and random forests.
Practical session: Implementing and visualizing decision trees and random forests.

Activity

Analyze a given dataset and select an appropriate supervised learning model. Document the rationale behind the choice.

06.10: Unsupervised Learning and Introduction to Neural Networks

Clustering and Dimensionality Reduction (2 hrs)

K-means clustering: theory and applications.
Principal Component Analysis (PCA)

Neural Networks: A Glimpse into Deep Learning (2 hrs)

Recap: Historical and theoretical introduction to neural networks.
Introduction to TensorFlow and its relevance in the ML ecosystem.

Activity

Use PCA on a high-dimensional dataset and discuss potential use cases of the transformed data in research.

01.11: Advanced TensorFlow Applications

Fashion-MNIST and Neural Network Training (1.5 hrs)

A detailed look at the Fashion-MNIST dataset.
The theory behind training a neural network: Loss functions, gradient descent, and optimizers.
Practical session on training a neural network using TensorFlow.

Tabular Data and Neural Network Applications (2 hrs)

Introduction to the Titanic dataset and its historical significance.
Practical session: Feature engineering, ti.data API, and model evaluation.

Closing Session (30 mins)

Reflection on learnings and applications in academic research.
Feedback and open discussion on potential machine learning research collaborations.

02.11: Real-world Application and Forward Look

Comprehensive Exercise: Kaggle Challenge (1,5-2 hrs)

Students are presented with a popular dataset from Kaggle (or similar platforms). The aim: allowing all students to apply some core concepts they've learned over the past course days.

- Data cleaning and manipulation with pandas.
- Visualization and exploratory data analysis.
- Machine learning modeling: Both supervised and unsupervised methods as relevant.
- Evaluation of their machine learning models.

Presentations (0,5-1 hrs)

- Each team presents their approach, findings, and results.

The Future of Data Science and Machine Learning (1 hr)

A presentation discussing emerging trends, tools, and methodologies.

- Introduction to other libraries and frameworks beyond pandas and scikit-learn.
- Evolution of neural networks: Deep learning, Transformers, and GPT-like architectures.
- Real-world applications that are pushing the boundaries: Health, finance, environment, etc.

Next Steps and Resources (0,5-1 hrs)

Guiding students on where to go from here:

- Advanced courses and certifications.
- Online platforms for continuous learning: Coursera, edX, Udemy, and others.
- Recommending books, research papers, and prominent ML researchers to follow.
- Importance of community engagement: Attending conferences, webinars, and meetups.

12.12: (option based): Exploring the Orange Library for Data Mining

Introduction and Basics (1 hr)

Orange

Introduction to the Orange Library in Python
Overview of Orange and its place in data analysis.
The GUI vs. the library: Different approaches, same power.
Setting Up the Environment (20 minutes)
Installing Orange.
Common issues and their resolutions.

Basic Data Handling with Orange (40 minutes)

Loading datasets.
Basic data exploration and statistics.

Activity

Participants will set up their environment, load a sample dataset, and explore its basic statistics.

Classification Techniques (1.5 hrs)

Building Classification Models (45 minutes)

Introduction to the Orange classification module.
Building and evaluating a classification tree.

Model Inspection and Cross-Validation (45 minutes)

Extracting insights from models.
Performing cross-validation and understanding results.

Activity

Participants will use a provided dataset, create a classification model, inspect it, and evaluate its performance using cross-validation.

Clustering and Dimensionality Reduction (1 hr)

Data Clustering with k-Means (30 minutes)

Introduction to clustering in Orange.
Evaluating clusters.

Dimensionality Reduction with PCA (30 minutes)

Basics of PCA.
Applying and interpreting PCA in Orange.

Activity

Participants will apply k-means clustering to a dataset, followed by dimensionality reduction using PCA.

Regression and Model Evaluation (1.5 hrs)

Building Regression Models (45 minutes)

Overview of regression techniques in Orange.
Building and evaluating a linear regression model.

Understanding Overfitting and Regularization (45 minutes)

Introduction to the concept of overfitting.
Applying regularization in Orange.

Activity

Participants will work on a regression task, evaluate the model, and implement regularization techniques.

14.12 (option based): Orange Visual

Orange is a very visual tool and designed for those who prefer a GUI-based approach to data analysis.

Getting Started with Orange (1 hr)

Exploring the Orange Canvas (20 minutes)

Navigating the interface.
Introduction to widgets and connections.

Basic Data Exploration and Workflows (30 minutes)

How to create a basic workflow.
Visualization tools and data statistics.

Saving Your Work (10 minutes)**

Saving and loading workflows for future use.

Activity

Participants create their first workflow, explore a provided dataset, and save their workflow.

Diving into Classification (1.5 hrs)

Loading another provided Dataset (15 minutes)

Formats supported and importing datasets.

Building Classification Trees (30 minutes)

Basics of classification trees.
Creating a classification tree in Orange.

Model Inspection and Accuracy (30 minutes)

Understanding classification results.
Cross-validation in Orange.
Introduction to other classifiers.

Decision Thresholds and Model Scoring (15 minutes)

How to adjust decision thresholds.
Evaluating models using different metrics.

Activity

Participants load a given dataset, build a classification tree, evaluate its accuracy, and experiment with different classifiers.

Advanced Classification Techniques (1 hr)

Correct Test and Score Procedure (20 minutes)

Importance of separate test sets.
Ensuring unbiased evaluation.

More on Model Scoring (20 minutes)

Diving deeper into evaluation metrics.
Interpretation of results.

Choosing the Decision Threshold (20 minutes)

The balance between precision and recall.
Adjusting thresholds in Orange.

Activity

Using a different dataset, participants split data into training and test sets, evaluate models, and adjust decision thresholds.

Introduction to Regression in Orange (1.5 hrs)

Linear Regression (30 minutes)

Another Recap: Basics of linear regression.
Building and evaluating regression models in Orange.

Regularization (30 minutes)

Lasso and Ridge regression.
The importance of regularization and its effect.

Accuracy on the Test Set (30 minutes)

Evaluating regression models.
Comparing results with and without regularization.

Activity

Participants work on a regression problem, apply regularization techniques, and evaluate model performance on a test set.