ML & AI +Masters: Individual Assignment 1

Introduction

In this assignment, you will work to identify a real-world machine learning problem that interests you. This problem will serve as the foundation for your project throughout the course. You are encouraged to think critically and creatively about the problems around you that can be addressed using ML techniques.

The aim is to provide a practical context to your learning and to see how all techniques that you will learn throughout this course can be applied to solve meaningful problems.

Examples

To help you get started, here are a few examples of real-world problems that could be addressed with machine learning.

- One example is developing a model to predict house prices based on various features such as location, size, and number of rooms, which can help real estate professionals make informed decisions.
- Another example is creating a sentiment analysis tool that classifies movie reviews as positive or negative, providing valuable insights for film producers and marketers.
- You could also work on a project to build a recommendation system for an online store, suggesting products to users based on their browsing and purchasing history.

These examples are practical, relevant, and provide a great opportunity to apply machine learning techniques to solve real-world problems.

Dataset Selection

Once you have identified your problem, the next step is to find a dataset that aligns with your chosen problem and explore the feature variables.
Several online platforms provide high-quality datasets suitable for machine learning projects. Kaggle is a great starting point, offering a wide range of datasets across

various domains. You can also explore datasets on UCI Machine Learning Repository, which provides well-documented datasets for academic purposes.

Another excellent source is Data.gov, which offers a vast collection of government datasets. For more domain-specific datasets, you might consider websites like IMDb for movie data or OpenWeather for weather data.

Selecting the right dataset is crucial as it will directly impact the effectiveness and relevance of your machine learning model. Be sure to explore the feature variables within your dataset to understand what data you have and how it can be used to solve your problem.

Data Preprocessing

Now that you hopefully have the data you need, it's time to explore it further and apply cleansing and preprocessing techniques to make sure it's suitable to be used in the next stages.

Data cleansing entails dealing with null values, addressing class imbalances, and encoding categorical variables. For instance, you might handle missing values by imputing them with the mean or median of the column, or by removing rows with missing data if they are few. Class imbalance, where one class significantly outnumbers another, can be addressed using techniques such as oversampling the minority class or undersampling the majority class. We will cover these techniques and more throughout our course.

Data preprocessing is a critical step in preparing your data for further analysis and making sure the algorithms you choose will perform well. No model performs well with poor quality data ("garbage in, garbage out").

Exploratory Data Analysis

The final task for this week is exploratory data analysis. While not directly ML-related, EDA helps you understand the underlying patterns and relationships within your dataset, guiding you on how to approach your problem.
Through visualisations like histograms, scatter plots, and box plots, you can identify trends, outliers, and correlations between variables. This step is crucial because it informs the feature selection process, helping you choose the most relevant

variables for your model. EDA also provides insights that can influence the choice of algorithms and parameter settings, ultimately improving the performance of your machine learning model.

By the end of this, you should have a clean, well-understood dataset ready for the next stages of your project.

What to Submit:

- Report including the following:
    - Problem Identification
    - Dataset Selection and Exploration (include a link if applicable)
    - Data Preprocessing
    - Exploratory Data Analysis
- Code Submission (Python Notebook or Script)