**MODUL 1: Introduction to Artificial Intelligence and Machine Learning**

Artificial Intelligence is a branch of computer science dealing with the simulation of intelligent behavior in computers. Machines mimic cognitive functions such as learning and problem solving.

Machine learning is the study of programs that are not explicitly programmed, but instead these algorithms learn patterns from data.

Deep learning is a subset of machine learning in which multilayered neural networks learn from vast amounts of data.

**History of AI**

AI has experienced cycles of AI winters and AI booms.

AI solutions include speech recognition, computer vision, assisted medical diagnosis, robotics, and others.

**Modern AI**

Factors that have contributed to the current state of Machine Learning are: bigger data sets, faster computers, open source packages, and a wide range of neural network architectures.

**Machine Learning Workflow**

The machine learning workflow consists of:

* Problem statement
* Data collection
* Data exploration and preprocessing
* Modeling
* Validation
* Decision Making and Deployment

This is a summary of the common taxonomy for data in open source packages for Machine Learning:

* target: category or value you are trying to predict
* features: explanatory variables used for prediction
* example: an observation or single data point within the data
* label: the value of the target for a single data point

**MODUL 2: Retrieving Data**

You can retrieve data from multiple sources:

* SQL databases
* NoSQL databases
* APIs
* Cloud data sources

The two most common formats for delimited data flat files are comma separated (csv) and tab separated (tsv). It is also possible to use special characters as separators.

SQL represents a set of relational databases with fixed schemas.

**Reading in Database Files**

The steps to read in a database file using the sqlite library are:

* create a path variable that references the path to your database
* create a connection variable that references the connection to your database
* create a query variable that contains the SQL query that reads in the data table from your database
* create an observations variable to assign the read\_sql functions from pandas package
* create a tables variable to read in the data from the table sqlite\_master

JSON files are a standard way to store data across platforms. Their structure is similar to Python dictionaries.

NoSQL databases are not relational and vary more in structure. Most NoSQL databases store data in JSON format.

**Data Cleaning**

Data Cleaning is important because messy data will lead to unreliable outcomes.Some common issues that make data messy are: duplicate or unnecessary data, inconsistent data and typos, missing data, outliers, and data source issues.

You can identify duplicate or unnecessary data Common policies to deal with missing data are:remove a row with missing columns, impute the missing data, and mask the data by creating a category for missing values.

Common methods to find outliers are: through plots, statistics, or residuals.

Common policies to deal with outliers are: remove outliers, impute them, use a variable transformation, or use a model that is resistant to outliers.

**MODUL 3: Exploratory Data Analysis**

EDA is an approach to analyzing data sets that summarizes their main characteristics, often using visual methods. It helps you determine if the data is usable as-is, or if it needs further data cleaning.

EDA is also important in the process of identifying patterns, observing trends, and formulating hypothesis.

Common summary statistics for EDA include finding summary statistics and producing visualizations.

Feature Engineering and Variable Transformation

Transforming variables helps to meet the assumptions of statistical models. A concrete example is a linear regression, in which you may transform a predictor variable such that it has a linear relation with a target variable.

Common variable transformations are: calculating log transformations and polynomial features, encoding a categorical variable, and scaling a variable.

**MODUL 4: Estimation and Inference**

Inferential Statistics consist in learning characteristics of the population from a sample. The population characteristics are parameters, while the sample characteristics are statistics. A parametric model, uses a certain number of parameters like mean and standard deviation.

The most common way of estimating parameters in a parametric model is through maximum likelihood estimation.

Through a hypothesis test, you test for a specific value of the parameter.

Estimation represents a process of determining a population parameter based on a model fitted to the data.

The most common distribution functions are: uniform, normal, log normal, exponential, and poisson.

A frequentist approach focuses in observing man repeats of an experiment. A bayesian approach describes parameters through probability distributions.

Hypothesis Testing

A hypothesis is a statement about a population parameter. You commonly have two hypothesis: the null hypothesis and the alternative hypothesis.

A hypothesis test gives you a rule to decide for which values of the test statistic you accept the null hypothesis and for which values you reject the null hypothesis and accept he alternative hypothesis.

A type 1 error occurs when an effect is due to chance, but we find it to be significant in the model.

A type 2 error occurs when we ascribe the effect to chance, but the effect is non-coincidental.

Significance level and p-values

A significance level is a probability threshold below which the null hypothesis can be rejected. You must choose the significance level before computing the test statistic. It is usually .01 or .05.

A p-value is the smallest significance level at which the null hypothesis would be rejected. The confidence interval contains the values of the statistic for which we accept the null hypothesis.

Correlations are useful as effects can help predict an outcome, but correlation does not imply causation.

When making recommendations, one should take into consideration confounding variables and the fact that correlation across two variables do not imply that an increase or decrease in one of them will drive an increase or decrease of the other.

Spurious correlations happen in data. They are just coincidences given a particular data sample.