Week 1: Introduction to Machine Learning

KEY CONCEPTS

- To give examples of Machine Learning
- To demonstrate the Python libraries for Machine Learning
- To classify Supervised vs. Unsupervised algorithms

What is Machine Learning?

Use ML for recommendation systems, customer segmentation, bank decisions and healthcare industry.

Skills:

- Regression
- Classification
- Clustering
- Scikit Learn
- Scipy

Projects:

- Cancer detection
- Predicting economic trends
- Predicting customer churn
- Recommendation engines

Introduction to Machine Learning

Machine learning helps with predictions.

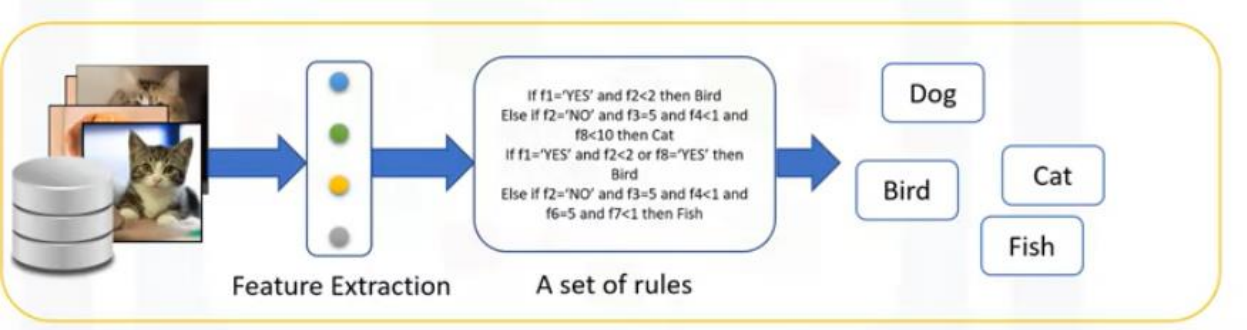
Cancer Detection Example:

- Clean your data
- Select a proper algorithm for building a prediction model
- Train the model to understand patterns of benign or malignant cells within the data.

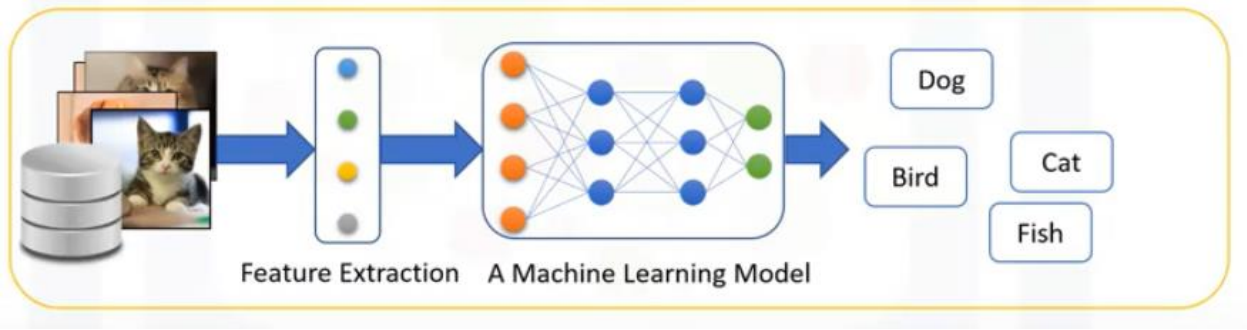
Once the model has been trained by going through data iteratively, it can be used to predict your new or unknown cell with a rather high accuracy. This is machine learning!

Definition: Machine learning is the subfield of computer science that gives "**computers the ability to learn without being explicitly programmed.**"

Traditionally, we write down some rules or methods in order to get computers to be intelligent and detect the animals. But it was a failure since it needed a lot of rules, highly dependent on the current dataset, and not generalized enough to detect out-of-sample cases.



Using machine learning allows us to build a model that looks at all the feature sets, and their corresponding types od animals, and it learns the pattern of each animal. It is a model built by machine learning algorithms.



Examples of Machine Learning

- Netflix and Amazon recommend videos, movies, and TV shows to its users using Machine Learning to produce suggestions that you might enjoy.
- Banks decide when approving a loan application using machine learning to predict the probability of default for each applicant, and then approve or refuse the loan application based on that probability.
- Telecommunication companies use their customers' demographic data to segment them or predict if they will unsubscribe from their company the next month.
- There are many other applications of machine learning that we see every day in our daily life, such as chatbots, logging into our phones or even computer games using face recognition. Each of these use different machine learning techniques and algorithms.

Major Machine Learning Techniques

- **Regression/Estimation technique** is used for **predicting a continuous value**. For example, predicting things like the price of a house based on its characteristics, or to estimate the Co2 emission from a car's engine.
- **Classification technique** is used for **Predicting the class or category of a case**, for example, if a cell is benign or malignant, or whether a customer will churn.
- **Clustering** is used for **Finding the structure of data; summarization**. It groups similar cases, for example, can find similar patients, or can be used for customer segmentation in the banking field.

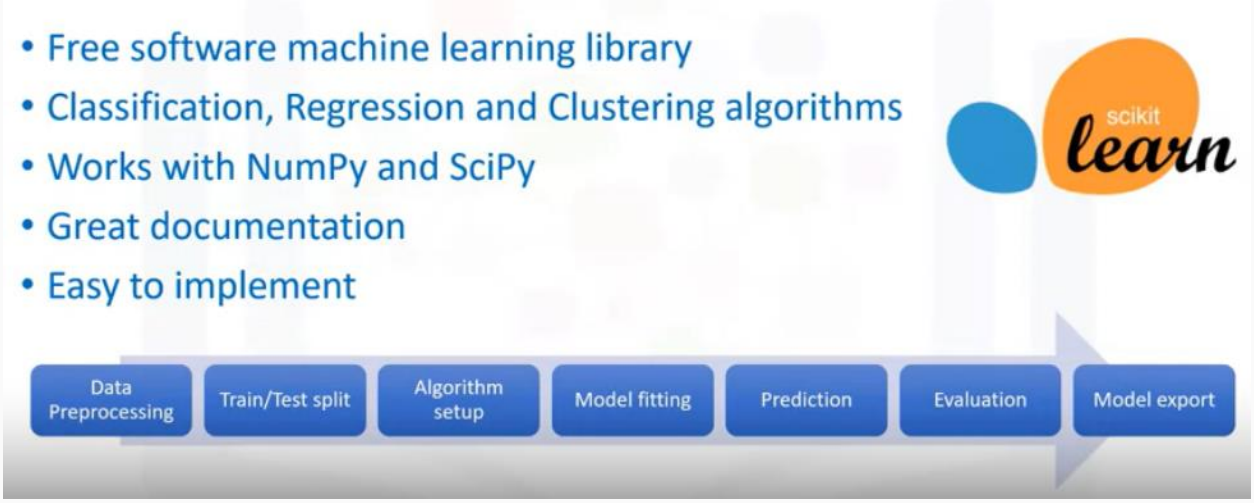
- **Association technique** is used for **finding items or events that often co-occur**, for example, grocery items that are usually bought together by a customer.
- **Anomaly detection** is used to **discover abnormal and unusual cases**, for example, it is used for credit card fraud detection.
- **Sequence mining** is used for **predicting the next event**, for instance**, the click-stream in websites (Markov Model, HMM)**.
- **Dimension reduction** is used to **reduce the size of data (PCA).**
- **Recommendation systems**, this associates people's preferences with others who have similar tastes, and recommends new items to them, such as books or movies.

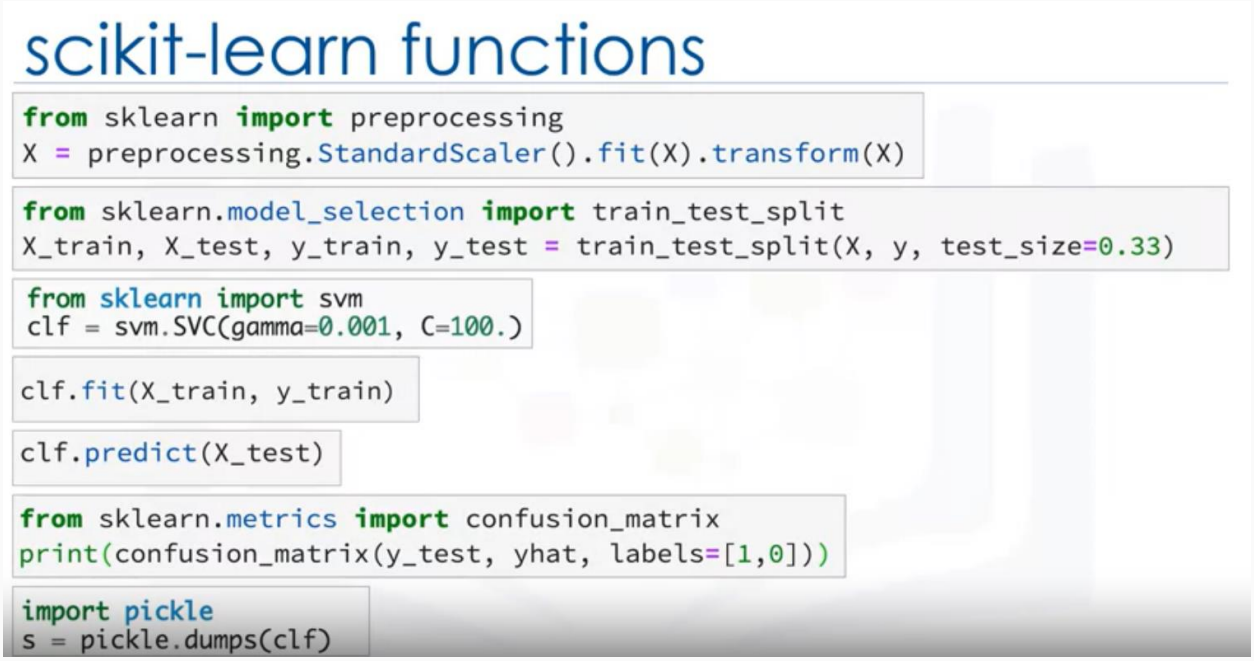
Difference between Artificial Intelligence, Machine Learning, and Deep Learning

- **AI** tries to make computers intelligent in order to mimic the cognitive functions of humans. So, Artificial Intelligence is a general field with a broad scope including: **Computer Vision, Language Processing, Creativity, and Summarization.**
- **Machine Learning** is the branch of AI that covers the statistical part of artificial intelligence. It teaches the computer to solve problems by looking at hundreds or thousands of examples, learning from them, and then using that experience to solve the same problem in new situations. It includes **Classification, Clustering, and Neural Network**.
- **Deep Learning** is a very special field of Machine Learning where computers can learn and make intelligent decisions on their own. Deep learning involves a deeper level of automation in comparison with most machine learning algorithms.

Python for Machine Learning

- **NumPy** is a math library to work with N-dimensional arrays in Python. It enables you to do computation efficiently and effectively. It is better than regular Python because of its amazing capabilities. For example, for **working with arrays, dictionaries, functions, datatypes and working with images** you need to know NumPy.
- **SciPy** is a collection of numerical algorithms and domain specific toolboxes, including **signal processing, optimization, statistics** and much more. SciPy is a good library for scientific and high-performance computation.
- **Matplotlib** is a very popular plotting package that provides 2D plotting, as well as 3D plotting.
- **Pandas library** is a very high-level Python library that provides high performance easy to use data structures. It has many functions for **data importing, manipulation and analysis**. It offers data structures and operations for manipulating numerical tables and timeseries.
- **SciKit Learn** is a collection of algorithms and tools for machine learning. Most of the tasks that need to be done in a machine learning pipeline are implemented already in Scikit Learn including **pre-processing of data, feature selection, feature extraction, train test splitting, defining the algorithms, fitting models, tuning parameters, prediction, evaluation, and exporting the model**.





Basically, machine-learning algorithms benefit from standardization of the dataset. If there are some outliers or different scales fields in your dataset, you must fix them. The pre-processing package of SciKit Learn provides several common utility functions and transformer classes to change raw feature vectors into a suitable form of vector for modeling. You must split your dataset into train and test sets to train your model and then test the model's accuracy separately. SciKit Learn can split arrays or matrices into random train and test subsets for you in one line of code. Then you can set up your algorithm. For example, you can build a classifier using **a support vector classification algorithm**. We call our estimator instance CLF and initialize its parameters. Now you can train your model with the train set by passing our training set to the fit method, the CLF model learns to classify unknown cases. Then we can use our test set to run predictions, and the result tells us what the class of each unknown value is. Also, you can use the different metrics to evaluate your model accuracy. For example, using a confusion matrix to show the results. And finally, you save your model.

Supervised vs Unsupervised

Supervised Learning

- We “teach the model” then with that knowledge, it can predict unknown or future instances.
- Teaching the model with **labeled** data
- **Classification** (the process of predicting discrete class labels or categories) and **Regression** (the process of predicting continuous values, ex. trend)
- It has more evaluation methods than unsupervised learning.
- Controlled environment

Unsupervised Learning

- The model works on its own to discover information.
- All the data is **unlabeled**.
- Unsupervised learning has more difficult algorithms than supervised learning since we know little to no information about the data, or the outcomes that are to be expected.
- It has **fewer models and evaluation methods** than supervised learning
- Dimension reduction, Density estimation, Market basket analysis and **Clustering** (find patterns and grouping of data points or objects that are somehow similar by discovering structure, summarization, anomaly detection)
- **Less controlled environment**