Trainer: Faris Hassan
Email: farishassan2012@gmail.com
Github: github.com/Faris137

DevCon:{ACADEMIA}

Machine Learning Notes

Pre-Intro:

What is Machine Learning ?

Machine learning in the simple words of Samuel, Arthur (1959).is a subfield of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.

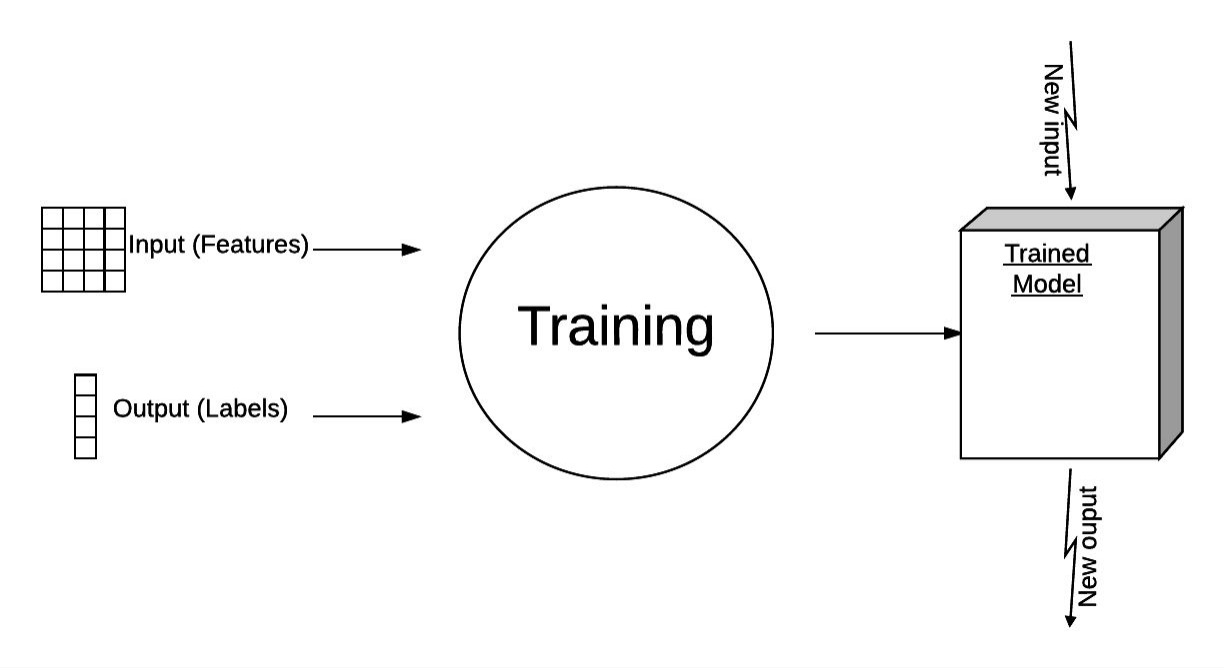
Tom Mitchell (1997) is a machine learning pioneer came up with a formal definition "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E."

Artificial Intelligence moved from the rules based programming and if-else methodology into the concept of learning. Machine learning is a shift in the paradigm moving away from the typical thinking:

The shift happens when we realize that machines learning is actually produces a model that is trained to perform certain tasks. The primitive intuition of intelligence started long time ago as a philosophical conundrum and Alan Turing proposed the famous question:

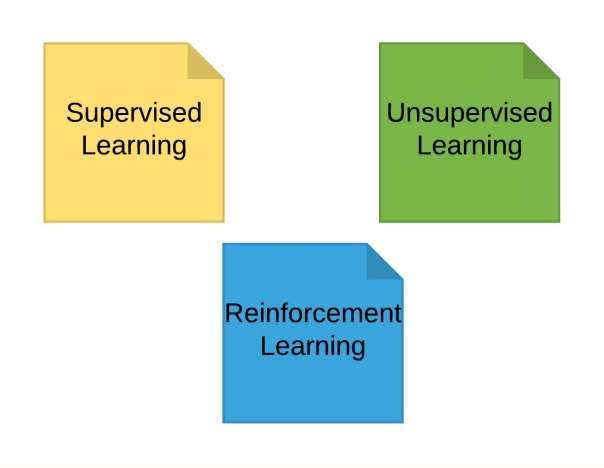
" Can Machines think ? "

The diagram below describe the machine learning paradigm:



What drove the development in computer science and artificial intelligence is the evolution of computing powers and storage. This have had led us to have abundance of structured and unstructured data in abundance. Instead of requiring humans to manually derive rules and build models from analyzing large amounts of data, machine learning offers a more efficient alternative for capturing the knowledge in data to gradually improve the performance of predictive models, and make data-driven decisions.

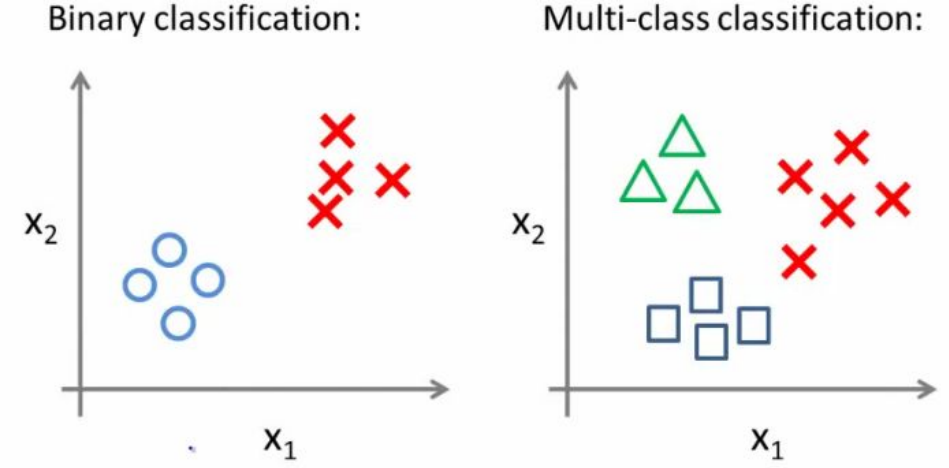
Types of Machine Learning problems



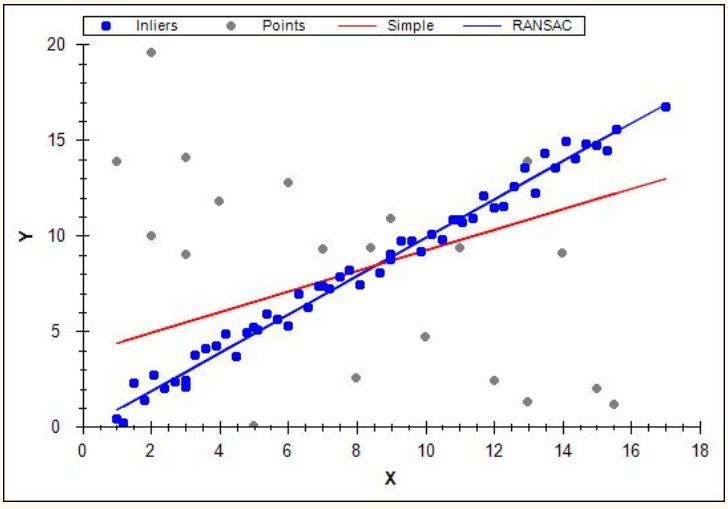
Supervised Learning

The purpose of this method is to perform learning for a model based on labeled data that will facilitate making predictions for unseen data. The notion of supervision refers to a data that is already labeled and known. Supervised learning algorithms focus on establishing relationship between inputs (features) and outputs (labels). There are two types of learning falling under this category:

1. Classification: is a subcategory of supervised learning where the goal is to predict the categorical class labels of new instances based on past observations. Those class labels are discrete, unordered values that can be understood as the group memberships of the instances. For instance in an e-mail spam detection we will have the machine learning algorithm distinguish between the two class: spam and non-spam email.

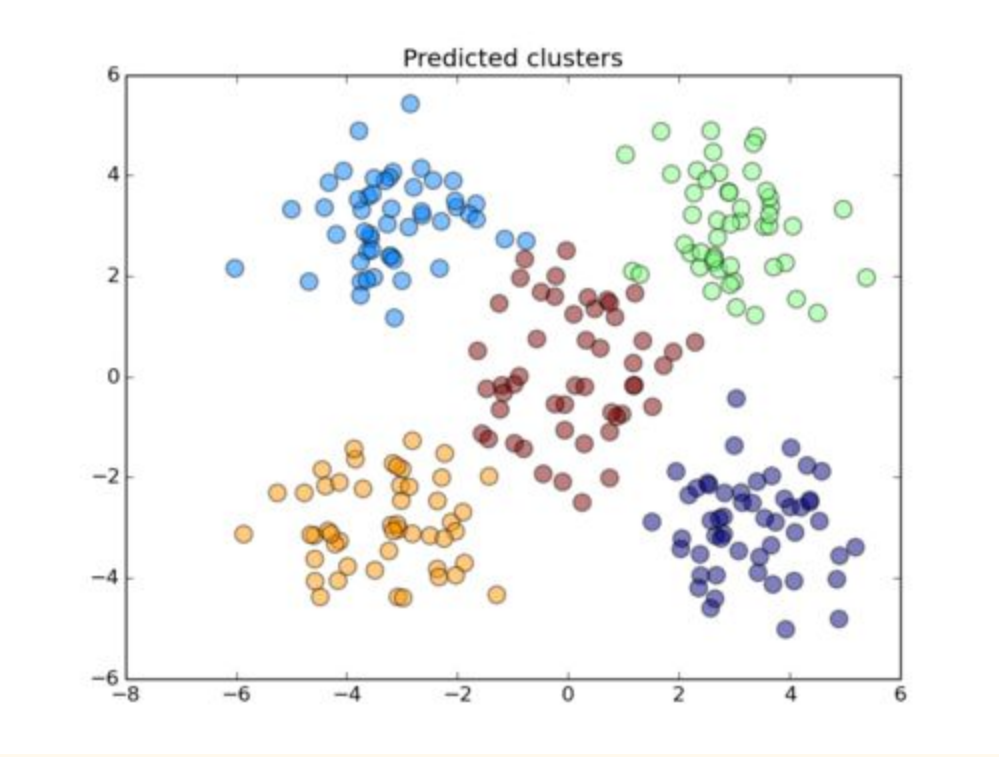


2. Regression: which is a prediction of continuous values. For example prediction of sales, prices and amounts. The outcome that is expected from the model is always something of a continuous value.

3. 

Unsupervised Learning

In supervised learning we know the right answer and we include that knowledge when we learn the model. The unsupervised learning is dealing with data of unknown structure and we rely on the intelligence of the machine to explore the structure of the data to derive meaningful information independent from human guidance or reward function. One of the famous unsupervised methods is clustering. Clustering is an exploratory data analysis technique that allows us to organize a pile of information into meaningful subgroups (clusters) without having any prior knowledge of their group memberships. Each cluster that may arise during the analysis defines a group of objects that share a certain degree of similarity but are more dissimilar to objects in other clusters, which is why clustering is also sometimes called "unsupervised classification." Clustering is a great technique for structuring information and deriving meaningful relationships among data, For example, it allows marketers to discover customer groups based on their interests in order to develop distinct marketing programs.

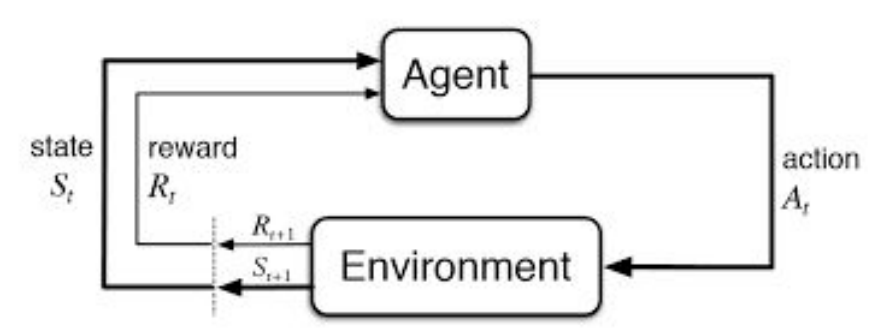


Reinforcement Learning

Another type of machine learning is reinforcement learning. In reinforcement learning, the goal is to develop a system (agent) that improves its performance based on interactions with the environment. Since the information about the current state of the environment typically also includes a so-called reward signal, we can think of reinforcement learning as a field related to supervised learning. However, in reinforcement learning this feedback is not the correct

ground truth label or value, but a measure of how well the action was measured by a reward function. Through the interaction with the environment, an agent can then use reinforcement learning to learn a series of actions that maximizes this reward via an exploratory trial-and-error approach or deliberative planning.

A popular example of reinforcement learning is a chess engine. Here, the agent decides upon a series of moves depending on the state of the board (the environment), and the reward can be defined as win or lose at the end of the game:



Types of machine learning algorithms

Linear Classifiers: In machine learning, the goal of classification is to group items that have similar feature values, into groups. Timothy et al (Timothy Jason Shepard, 1998) stated that a linear classifier achieves this by making a classification decision based on the value of the linear combination of the features. If the input feature vector to the classifier is a real vector , then the output score is:

$$y = f(w.x) = f(\sum_j wj.xj)$$

where *w* is a real vector of weights and f is a function that converts the dot product of the two vectors into the desired output. The weight vector *w* is learned from a set of labelled training samples. Often f is a simple function that maps all values above a certain threshold to the first class and all other values to the second class.A linear classifier is often used in situations where the speed of classification is an issue, since it is often the fastest classifier, especially when is sparse. However, decision trees can be faster.

K-means Clustering: The basic step of k-means clustering is uncomplicated. In the beginning we determine number of cluster K and we assume the centre of these clusters. We can take any random objects as the initial centre or the first K objects in sequence can also serve as the initial centre. Then the K means algorithm will do the three steps below until convergence. Iterate until stable (= no object move group):

1. Determine the centre coordinate
2. Determine the distance of each object to the centre
3. Group the object based on minimum distance

Decision trees: Decision trees are known to be one of the most powerful and widely used modeling techniques in the field of Machine learning. Decision trees naturally induce rules that can be used in data classification and prediction. Following is an example of a rule definition derived from building a Decision tree: If (laptop model is x) and (manufactured by y) and (is z years old) and (with some owners being k) then (the battery life is n hours). When closely observed, these rules are expressed in simple, human readable, and comprehensible formats. Additionally, these rules can be stored for later reference in a data store.

- Refer to Types of machine learning algorithm for more details.

Machine Learning Life cycle: