Machine Learning – An introduction

Overview

What is Machine Learning?
Tom M. Mitchell provided a widely quoted, more formal definition of the algorithms studied in the machine learning field:

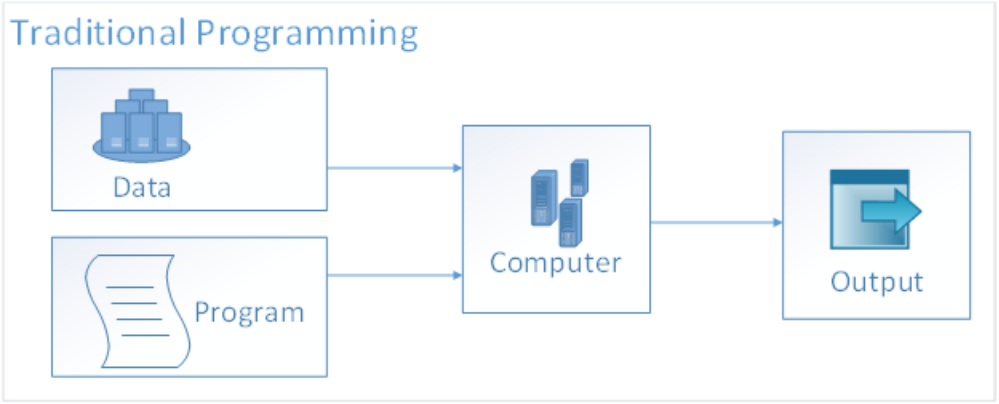
"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E."

In simple terms

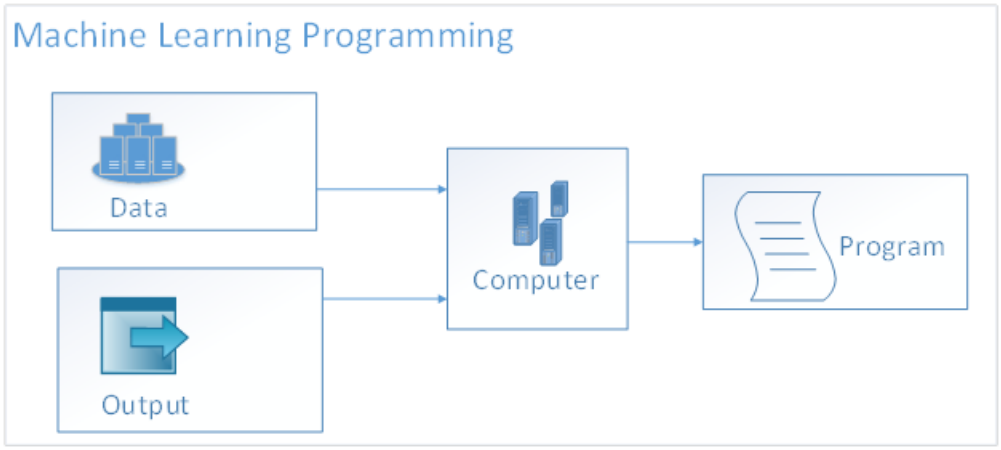
Machine learning is a type of artificial intelligence (AI) that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build algorithms that can receive input data and use statistical analysis to predict an output value within an acceptable range.

Here we can see the difference between traditional programming and machine learning programming model.

Traditional Programming Model
Under traditional programming models, programs and data are processed by the computer to produce a desired output, such as using programs to process data and produce the result.



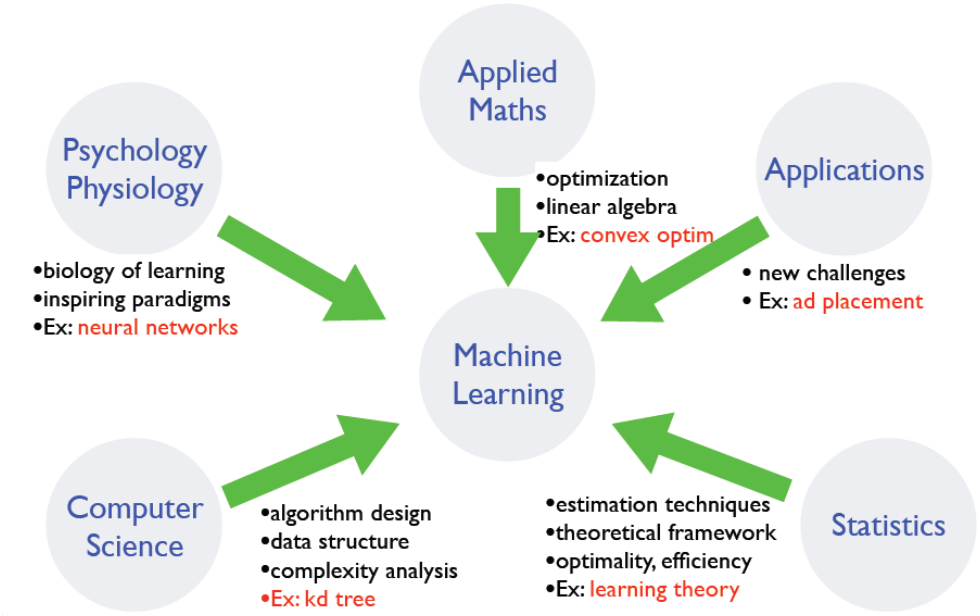
Machine Learning Programming Model
The data and the desired output are reverse-engineered by the computer to produce a new program.



The power of this new program is that it can effectively **predict** the output, based on the supplied input data. The primary benefit of this approach is that the resulting **program** that is developed has been trained (via massive quantities of learning data) and finely tuned (via feedback data about the desired output) and is now capable of predicting the likelihood of a desired output based on the provided data.

This pattern of using conjoined data to infer additional data attributes is where the science of data really takes off, and it has serious financial benefits to organizations that know how to leverage this technology effectively. This is where data scientists can add the most value, by aiding the machine learning process with valuable data insights and inferences that are (still) more easily understood by humans than computers.
This is also where it becomes most critical to have the ability to rapidly test a hunch or theory to either **fail-fast** or confirm the logic of your prediction algorithms, and really fine-tune a prediction model.

Where ML fit in?



Slide credit: Dhruv Batra, Fei Sha

Machine Learning Algorithms
Machine learning algorithms are classified as

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Supervised Learning
Supervised learning is a type of machine learning algorithm that uses known datasets to create a model that can then make predictions. The known data sets are called and include input data elements along with known response values. From these training datasets, supervised learning algorithms attempt to build a new model that can make predictions based on new input values along with known outcomes.

– **Given: training data, desired outputs (labels)**

Fit a model that relates response to the feature tuples, with the aim of accurately predicting the response for future observation or better understanding the relationship between response and features.

Regression
Regression problems are supervised learning problems where target / response is a continuous variable (any real number).
These algorithms can predict one or more continuous variables, such as profit or loss, based on other columns in the data set.

Types of Regression Techniques

- Linear regression
- Polynomial regression
- Stepwise regression
- Ridge regression
- Lasso regression
- ElasticNet regression

Regression Example (Predicting House Price)
Housing Price Prediction is a typical regression example. The output or the continues variable (**Y)** would be predicted housing price. Input or dependent variables **(X)** are House Type, SQ Ft, Location, City, and Facilitates.

X						Y
House Type	SQ Feet	Location	City	Type	Facilities	House Price
Apartment	1350	Koramangala	Bangalore	Bangalore Urban	Swimming Pool, Club House	1,23,00,000
Villa	1400	Koramangala	Bangalore	Bangalore Urban	Swimming Pool, Club House	31,23,00,000
Apartment	1300	Electronic City	Bangalore	Bangalore Urban	Swimming Pool Tennis Court,	72,00,000
Apartment	1200	Chandapura	Bangalore	Bangalore Rural	Swimming Pool Tennis Court,	45,00,000

Few examples of regression use cases

- Advertising Popularity Prediction
- Weather Forecasting
- Market Forecasting
- Estimation life expectancy
- Popular Growth Predictions

Classification
These algorithms are used for predicting responses that can have just a few known values such as married, single, or divorced—based on the other columns in the dataset. Classification problems are supervised Learning problems where target/response variables take only discrete (finite/countable) values.

Types of Classification Techniques

- Logistic Regression
- Naïve Bayes
- Stochastic Gradient Descent
- K-Nearest Neighbours
- Decision Tree
- Random Forest
- Support Vector Machine

Classification Example (Predicting Housing Loan Approval)
Housing loan approval Prediction is a typical classification example. The output or discrete variable (**Y)** would be predicted housing loan approval **(Yes or No)**. Input or dependent variables (X) are Monthly income, Income Type, Dependent, Credit Score, Savings, Loan amount.

X						Y
Monthly Income	Income Type	Dependent	Credit Score	Savings	Loan Amount	Housing Loan Approval
1,00,000	salaried	3	700	5,00,000	35,00,000	Yes
2,00,000	Business	6	600	3,00,000	65,00,000	No
50,000	salaried	1	740	1,00,000	10,00,000	Yes
25,00	Business	2	600	45,000	2,00,000	Yes

Few examples of classification use cases

- Identity Fraud Deduction
- Image Classification
- Customer Retention
- Diagnostics

Difference between Regression and Classification

Regression	Classification
Supervised Learning	Supervised Learning
There are predefined labels assigned to each input instances according to their properties used to find the dependent values	There are predefined labels assigned to each input instances according to their properties used to find the dependent values
In Regression, the output variable must be of continuous nature or real value.	In Classification, the output variable must be a discrete (categorical) value.
Training and Testing datasets are mandatary	Training and Testing datasets are mandatary
Regression can be evaluated using root mean square error.	Classification is evaluated by measuring confusion matrix.
In Regression, we try to find the best fit line, which can predict the output more accurately.	In Classification, we try to find the decision boundary, which can divide the dataset into different classes.
Regression algorithms can be used to solve the regression problems such as Weather Prediction, House price prediction, etc.	Classification Algorithms can be used to solve classification problems such as Identification of spam emails, disease prediction