Machine Learning

Machine Learning

- Information from observable environment can be systematically recorded and processed in a way different from a traditional human processing.
- Electronic spreadsheets contributed to an increase in richness of recorded data.
- A study of developing computer algorithms for transforming data into intelligent action is known as '*Machine learning'.*
- Originates at an intersection of statistics,

Why Machine Learning & Data Analytics?

The Context of Data:

- Multi-Model and heterogeneous
- Noisy and incomplete
- Time and location dependent
- Dynamic and varies in quality
- Data can be biased

Solution:

Powerful analytics will play a vital role in

Data Vs Machine Learning

- Human life is always inundated with data.

Ex; Body parts like ear, nose, eyes etc all are continually assailed with raw data that the brain translates into sounds, smells, sights etc.

- Astronomers recorded patterns of planets and stars
- Biologists noted results from experiments on plants and animals
- The deluge of data has led to an era of Big Data where larger data sets can be accessed

Uses of Machine Learning

- Machine learning is used to make sense of complex data. It is widely used to:

1. Predict the outcomes of elections.
2. Identify and filter spam messages from e-mail.
3. Foresee criminal activity.
4. Automate traffics signals according to road conditions.
5. Examine customer churn.
6. Create auto-piloting planes and auto-driving cars.
7. Target advertising to specific types of consumers.

How do Machines Learn?

- A machine learns when it is able to take experience and utilize it such that its performance improves on similar experiences.
- Basic learning process includes:
  - ➢ Data Input
  - ➢ Abstraction
  - ➢ Generalization

Steps to Apply Machine Learning

Types of Machine Learning

- Machine learning algorithms can be divided into two main groups:
  - ✓ Supervised
  - ✓ Unsupervised
- Supervised learners are used to construct predictive models.
- Unsupervised learners are used to build descriptive models

How to Choose a Machine Learning Algorithm?

Match the characteristics of the data to be learned to the biases of the available approaches.

Supervised Learning Algorithms

Algorithm	Task
Classification Learners	Classification
Linear Regression	Numeric Prediction
Nearest Neighbor	Classification
Decision Trees	Classification
Regression Trees	Numeric Prediction
Model Trees	Numeric Prediction
Neural Networks	Dual use

Unsupervised Learning Algorithms

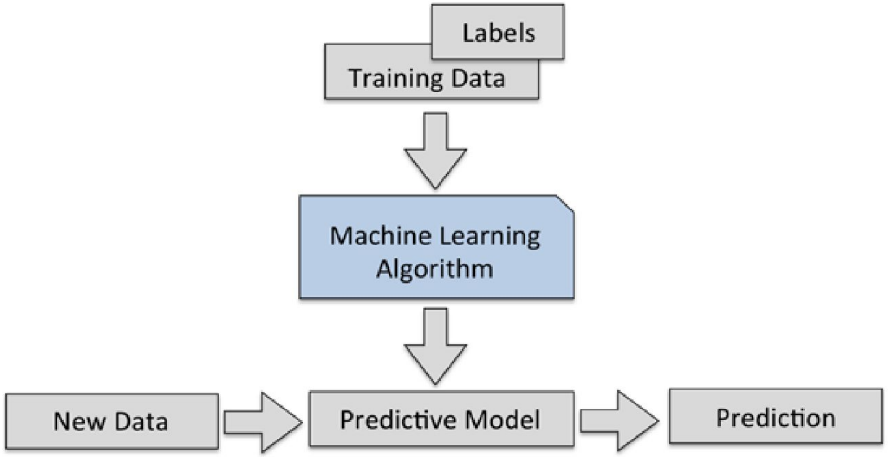
Algorithm	Task
Association Rules	Pattern detection
K-means Clustering	Clustering

One can actually use a classification, numeric prediction, pattern detection, or clustering depending on type of problem

Assessing the success of learning

- Most times learning can be biased and may have its weaknesses.
- After a model is trained on an initial dataset, it is tested on a new dataset to see how well the characterization of the training data generalizes to the new data.
- Usually models fail to generalize due to unexplained variations in data called ‘Noise’.

- Machine learning is where these computational and algorithmic skills of datascience meet the statistical thinking of data science, and the result is a collection of approaches to inference and data exploration that are not about effective theory so much as effective computation.
- It is the application and science of algorithms that makes sense of data.Fundamentally, machine learning involves building mathematical models to help understand data.
- Machine learning can be categorized into two main types: supervised learning and unsupervised learning.
- Three types of machine learning are available: *supervised learning, unsupervised learning, and reinforcement learning.*
- The main go l from labelled training data that all or future data.



Supervised Learning

Supervised learning involves somehow modeling the relationship between measured features

of data and some label associated with the data; once this model is determined, it can be used

to apply labels to new, unknown data.

- It is subdivided into classification tasks and regression tasks: in classification, the labels are discrete categories, while in regression, the labels are continuous quantities.
- Unsupervised learning involves modeling the features of a dataset without reference to any label,

  and is often described as "letting the dataset speak for itself. These models include tasks such as clustering and dimensionality reduction. Clustering algorithms identify distinct groups of data,

  while dimensionality reduction algorithms search for more

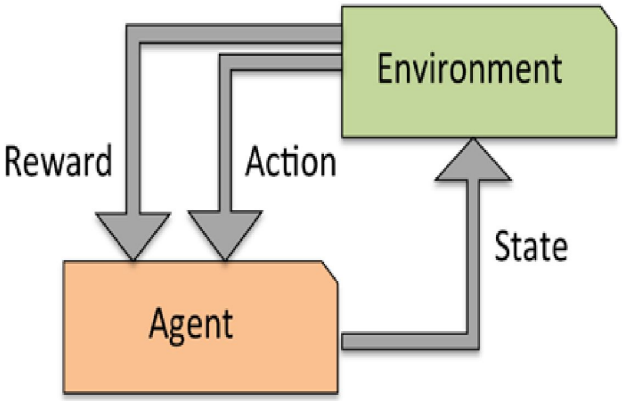
- In unsupervised learning, however, we are dealing with unlabeled data or data of *unknown structure*.

Ex:

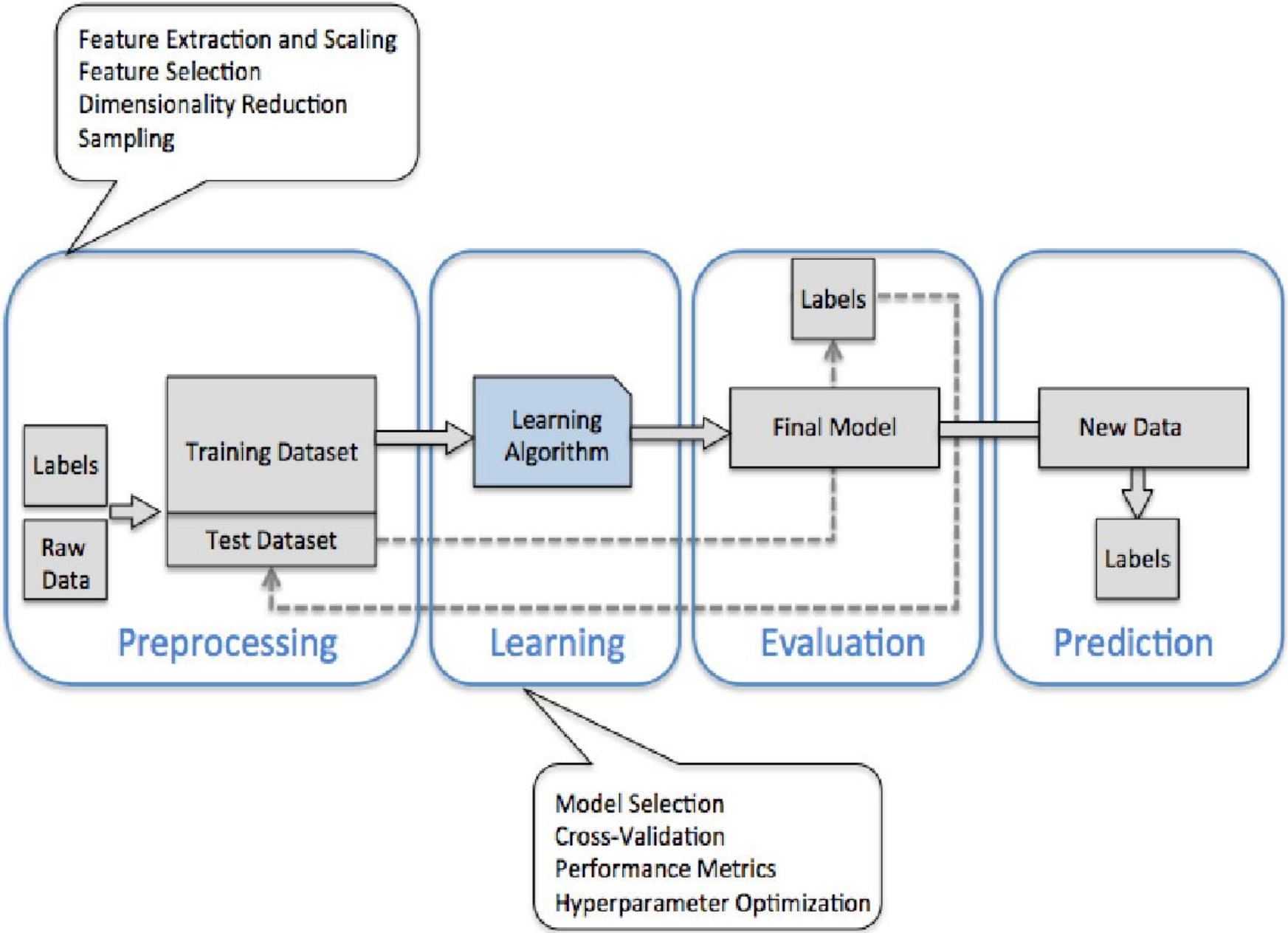
- *Clustering* is an exploratory data analysis technique that allows us to organize a pile of information into meaningful subgroups (*clusters*) without having any prior knowledge of their group memberships.
- Each cluster that may arise during the analysis defines a group of objects that share a certain degree of similarity but are more dissimilar to objects in other clusters, which is why clustering is also sometimes called "unsupervised classification."

- Another subfield of unsupervised learning is *dimensionality reduction*.
- Often we may have data of high dimensionality—each observation comes with a high number of measurements—that can present a challenge for limited storage space and the computational performance of machine learning algorithms.
- Unsupervised dimensionality reduction is a commonly used approach in feature preprocessing to remove noise from data, which can also degrade the predictive performance of certain algorithms, and compress the data onto a smaller dimensional subspace while retaining most of the relevant information.

- 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.
- 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.
- An agent can to learn a series of actions that maximizes th :ory trial-and-error approach or deliberative pl



The Roadmap for building machine



Preprocessing

- Raw data rarely comes in the form and shape that is necessary for the optimal performance of a learning algorithm.
- *Preprocessing* of the data is one of the most crucial steps in any machine learning application.
- Machine learning algorithms also require that the selected features are on the same scale for optimal performance, which is often achieved by transforming the features in the range [0, 1] or a standard normal distribution with zero mean and unit variance .
- To determine whether our machine learning algorithm not only performs well on the training set but also generalizes well to new data, we also want to randomly divide the dataset into a separate training and test set.
- We use the training set to train and optimize our machine learning model, while we keep the test set until the very end to evaluate the final model.

Training and Selecting a Predictive Model

- Different machine learning algorithms have been developed to solve different problem tasks.

How do we know which model performs well on the final test dataset and real-world data if we don't use this test set for the model selection but keep it for the final model evaluation?

➢ In practice, it is therefore essential to compare at least a handful of different algorithms in order to train and select the best performing model.

➢ One commonly used metric is classification accuracy, which is defined as the proportion of correctly classified instances.

Evaluating Models and predicting unseen data instances

- After selecting a model that has been fitted on the training dataset, we can use the test dataset to estimate how well it performs on this unseen data to estimate the generalization error.
- If satisfied with performance on unseen data, the model can be used to predict new, future data.

Supervised Learning: Linear Regression

- Linear regression is a prediction method
- Linear regression assumes a linear or straight line relationship between the input variables (X) and the single output variable (y).
- More specifically, that output (y) can be calculated from a linear combination of the input variables (X). When there is a single input variable, the method is referred to as a simple linear regression.
- In simple linear regression we can use statistics on the training data to estimate the coefficients required by the model to make predictions on new data.
- The line for a simple linear regression model can be written as:

$$y = b+ax$$

where b is known as *intercept* and a is known as *slope* are the coefficients we must estimate from the training data.

Forecasting- Working with Regression Methods

- Mathematical relationships provide information related to various aspects. Statistics describes techniques for estimating such numeric relationships among data elements through 'Regression Analysis'.
- Regression methods can be used for forecasting numeric data and quantifying the size and strength of relationship between an outcome and its predictor.
- Regression methods are used for hypothesis testing to indicate whether a presupposition is likely to be true or false.

What is Regression?

*Specifying a **relationship between single numeric dependent variable**(value to be predicted) and **one or more independent variables**(the predictors)*

Mathematically

$y = a+bx$ *;y---dependent variable ; x- independent variable*

;b--- slope(indicates how much the line rises for each increase in 'x'

Linear Regression

Linear regression is a statistical model that examines the linear relationship between two (Simple Linear Regression) or more (Multiple Linear Regression) variables — a dependent variable and independent variable(s).

Simple linear

Involves a single independent variable and a dependent variable.

Multiple regression

Involves several independent variables and a dependent variable.

Simple Linear Regression

- Defines relationship between a dependent variable and a single independent predictor variable.
- Denoted using the equation y = a + bx

;a--- intercept: *describes where line crosses y-axis*

*;*b--- slope: *describes changes in y given an increase of x*

;y--- *dependent variable*

;x--- *independent variable*

- Regression equation models data using a slope-intercept format.
- Machine learner's job is to identify values of 'a' and 'b' to enable the specified line to best relate the supplied

Case Study : Risk Analysis of Space Shuttle

Seven crew members of a US space shuttle challenger were killed due to O-rings failure where O-rings are responsible for sealing the joints of the rocket booster . This caused a catastrophic explosion. The Rogers Commission report on the space shuttle Challenger accident concluded that the accident was caused by a combustion gas leak through a joint in one of the booster rockets, which was sealed by a device called an O-ring. The commission further concluded that O-rings do not seal properly at low temperatures.

Critical Thinking:

How low temperature forecast might affect the safety of the launch.

Challenge:

Shuttle components have never been tested in cold weather making it unclear to withstand the strain from freezing temperatures.

Presupposition(Hypothesis):

Rocket engineers believed that cold temperatures could make the components more brittle and less able to seal properly resulting in higher chance of dangerous fuel leak.

The analysis of this article demonstrates that statistical science can play an

- How to estimate statistical quantities from training data.
- How to estimate linear regression coefficients from data.
- How to make predictions using linear regression for new data.

Multiple Linear Regression

- Real-world problems involve more than one independent variable.
- Multiple linear regression is used for numeric prediction task in this scenario.
- Mathematically represented as :

$$y = a+b_1x_1 +b_2x_2 + b_3x_3 \ldots\ldots + b_nx_n$$

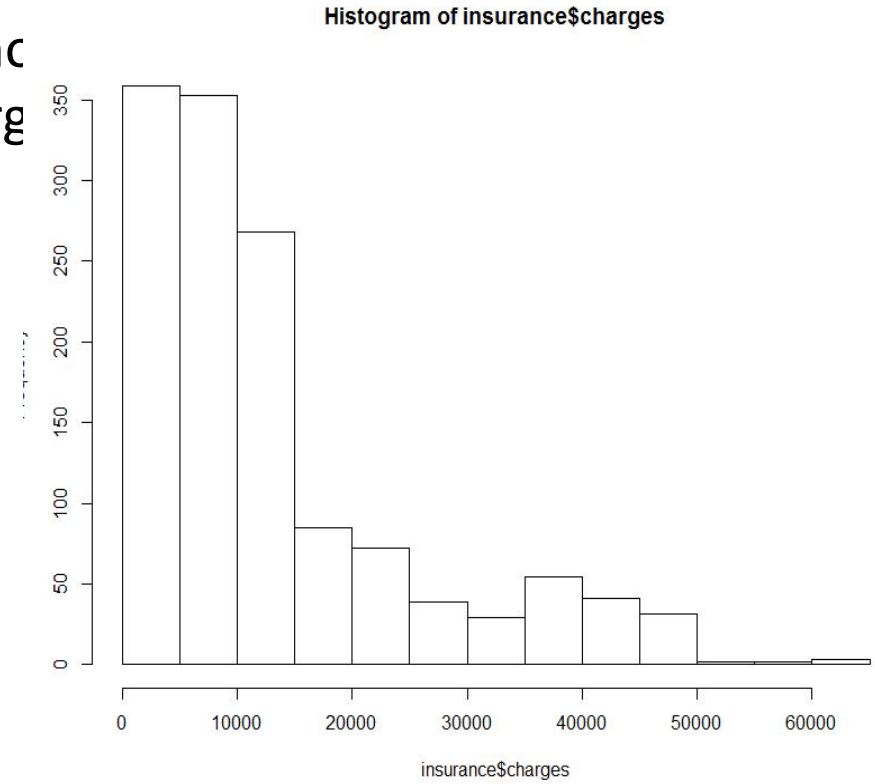
Case Study : Predicting Medical Expenses

In order for an insurance company to make money, it needs to collect more in yearly premiums than it spends on medical care to its beneficiaries. As a result, insurers invest a great deal of time and money to develop models that accurately forecast medical expenses.

Medical expenses are difficult to estimate because the most costly conditions are rare and seemingly random. Still, some conditions are more prevalent for certain segments of the population. For instance, lung cancer is more likely among smokers than non-smokers, and heart disease may be more likely among the obese.

The goal of this analysis is to use patient data to estimate the average medical care expenses for such population segments. These estimates could be used to create actuarial tables which set the price of yearly premiums higher or lower depending on the expected treatment costs.

- Import insurance dataset into pythc
- Plot a histogram on insurance charg



Majority of the individuals
Have yearly medical expenses
Between zero and $15000.

Key points before analysis:

1. Linear regression assumes a normal distribution of dependent variable.
2. Regression models require every feature to be numeric .

Explore relationships among features

Perform initial assessment of how independent variables are related to dependent variable using correlation matrix.
are strong