IT496: Introduction to Data Mining

Lecture 07

Fundamentals of Predictive Analytics

[Representation, Evaluation, and Optimization]

Arpit Rana

8th August 2023

Data Mining Tasks

Disclaimer: Most images incorporated within the presentation slides have been sourced from different sources on the web and ML books.

Data Mining Tasks

Data Mining Tasks

The actual data mining task is the semi-automatic or automatic analysis of large quantities of data to extract interesting patterns.

Descriptive

Find human-interpretable patterns that describe the data.

- Cluster Analysis
- Outlier Analysis
- Association Rule Mining
- Sequence Pattern Mining

Predictive

Use some variables to predict future or unknown values of other variables.

- Regression
- Classification

In Machine Learning terminology, these tasks are categorised as "***Unsupervised Learning***".

In Machine Learning terminology, these tasks are categorised as "***Supervised Learning***".

Data Mining Tasks

Data Mining Tasks

The actual data mining task is the semi-automatic or automatic analysis of large quantities of data to extract interesting patterns.

Descriptive

Find human-interpretable patterns that describe the data.

- Cluster Analysis
- Outlier Analysis
- Association Rule Mining
- Sequence Pattern Mining

In Machine Learning terminology, these tasks are categorised as "***Unsupervised Learning***".

Predictive

Use some variables to predict future or unknown values of other variables.

- Regression
- Classification

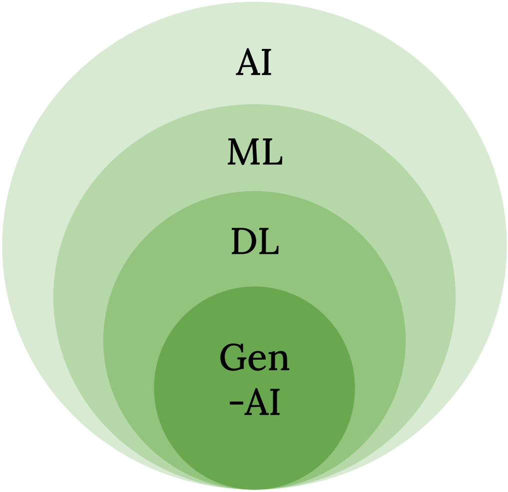
In Machine Learning terminology, these tasks are categorised as "***Supervised Learning***".

Machine Learning: Definition

Machine Learning is

- the science (and art) of programming computers
- so they can learn from data.

– *Aurelien Geron, Google*

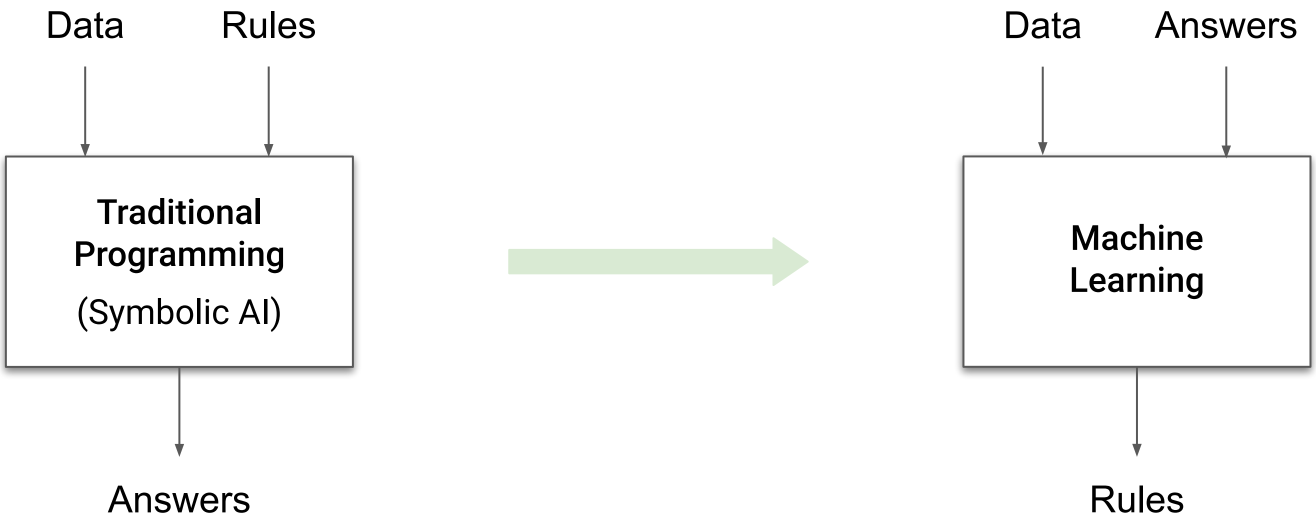


Machine Learning: Example

A Spam Filter,

- a Machine Learning Program, given
  - examples of “spam” emails (e.g. flagged by users), and
  - examples of “ham” (i.e. regular) emails
- can learn to flag spam

Machine Learning: A New Programming Paradigm



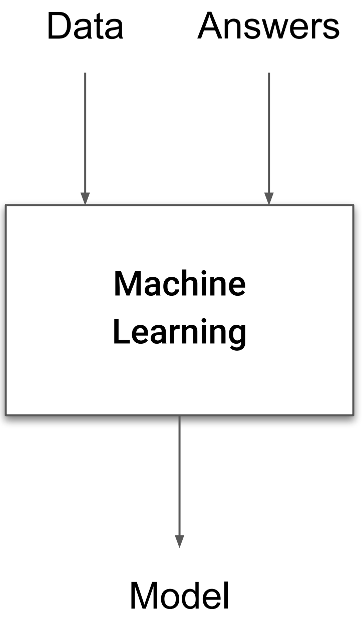
- A long list of complex (hard coded) rules
- Keep writing new rules as the new phrases are introduced by spammers

- Automatically learns which words or phrases are good predictors of spam

Machine Learning: Definition Revisited

Machine Learning is the training of a model from data that generalises a decision against a performance measure.

- Training a model suggests training examples.
- A model suggests state acquired through experience.
- Generalises a decision suggests the capability to make a decision based on inputs and anticipating unseen inputs in the future for which a decision will be required.
- against a performance measure suggests a targeted need and directed quality to the model being prepared.

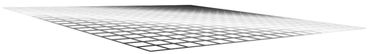


Learning = Representation + Evaluation + Optimization

Representation

Choosing a representation of the learner: the *hypotheses space* or *the model class* — the set of models that it can possibly learn.

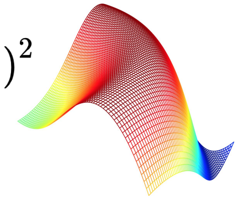
$$h_\beta(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_m X_m$$
$$= \sum_{i=1}^{m} \beta_i X_i$$



Evaluation

Choosing an evaluation function (also called objective function, utility function, loss function, or scoring function) is needed to distinguish good classifiers from bad ones.

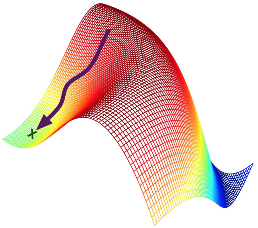
$$J(\beta) = \sum_{i=1}^{m} (h_\beta(X_i) - y_i)^2$$



Optimization

Choosing a method to search among the models in the hypothesis space for the highest-scoring one.

$$min\ J(\beta)$$
��



Learning = Representation + Evaluation + Optimization

Representation	Evaluation	Optimization
Instances	Accuracy/Error rate ✔	Combinatorial optimization
K-nearest neighbor ✔	Precision and recall ✔	Greedy search ✔
Support vector machines ✔	Squared error ✔	Beam search ✔
Hyperplanes	Likelihood	Branch-and-bound
Naive Bayes	Posterior probability	Continuous optimization
Logistic regression ✔	Information gain ✔	Unconstrained
Decision trees ✔	K-L divergence	Gradient descent ✔
Sets of rules	Cost/Utility ✔	Conjugate gradient
Propositional rules	Margin ✔	Quasi-Newton methods
Logic programs		Constrained
Neural networks ✔		Linear programming
Graphical models		Quadratic programming ✔
Bayesian networks		
Conditional random fields		

Supervised Learning

Problem Settings and Examples

Supervised Learning: A Formal Model

The learner's input:

- Domain set
  An arbitrary set (instance space), X, the set of objects (a.k.a. instances, domain points) we may wish to label.

- Label set
  A set of possible labels, Y. e.g., {0, 1}, {-1, 1}.

- Training data
  $S = ((x_1, y_1) \ldots (x_m, y_m))$ is finite <u>*sequence*</u> of pairs in X x Y, i.e., a sequence of labeled domain points.

The learner's output:

- A prediction rule, $h : X \rightarrow Y$, also called a *predictor*, a *hypothesis*, or a *classifier*.
  - The learner returns h upon receiving the training sequence S.
  - It can be used to predict the label of new domain points (*like the past ones*).

Supervised Learning: A Formal Model

Data-generation Model:

- Let D be a probability distribution over X x Y, i.e., D is joint probability distribution over domain points and labels.
  - A distribution D_x over unlabeled domain points (sometimes called *marginal distribution*),
  - A conditional probability over labels for each domain point, $D((x, y) \mid x)$.

Independent and Identically Distributed (I.I.D.) Assumption

- Each domain point x has the same prior probability distribution (to be sampled):

  $P(x_i) = P(x_{i+1}) = P(x_{i+2}) = \cdots,$

  and is independent of the previous examples:

  $P(x_i) = P(x_i \mid x_{i-1}, x_{i-2}, \ldots) .$

Supervised Learning: A Formal Model

More formally, the task of supervised learning can be defined as -

Given a training set (S) of m example input-output pairs,

$$S = (X,\, y)$$

$$X = \begin{pmatrix} \left(x^{(1)}\right)^T \\ \left(x^{(2)}\right)^T \\ \cdot \\ \cdot \\ \cdot \\ \left(x^{(m-1)}\right)^T \\ \left(x^{(m)}\right)^T \end{pmatrix},\ y = \begin{pmatrix} y^{(1)} \\ y^{(1)} \\ \cdot \\ \cdot \\ \cdot \\ y^{(m-1)} \\ y^{(m)} \end{pmatrix}$$

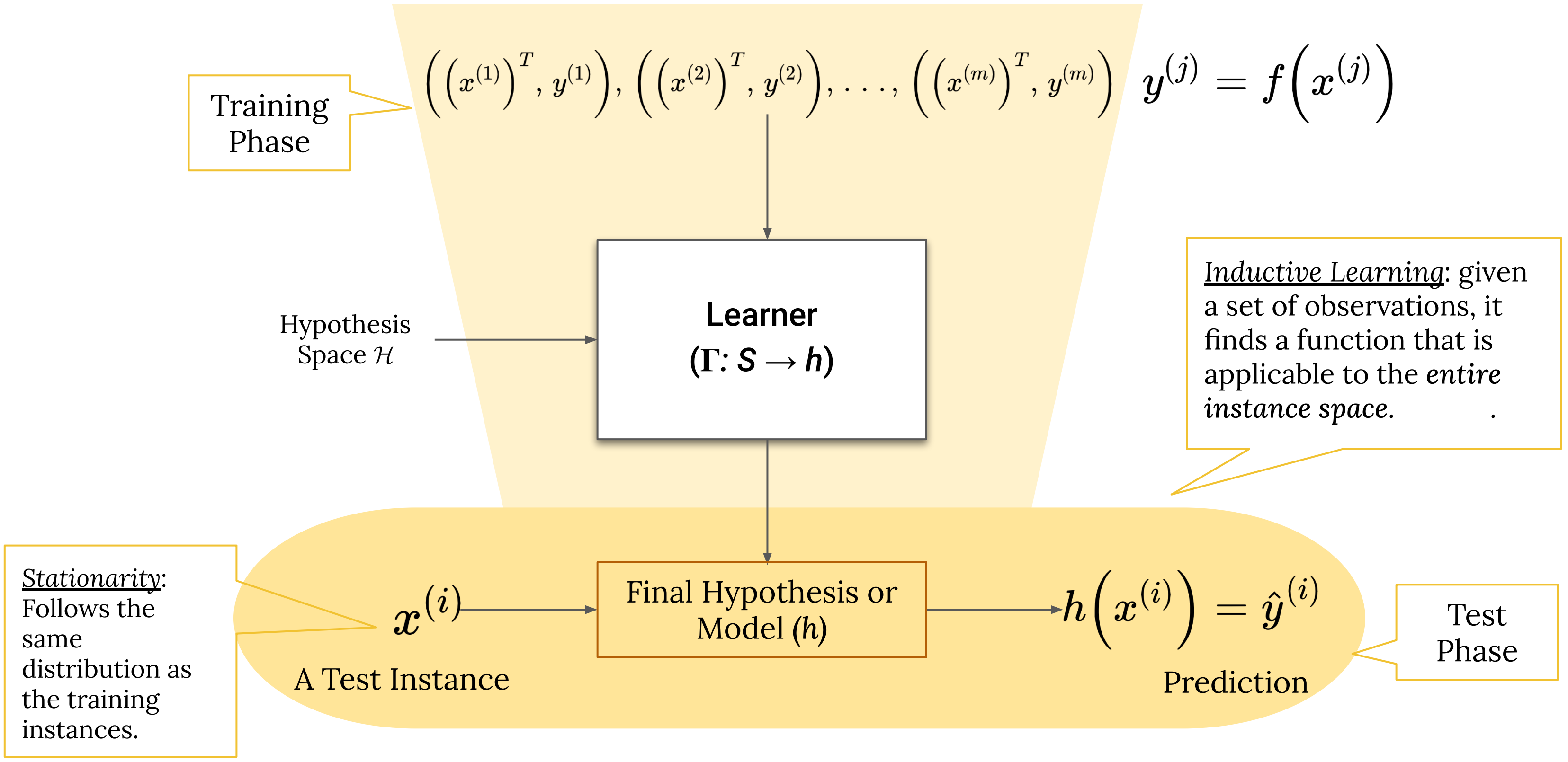
We call the output $\boldsymbol{y}^{(i)}$ the *ground truth* — the true answer we are asking our model to predict.

$$\left(\left(x^{(1)}\right)^T,\, y^{(1)}\right),\ \left(\left(x^{(2)}\right)^T,\, y^{(2)}\right),\ \ldots,\ \left(\left(x^{(m)}\right)^T,\, y^{(m)}\right)$$

where each pair was generated by an unknown function $y = f(x)$,

discover a function h that approximates the true function f .

Supervised Learning Process



Next lecture

Choosing a Hypothesis Space

10th August 2023