Welcome to:

Learning deterministic models

Unit objectives

After completing this unit, you should be able to:

- Understand the concept of Supervised Learning
- Gain insight on Regression and its variants along with real world problems solved using regression
- Learn about the estimation of parameters
- Acquire knowledge on decision trees and their role in classification
- Develop skills on clustering and its importance in machine learning
- Gain knowledge on the reinforcement learning and their significance
- Learn about Structured Learning problem and Casual Learning

Introduction

- Deterministic models are mathematical models in which outcomes are precisely determined through known relationships among system states and events.

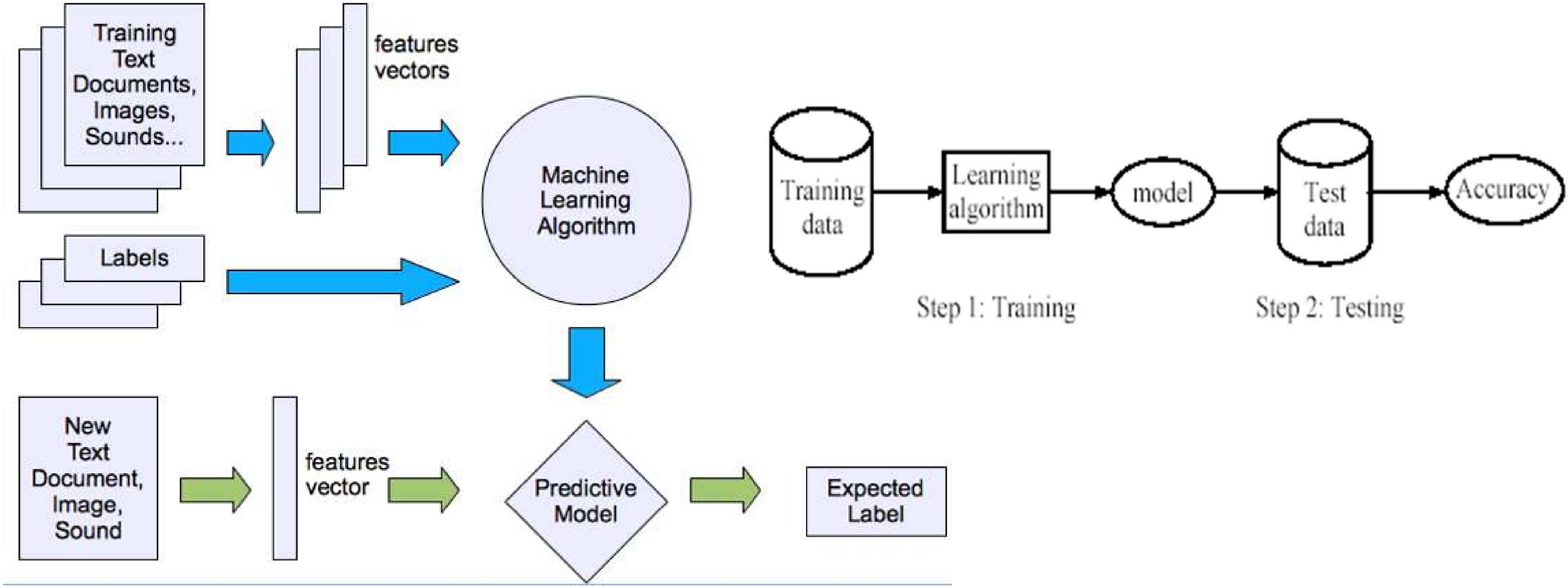
- No random variation in the outcomes
  - given input will always produce the same output.
  - a known chemical reaction which always gives known chemical compounds
  - a computer program which gives expected output for a given set of inputs.

- Stochastic models give ranges of values for variables in the form of probability distributions.

- Learning the deterministic models is a process of learning a deterministic function which maps each input variable to an output variable.

- Two main paradigms in deterministic learning
  - supervised learning
  - unsupervised learning
  - Third paradigm : reinforcement learning.

Supervised learning



- Learning (training): Learn a model using the training data
- Testing: Test the model using unseen test data to assess the model accuracy

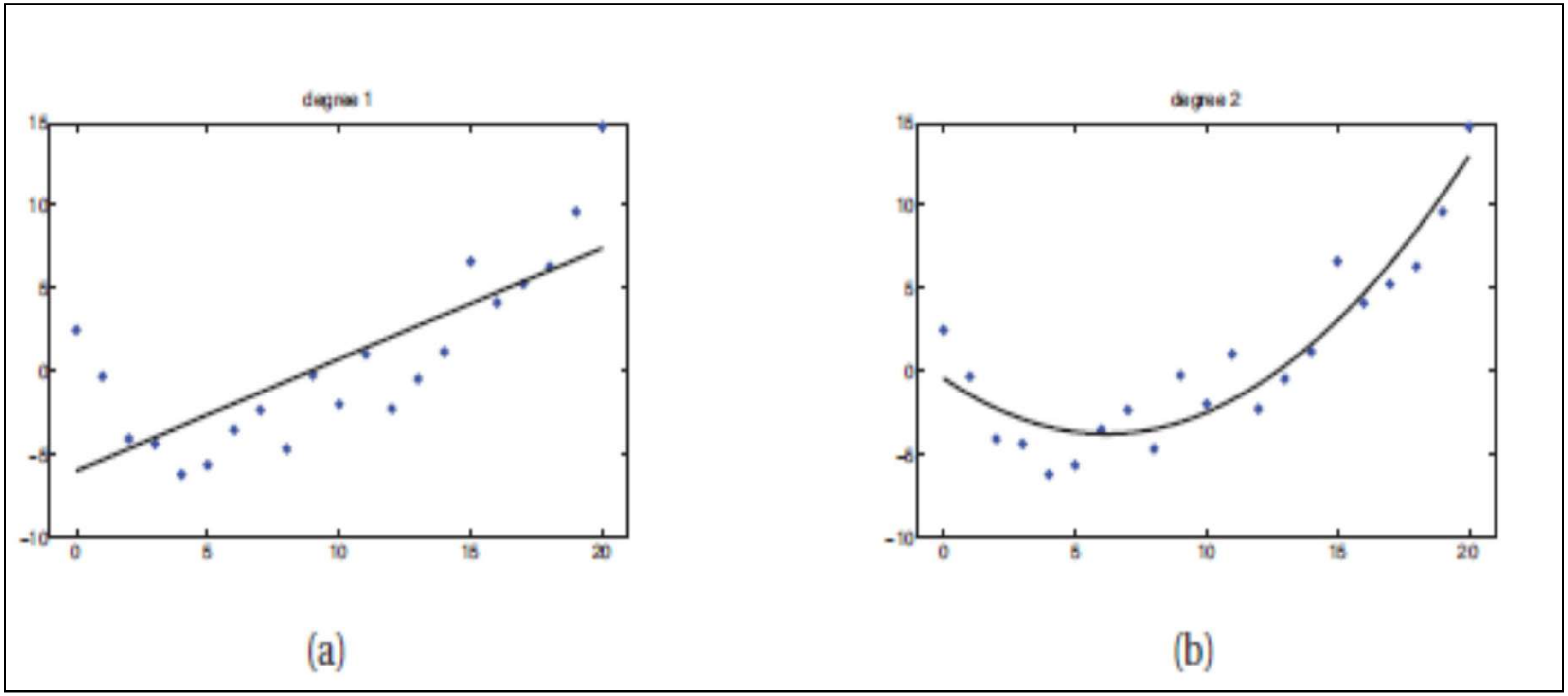
$$\text{Accuracy} = \frac{\text{Number of correct classifica tions}}{\text{Total number of test cases}}$$

Regression

- Regression is a standard statistical technique for performing supervised learning when all variables are continuous.

- It is just like classification except that the response variable is continuous.

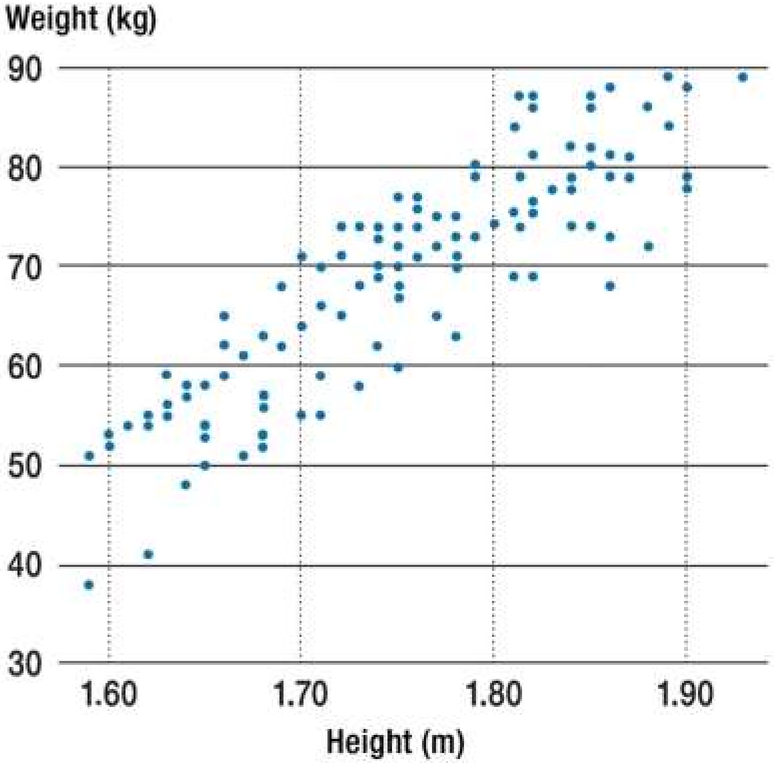
- Basically regression is given with y values corresponding to x values to fit a curve which can represent most of the y values (best fit).



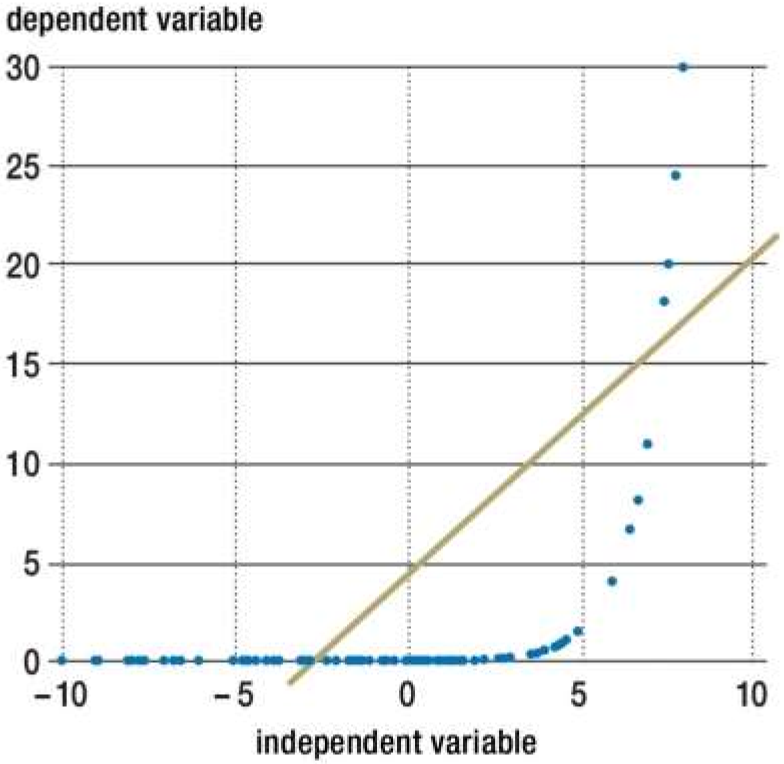
Linear regression

- Linear regression is used to study the linear relationship between a dependent variable Y (blood pressure) and one or more independent variables X (age, weight, sex).

- The dependent variable Y must be continuous, while the independent variables may be either continuous (age), binary (sex), or categorical (social status).



A scatter plot showing a linear relationship



A scatter plot showing an exponential relationship.

Multiple linear regression

- Linear regression model can also be extended to the models with more than one input vectors. This model is called multiple regression model. Geometrically, multiple regression is fitting a hyper plane in the d- dimensional space.

- Let x be a d dimensional vector, $\boldsymbol{x} = (x_1, \ldots, x_d)$, then, simple regression can be extended to higher dimensional data as follows.

$$y(x) = w^T x + \epsilon = \sum_{i=1}^{d} \omega_i x_i + \epsilon$$

- Where ω_i are the parameters to be tuned during the training phase and ϵ is the error factor which is the difference between true response and the estimated value. Parameters are tuned so that the error function ϵ is minimised.

A multiple regression analysis

- A multiple regression analysis involves estimation, testing, and diagnostic procedures designed to fit the multiple regression model to a set of data.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- The Method of Least Squares

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

- The prediction equation is the line that minimizes SSE, the sum of squares of the deviations of the observed values y from the predicted values

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{i,k} + \varepsilon_i \\ &= \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i \qquad i = 1,\ldots,n \end{aligned}$$

The analysis of variance for multiple regression

- The analysis of variance divides the total variation in the response variable y, into two portions:

$$\mathrm{SS_T} = \sum y_i^2 - \frac{{\sum y_i}^{\ 2}}{n}$$

  - – SSR (sum of squares for regression) measures the amount of variation explained by using the regression equation.
  - – SSE (sum of squares for error) measures the residual variation in the data that is not explained by the independent variables.

- The values must satisfy the equation Total SS = SR + SSE.
- There are (n - 1) degrees of freedom.
- There are k regression degrees of freedom.
- There are (n – p) degrees of freedom for error.
- MS = SS / d f

Examples for multiple regression

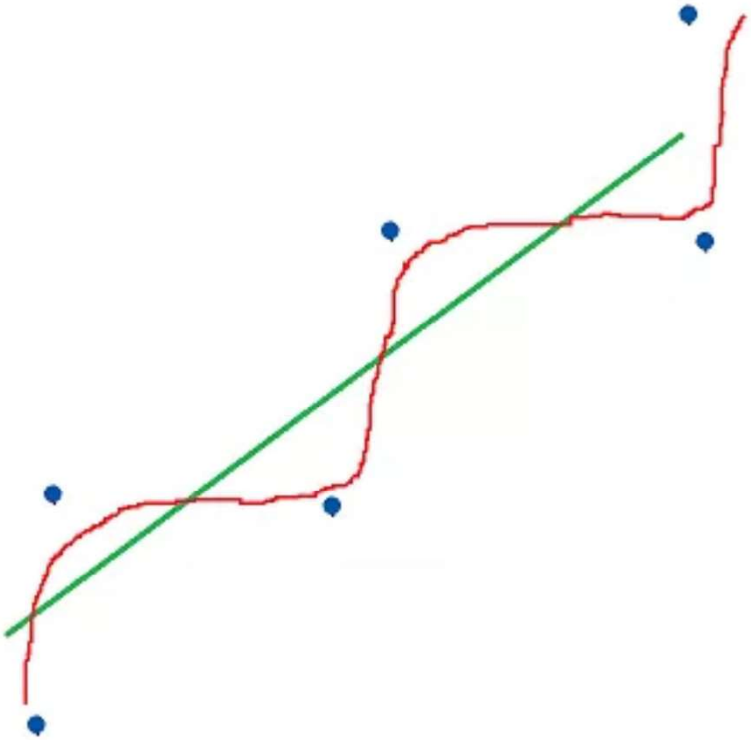
- Example 1: Consider the problem of finding whether the patient is suffering from a particular disease. In this case, the input vector x might have multiple dimensions such as, age, genetic information and medical diagnosis.

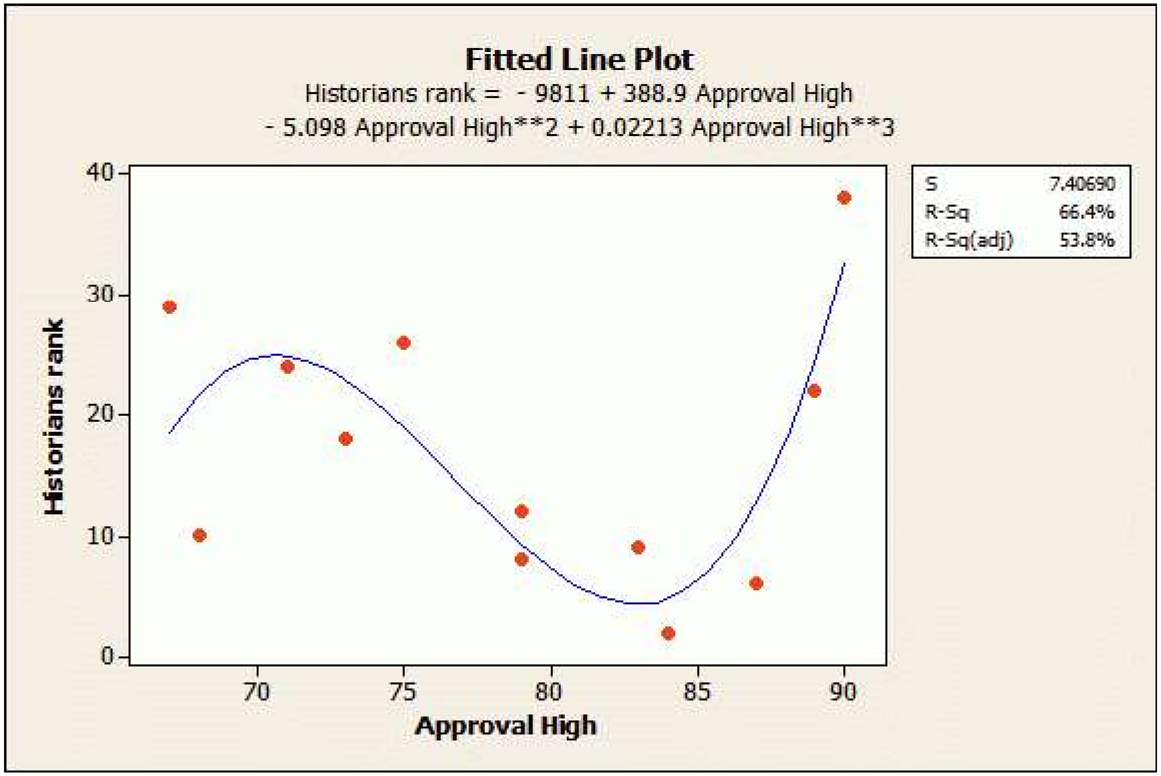
- Example 2: Consider a 3-dimensional system with a linear regression model, y=a x_1+b x_2+c x_3+d , with the parametric values, a =1, b=2, c=3 and d = 1. Then for a given input value x = (1,2,1), y is estimated as, y = 1 + 4 + 3 + 1 = 9. If we consider a nonlinear regression model say, =a x_(1)^3+b x_(2)^2+cx_3 +d , with the same parameters and input value then we estimate y as, y = 1 + 8 + 3 + 1 = 13.

Overfitting

- A statistical model begins to describe the random error in the data rather than the relationships between variables
  - Produce misleading R-squared values, regression coefficients, and p-values

- Graphical Illustration of Overfitting Regression Models





Detecting overfit models- Cross validation

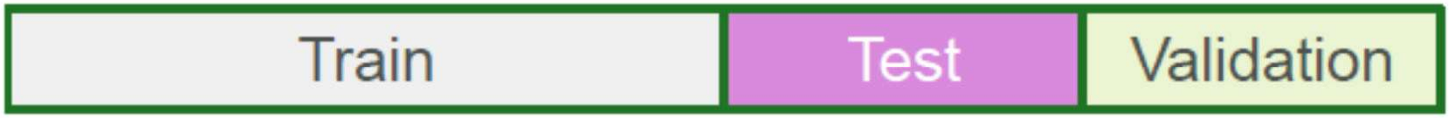
- Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.

- The general procedure:
  - Shuffle the dataset randomly.
  - Split the dataset into k groups
  - For each unique group:
    - Take the group as a hold out or test data set
    - Take the remaining groups as a training data set
    - Fit a model on the training set and evaluate it on the test set
    - Retain the evaluation score and discard the model
  - Summarize the skill of the model using the sample of model evaluation scores

Cross validation- The ideal procedure

- Divide data into three sets, training, validation and test sets

- Find the optimal model on the training set, and use the test set to check its predictive capability



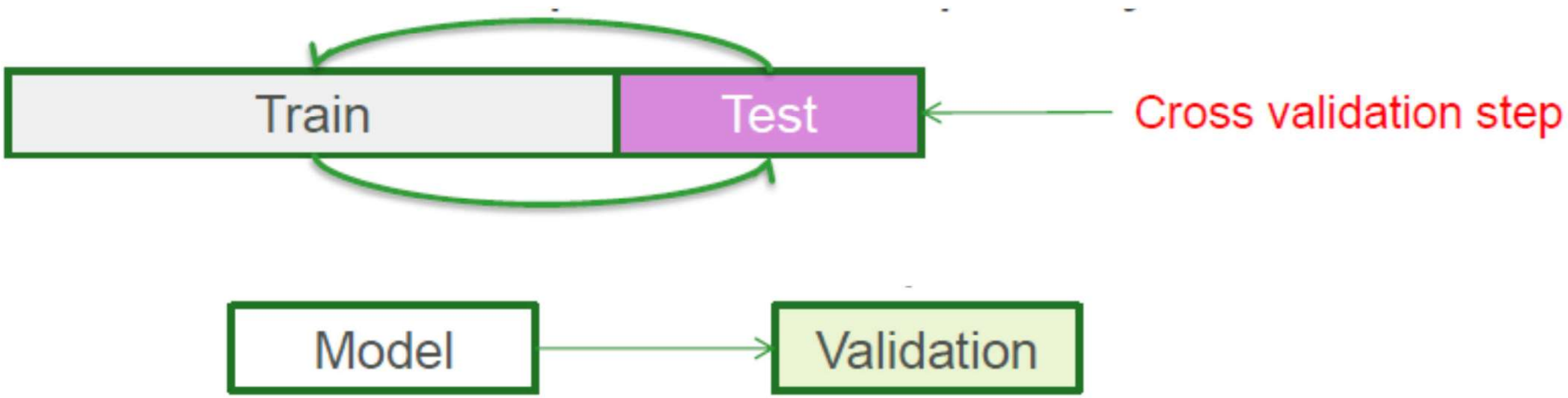
Train

Test

Validation

- See how well the model can predict the test set

- The validation error gives an unbiased estimate of the predictive power of a model



Train

Test

Cross validation step

Model

Validation

Parameter estimation

- The parameters are to be estimated based on the training set.
- The foremost criterion is to minimise the estimation error.
- Use least square error or maximum likelihood estimation (MLE) which is given by the error function

$$E = \sum_{i=1}^{N} (\boldsymbol{y_i} - w^T \boldsymbol{x_i})^2$$

- w^T is calculated so that the error function E is minimum.

Logistic regression

Linear regression can be generalised for two class classifiers (binary classifiers).

- We can recall, if the output y belongs to the set {0, 1}, it is called binary classifier.

- The classification procedure in which, linear regression is generalised for binary classification, by replacing the Gaussian distribution to Bernoulli distribution, is called logistic regression.

- It can be modelled as,

$$P(y|x,w) = \ Ber(y|\mu(x)$$

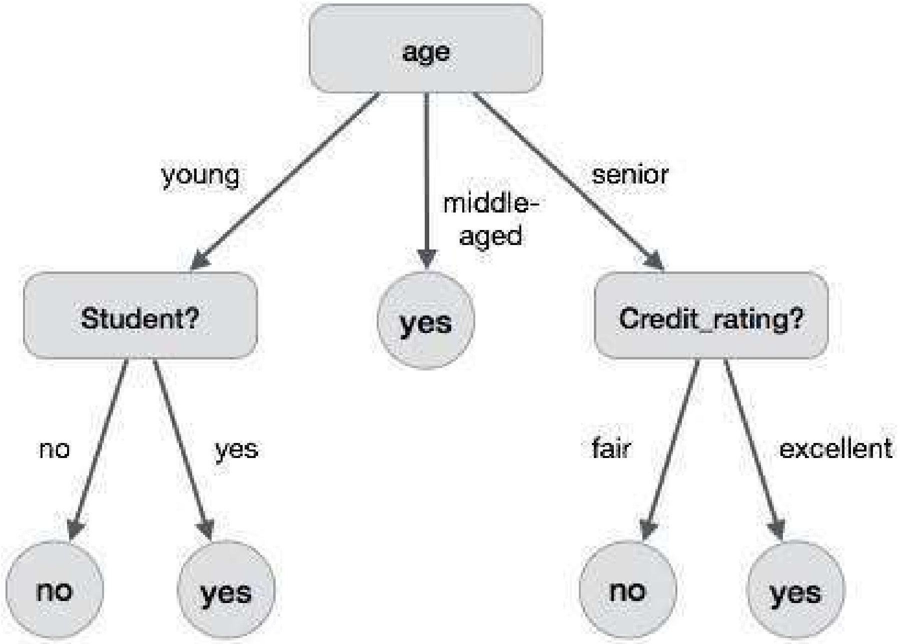
- Ber() stands for Bernoulli distribution which is more appropriate when the response is binary.

Decision trees- Background

Goal: Categorization

- Given an event, predict is category. Examples:
  - Who won a given ball game?
  - How should we file a given email?
  - What word sense was intended for a given occurrence of a word?

- Event = list of features. Examples:
  - Ball game: Which players were on offense?
  - Email: Who sent the email?
  - Disambiguation: What was the preceding word?

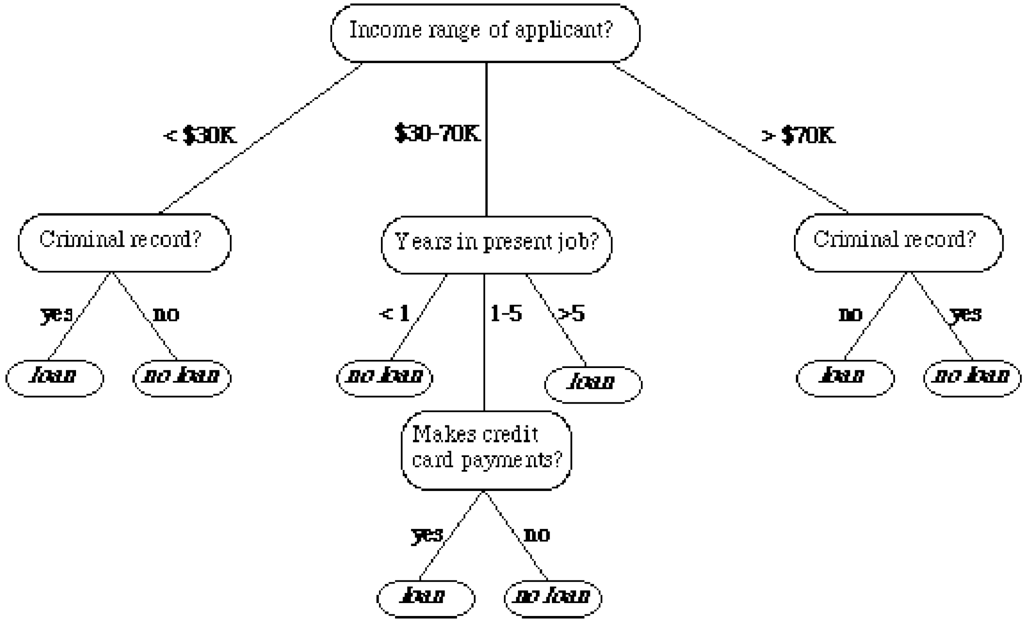


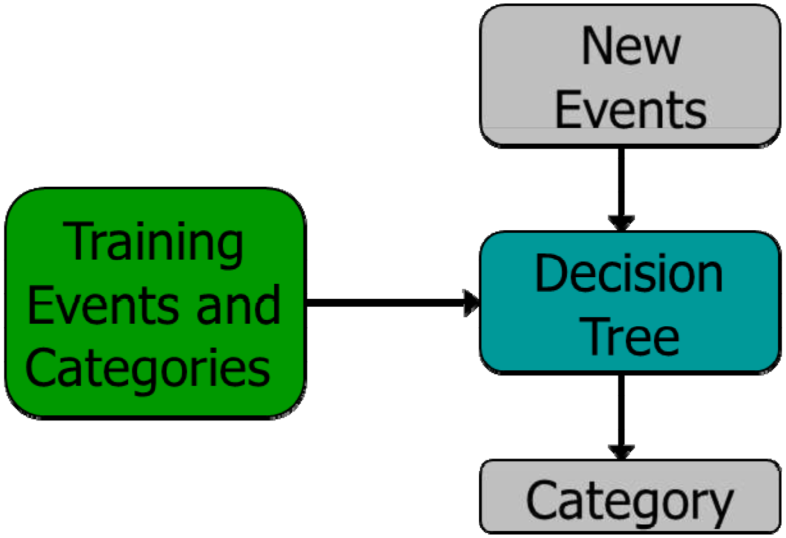
Decision trees

- A decision tree is a flowchart-like tree structure, where each internal node (nonleaf node) denotes a test on attribute, each branch represents an outcome of the test, and leaf node (or terminal) holds a class label.

- Decision tree induction is the learning of decision trees from class-labeled training tuples(instances).

- Decision tree is used to predict categories for new events.
  - The training data is used to build the decision tree.
  - The topmost node in a tree is the root node




New Events

Training Events and Categories

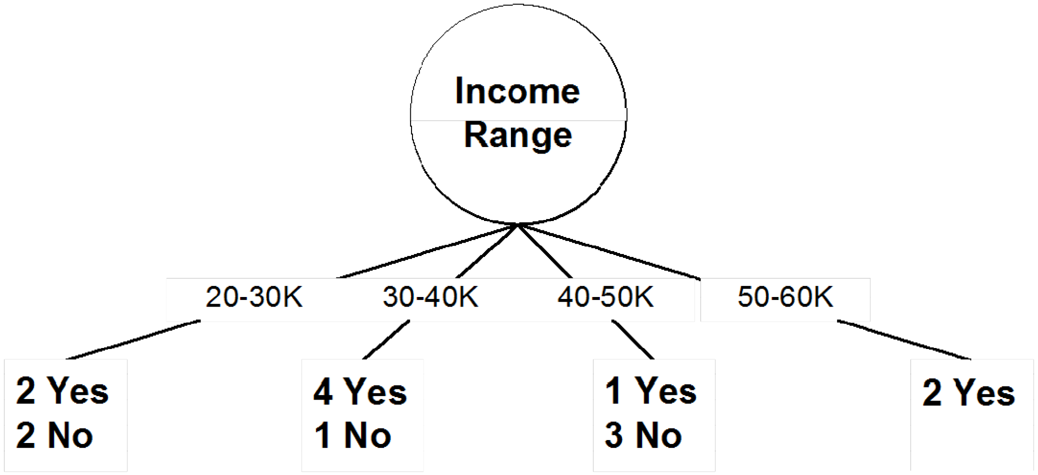
Decision Tree

Category

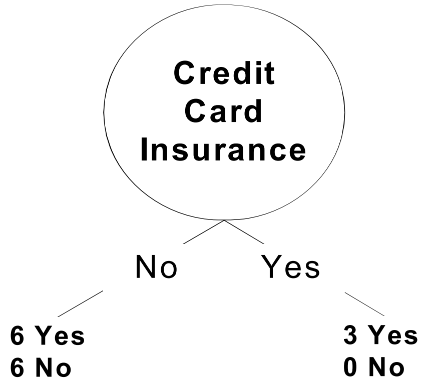
Decision trees for credit card promotion

Table: **The Credit Card Promotion Database**

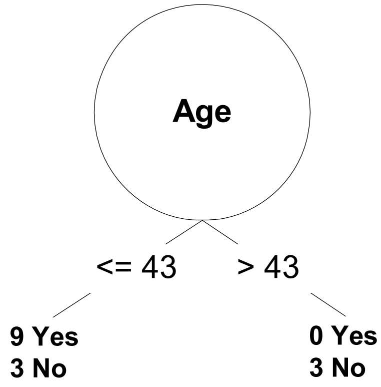
Income Range	Life Insurance Promotion	Credit Card Insurance	Sex	Age
40–50K	No	No	Male	45
30–40K	Yes	No	Female	40
40–50K	No	No	Male	42
30–40K	Yes	Yes	Male	43
50–60K	Yes	No	Female	38
20–30K	No	No	Female	55
30–40K	Yes	Yes	Male	35
20–30K	No	No	Male	27
30–40K	No	No	Male	43
30–40K	Yes	No	Female	41
40–50K	Yes	No	Female	43
20–30K	Yes	No	Male	29
50–60K	Yes	No	Female	39
40–50K	No	No	Male	55
20–30K	Yes	Yes	Female	19



A partial decision tree with root node=income range



A partial decision tree with root node=CreditCardInsurance



A partial decision tree with root node=age

An algorithm for building decision trees

- Let T be the set of training instances.

- Choose an attribute that best differentiates the instances in T.

- Create a tree node whose value is the chosen attribute.
  - Create child links from this node where each link represents a unique value for the chosen attribute.
  - Use the child link values to further subdivide the instances into subclasses.

- For each subclass created in step 3:
  - If the instances in the subclass satisfy predefined criteria or if the set of remaining attribute choices for this path is null, specify the classification for new instances following this decision path.
  - If the subclass does not satisfy the criteria and there is at least one attribute to further subdivide the path of the tree, let T be the current set of subclass instances and return to step 2.

Attribute selection measure- Information gain

- Select the attribute with the highest information gain
- Let p_i be the probability that an arbitrary tuple in D belongs to class C_i, estimated by $|C_{i,D}|/|D|$

- Expected information (entropy) needed to classify a tuple in D:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

- Information needed (after using A to split D into v partitions) to classify D:

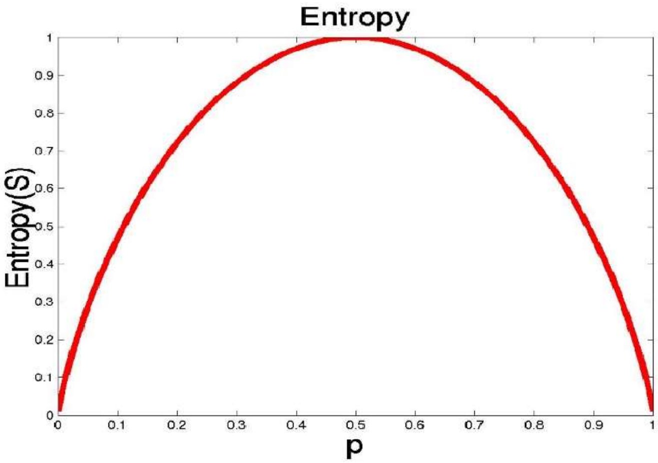
$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j)$$

- Information gained by branching on attribute A

$$Gain(A) = Info(D) - Info_A(D)$$

Entropy

- S is a sample of training examples
  - p+ is the proportion of positive examples
  - p- is the proportion of negative examples



- Entropy measures the impurity of S
  - Entropy(S) = -p+ log2 p+ - p- log2 p-

- Entropy(S)= expected number of bits needed to encode class (+ or -) of randomly drawn members of S.

- Information theory optimal length code assign –log2 p bits to messages having probability p.

- So the expected number of bits to encode (+ or -) of random member of S:
  - -p+ log2 p+ - p- log2 p-

Decison Tree- Weekend example

Step-1: Let *T* be the set of training instances.

Weekend (Example)	Weather	Parents	Money	Decision (Category
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay in
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

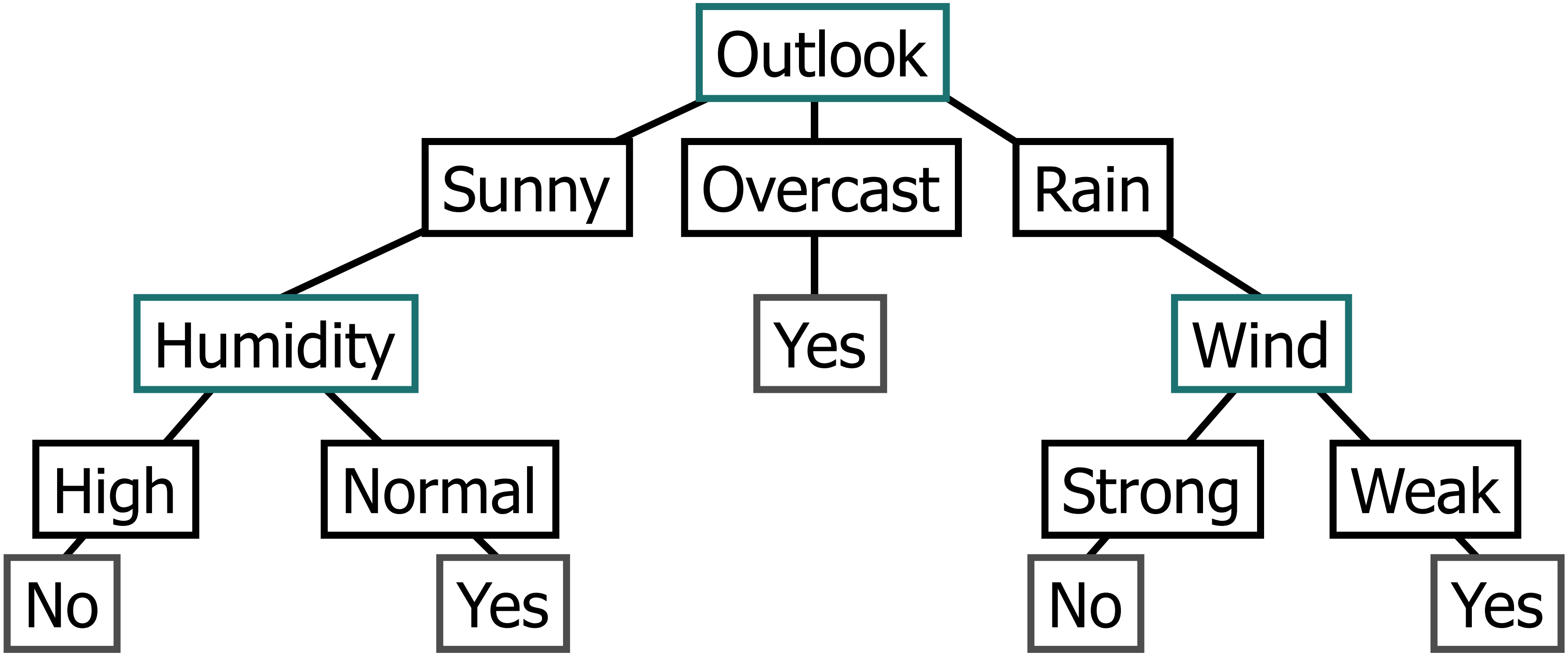
Occam’s Razor

- ”If two theories explain the facts equally weel, then the simpler theory is to be preferred”

- Arguments in favor:
  - Fewer short hypotheses than long hypotheses
  - A short hypothesis that fits the data is unlikely to be a coincidence
  - A long hypothesis that fits the data might be a coincidence

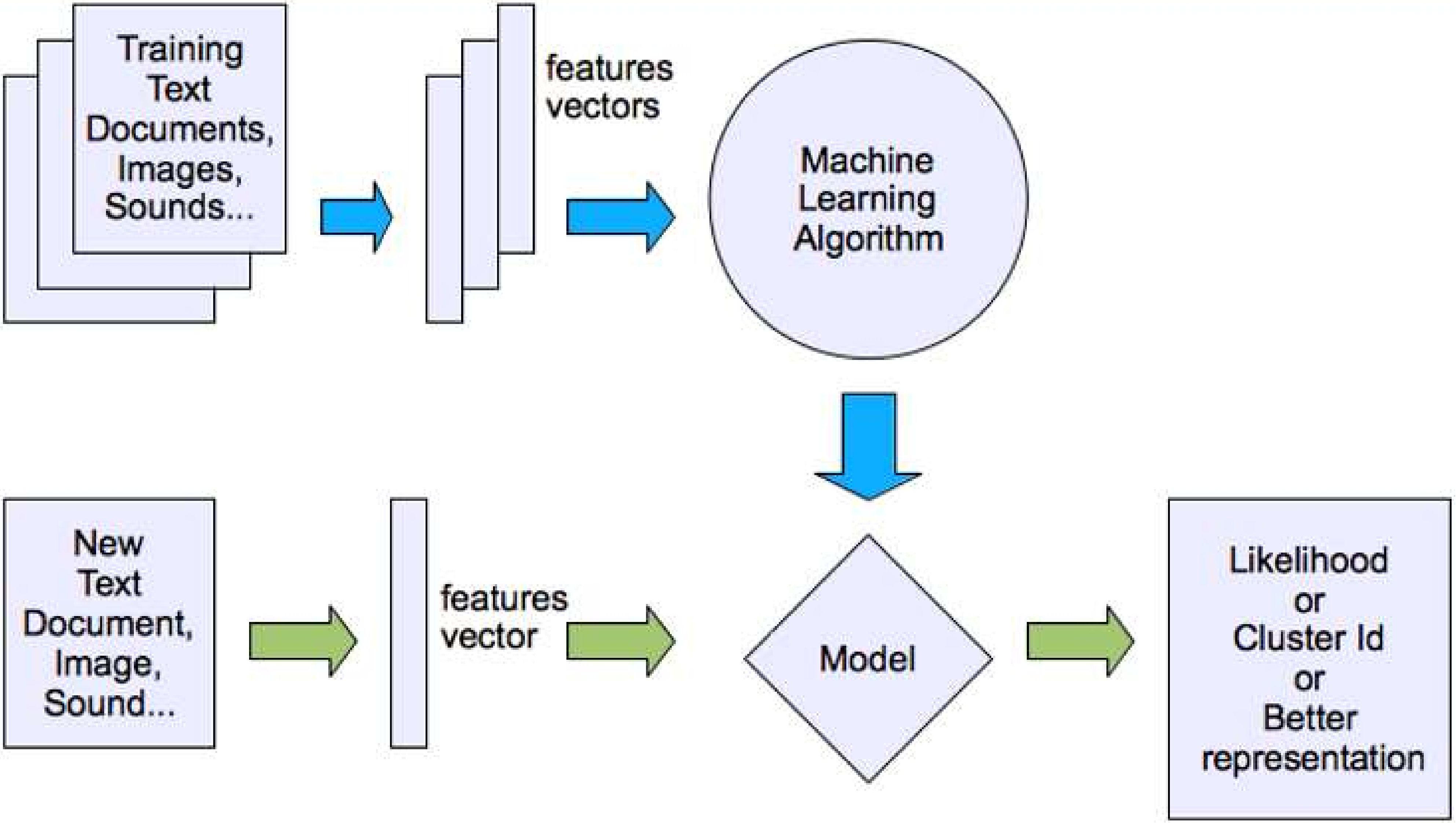
- Arguments opposed:
  - There are many ways to define small sets of hypotheses

Converting a tree to rules

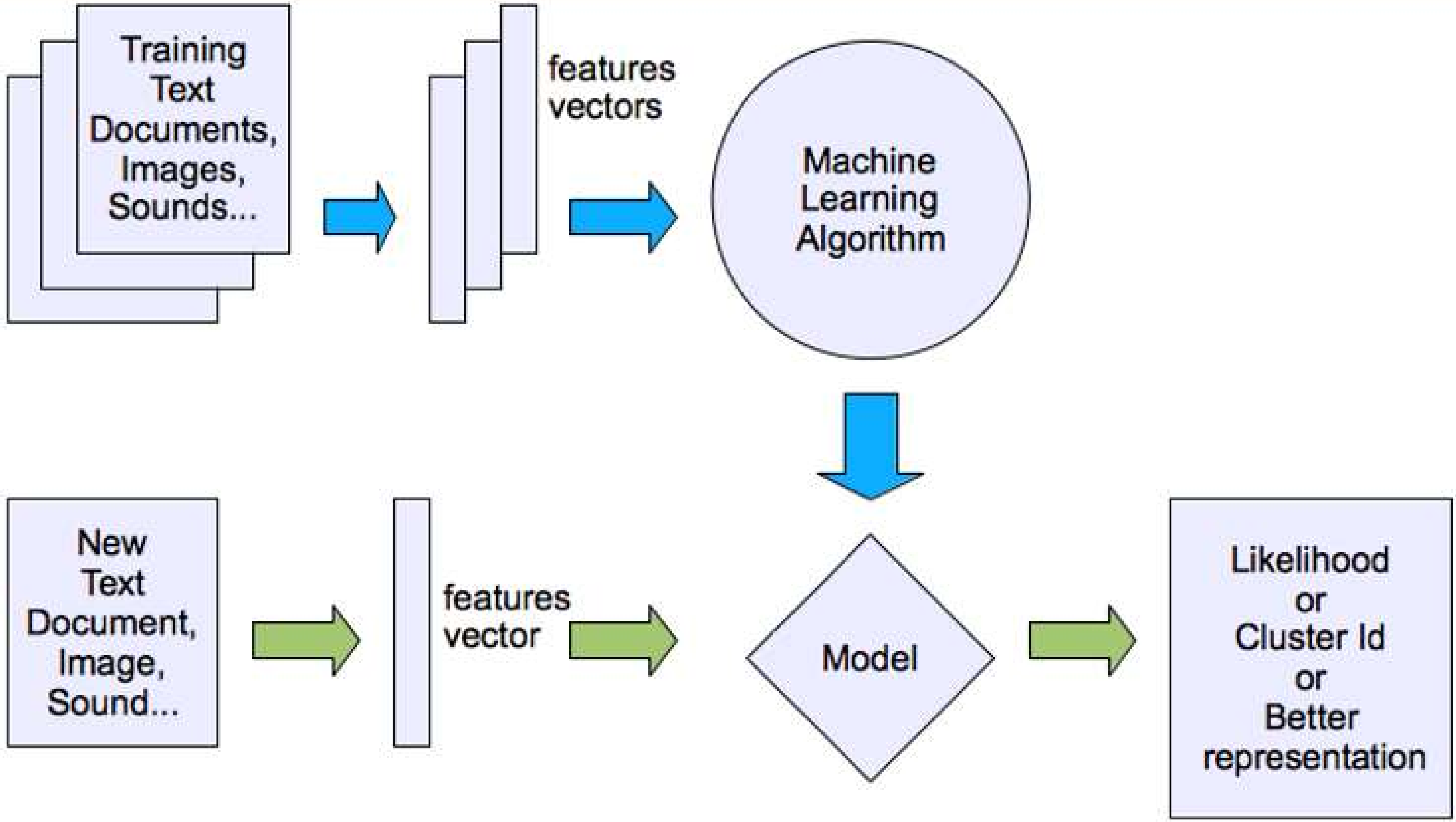


R_1: If (Outlook=Sunny) $\wedge$ (Humidity=High) Then PlayTennis=No
R_2: If (Outlook=Sunny) $\wedge$ (Humidity=Normal) Then PlayTennis=Yes
R_3: If (Outlook=Overcast) Then PlayTennis=Yes
R_4: If (Outlook=Rain) $\wedge$ (Wind=Strong) Then PlayTennis=No
R_5: If (Outlook=Rain) $\wedge$ (Wind=Weak) Then PlayTennis=Yes

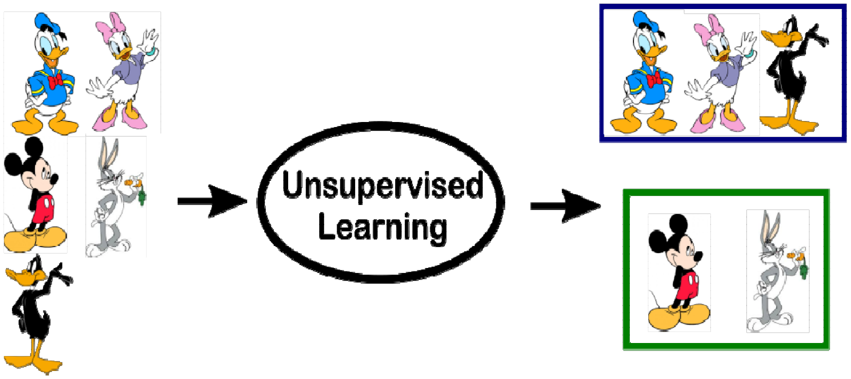
Unsupervised learning

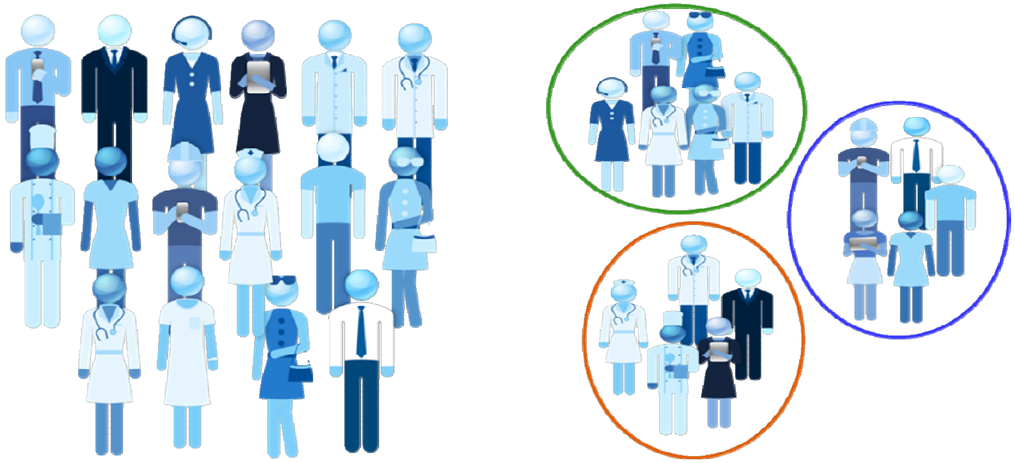

Training Text Documents, Images, Sounds...
features vectors
Machine Learning Algorithm
New Text Document, Image, Sound...
features vector
Model
Likelihood or Cluster Id or Better representation

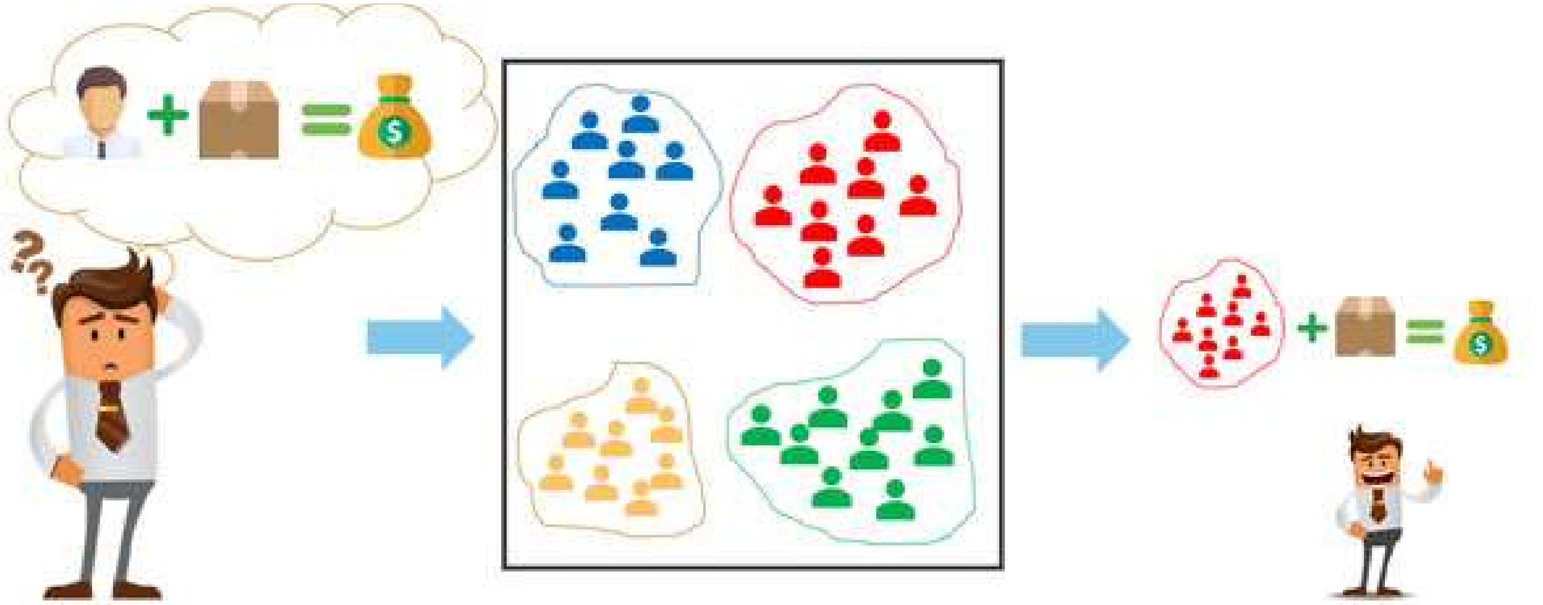
Semi-Supervised learning


Training Text Documents, Images, Sounds...
features vectors
Machine Learning Algorithm
New Text Document, Image, Sound...
features vector
Model
Likelihood or Cluster Id or Better representation

Clustering







Trying to determine the appropriate audience for the product

Using clustering algorithms on the customer base

Selling the product to the targeted audience

K – means clustering

- K-means clustering is a type of unsupervised learning, which is used when we have unlabelled data (i.e., data without defined categories or groups).

- K means algorithm will divide the given data into K clusters.

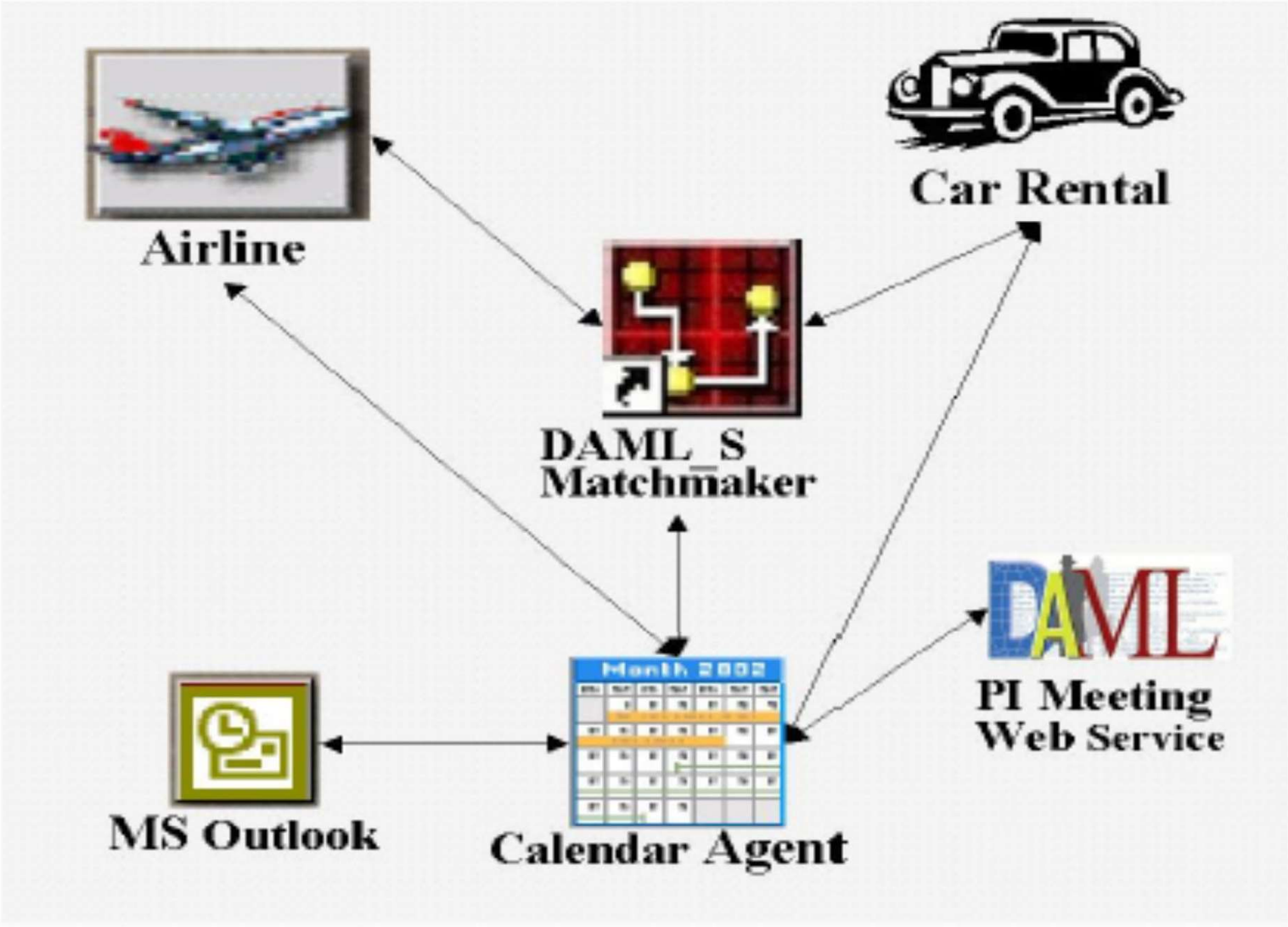
- This algorithm works iteratively to assign each data point to one of K groups.

- In order to divide the data into groups it utilises the unique feature of data objects.

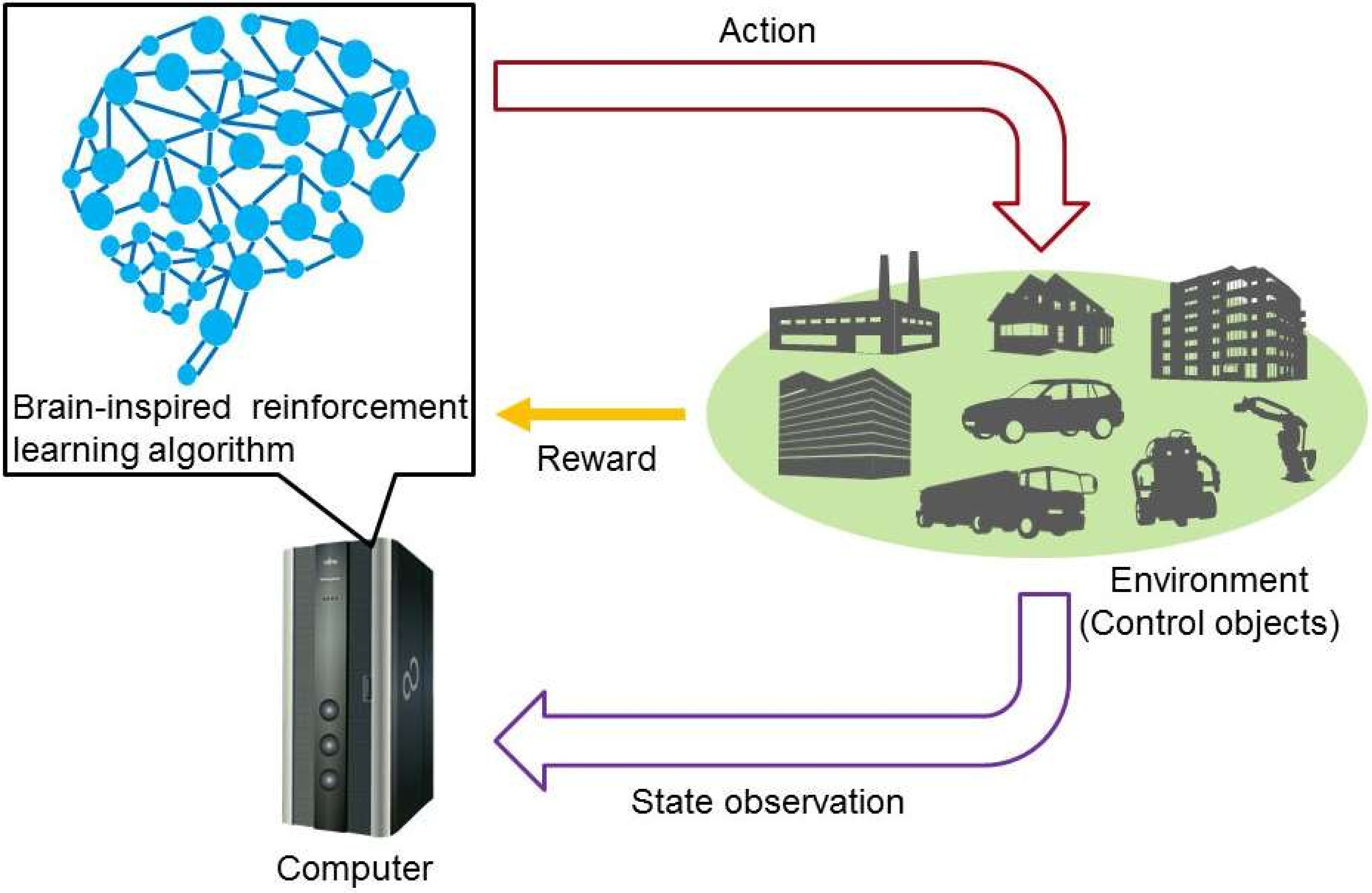
- Data points are clustered based on feature match score. Match score is a measure of similarity or dissimilarity.

- The output of K –means algorithms are the K clusters and the centroids of each of the clusters.

Automated discovery



Reinforcement learning



Brain-inspired reinforcement learning algorithm

Action

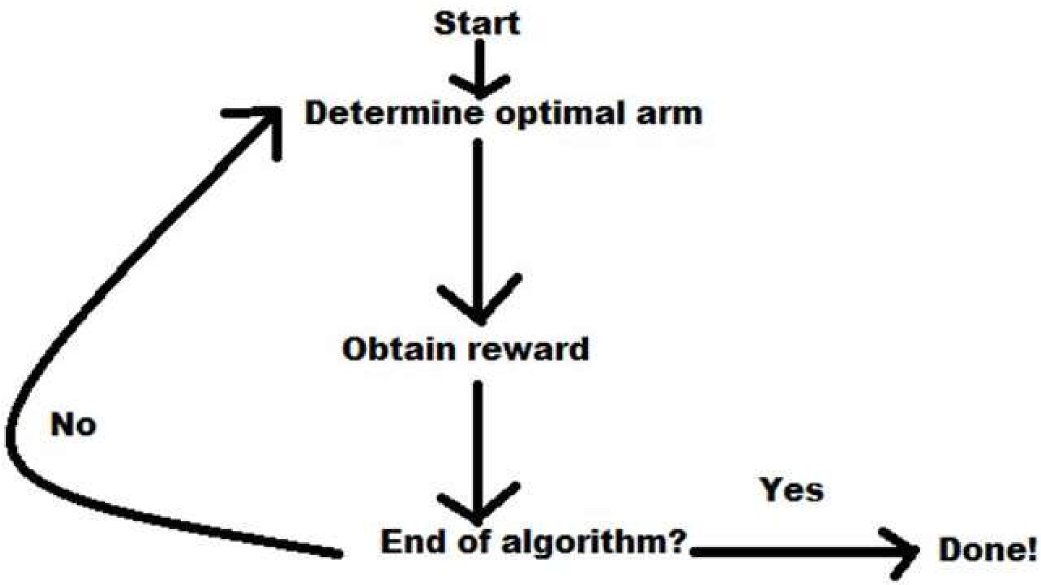
Reward

Environment (Control objects)

State observation

Computer

Multi-armed bandit algorithms



Influence diagrams

- Like decision tress, influence diagrams are also graphical representations of a decision situation.

- These diagrams show the dependences between the variables.

- They are similar to decision trees; in fact, influence diagram gives the abstract of the information contained in a decision tree.

- These are used to evaluate decision trees.

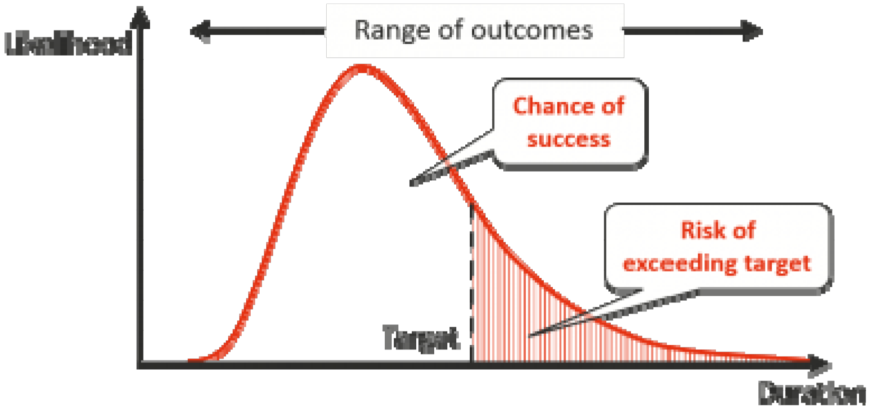
- An influence diagram is a directed acyclic graph with three types of nodes. A decision node corresponding to a decision distinction is drawn as a rectangle. An uncertainty node corresponding to an uncertainty distinction is drawn as an oval.

- Finally, a value node corresponding to a component of an additively separable utility function is drawn as an octagon, and solid lines are used to denote influence.

- The distinction and its corresponding node can interchangeable

Risk modelling

- Risk modelling is to develop a mathematical and statistical model for risk analysis.
- In the last few years there is significant increase in financial markets and services.
- There are a huge number of people on the financial markets taking risky positions and to evaluate their positions properly they need quantitative tools from risk management.
- Large losses on the financial market are mainly due to the absence of proper risk control.



Sensitivity analysis

- It is a technique of studying the behaviour of a model by changing the input setup. i.e., it is to observe the changes in the output for a small change in the input.

- Change in the input may be with one parameter or with multiple parameters. i.e. the changes in the input setup may include, changes in the coefficients of the objective function or in the input constraints.

Casual learning

- Casual learning can be better explained using casual maps. A casual map is explained in the following example.

- Example of a casual map:

- Let X, Y and Z be three random variables we can draw two simple casual graphs as follows:

- Graph 1: $Z \rightarrow X \rightarrow Y$ and Graph 2 : $X \leftarrow Z \rightarrow Y$

- Graph 1 implies that Y is depending on the independent random variable X, and X in turn is depending on Z. Graph 2 implies that X and Y are both depending on Z separately.

- The information learn is in case of Graph 1, Y occurs whenever X occurs regardless of the occurrence of Z. where as, in case of Graph 2, Y occurs whenever Z occurs regardless of occurrence of X.