Chapter 2:

Prediction methods and models for BI

[all topics in scanned copy “Adaptive business Intelligence” by ZbigniewMichlewicz martin Schmidt]

Chapter 2:
Prediction methods and models for BI

- Data preparation
- Prediction methods-Mathematical method, Distance methods, Logic method, heuristic method-local optimization technique, stochastic hill climber
- Evaluation of models

Chapter 2
Data Preprocessing

To make data more suitable for data mining.

To improve the data mining analysis with respect to time, cost and quality.

Prediction methods

Quantitative methods

- The Quantitative methods assume that sufficient amount of data exists about the past and this data can be quantified in form of numerical data and past patterns will continue in the future.
- amount of stored data
  - No. of cases
  - No. of variables

- more data the better data mining can produce better results when performed on large datasets, and the resulting prediction model are more accurate.

Qualitative methods

applied in situations where very little quantitative data is available but where sufficient qualitative knowledge exists

The process of building a prediction model usually consists of following steps:

- Data preparation:
  - "garbage in, garbage out“
  - data transformation
  - normalization
  - creation of derived attributes
  - variable selection
  - elimination of noisy data
  - supplying missing values
  - data cleaning (80% of data mining effort)
- Model building :
  - Complete analysis of data
  - Selection of best prediction methods
    - Explaing the variability of question
    - Producing consistent result
- Deployment and evaluation
  - Implementing the best prediction model
  - Applying it to new data for prediction
  - As new data arrives measure the prediction models performance and tune it accordingly

Data preparation:

- Data preparation:
  - "garbage in, garbage out“
  - data transformation
    - DOB converted to age
    - Code given to nominal data
    - Imposing a natural order very light ,light , medium, heavy, very heavy
  - normalization
  - creation of derived attributes
  - variable selection
  - elimination of noisy data
  - supplying missing values
  - data cleaning (80% of data mining effort)

Why Data Preprocessing?

- Data in the real world is dirty
  - incomplete: missing attribute values, lack of certain attributes of interest, or containing only aggregate data
    - e.g., occupation=“”
  - noisy: containing errors or outliers
    - e.g., Salary=“-10”
  - inconsistent: containing discrepancies in codes or names
    - e.g., Age=“42” Birthday=“03/07/1997”
    - e.g., Was rating “1,2,3”, now rating “A, B, C”
    - e.g., discrepancy between duplicate records

Why Is Data Preprocessing Important?

- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
    - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
- Data preparation, cleaning, and transformation comprises the majority of the work in a data mining application (90%).

Major Tasks in Data Preprocessing

- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers and noisy data, and resolve inconsistencies
- Data integration
  - Integration of multiple databases, or files
- Data transformation
  - Normalization and aggregation
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results
- Data discretization (for numerical data)

Data Cleaning

- Importance
  - "Data cleaning is the number one problem in data warehousing"
- Data cleaning tasks – this routine attempts to
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data
  - Resolve redundancy caused by data integration

Missing Data

- Data is not always available
  - E.g., many tuples have no recorded values for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data

How to Handle Missing Data?

1. Ignore the tuple
   - Class label is missing (classification)
   - Not effective method unless several attributes missing values
2. Fill in missing values manually: tedious (time consuming) + infeasible (large db)?
3. Fill in it automatically with
   - a global constant : e.g., “unknown”, a new class?! (misunderstanding)

Cont'd

4. the attribute mean
   - Average income of *AllElectronics* customer $28,000 (use this value to replace)
5. The attribute mean for all samples belonging to the same class as the given tuple
6. the most probable value
   - determined with regression, inference-based such as Bayesian formula, decision tree. (most popular)

Noisy Data

- Noise: random error or variance in a measured variable.
- Incorrect attribute values may due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - etc
- Other data problems which requires data cleaning
  - duplicate records, incomplete data, inconsistent data

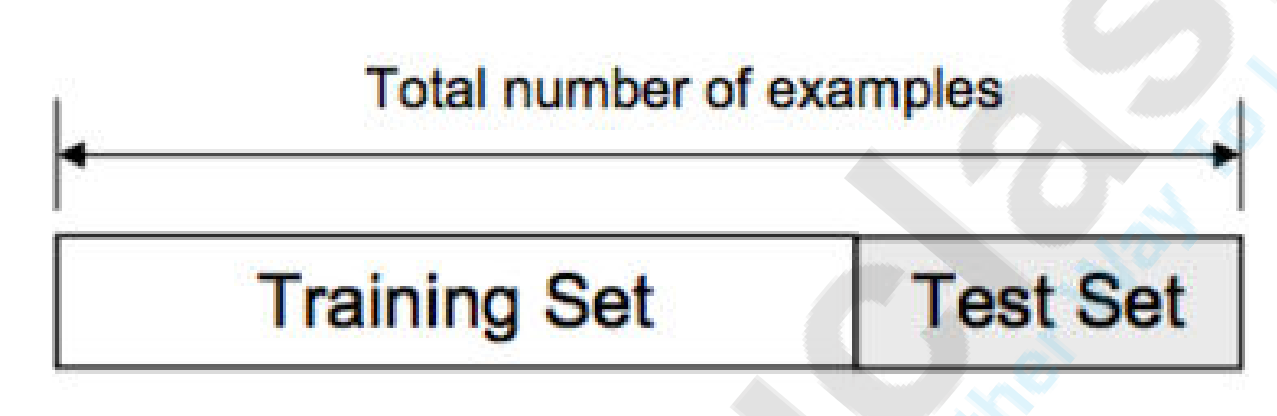
How to Handle Noisy Data?

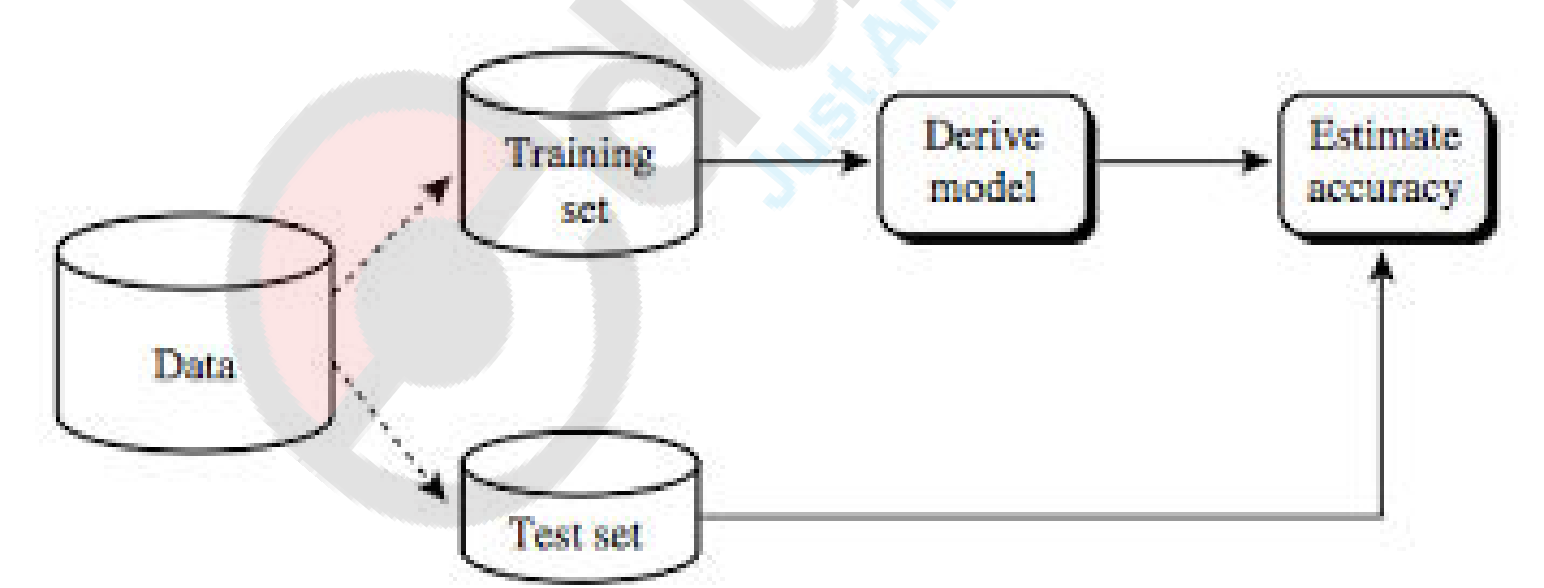
- Binning method:
  - first sort data and partition into (equi-depth) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Clustering
  - Similar values are organized into groups (clusters).
  - Values that fall outside of clusters considered outliers.
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)
- Regression
  - Data can be smoothed by fitting the data to a function such as with regression. (linear regression/multiple linear regression)

- Prediction methods
  - Mathematical method
  - Distance methods
  - Logic method
  - heuristic method
    - local optimization technique
    - stochastic hill climber
- Evaluation of models

Data mining

- Is about explaining the past and predicting the future by means of data analysis
- Is able to extract very valuable knowledge from this data.
- Involves
  - Statistics - collect , classify, summaries, interpret data
  - AI - study of computer algorithms, simulation of intelligent behavior to perform activities that need intelligence
  - Machine learning –improve automatically through experience
  - Databases –store, collect, manage data
  - Data warehousing – multidimensional reporting services in support of decision making process


Total number of examples
Training Set
Test Set

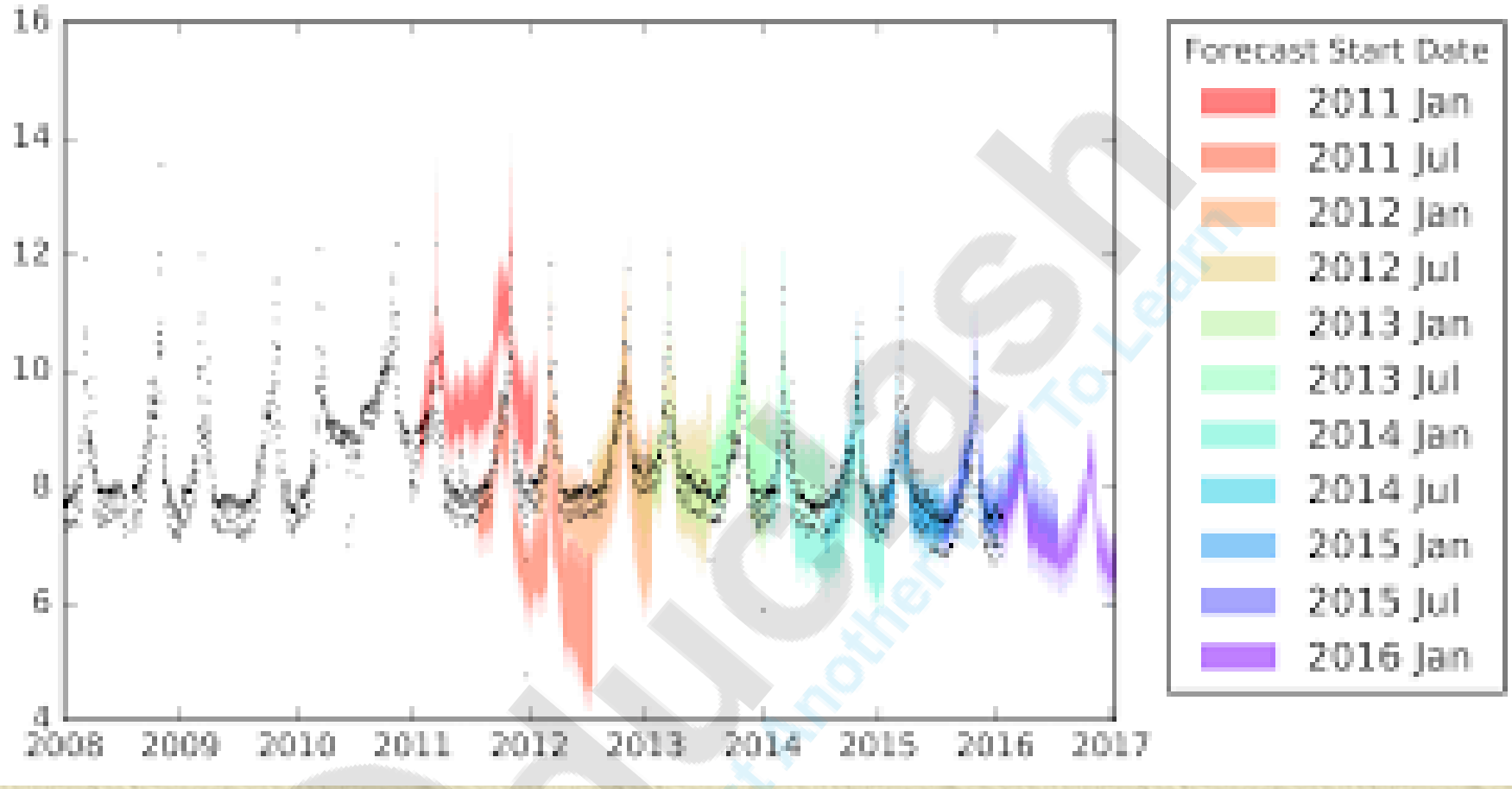

Training
set
Derive
model
Estimate
accuracy
Data
Test set

Prediction methods

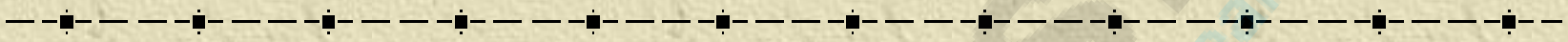
- After data is prepared we can begin our search for the right prediction method
- The goal is to build a prediction model that will predict the “outcome” of a new case.
- Eg.
  - Classification of a loan application
  - Prediction of price of a used car
  - Assignment of a new customer to an appropriate cluster
- Broad Categories
  - Mathematical method – linear regression, statistical method
  - Distance methods – instance based learning, clustering
  - Logic method – decision table decision trees, classification rules
  - heuristic method – neural network, evolutionary algo, fussy logic
    - local optimization technique
    - stochastic hill climber

Mathematical methods

- Regression
  - Output exhibits some explanatory relationship with some other variables
  - Eg, salary is a function of education, experience, industry & location
  - Linear regression
    - Saleprice=a+(b*mileage)+(c*year)+(d*color)….
    - Challenge is to find value of a,b,c,d that gives model best performance
  - Nonlinear regression
- Time series
  - goal is not to discover relation between variables but to purely predict
  - Eg. Neural network may not understand the connection between the weights yet provide quite accurate prediction
  - Data collected over time (generally equi distant)
  - Price of stock everyday, heart beat recorded every minute
  - Plotted data will have a pattern


16
14
12
10
8
6
4
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Forecast Start Date
2011 Jan
2011 Jul
2012 Jan
2012 Jul
2013 Jan
2013 Jul
2014 Jan
2014 Jul
2015 Jan
2015 Jul
2016 Jan

Distance methods

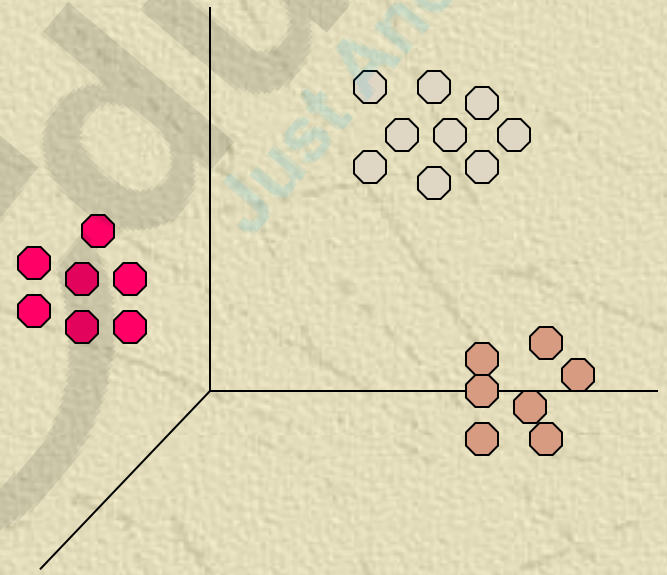


Clustering

- Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized



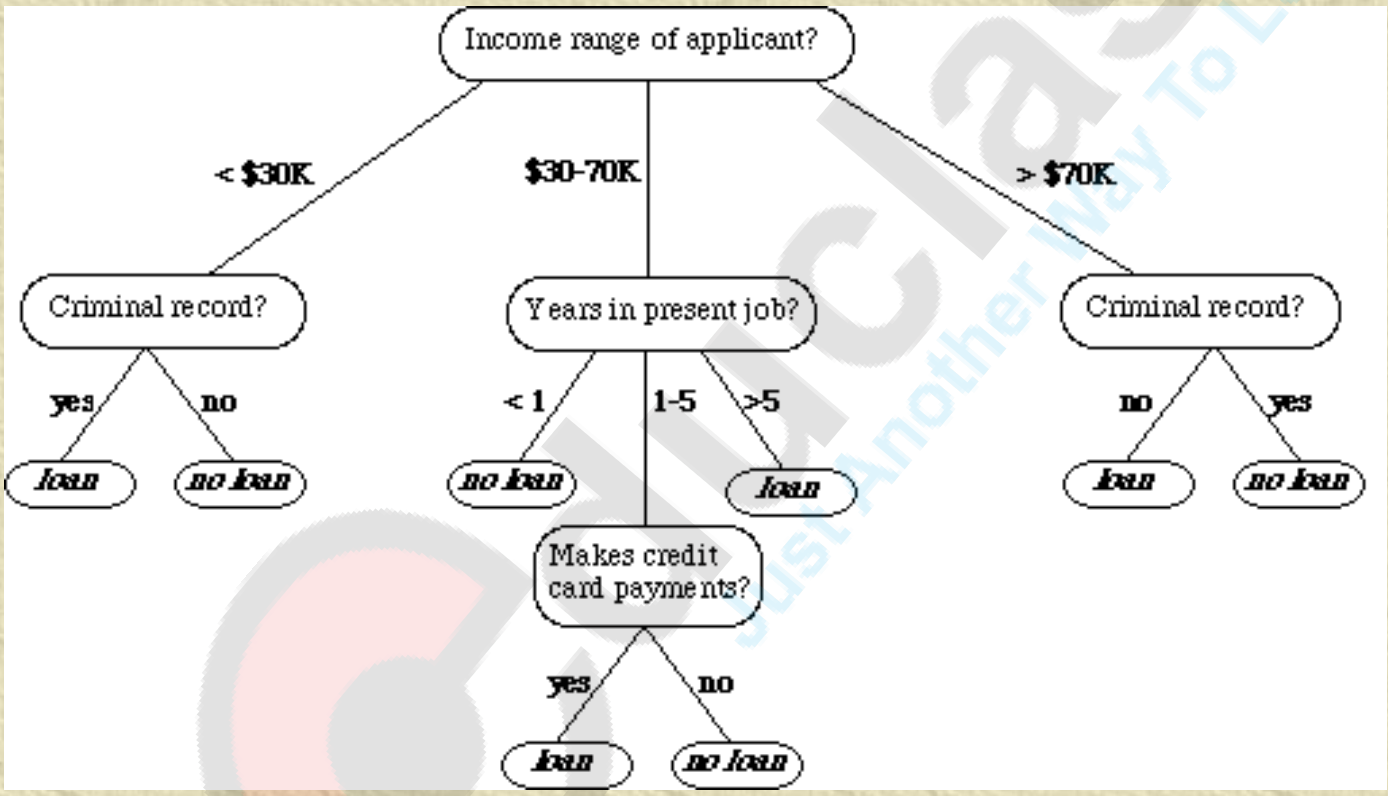
Clustering: Application 1

- Market Segmentation:
  - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - Approach:
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering: Application 2

- Document Clustering:
  - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
  - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
  - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

Logic method – decision table decision trees, classification rules

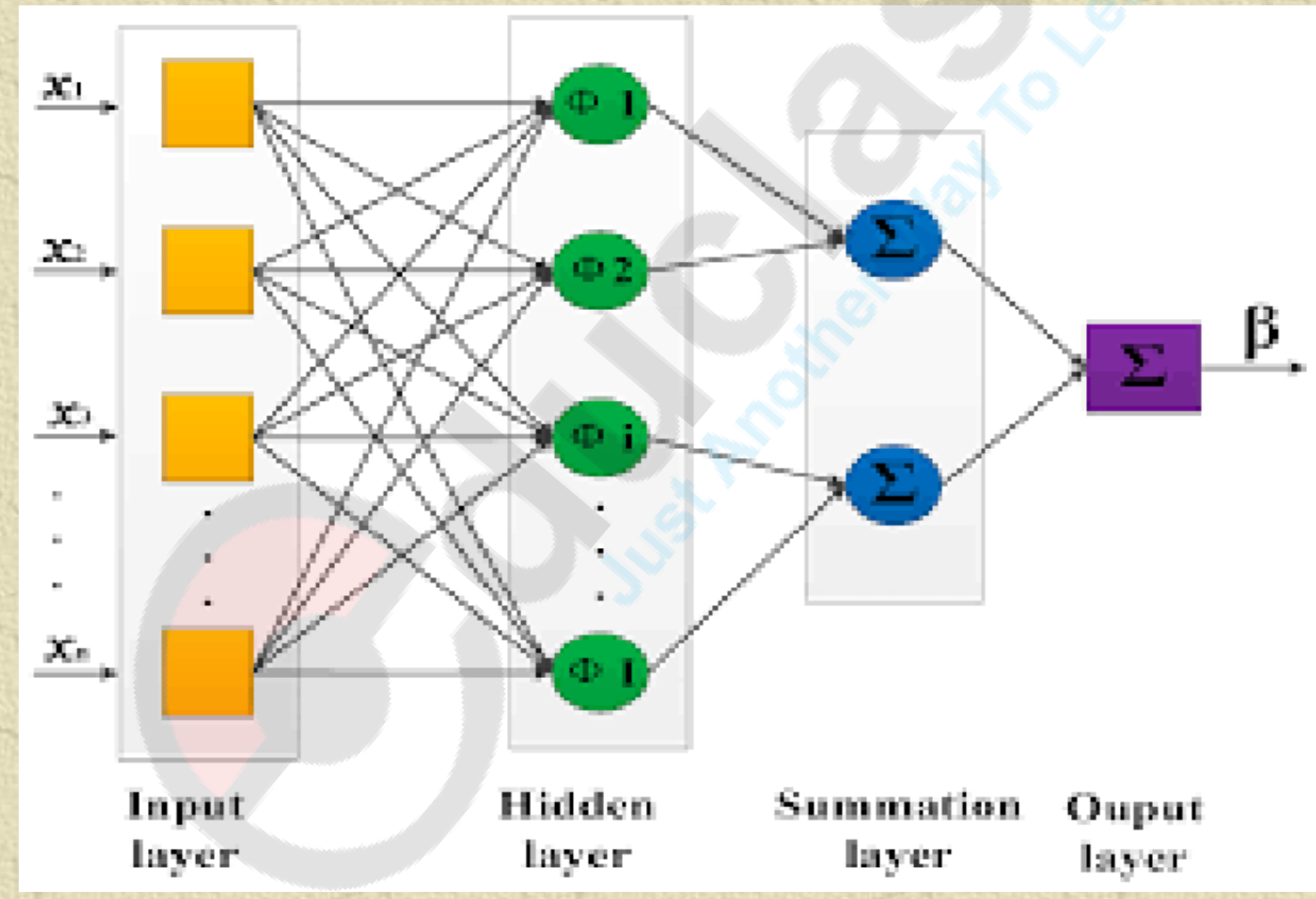


Logic method – decision table decision trees, classification rules

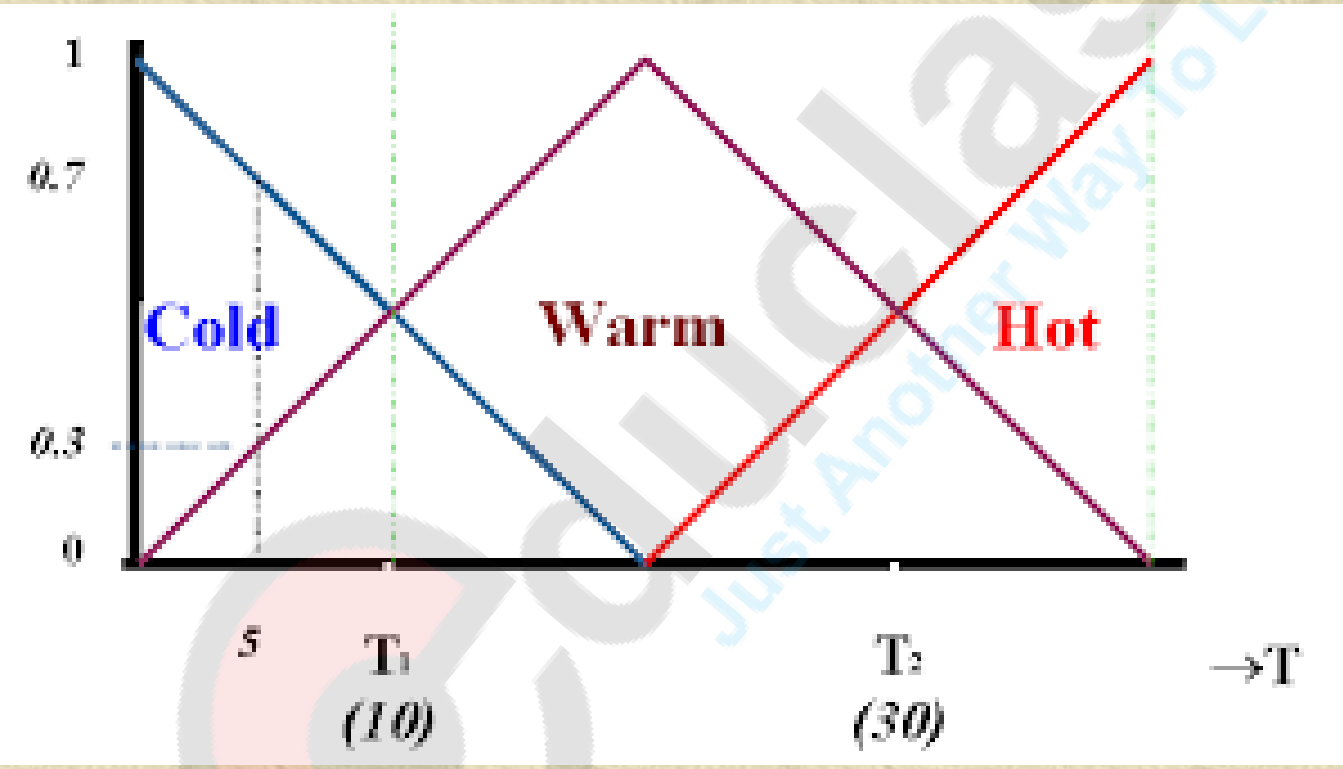
Rule-Based Classifier

- Classify records by using a collection of "if…then…" rules
- Rule: Represent the knowledge in the form of IF-THEN rules
  - (*Condition*) → *y*, where
    - *Condition* is a conjunctions of attributes
    - *y* is the class label
  - Examples of classification rules:
    - (Blood Type=Warm) ∧ (Lay Eggs=Yes) → Birds
    - (Taxable Income < 50K) ∧ (Refund=Yes) → Cheat=No
    - (rule antecedent or condition) → (rule consequent)

Heuristic method – neural network, evolutionary algo, fussy logic



Heuristic method – neural network, evolutionary algo, fussy logic



Heuristic

- A **heuristic** technique, often called simply a **heuristic**, is any approach to problem solving, learning, or discovery that employs a practical **method** not guaranteed to be optimal or perfect, but sufficient for the immediate goals.
- Heuristic method
  - local optimization technique
  - stochastic hill climber

Optimization

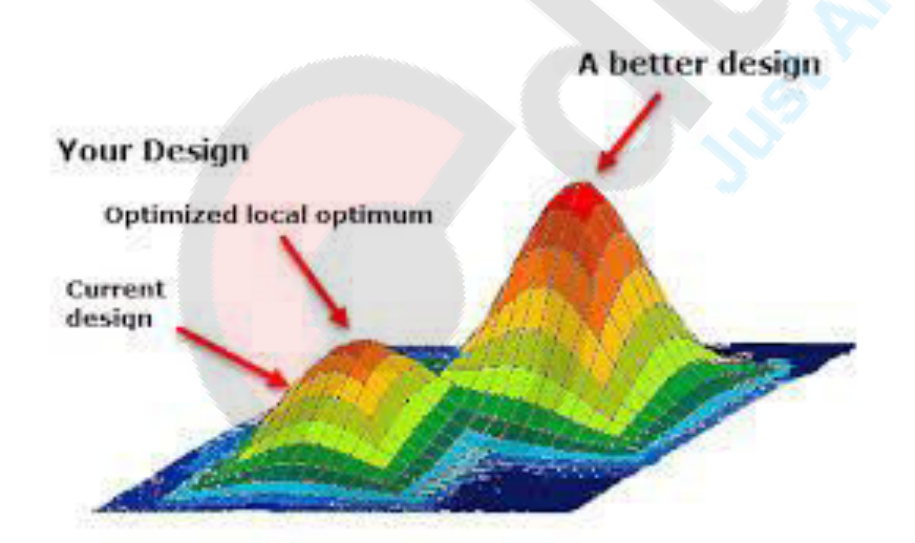
- Is the process of getting the best result under a given circumstances
  - Eg. Work done should be max in min time
  - Optimization can be finding max or min of a function.
- Three things always need to be specified before searching for any solution:
  - Representation of the solution : determines the search space and its size *(travelling sales man problem in Mumbai)*
  - Objective : task to be achieved *(minimize total distance of the route)*
  - Evaluation function: allows you to compare the quality of different solution *(distance is the evaluation function)*

Global optimum

- The goal is to find a solution that is feasible and better than any other solution present in the entire search space.
- The solution that satisfies these two conditions is called a global optimum.
- Finding global optimum is difficult , a much easier approach is to search the neighborhood of the that solution.

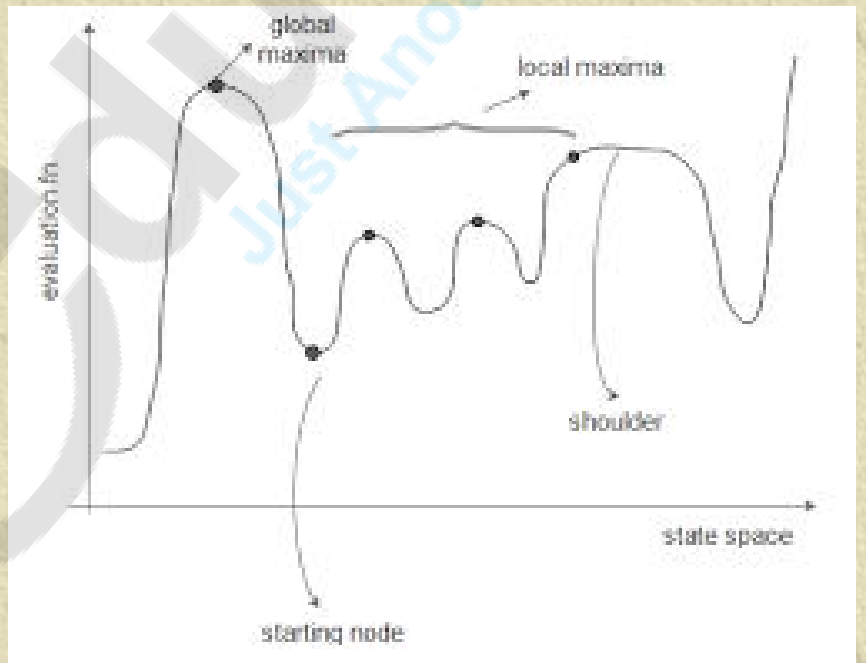
Local optimization technique

- Finding global optimum is difficult , a much easier approach is to search the neighborhood of the that solution.
- The problem of finding a solution with the highest quality measure score is similar to searching for a peak in a foggy mountain range.



Hill climbing

- **Hill climbing**: it is a graph search algorithm where the current path is extended with a successor node which is closer to the solution than the end of the current path.

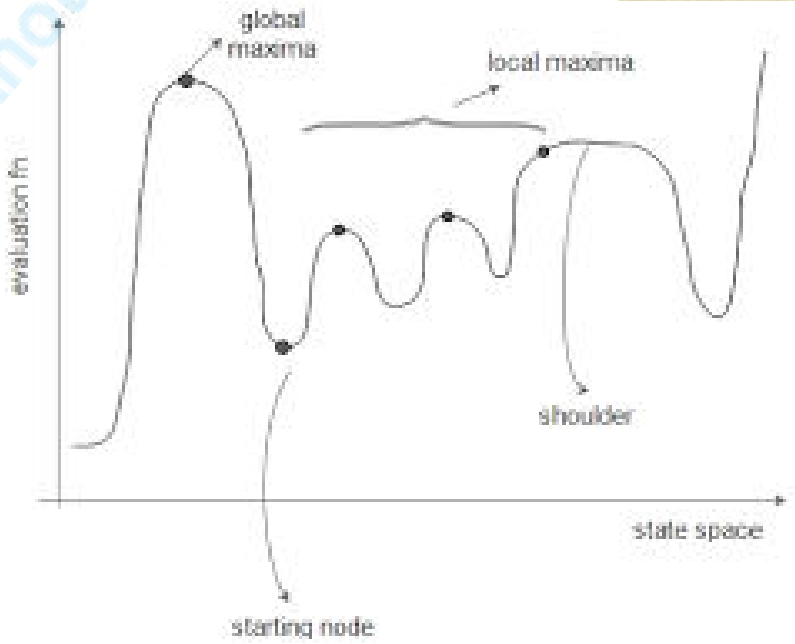


Hill climbing

- In **simple hill climbing**, the first closer node is chosen whereas in **steepest ascent hill climbing** all successors are compared and the closest to the solution is chosen.
- Both forms fail if there is no closer node. This may happen if there are local maxima in the search space which are not solutions.
- Hill climbing is used widely in artificial intelligence fields, for reaching a goal state from a starting node.
- Choice of next node/ starting node can be varied to give a number of related algorithms.

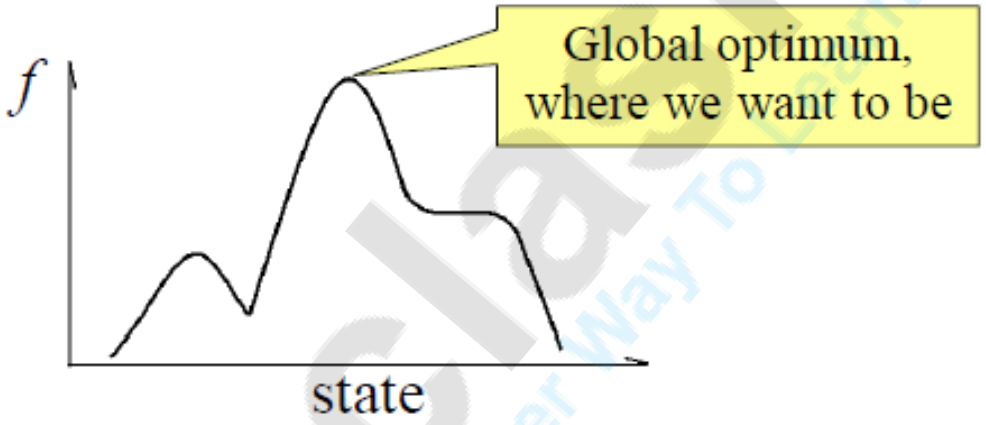
Hill Climbing Algorithm

1. Pick a random point in the search space
2. Consider all the neighbours of the current state
3. Choose the neighbour with the best quality and move to that state
4. Repeat 2 through 4 until all the neighbouring states are of lower quality
5. Return the current state as the solution state

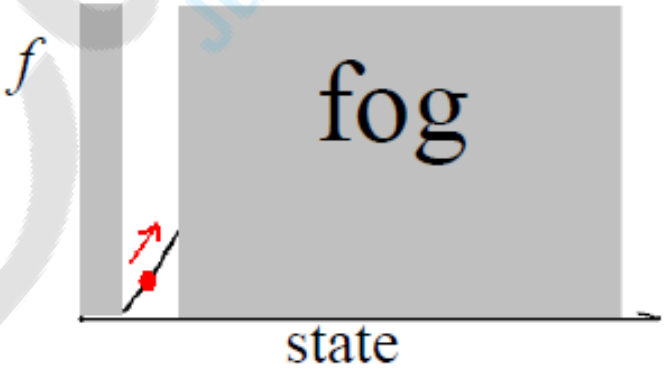


Local optima in hill climbing

- Useful conceptual picture: *f* surface = 'hills' in state space

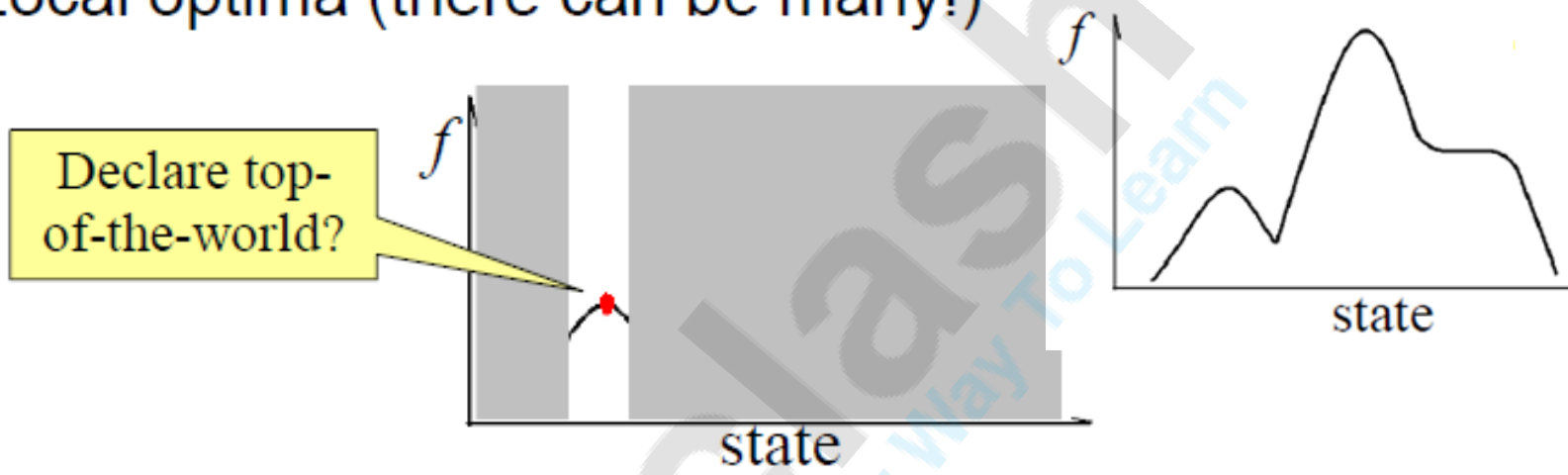


- But we can't see the landscape all at once. Only see the neighborhood. Climb in fog.

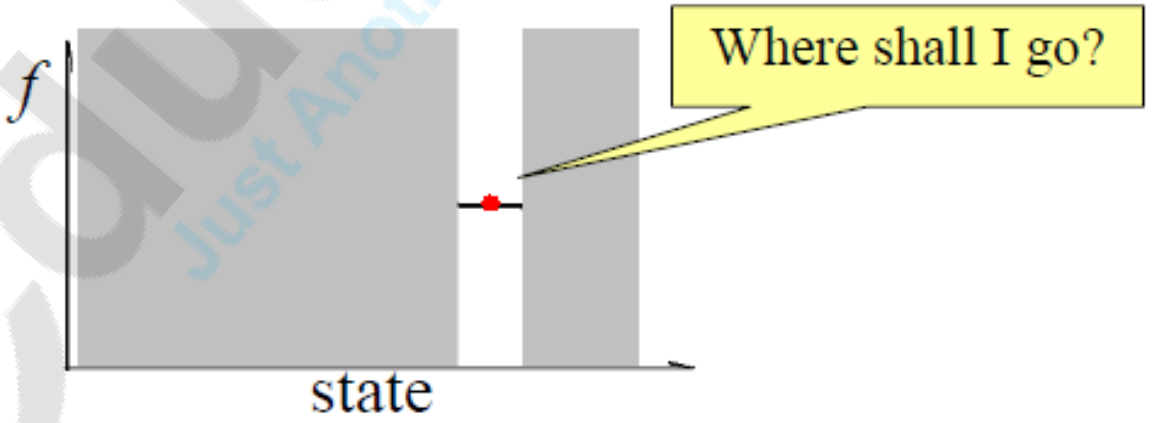


Local optima in hill climbing

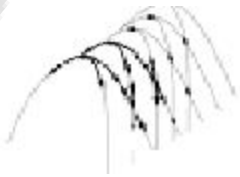
- Local optima (there can be many!)



- Plateaux



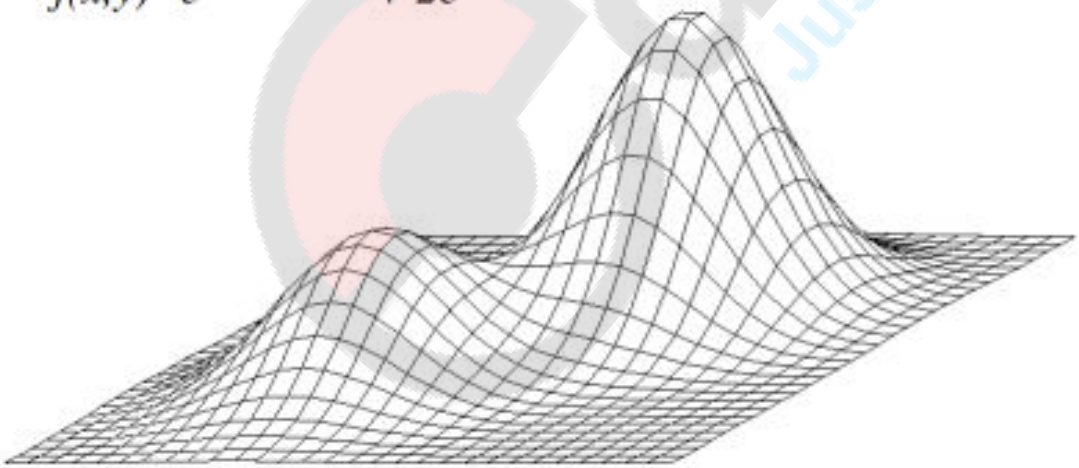
- Ridges

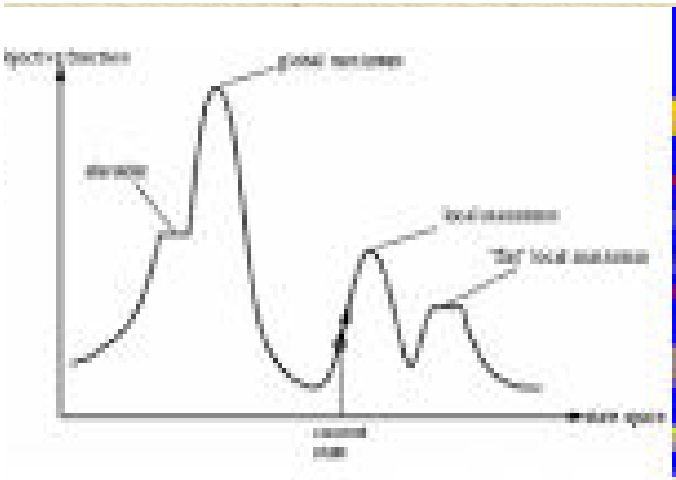


The Problem with Hill Climbing

- Gets stuck at local minima
- Possible solutions:
  - Try several runs, starting at different positions
  - Increase the size of the neighbourhood (e.g. in TSP try 3-opt rather than 2-opt)
  - Stochastic Hill-Climbing
    - Only one solution from neighbourhood is selected
    - This solution will be accepted for the next iteration with some probability, which depends from the difference between current solution and selected solution

$$f(x,y)=e^{-(x^2+y^2)} + 2e^{-((x-1.7)^2+(y-1.7)^2)}$$





Stochastic hill climber

- Proper choice is always dependent on the problem.
- May accept a inferior solution as a new current solution
- A new solution is accepted with some probability *p*.

Stochastic Hill Climbing

- The neighborhood of a current solution v_c consist from only one solution v_n
- The probability of acceptance of the solution v_n depends on:
  - Difference in merit between v_c and v_n
  - Parameter T

$$p = \frac{1}{1 + e^{\frac{eval(v_c) - eval(v_n)}{T}}}$$

- T remains constant during the execution of algorithm

Role of parameter T

- Example:
  - eval(v_c)=107, eval(v_n)=120
  - maximization problem

$$p = \frac{1}{1 + e^{\frac{-13}{T}}}$$

T	p
1	1.00
5	0.93
10	0.78
20	0.66
50	0.56
10^{10}	0.5…

Role of parameter T

$$p = \frac{1}{1 + e^{\frac{-13}{T}}}$$

T	p
1	1.00
5	0.93
10	0.78
20	0.66
50	0.56
10^{10}	0.5...

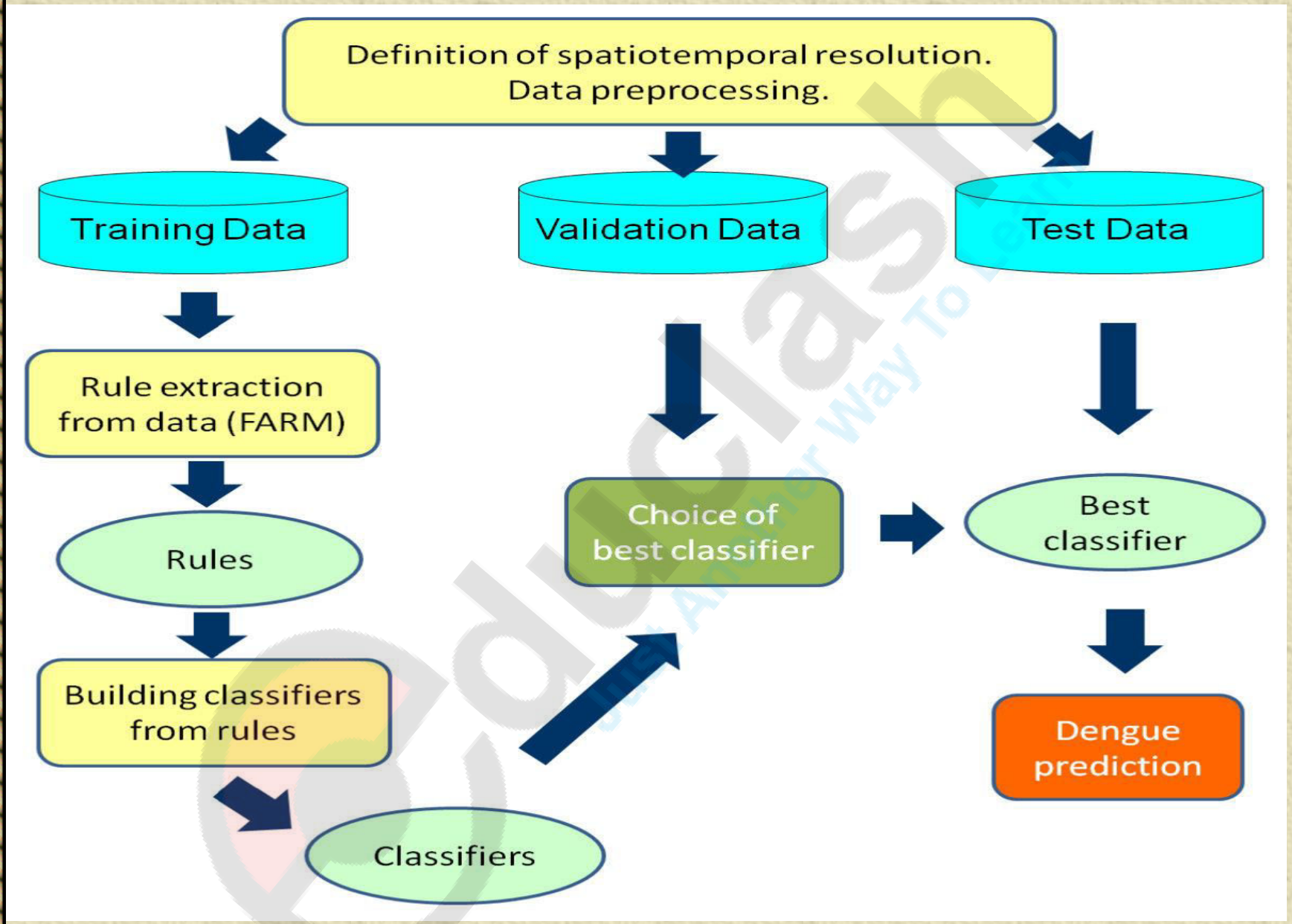
The greater the parameter T, the smaller the importance of the relative merit of the competing points v_c and v_n

- If T is huge -> search becomes random
- T is very small -> stochastic hill-climber reverts into ordinary hill climber

Evaluation of Models

- Varity of different prediction methods
- Which method should be applied to a particular problem?
- Therefore necessary to evaluate and compare different models
- Evaluation methodology
  - Fair & just
  - After we complete and train a few models we can test them on the data and measure the prediction error
  - Issues
    - The amount of available data might not be that large
    - Performance of a prediction model on the training data might be very different from the performance of the same model on an independent set of data
    - Prediction models that provide different outcomes requires different techniques for error measurement
    - Take into account cost of the potential error (consequences of error)

- Once a prediction model is created on the basis of the training data set, it can be fairly evaluated for performance on the test data set.
- Data set is split
  - Training set : to build the model
  - Validation set : tuning the parameters of the model
  - Test data set : evaluate the performance of the model


Definition of spatiotemporal resolution.
Data preprocessing.
Training Data
Validation Data
Test Data
Rule extraction from data (FARM)
Rules
Building classifiers from rules
Classifiers
Choice of best classifier
Best classifier
Dengue prediction