Pattern Recognition
Course Code: CSE423

By
Sudarshan Nandy, Ph. D(Engg. & Tech.),
Department of Computer Science and Engineering
Amity School of Engineering and Technology (ASETK),
Amity University Kolkata, West Bengal, India.

Syllabus

Module I : Introduction

Basics of pattern recognition, Design principles of pattern recognition system, Learning and adaptation, Pattern recognition approaches, Mathematical foundations – Linear algebra, Probability Theory, Expectation, mean and covariance, Normal distribution, multivariate normal densities, Chi squared test.

Module II: ***Statistical Patten Recognition:***

Bayesian Decision Theory, Classifiers, Normal density and discriminant functions

Module III : Parameter estimation methods:

Maximum-Likelihood estimation, Bayesian Parameter estimation, Dimension reduction methods - Principal Component Analysis (PCA), Fisher Linear discriminant analysis, Expectation-maximization (EM), Hidden Markov Models (HMM), Gaussian mixture models.

Module IV : Nonparametric Techniques:

Density Estimation, Parzen Windows, K-Nearest Neighbor Estimation, Nearest Neighbor Rule, Fuzzy classification.

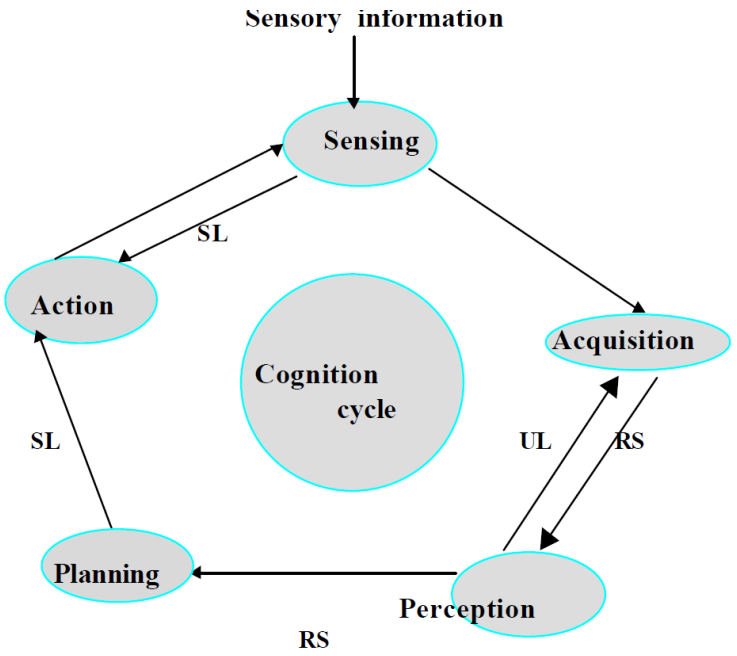
Module V: Unsupervised Learning & Clustering:

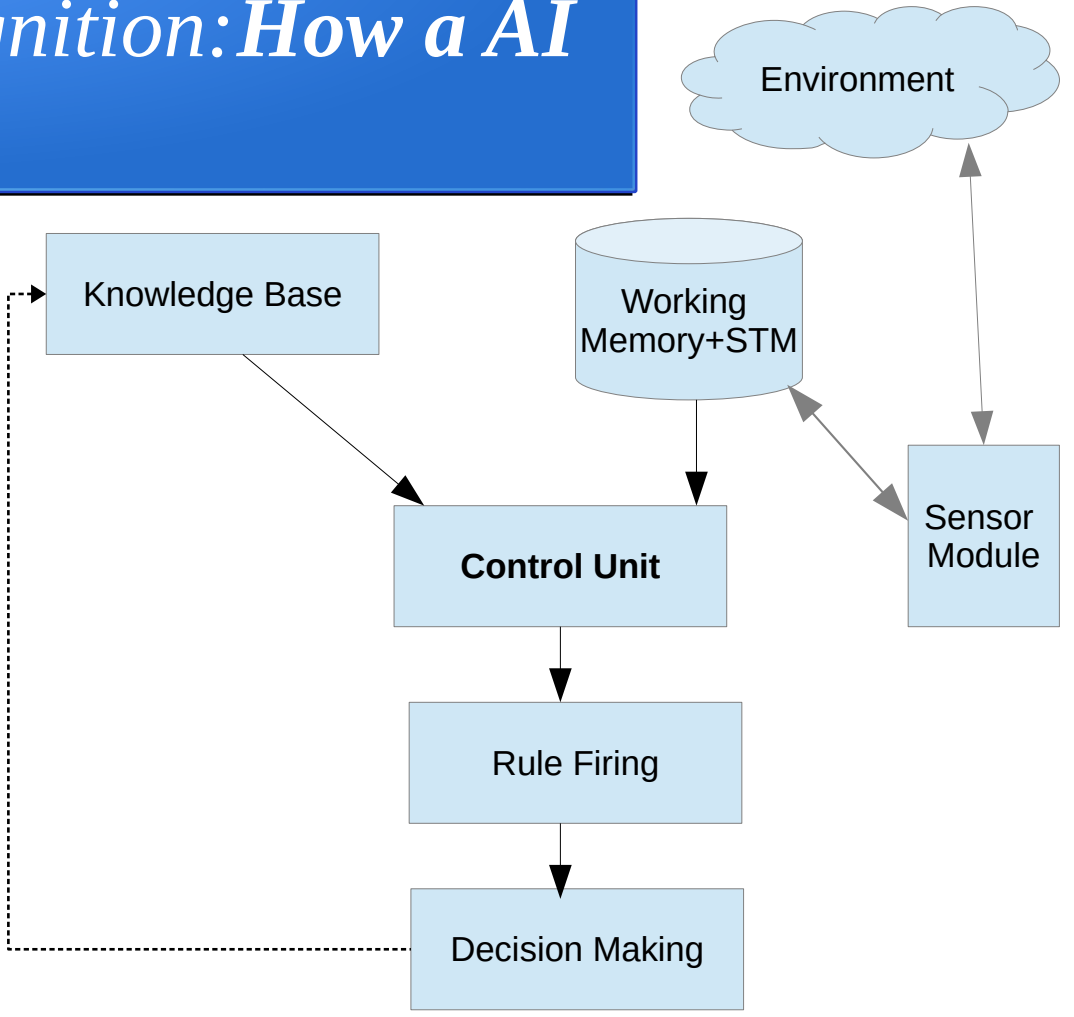
Criterion functions for clustering, Clustering Techniques: Iterative square - error partition clustering – K means, agglomerative hierarchical clustering, Cluster validation.

Module-I:*Basics of pattern recognition*

- **What is AI --> artificial intelligence?**
  -> Natural Language understanding
  ->Learning
  ->adaptive
- Artificial INTELLIGENCE:
  >System that acts like a human
  > System that thinks like a human
  >System that acts rational
  > System that think rational
- What is soft computing?
  Soft-computing : TOOLS ---> 4
  A . Fuzzy system ----> degree of truth
  B. Neural Network ----> a = f(net)=f(wp+b) non-linear model ----------> classify the patterns
  C. Genetic Algorithm
  D. Swarm Intelligence
- **How a AI system is created?**
- **What is Machine Learning?**
- **What is pattern Recognition?**

Basics of pattern recognition: ***How a AI system is created?***





PRDUCTION SYSTEM /EXPERT SYSTEM

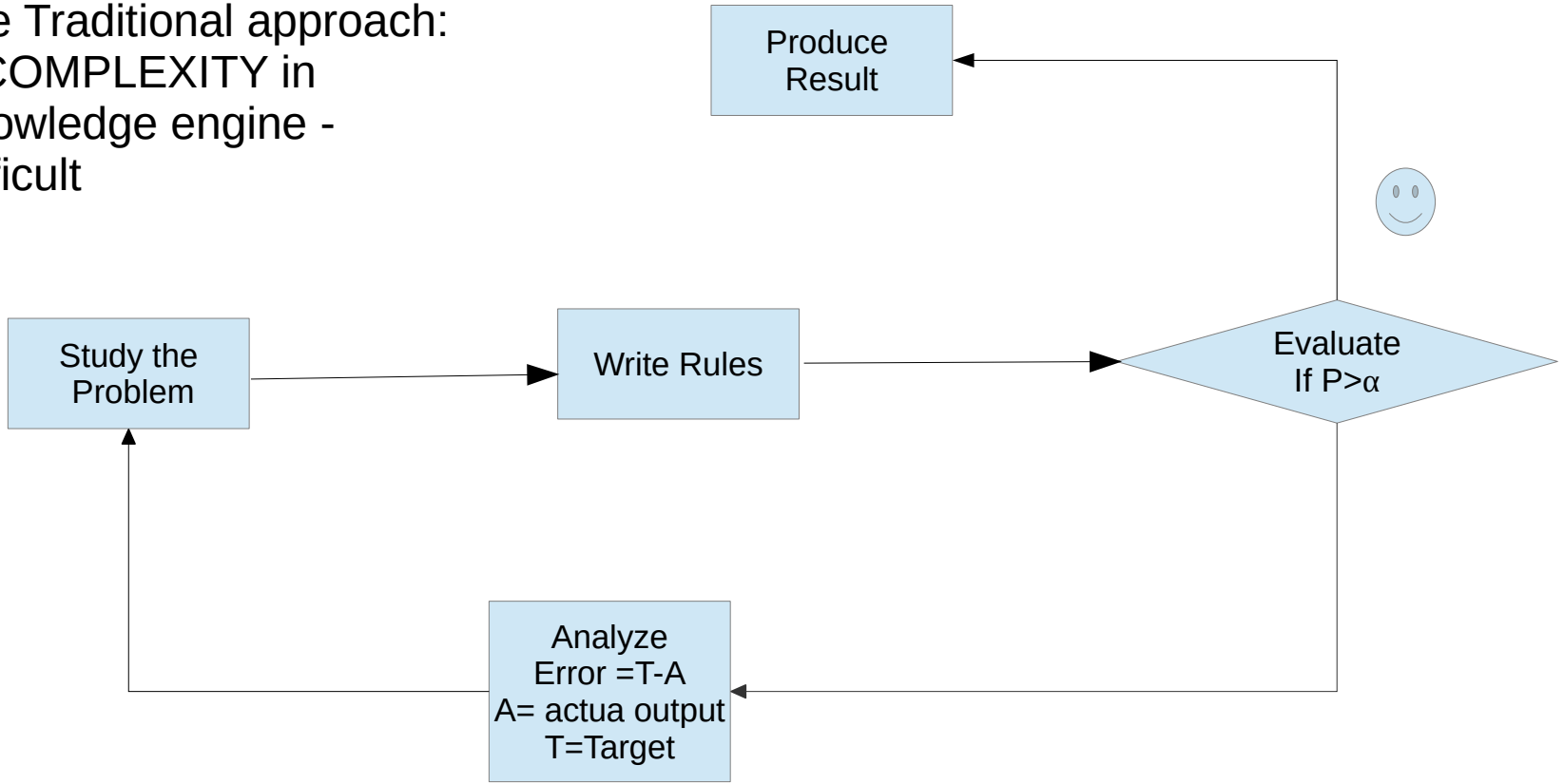
Basics of pattern recognition: ***What is Machine Learning?***

- Field of study that gives the computers the ability to learn without being explicitly programmed. *.....Aurther Samuel, 1959*
- A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

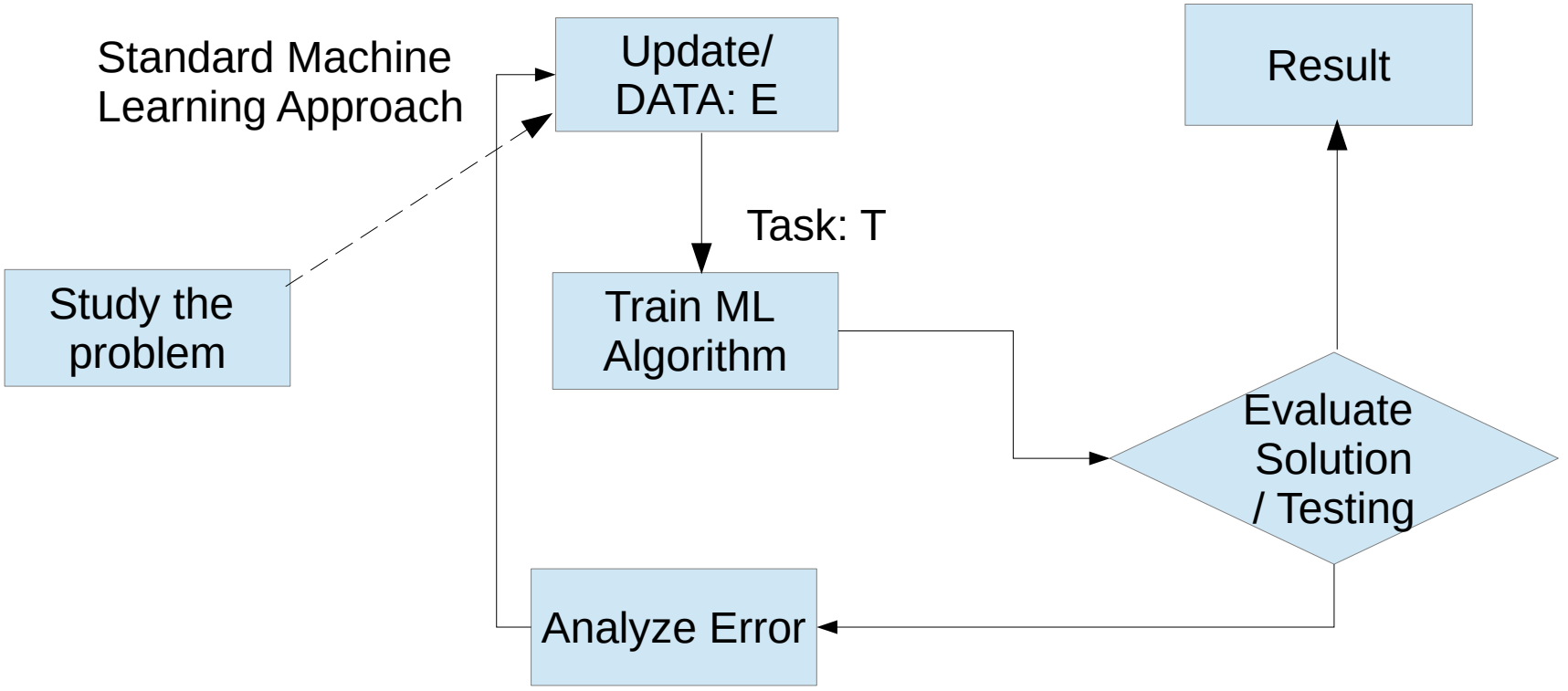
  Example: machine learning program that learns to flag spam.

Basics of pattern recognition: ***What is Machine Learning?***

The Traditional approach:
->COMPLEXITY in Knowledge engine - Difficult



Basics of pattern recognition: ***What is Machine Learning?***



Basics of pattern recognition: ***What is Machine Learning?***

Process(traditional approach):

1. Identify of the phrase like: “Credit Card”, “Job”, “4u”, “free shop”, ...

2. An algorithm of rules which will identify the patterns and flag emails as a spam

3. Testing of the result.

Basics of pattern recognition: ***What is Machine Learning?***

- An explicit algorithm to find the solution for a problem which requires lot of hand tuning and a long list of rules
- Machine Learning approach can adapt to a new data
- Process complex data and provides an insight of the complex problem.
- Mainly categorised into three types:

  >Supervised algorithm

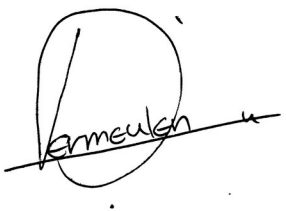
  >Un-supervised agorithm

  >Reinforcement algorithm

• *Basics of pattern recognition:* ***What is pattern Recognition?***

- Automatic (machine) recognition, description, classification, and grouping of patterns are important problems in a variety of engineering and scientific disciplines such as biology, psychology, medicine, marketing, computer vision, artificial intelligence, and remote sensing.
- The more relevant patterns at your disposal, the better your decisions will be. This is hopeful news to proponents of artificial intelligence, since computers can surely be taught to recognize patterns.

Genuines Signature



Forgeries

What is pattern?

- The best pattern recognizers in most instances are humans, yet we do not understand how humans recognize patterns.

P1: (4,1) - > snandy

P2: (4.5, 0.5) -> xyz

	1st	2nd
Line	4	4.5
Curve	1	0.5

- as opposite of a chaos; it is an entity, vaguely defined, that could be given a name.

What is pattern?

- Given a pattern, its recognition/classification may consist of one of the following two tasks

  1) supervised

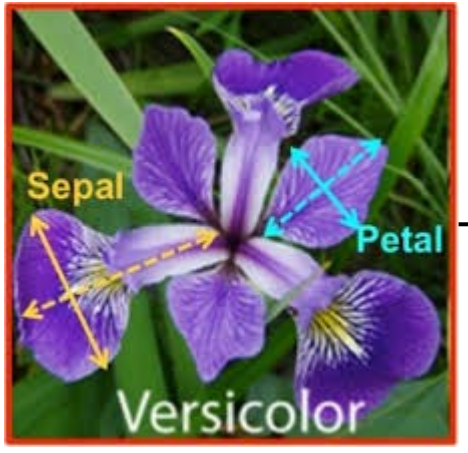
     classification (e.g., discriminant analysis) in which the input pattern is identified as a member of a predefined class,

  2) unsupervised classification (e.g., clustering)

     in which the pattern is assigned to an unknown number of classes.

- Note that the recognition problem here is being posed as a classification or categorization task, where the classes are either defined by the system designer (in supervised classification) or are learned based on the similarity of patterns (in unsupervised classification).

Basics of pattern recognition: Creating pattern

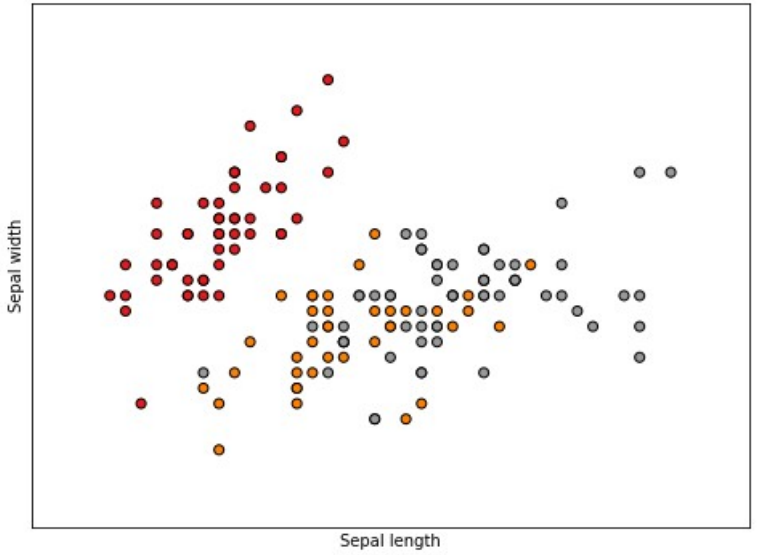


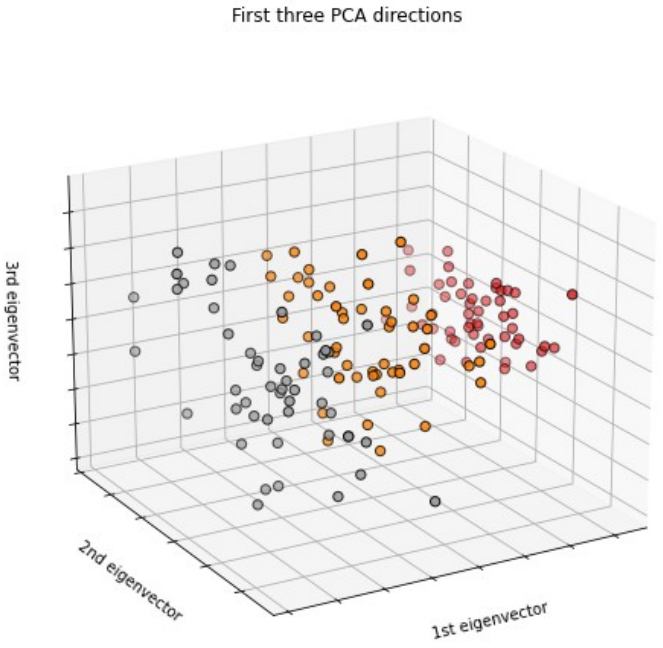
p1->6.2,2.9,4.3,1.3,Iris-versicolor
P2->5.7,2.8,4.1,1.3,Iris-versicolor
7.1,3.0,5.9,2.1,Iris-virginica
7.6,3.0,6.6,2.1,Iris-virginica
5.8,4.0,1.2,0.2,Iris-setosa
5.4,3.9,1.3,0.4,Iris-setosa

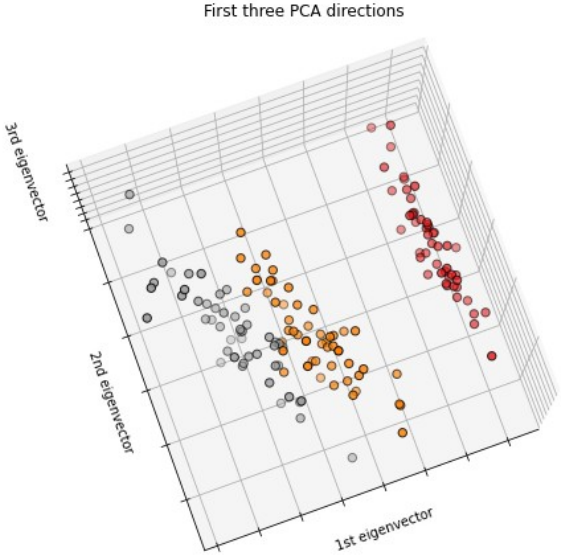
Dataset

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm

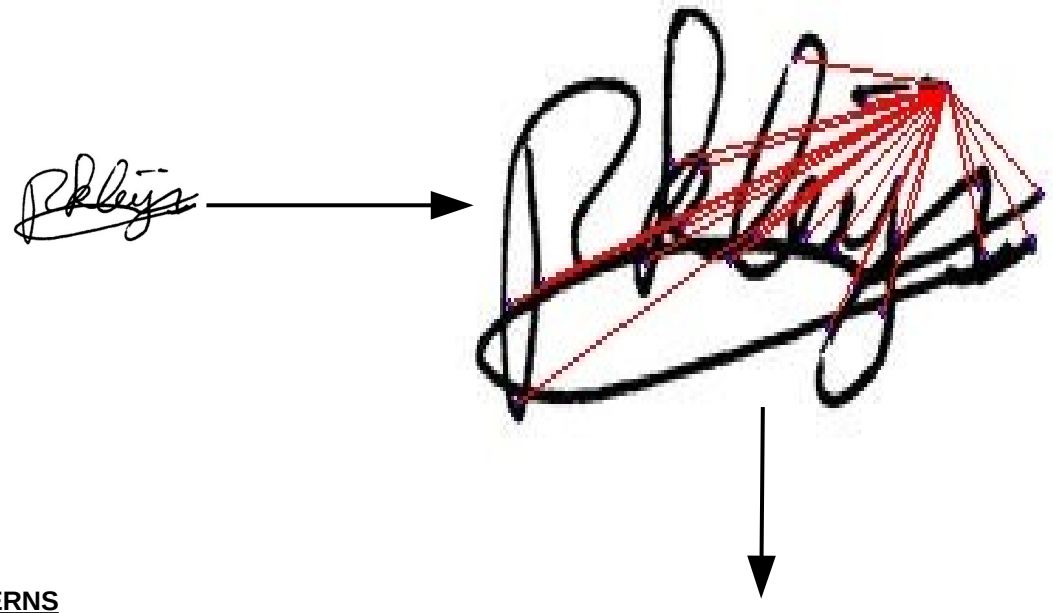
Basics of pattern recognition: Creating pattern







Basics of pattern recognition: Creating pattern



PATTERNS

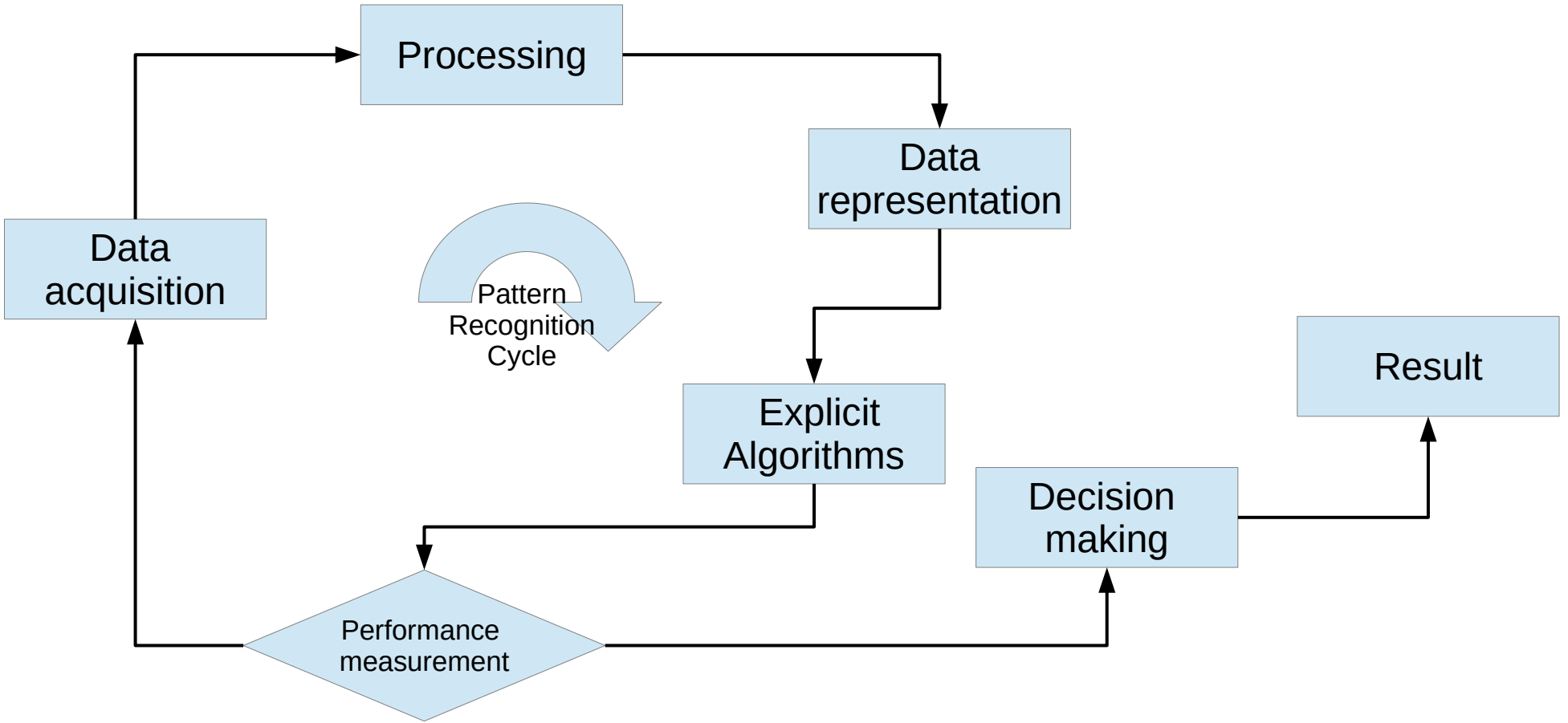
P1 : 9.6005,16.322,16.2984,26.4433356444,18.1024860171,18.6935817863,5.94138031101,29.6204321373,29.3393933134,10.8226614102 ---- > Class1
P2: 4.4102,11.4978,8.0231,38.815589,26.742,16.613849,32.9479893165,29.8303536687,4.15932686862,37.6750580623 ------> Class 1

Basics of pattern recognition: Applications

Examples of Pattern Recognition Applications

Problem Domain	Application	Input Pattern	Pattern Classes
Bioinformatics	Sequence analysis	DNA/Protein sequence	Known types of genes/ patterns
Data mining	Searching for meaningful patterns	Points in multi-dimensional space	Compact and well-separated clusters
Document classification	Internet search	Text document	Semantic categories (e.g., business, sports, etc.)
Document image analysis	Reading machine for the blind	Document image	Alphanumeric characters, words
Industrial automation	Printed circuit board inspection	Intensity or range image	Defective / non-defective nature of product
Multimedia database retrieval	Internet search	Video clip	Video genres (e.g., action, dialogue, etc.)
Biometric recognition	Personal identification	Face, iris, fingerprint	Authorized users for access control
Remote sensing	Forecasting crop yield	Multispectral image	Land use categories, growth pattern of crops
Speech recognition	Telephone directory enquiry without operator assistance	Speech waveform	Spoken words

Basics of pattern recognition: Design of pattern recognition system

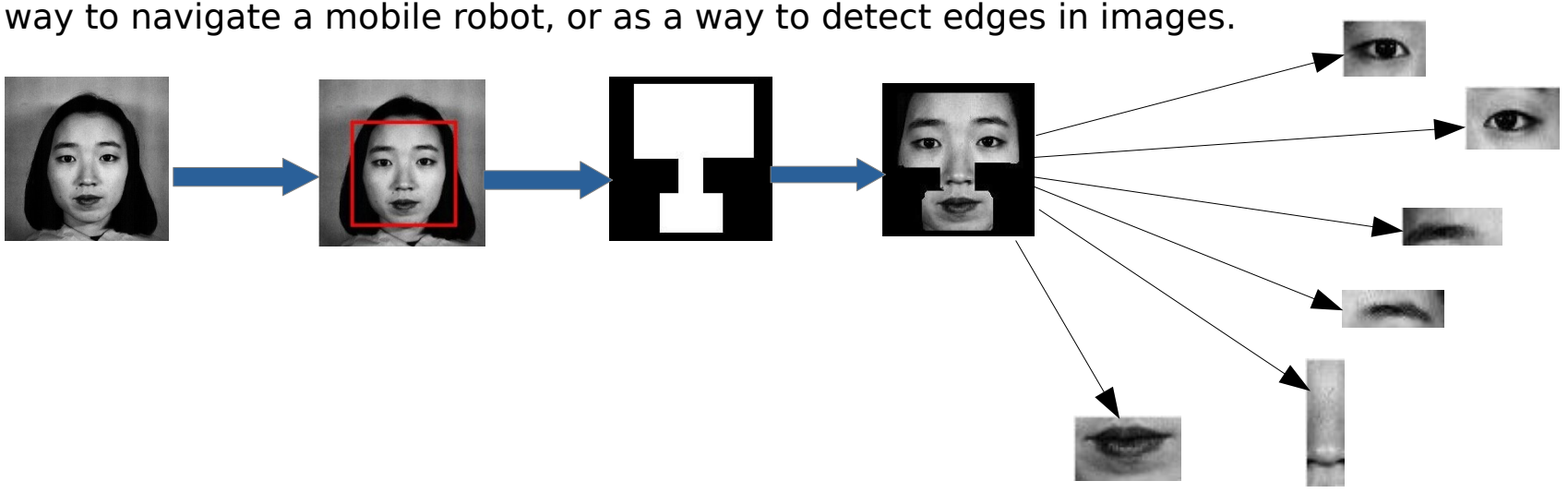


Basics of pattern recognition: Different Types

The four best known approaches for pattern recognition are:

1) Template matching:

Template matching is a technique in digital image processing for finding small parts of an image which match a template image. It can be used in manufacturing as a part of quality control, a way to navigate a mobile robot, or as a way to detect edges in images.



Basics of pattern recognition: Different Types

- 2) Statistical Approach:

  In statistics, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.

- 3) Syntactic or structural matching:

  In many recognition problems involving complex patterns, it is more appropriate to adopt a hierarchical perspective where a pattern is viewed as being composed of simple subpatterns which are themselves built from yet simpler subpatterns.

- 4) Neural networks:

  A non-linear model help us to classify the dataset.

Basics of pattern recognition: Different Types

Pattern Recognition Models

Approach	Representation	Recognition Function	Typical Criterion
Template matching	Samples, pixels, curves	Correlation, distance measure	Classification error
Statistical	Features	Discriminant function	Classification error
Syntactic or structural	Primitives	Rules, grammar	Acceptance error
Neural networks	Samples, pixels, features	Network function	Mean square error

Statistics in Pattern

- Population and sample

  mean, standard deviation, and variance
- Standard normal Distribution
- Gussian normal distribution
- Probability density Function

Statistics in Pattern: population and Sample

- Population: is set of similar items or events which we want to study for our experience or experiment.

  Population --- N

- Sample: is a selection specific group or the subset of individuals within a population to evaluate the characteritics of the whole population.

  Sample --- n

- Types of Sampling:
  - Simple random Sample
  - Stratified Sample
  - Systematic sampling

Statistics in Pattern: population and Sample

- Simple random Sampling: every element in the population (X) has an equal oppertunity to get selected for the sample(r).
- Stratified sampling: When the population embraces a number of distinct categories, the frame can be organized by these categories into separate "strata." Each stratum is then sampled as an independent sub-population, out of which individual elements can be randomly selected
- Systematic Sampling: It is a sampling which is performed on the basis of some order and then selection of individual or elements based on that order.

Statistics in Pattern: Arithmetic Mean

- Mean – sum of values divided by number of values
- Is the measurement of central tendency in dataset
- Cetral tendency is the measure of single value to represent that dataset.
- The measure of central tendency is performed with mean, median and mode.

  Formule to calculate Mean= $\mu = \frac{\sum x}{N}$

  Ex: consider a age group = [20, 21, 23,24]
- Then $\mu = \frac{20+21+23+20+24}{5} = 108/5 = 21.6$

- Population Mean: $\mu = \frac{\sum x}{N}$
- Sample Mean: $\overline{X} = \frac{\sum x}{n}$

Statistics in Pattern: Arithmetic Mean

- Median: it is the central or middle value of a list of numbers. But for this we need to arrange the values in an order then we can find the values which lies in the middle among them.

  Example: 11,11,12,12,12,13,14,15

  $$Median = \frac{12+12}{2} = 12$$

- Mode: highest frequecies in the occurance

  Example: 11,11,12,12,12,13,14,15 So, mode = 12

- If two highest occurance --> Bi-modal

  Example: 11, 11, 12, 12, 12,13,13,13,14,15

  So, Mode = 12, 13

- More then two highest Occurance ---- > Multi-modal

Statistics in Pattern: Arithmetic Mean

- What to use among mean, median, mode and when?

  Outlier --> is causing the problem in observing mean value.

  Outlier: is value or a group of two or three value which are very much numerically different from other values.

  Example = 10,10, 11,11,12,12,12,13,13

  Mean = (10+10+11+11+12+12+12+13+13) = 104/9 =11.56

  Median = (12+12)/2 =12

  Mode = 10,10, 11,11,12,12,12,13,13 = 12

  But if Example = 10,10, 11,11,12,12,12,13,13, 600

  Mean = (10+10+11+11+12+12+12+13+13+600) = 704/9 =70.4

  Median = (12+12)/2 =12

  Mode = 10,10, 11,11,12,12,12,13,13 = 12

Statistics in Pattern

Now consider an example:

- Item Set 1 = (10,10,11,11,12,12,12,13,13,13)

  mean1 = (10+10+11+11+12+12+12+13+13+13)/10 = 11.7
- Item set 2 = (1,2,3,4,4,5,5,31,42,20)

  mean2 = (1+2+3+4+5+31+42+20+5+4)/10 =11.7
- So it is very difficult for us to understand the difference between these two attributes of patterns.

Statistics in Pattern

- Understanding of difference or scatterness mesurment or the mesurment of dispersion is possible in statistics by means of:

  - Variance

  -Standard deviation

- Population variance

$$\sigma^2 = \frac{\sum (x-\mu)^2}{N}$$

- Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$$

Statistics in Pattern

- Sample variance

$$S^2 = \frac{\sum_{i=1}^{n} (x - \bar{x})}{(n-1)}$$

- Sample Standard Deviation

$$S = \sqrt{\frac{\sum_{i=1}^{n} (x - \bar{x})}{(n-1)}}$$

Statistics in Pattern- partition values (Percentile, quartiles)

- In statistics, a percentile is a score below which a given percentage of scores in its frequency distribution fall or a score at or below which a given percentage fall.
- The percentile (or percentile score) and the percentile rank are related terms.
- Example: 2,2,3,4,5,6,7,7,7,7,8,8,9,9,10,11,11,11,12,12,13

  what is the percentile score/rank of ‘11’?

  2,2,3,4,5,6,7,7,7,7,8,8,9,9,10,11,11,11,12,12,13

Statistics in Pattern- partition values (Percentile, quartiles)

what is the percentile score of ‘11’?

2,2,3,4,5,5,6,7,7,7,8,8,9,9,10,10,11,11,11,12

$$Percentile\,score\,of\,X = \frac{no\,of\,value\,below\,X}{total\,number\,of\,values} * 100$$

$$Percentile\,score\,of\,11 = \frac{16}{20} * 100 = 80\,\%$$

What value exists at percentile score of 25% or 50% or 75%?

$$value = \frac{Percentile}{100}(n+1) \qquad value = \frac{25}{100}(20+1) = \frac{21}{4} = 5.25$$

So, values lies in between 5th and 6th number and hence,

$$\frac{(5+5)}{2} = 5$$

Statistics in Pattern- partition values (Percentile, quartiles)

Quartiles: a quartile is a type of quantile which divides the number of data points into four parts, or quarters, of more-or-less equal size.

- The first quartile (Q1) is defined as the middle number between the smallest number (minimum) and the median of the data set. It is also known as the lower or 25th empirical quartile, as 25% of the data is below this point.
- The second quartile (Q2) is the median of a data set; thus 50% of the data lies below this point.
- The third quartile (Q3) is the middle value between the median and the highest value (maximum) of the data set. It is known as the upper or 75th empirical quartile, as 75% of the data lies below this point.

Statistics in Pattern- partition values (Percentile, quartiles)

- We will usually calculate Q1 and Q3 because Q2 is median and in Q4 series is not divide

Q1: $Q1 = 1x\left(\frac{n+1}{4}\right)$ Example: Where Sample Size = 12 $(12+1)/4 = 13/4 = 3.25 ====>$ position

Q2: $Q2 = 2\left(\frac{n+1}{4}\right) = \frac{n+1}{2} = Median$

Q3: $Q3 = 3\left(\frac{n+1}{4}\right)$

Q4: $Q4 = 4\left(\frac{n+1}{4}\right) = n+1$

Statistics in Pattern- partition values (Percentile, quartiles)

- Find the Q1 and Q3 in the following series:

  7,5,6,3,9,2,4,8

  solution:- 2,3,4,5,6,7,8,9

  q1 = (n+1)/4 =(8+1)/4 = 9/4 = 2.25

Number	2	3	2.25	4	5	6	7	8	9
Position/ rank	1	2	Q1	3	4	5	6	7	8

Q1 = 2.25 = 2nd+0.25*(3rd -2nd) = 3+(0.25*1) =3.25

Q3 = 3(n+1)/4 = 3*2.25 = 6.75

Q3=6.75 = 6th+0.75*(7th-6th) = 7+0.75(8-7) =7.75

Statistics in Pattern- partition values (Percentile, quartiles)

- Find the two quartiles q1 and q3 from the following series:

  391,384,591,407,672,522,777,773,2488,1490

  solution: Q1 = (10+1)/4 =2.75 and Q3 = 3*2.75 = 8.25 ,Median = 11/2 = 5.5

- Q1= 2.75 = 2nd+0.75*(3rd - 2nd) = 391+0.75*(407-391)=391+12 = 403

Numbers	384	391	2.75	407	522	591	5.5	672	733	777	8.25	1490	2488
Position	1	2	q1	3	4	5	median	6	7	8	q3	9	10

Statistics in Pattern- partition values (Percentile, quartiles)

- Now, simple interpolation says,

x	y
2	391
x-->2.75	y---?
3	407

- $Q1 => \frac{\text{Partial difference in x}}{\text{Total differences in x}} = \frac{\text{Partial difference in y}}{\text{Total differences in y}} => \frac{(2.75-2)}{(3-2)} = \frac{(Q1-391)}{(407-391)} = 403$
- Q3 => 8.25 = 8th+0.25*(9th - 8th) =777+0.25*(1490-777)=955.25
- Q2 => 5.5=5th +0.5(6th -5th)=591+0.5*(672- 591)=631.5

Statistics in Pattern- partition values (Percentile, quartiles)

- So, what value we know from the previous Example?
- Quartile values = Q1, Q2, Q3
- Quartile range / Inter-quartile range = IQR
- Where, IQR = Q3 – Q1
- Benefit:

  -Help us to understand the outlier values

Statistics in Pattern- partition values (Percentile, quartiles)

- How to calculate the Outlier values Through IQR?

  Process: Example- 391,384,591,407,672,522,777,773,2488,1490

  -Q1 = 403, Q2=631.5, Q3 = 955.25

  - Calculate IQR= Q3 – Q1 = 552.25
  - Multiply = IQR *1.5 = 828.375
  - Lower limit = Q1 – (IQR*1.5) = -425.38
  - Upper Limit = Q3 + (IQR*1.5) = 955.25+552.25 = 1507.25
  - If any series value < lower limit then

    it is outlier

    If any series value > upper limit then

    It is outlier

    example: 391,384,591,407,672,522,777,773,2488,1490

Statistics in Pattern- partition values (Percentile, quartiles)

- Representing the data: BOX PLOT

  Understanding:

  - Understanding of the central tendencies through mean

  - Understanding of how the value is scattered or disperse

  - Understding the skewness

  - Normal distribution: the median should be in the center of the box

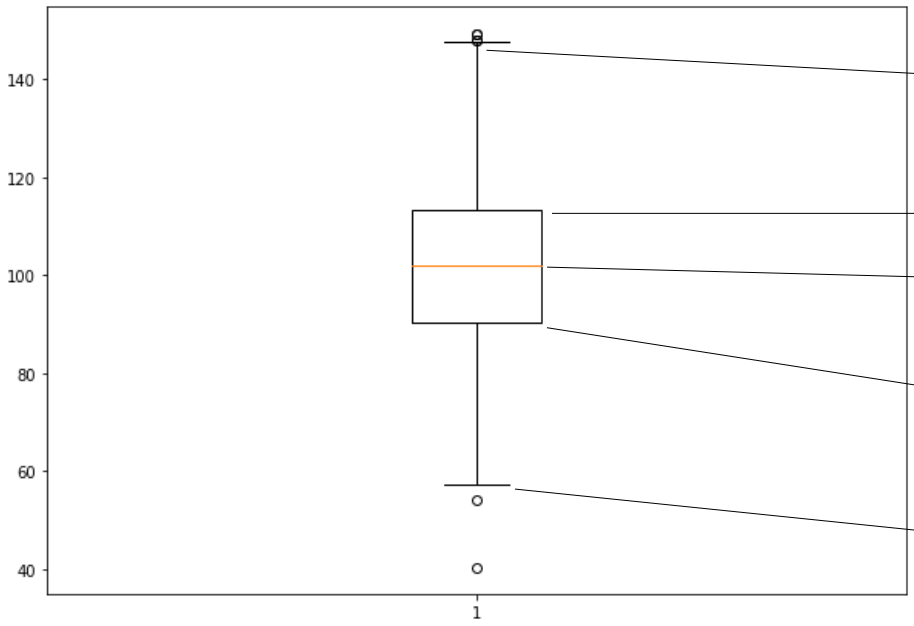
  -Right Skew: The right tail is longer; the mass of the distribution is concentrated on the left of the figure. The distribution is said to be right-skewed, right-tailed, or skewed to the right, despite the fact that the curve itself appears to be skewed or leaning to the left; right instead refers to the right tail being drawn out and, often, the mean being skewed to the right of a typical center of the data. A right-skewed distribution usually appears as a left-leaning curve

  > if the distance(median , Q1) < distance (median, Q3) and wishkers in lower end is shorter then +VE

  - Left Skew:The left tail is longer; the mass of the distribution is concentrated on the right of the figure. The distribution is said to be left-skewed, left-tailed, or skewed to the left, despite the fact that the curve itself appears to be skewed or leaning to the right; left instead refers to the left tail being drawn out and, often, the mean being skewed to the left of a typical center of the data. A left-skewed distribution usually appears as a right-leaning curve.

  > if the distance(median , Q1) > distance (median, Q3) and wishkers in upper end is shorter then -VE

Statistics in Pattern- partition values (Percentile, quartiles)



Statistics in Pattern- partition values (Percentile, quartiles)

- How to do the code ?

```
import matplotlib.pyplot as plt
import numpy as np
Data = [ ]
fig1 = plt.figure(figsize =(9, 6))
plt.boxplot(data)
plt.show()
```

Statistics in Pattern- partition values (Percentile, quartiles)

- Skewness is calculated as follows:
- $$Skewness = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)} = \frac{(Q_3 - 2Q_2 + Q_1)}{(Q_3 - Q_1)}$$
- Question: Find the suitable measure of Skewness from the following:

  range---->F-------->C.F.

  0-20 ----> 20 ----> 20

  20-50 ---->50 ---->20+50=70

  50-100 ---->69---->139

  100 -250 ---->30---->169

  250-500 ----> 25---->194

  500 – 1000 ---->19---->213
- N= Σf = (20+50+69+30+25+19)=213
- Q1 = N/4 = 213/4 = 53.25 ; Q2 = 213/2 =106.5; Q3 = 3*53.25 = 159.75

Statistics in Pattern- partition values (Percentile, quartiles)

- $Q1 = L_1 + (((N/4)-CF) / F)*I$
- $L_1=20$
- N/4 = 53.25
- F=50
- (Previous)CF = 20
- I = diff(20, 50) = 50-20 = 30

  Q1 = 20+(((53.25-20)/50)*30) = 39.95

  Q2 = 76.45

  Q3 = 203.75
- Skewness = (203.75-(2 x 76.45) +39.95) / (203.75-39.95) = +0.55