Introduction to Unsupervised Learning & k-means

Swarna Chaudhary
Asst. Professor
WILP Division, BITS-Pilani

BITS Pilani
Pilani Campus

Textbooks/Reference Books

Text Books

T1	Pattern Recognition and Machine Learning, Christopher M Bishop, 2006, Springer
T2	Data Mining: Concepts and Techniques, Third Edition by Jiawei Han and Micheline Kamber Morgan Kaufmann Publishers, 2006
T3	Artificial Intelligence - A Modern Approach, Russel & Norvig, Third Ed, PHI

Reference Book(s) & other resources

R1	Introduction to Machine Learning, Ethem Alpaydin, 3rd Ed, 2014, MIT Press

Modular Structure

No	Title of the Module
M1	Introduction
M2	K-Means Algorithm
M3	MoG and EM Algorithm
M4	Hierarchical Clustering
M5	Density Based Clustering
M6	Quality of Clustering
M7	Association Rule Mining
M8	HMM and Time Series Data

Unsupervised Learning

Learning from unlabelled data

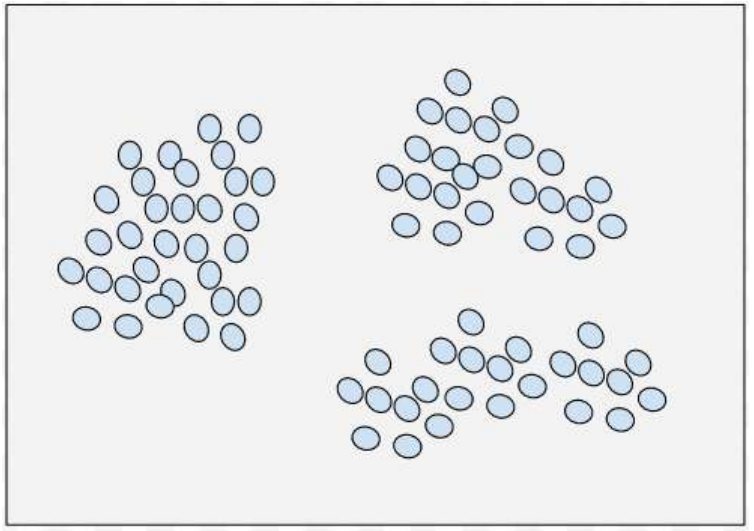
Let $X = \{x^{(1)}, x^{(2)}, x^{(3)}, \ldots, x^{(N)}\}$

- The points does not carry labels

Supervised vs. Unsupervised Learning

Objective:

- Find patterns / sub-groups among the the data points using data similarity

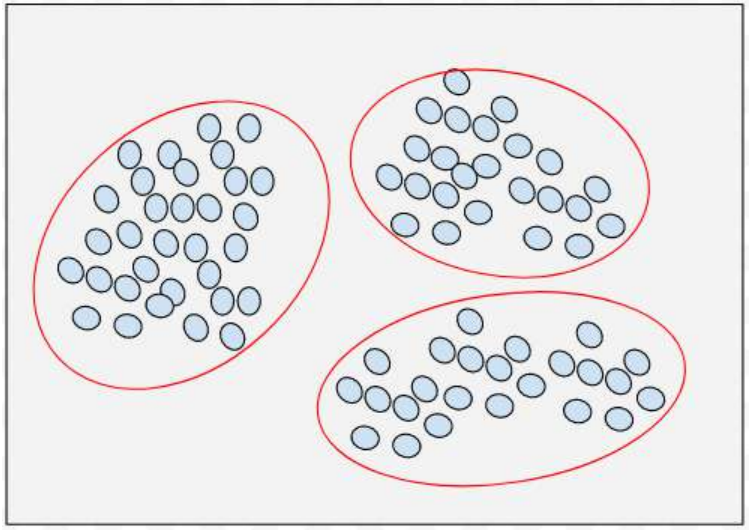


Unsupervised Learning - Find grouping based using data similarity

Clustering

Clustering aims to find groupings in data

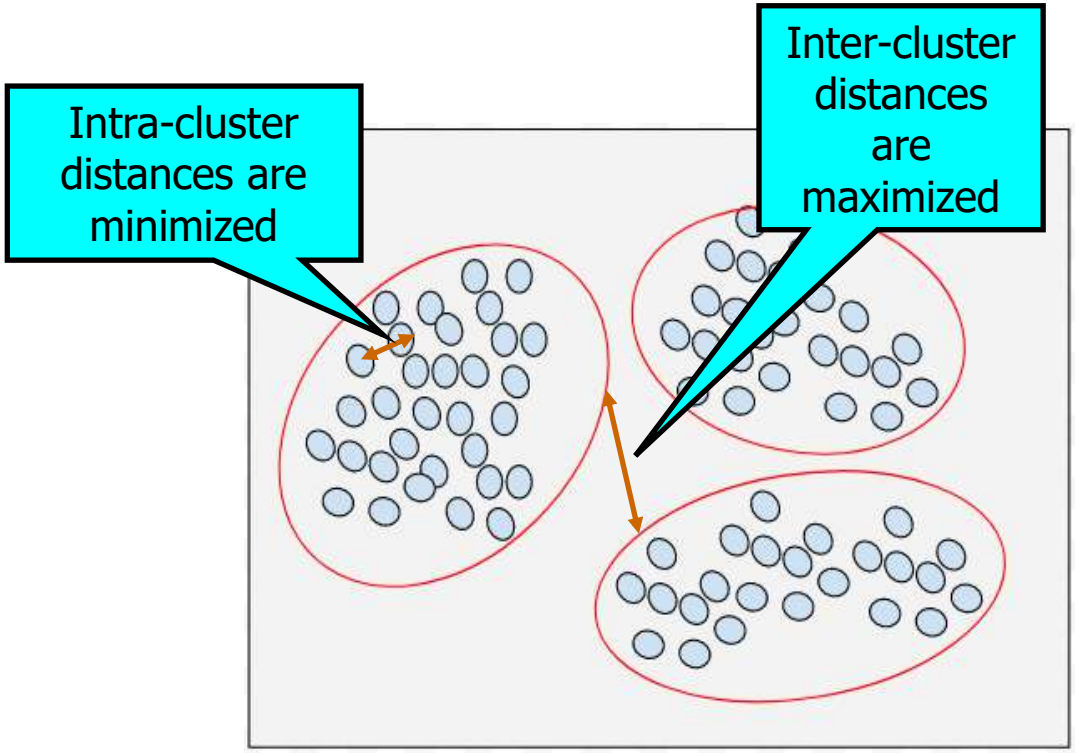
- Given a X, find K clusters using data similarity



Unsupervised Learning - Find grouping based using data similarity

Clustering

- Clustering aims to find groupings in data
  - Given a X, find K clusters using data similarity



Unsupervised Learning - Find grouping based using data similarity

Quality: What Is Good Clustering?

- A good clustering method will produce high quality clusters with
  - high intra-cluster similarity
  - low inter-cluster similarity
- The quality of a clustering result depends on both the similarity measure used by the method and its implementation
- The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns

What is not Cluster Analysis?

- Supervised classification
  - Have class label information
- Simple segmentation
  - Dividing students into different registration groups alphabetically, by last name
- Results of a query
  - Groupings are a result of an external specification

Applications: Customer Segmentation

Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs

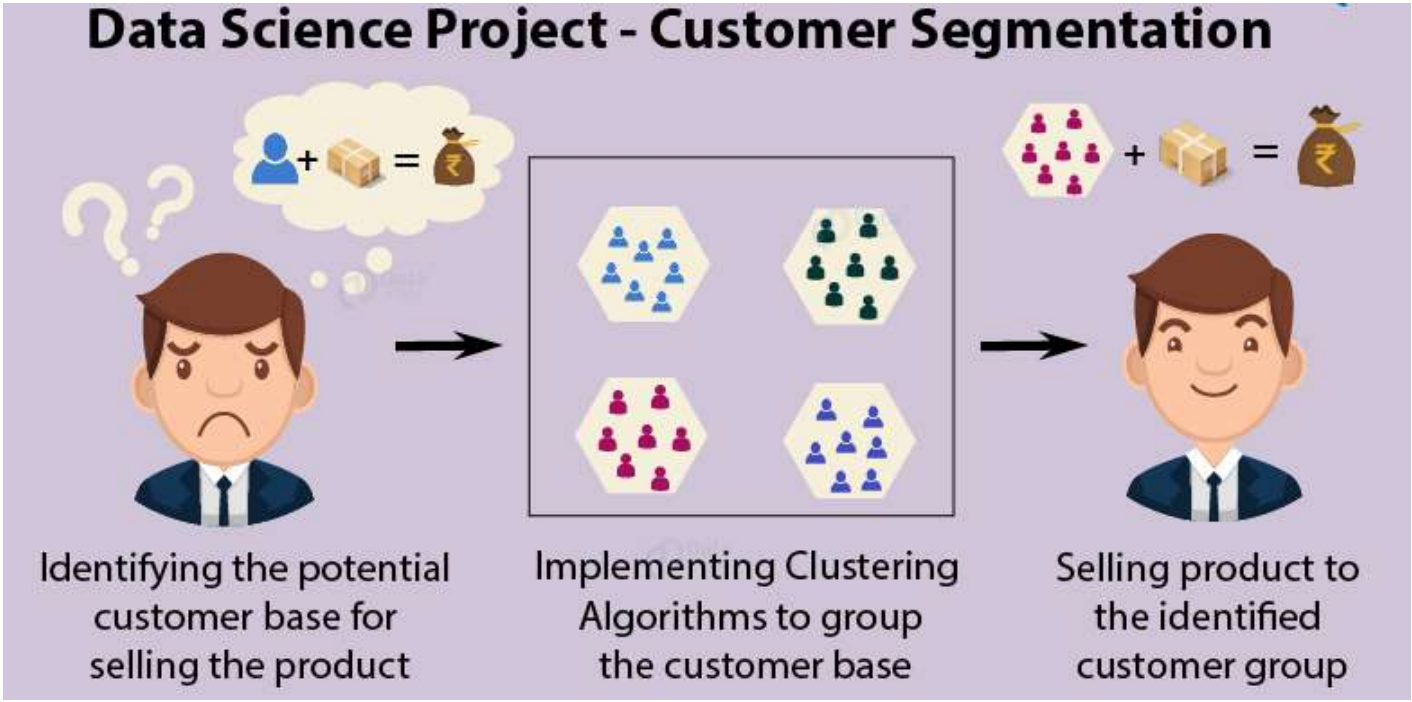


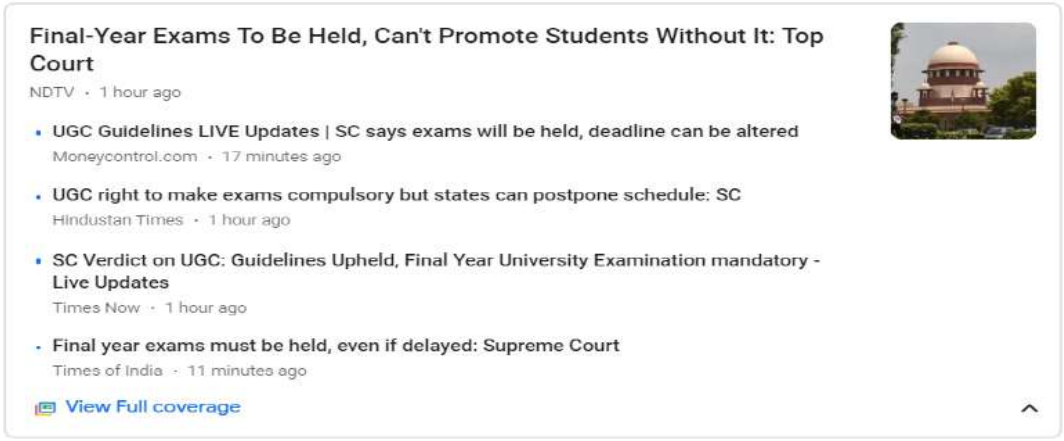
Image Compression using clustering

Input Image

output Image

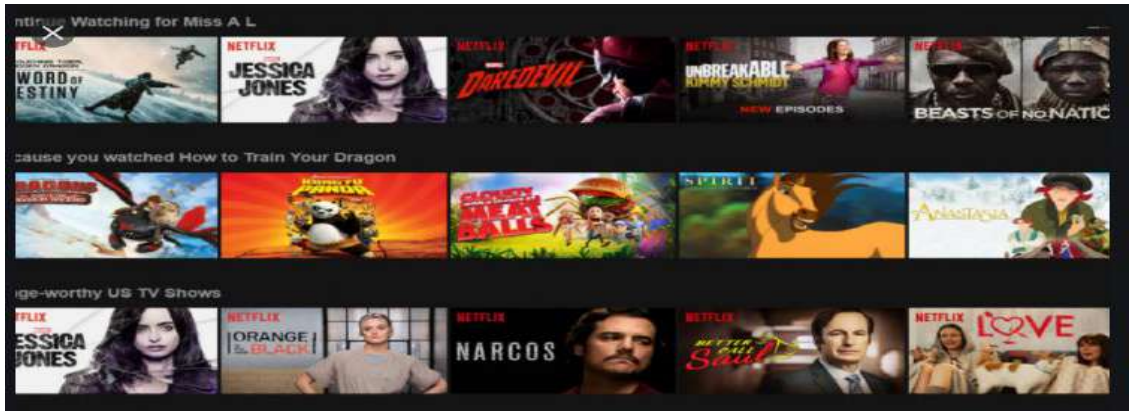
Document Clustering

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0



Recommendation System

company	country	director	genre	gross	name	rating	released	runtime	score	star
Columbia Pictures Corporation	USA	Rob Reiner	Adventure	52287414.0	Stand by Me	R	1986-08-22	89	8.1	Wil Wheaton
Paramount Pictures	USA	John Hughes	Comedy	70136369.0	Ferris Bueller's Day Off	PG-13	1986-06-11	103	7.8	Matthew Broderick
Paramount Pictures	USA	Tony Scott	Action	179800601.0	Top Gun	PG	1986-05-16	110	6.9	Tom Cruise
Twentieth Century Fox Film Corporation	USA	James Cameron	Action	85160248.0	Aliens	R	1986-07-18	137	8.4	Sigourney Weaver
Walt Disney Pictures	USA	Randal Kleiser	Adventure	18564613.0	Flight of the Navigator	PG	1986-08-01	90	6.9	Joey Cramer



https://ieeexplore.ieee.org/document/7019655

Other Applications

- Clustering on Gene Expression data
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3491664/

- Identifying Fake News
- http://snap.stanford.edu/mis2/files/MIS2_paper_2.pdf
- https://ai.intelligentonlinetools.com/ml/k-means-clustering-example-word2vec/

- COVID-19 Cluster Analysis
- https://towardsdatascience.com/covid-19-cluster-analysis-405ebbd10049

- Case Study
- https://medium.com/@msuginoo/three-different-lessons-from-three-different-clustering-analyses-data-science-capstone-5f2be29cb3b2

Requirements of Clustering

- Scalability
- Ability to deal with different types of attributes
- Ability to handle dynamic data
- Discovery of clusters with arbitrary shape
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noise and outliers
- Insensitive to order of input records
- High dimensionality
- Incorporation of user-specified constraints
- Interpretability and usability

Notion of a Cluster can be Ambiguous

How many clusters?

Six Clusters

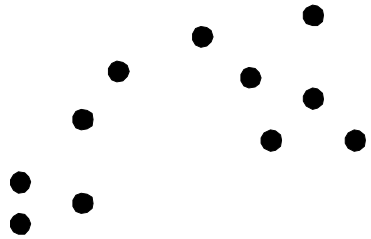
Two Clusters

Four Clusters

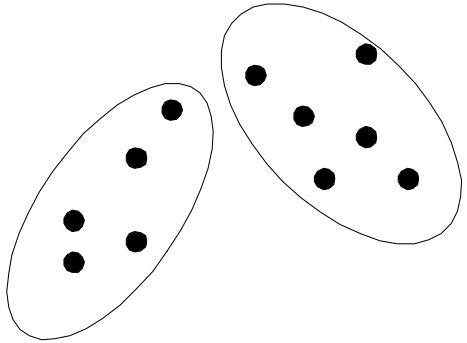
Types of Clustering

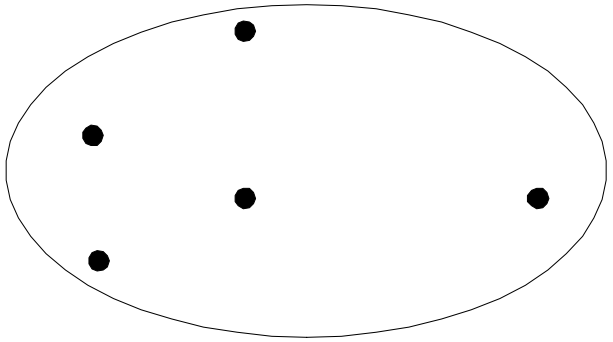
- A clustering is a set of clusters
- Partitional Clustering
  - A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset
- Hierarchical clustering
  - A set of nested clusters organized as a hierarchical tree
- Density based
  - identify distinctive groups/clusters in the data, based on the idea that a cluster in a data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density.
- Distribution Based
  - Idea is data generated from the same distribution, belongs to the same cluster if there exists several distributions in the dataset.

Partitional Clustering



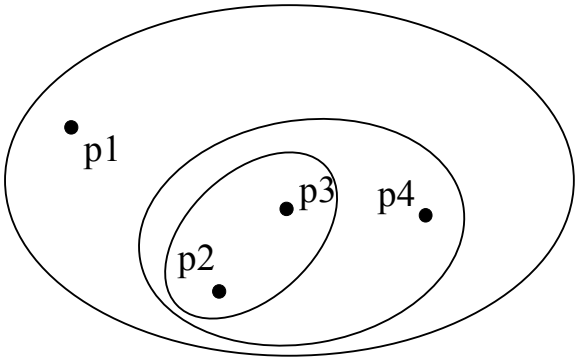
Original Points



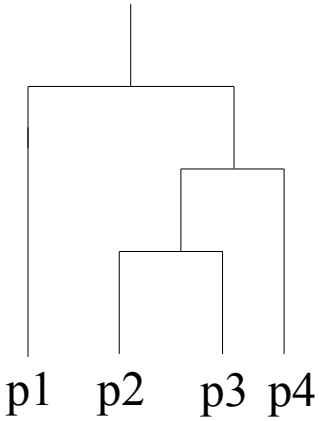


A Partitional Clustering

Hierarchical Clustering



Hierarchical Clustering

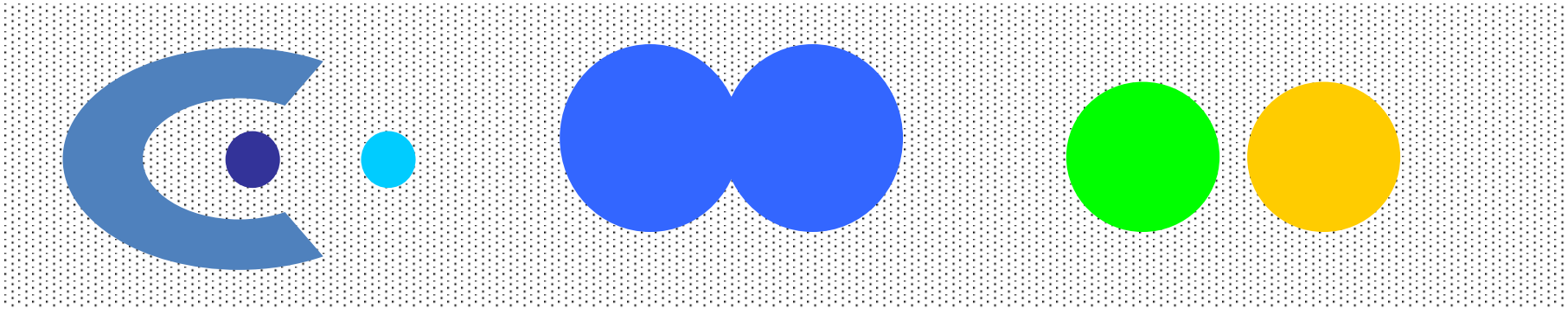


Dendrogram

Density-Based

Density-based

- A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density.
- Used when the clusters are irregular or intertwined, and when noise and outliers are present.



6 density-based clusters

Data Similarity/Dissimilarity

Similarity and Dissimilarity

- **Similarity**
  - Numerical measure of how alike two data objects are
  - Value is higher when objects are more alike
  - Often falls in the range [0,1]
- **Dissimilarity** (e.g., distance)
  - Numerical measure of how different two data objects are
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- **Proximity** refers to a similarity or dissimilarity

Data Matrix and Dissimilarity Matrix

- Data matrix
  - n data points with p dimensions

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- Dissimilarity matrix
  - n data points, but registers only the distance
  - A triangular matrix

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Proximity Measure for Nominal Attributes

- Can take 2 or more states, e.g., red, yellow, blue, green (generalization of a binary attribute)
- Simple matching
  - m: # of matches, p: total # of variables

$$d(i,j) = \frac{p - m}{p}$$

Proximity Measure for Binary Attributes

- A contingency table for binary data

Object *i* \ Object *j*	1	0	sum
1	q	r	$q+r$
0	s	t	$s+t$
sum	$q+s$	$r+t$	p

- Distance measure for symmetric binary variables:

$$d(i,j) = \frac{r+s}{q+r+s+t}$$

- Distance measure for asymmetric binary variables:

$$d(i,j) = \frac{r+s}{q+r+s}$$

- Jaccard coefficient (*similarity* measure for *asymmetric* binary variables):

$$sim_{Jaccard}(i,j) = \frac{q}{q+r+s}$$

Dissimilarity between Binary Variables

- Example

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

  - Gender is a symmetric attribute
  - The remaining attributes are asymmetric binary
  - Let the values Y and P be 1, and the value N 0

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

Distance on Numeric Data: Minkowski Distance

- *Minkowski distance*: A popular distance measure

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional data objects, and h is the order (the distance so defined is also called L-h norm)

- Properties
  - $d(i, j) > 0$ if $i \neq j$, and $d(i, i) = 0$ (Positive definiteness)
  - $d(i, j) = d(j, i)$ (Symmetry)
  - $d(i, j) \leq d(i, k) + d(k, j)$ (Triangle Inequality)
- A distance that satisfies these properties is a metric

Special Cases of Minkowski Distance

- $h = 1$: Manhattan (city block, L_1 norm) distance
  - E.g., the Hamming distance: the number of bits that are different between two binary vectors

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \ldots + |x_{i_p} - x_{j_p}|$$

- $h = 2$: (L_2 norm) Euclidean distance

$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \ldots + |x_{i_p} - x_{j_p}|^2)}$$

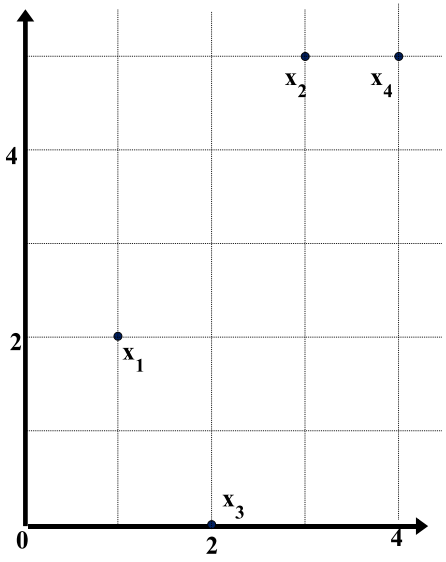
- $h \to \infty$. "supremum" (L_{max} norm, L_∞ norm) distance.
  - This is the maximum difference between any component (attribute) of the vectors

$$d(i, j) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_{f}^{p} |x_{if} - x_{jf}|$$

Example: Minkowski Distance

(Dissimilarity Matrices)

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



Manhattan (L_1)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L_2)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Supremum

L_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

Standardizing Numeric Data

- Z-score: $z = \frac{x - \mu}{\sigma}$
  - X: raw score to be standardized, μ: mean of the population, σ: standard deviation
  - the distance between the raw score and the population mean in units of the standard deviation
  - negative when the raw score is below the mean, "+" when above

Example

- The two tables above show the ‘area’ and ‘price’ of the same objects. Only the units of the variables change.

Calculate Euclidean distance in both the cases.

Area (sq.ft)	Price ($ 1000's)	Area (acre)	Price ($M)
2400	156000	0.0550944	156
1950	126750	0.0447642	126.75
2100	105000	0.0482076	105
1200	78000	0.0275472	78
2000	130000	0.045912	130
900	54000	0.0206604	54

Ordinal Variables

- An ordinal variable can be discrete or continuous
- Order is important, e.g., rank
- Can be treated like interval-scaled
  - replace x_{if} by their rank $r_{if} \in \{1, \ldots, M_f\}$
  - map the range of each variable onto [0, 1] by replacing *i*-th object in the *f*-th variable by

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

  - compute the dissimilarity using methods for interval-scaled variables

Attributes of Mixed Type: Gower distance

- A database may contain all attribute types
  - Nominal, symmetric binary, asymmetric binary, numeric, ordinal

$$d(i,j) = \frac{\sum_{c=1}^{n} \omega_c \delta_{ij}^{(c)} d_{ij}^{(c)}}{\sum_{c=1}^{n} \omega_c \delta_{ij}^{(c)}}$$

$d(i,j)$ = dissimilarity between row i and row j

c = the cth column

n = number of columns in the dataset

ω_c = weight of cth column = $\frac{1}{\text{nrows in dataset}}$

δ_{ij}^{c} =
- 0 if column c is missing in row i or j
- 0 if column c is asymmetric binary and both values in row i and j are 0
- 1 otherwise

d_{ij}^{c}(categorical) =
- 0 if i and j are equal in column c
- 1 otherwise

d_{ij}^{c}(continuous/ordinal) = $\frac{|\text{row i in column c - row j in column c}|}{\max(\text{column c}) - \min(\text{column c})}$

https://healthcare.ai/clustering-non-continuous-variables/

Example

Based on the information given in the table below, find most similar and most dissimilar persons among them. Apply min-max normalization on income to obtain [0,1] range. Consider profession and mother tongue as nominal. Consider native place as ordinal variable with ranking order of [Village, Small Town, Suburban, Metropolitan]. Give equal weight to each attribute.

Name	Income	Profession	Mother tongue	Native Place
Ram	70000	Doctor	Bengali	Village
Balram	50000	Data Scientist	Hindi	Small Town
Bharat	60000	Carpenter	Hindi	Suburban
Kishan	80000	Doctor	Bhojpuri	Metropolitan

Solution

After normalizing income and quantifying native place, we get

Name	Income	Profession	Mother tongue	Native Place
Ram	0.67	Doctor	Bengali	1
Balram	0	Data Scientist	Hindi	2
Bharat	0.33	Carpenter	Hindi	3
Kishan	1	Doctor	Bhojpuri	4

d(Ram, Balram) = 0.67+1+1+(2-1)/(4-1)=3 d(Ram, Bharat) = 0.33+1+1+(3-1)/(4-1)=3
d(Ram, Kishan) = 0.33+0+1+(4-1)/(4-1) = 2.33 d(Balram, Bharat) = 0.33+1+0+(3-2)/(4-1)=1.67
d(Balram, Kishan) = 1+1+1+(4-2)/(4-1) = 3.67 d(Bharat, Kishan) = 0.67+1+1+(4-3)/(4-1) = 3

Most similar – Balram and Bharat; Most dissimilar – Balram and Kishan

Cosine Similarity

- A **document** can be represented by thousands of attributes, each recording the *frequency* of a particular word (such as keywords) or phrase in the document.

Document	*team*	*coach*	*hockey*	*baseball*	*soccer*	*penalty*	*score*	*win*	*loss*	*season*
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

- Other vector objects: gene features in micro-arrays, …
- Applications: information retrieval, biologic taxonomy, gene feature mapping, ...
- Cosine measure: If d_1 and d_2 are two vectors (e.g., term-frequency vectors), then

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / \|d_1\| \, \|d_2\| ,$$

where $\bullet$ indicates vector dot product, $\|d\|$: the length of vector d

Example: Cosine Similarity

- $\cos(d_1, d_2) = (d_1 \bullet d_2) / \|d_1\| \|d_2\|$,
  where $\bullet$ indicates vector dot product, $\|d\|$: the length of vector d

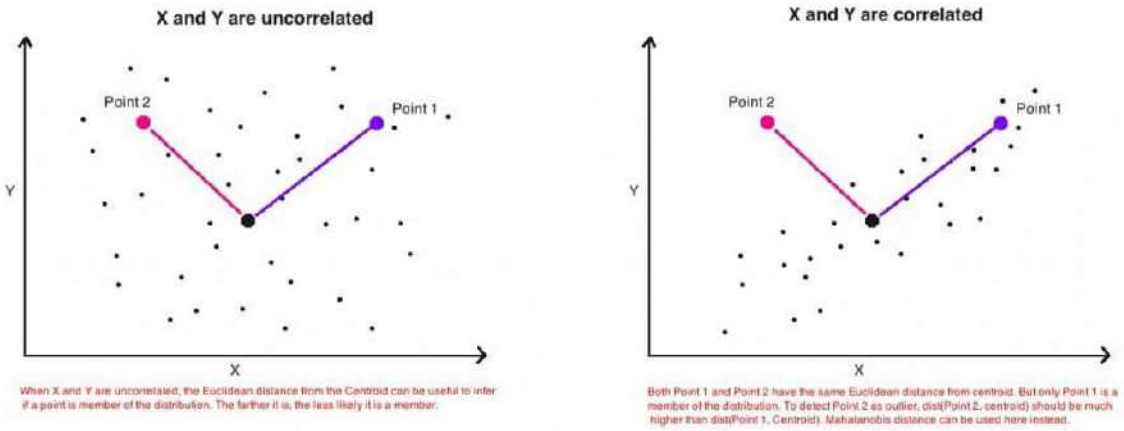
- Ex: Find the **similarity** between documents 1 and 2.

$d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$
$d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$

$d_1 \bullet d_2 = 5*3+0*0+3*2+0*0+2*1+0*1+0*1+2*1+0*0+0*1 = 25$
$\|d_1\| = (5*5+0*0+3*3+0*0+2*2+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$
$\|d_2\| = (3*3+0*0+2*2+0*0+1*1+1*1+0*0+1*1+0*0+1*1)^{0.5} = (17)^{0.5} = 4.12$
$\cos(d_1, d_2) = 0.94$

Drawback of Euclidean

- If the dimensions are correlated to one another, which is typically the case in real-world datasets, the Euclidean distance between a point and the center of the points (distribution) can give little or misleading information about how close a point really is to the cluster.



The two points above are equally distant (Euclidean) from the center. But only one of them (blue) is actually more close to the cluster, even though, technically the Euclidean distance between the two points are equal.

Mahalanobis Distance

- Mahalanobis distance is an effective multivariate distance metric that measures the distance between a point (vector) and a distribution.
- Mahalonobis distance is the distance between a point and a distribution. And not between two distinct points.
- The formula to compute Mahalanobis distance is as follows:

$$D^2 = (x - m)^T \cdot C^{-1} \cdot (x - m)$$

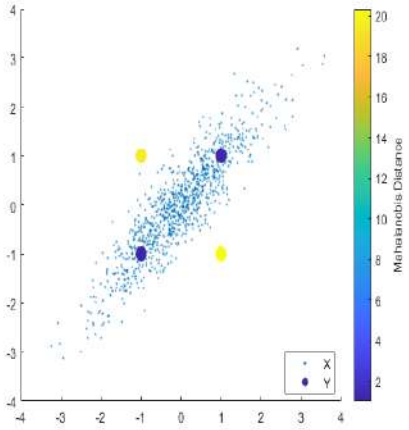
```
where,

 - D^2        is the square of the Mahalanobis distance.

 - x          is the vector of the observation (row in a dataset),

 - m          is the vector of mean values of independent variables
(mean of each column),

 - C^(-1)     is the inverse covariance matrix of independent
variables.
```

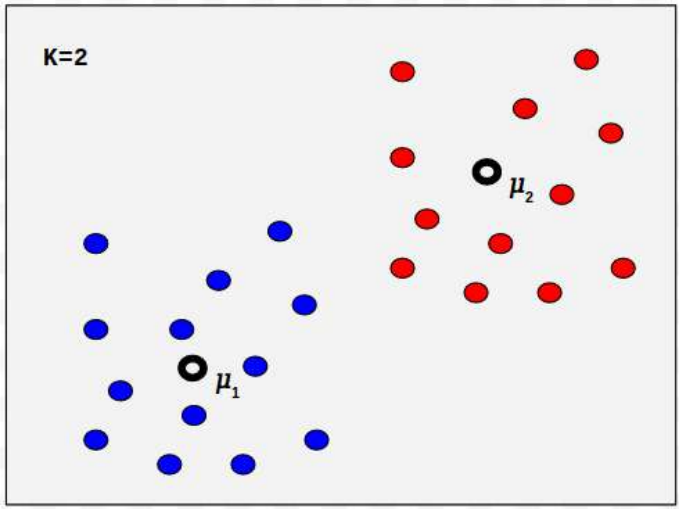


K-Means

K-Means Algorithm

- Works iteratively to find $\{\mu_k\}$and $\{r_{nk}\}$ such that J is minimized

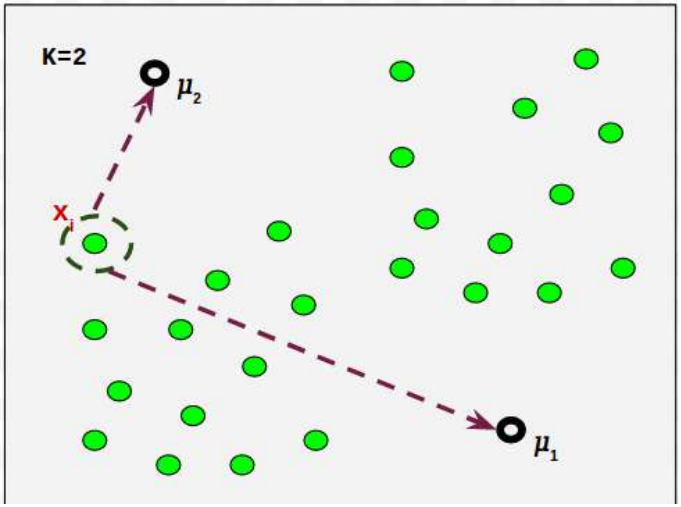
$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$



- Iteration involves two key steps
  - Find $\{r_{nk}\}$, fixing $\{\mu_k\}$ to minimize J
  - Find $\{\mu_k\}$, fixing $\{r_{nk}\}$ to minimize J
- Let us look at each of these steps

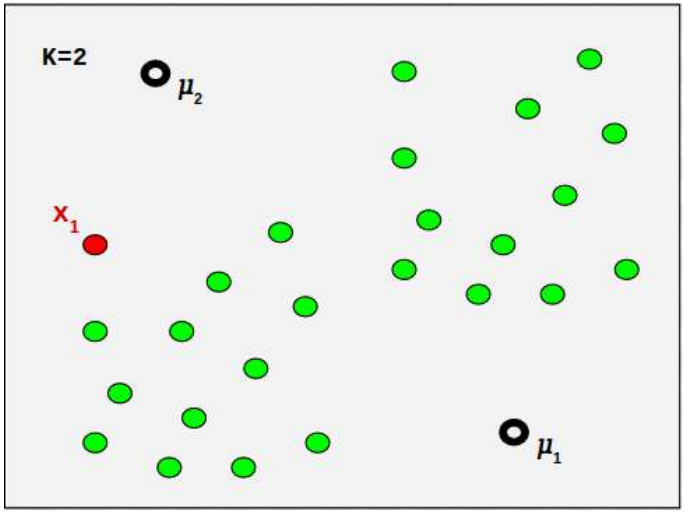
K-Means Algorithm

A sample E-Step



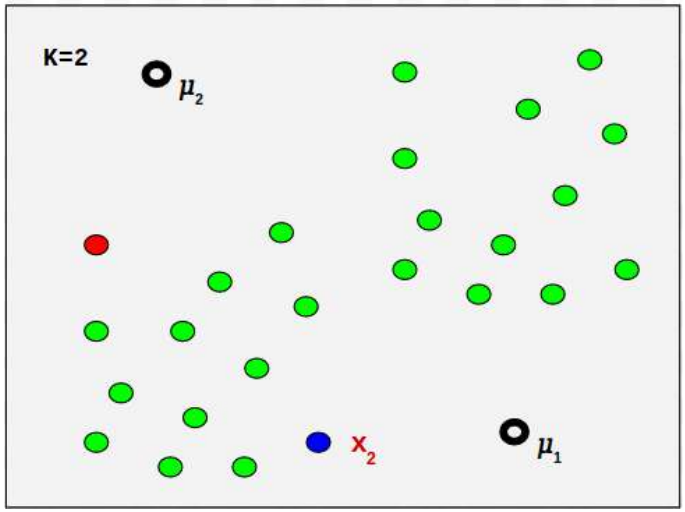
K-Means Algorithm

A sample E-Step



K-Means Algorithm

A sample E-Step

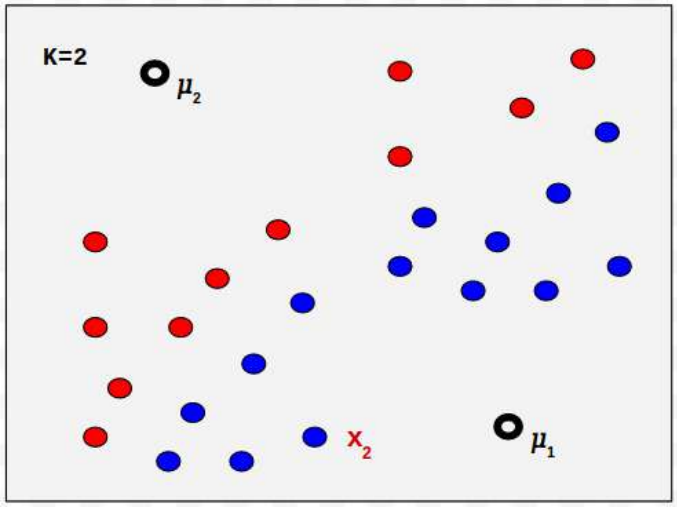


K-Means Algorithm

E-Step:

For all $x_t \in X$:

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$

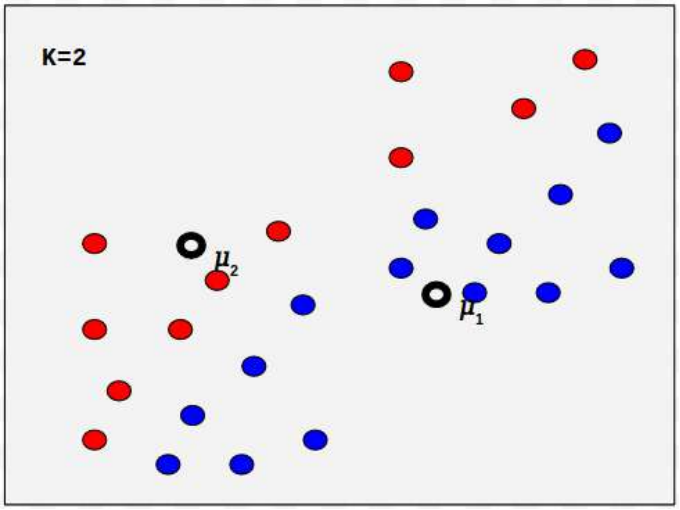


K-Means Algorithm

M-Step:

For all μ_k[where k = 1,2,...,K] :

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}$$



K-Means Algorithm

Algorithm:

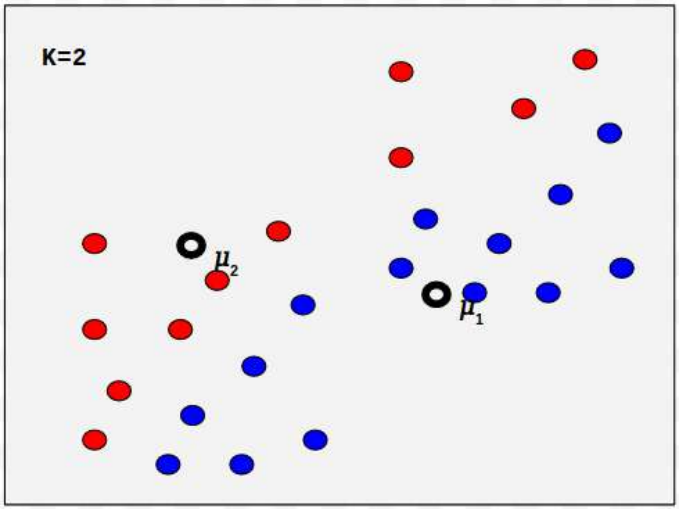
_____Initialize μ_k[where k = 1,2,...,K]

Repeat

E-Step [as defined earlier]

M-Step [as defined earlier]

Until convergence of μ_k .



K-Means Algorithm

Algorithm:

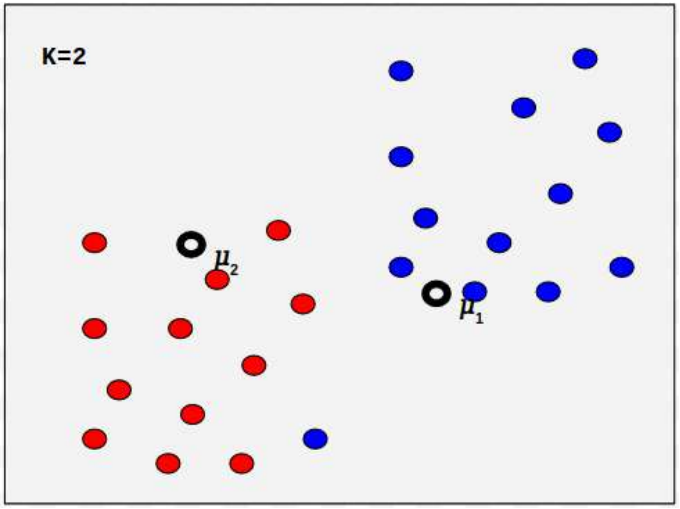
Initialize μ_k[where k = 1,2,...,K]

Repeat

E-Step [as defined earlier]

M-Step [as defined earlier]

Until convergence of μ_k .



E-Step in the second iteration

K-Means Algorithm

Algorithm:

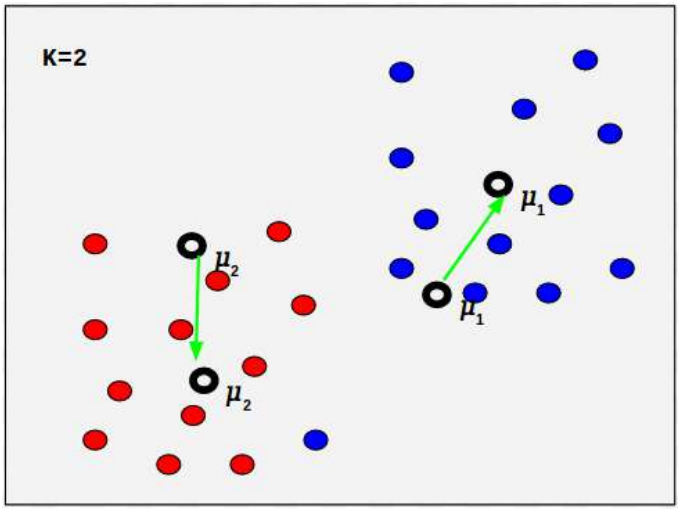
Initialize μ_k[where k = 1,2,...,K]

Repeat

E-Step [as defined earlier]

M-Step [as defined earlier]

Until convergence of μ_k .



M-Step in the second iteration

K-Means Algorithm

Algorithm:

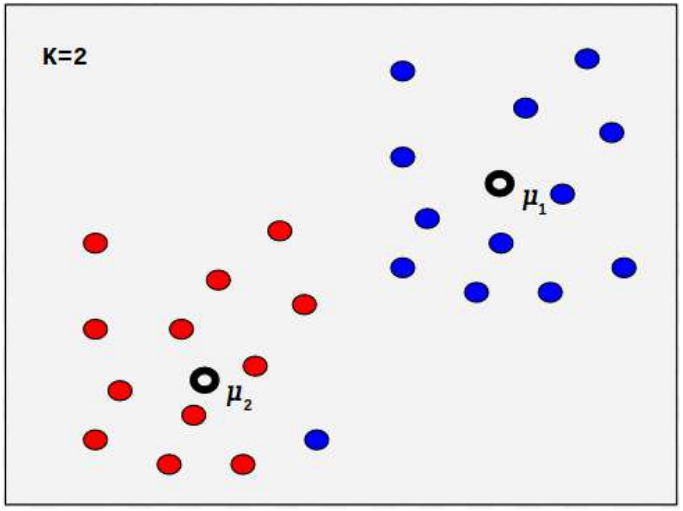
Initialize μ_k[where k = 1,2,...,K]

Repeat

E-Step [as defined earlier]

M-Step [as defined earlier]

Until convergence of μ_k .



M-Step in the second iteration

Example

Consider the analysis of weights of individuals and their respective blood glucose levels as given below:

a) Identify the clusters using K-means clustering (k=2) for the given data, assuming candidate 1 and 2 as initial centroids.

b) How many iterations does it take before termination?

Candidate	Weight	Glucose level
1	72	185
2	56	170
3	60	168
4	68	179
5	72	182
6	77	188
7	70	180
8	84	183

Example

Candidate	Cluster1(72, 185)	Cluster2(56, 170)
3(60, 168)	$=17^2+12^2$	$=2^2+4^2$
4(68 179)	$=6^2+4^2$	$=9^2+12^2$
5(72, 182)	$=3^2+0^2$	$=12^2+16^2$
6(77, 188)	$=3^2+5^2$	$=18^2+11^2$
7(70, 180)	$=5^2+2^2$	$=10^2+14^2$
8(183, 84)	$=2^2+12^2$	$=13^2+28^2$

After 1st iteration, the cluster groups are C1{1, 4, 5, 6, 7, 8} and C2{2, 3}

Re-computing centroids:

C1([72+68+72+77+70+84]/6, [185 +179+182+188+180+183]/6) =C1(73.83, 182.83,)

C2([56 +60]/2, [170+168]/2) =C2(58, 169)

2nd Iteration –computing distances from resulting centroids:

After 2nd iteration, the cluster groups are C1{1, 4, 5, 6, 7, 8} and C2{2, 3}; there is no change in cluster groups hence the algorithm terminates.

Candidate	Cluster1(182.83, 73.83)	Cluster2(169, 58)
1(185, 72)	$(182.83-185)^2+(73.83-72)^2$ $=2.17^2+1.83^2$	$(169-185)^2+(58-72)^2$ $=16^2+14^2$
2(170, 56)	$(182.83-170)^2+(73.83-56)^2$ $=12.83^2+17.83^2$	$(169-170)^2+(58-56)^2$ $=1^2+2^2$
3(168, 60)	$(182.83-168)^2+(73.83-60)^2$ $=14.83^2+13.83^2$	$(169-168)^2+(58-60)^2$ $=1^2+2^2$
4(179, 68)	$(182.83-179)^2+(73.83-68)^2$ $=3.83^2+5.83^2$	$(169-179)^2+(58-68)^2$ $=10^2+10^2$
5(182, 72)	$(182.83-182)^2+(73.83-72)^2$ $=0.83^2+1.83^2$	$(169-182)^2+(58-72)^2$ $=13^2+14^2$
6(188, 77)	$(182.83-188)^2+(73.83-77)^2$ $=5.17^2+3.17^2$	$(169-188)^2+(58-77)^2$ $=19^2+19^2$
7(180, 70)	$(182.83-180)^2+(73.83-70)^2$ $=2.17^2+1.83^2$	$(169-180)^2+(58-70)^2$ $=11^2+12^2$
8(183, 84)	$(182.83-183)^2+(73.83-84)^2$ $=0.17^2+10.17^2$	$(169-183)^2+(58-84)^2$ $=14^2+26^2$

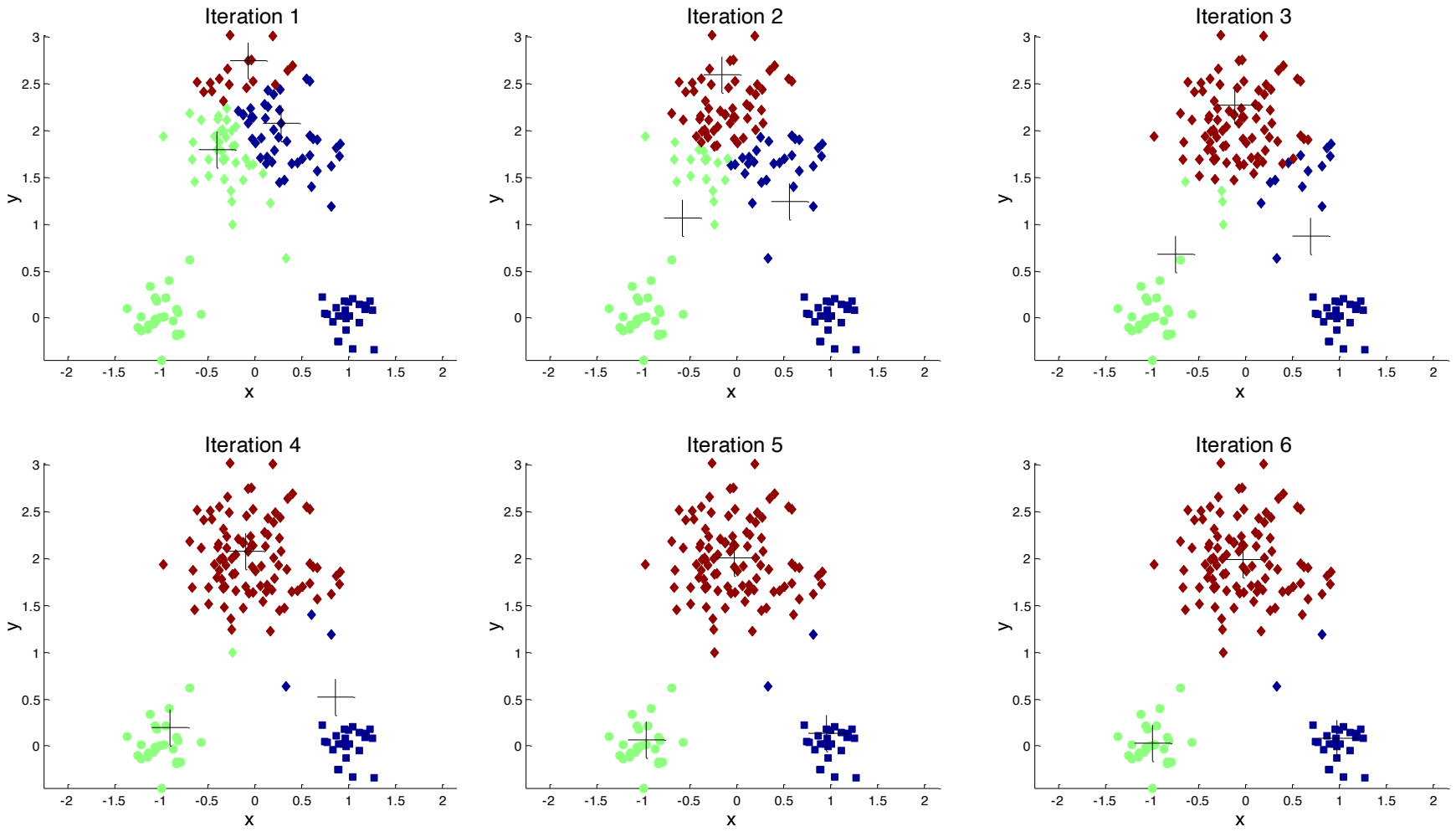
Evaluating K-means Clusters

- Most common measure is Sum of Squared Error (SSE)
  - For each point, the error is the distance to the nearest cluster
  - To get SSE, we square these errors and sum them.

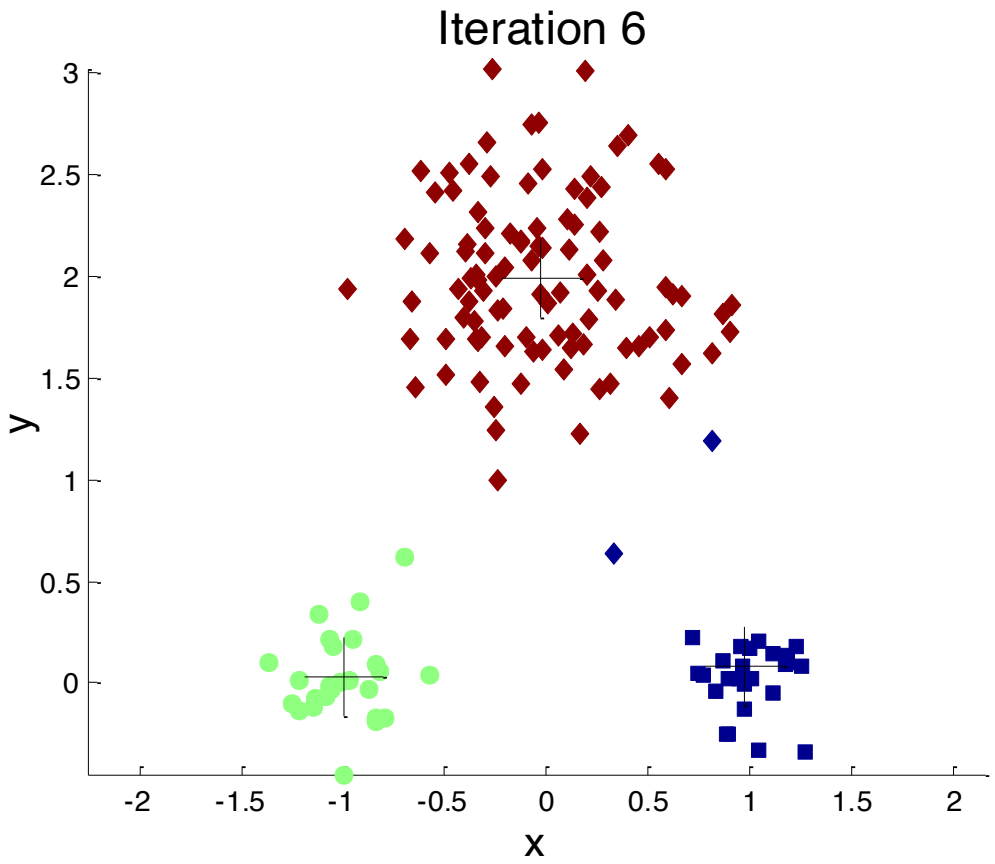
$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

  - x is a data point in cluster C_i and m_i is the representative point for cluster C_i
    - can show that m_i corresponds to the center (mean) of the cluster
  - Given two clusters, we can choose the one with the smallest error
  - One easy way to reduce SSE is to increase K, the number of clusters
    - A good clustering with smaller K can have a lower SSE than a poor clustering with higher K

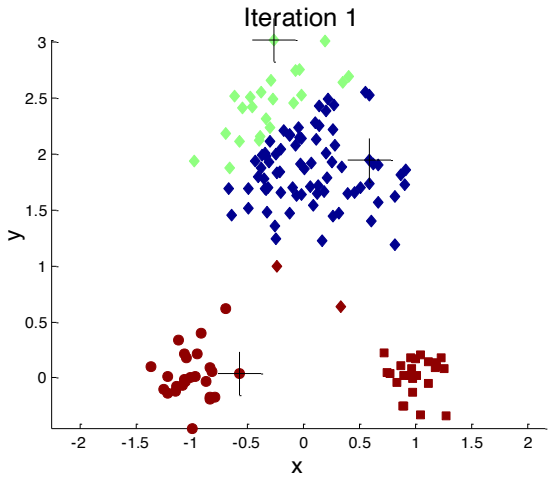
Importance of Choosing Initial Centroids

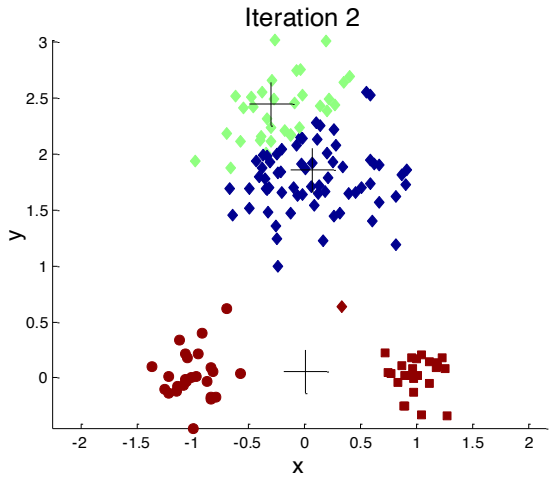


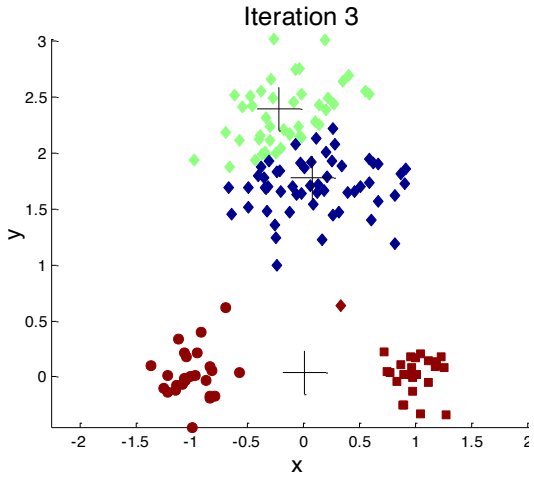
Importance of Choosing Initial Centroids

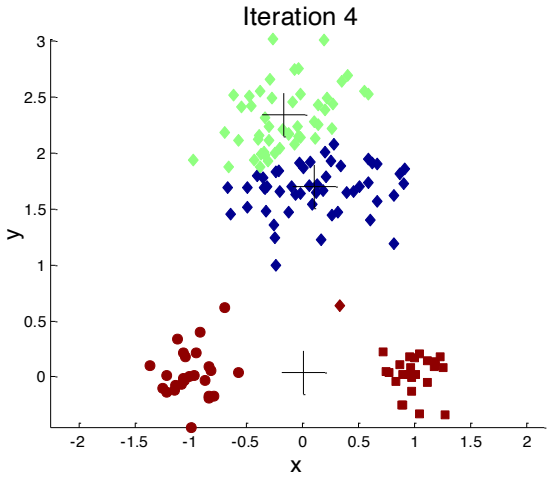


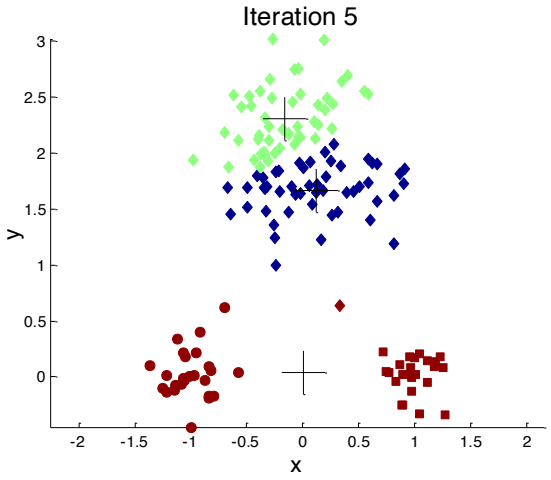
Importance of Choosing Initial Centroids …



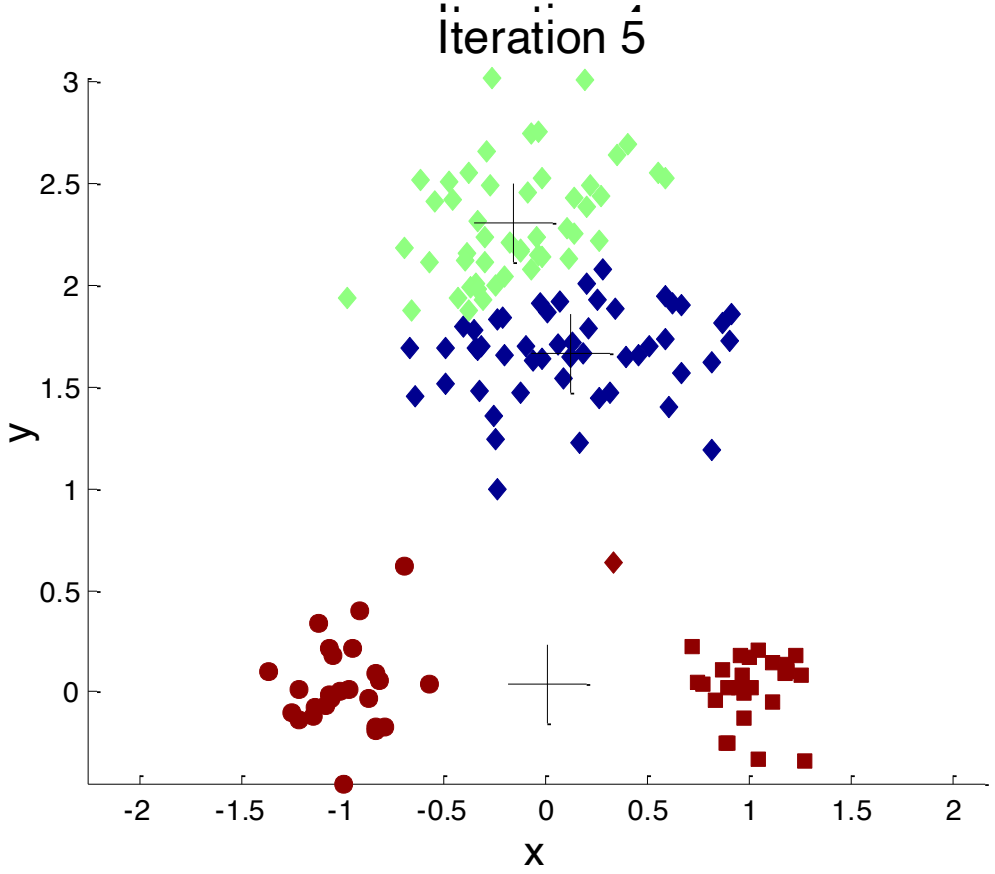








Importance of Choosing Initial Centroids …

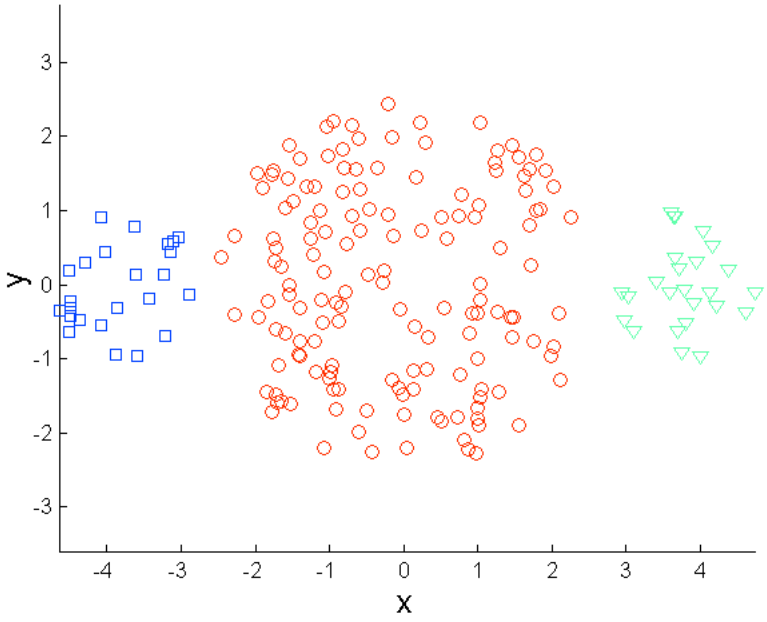


Limitations of K-means

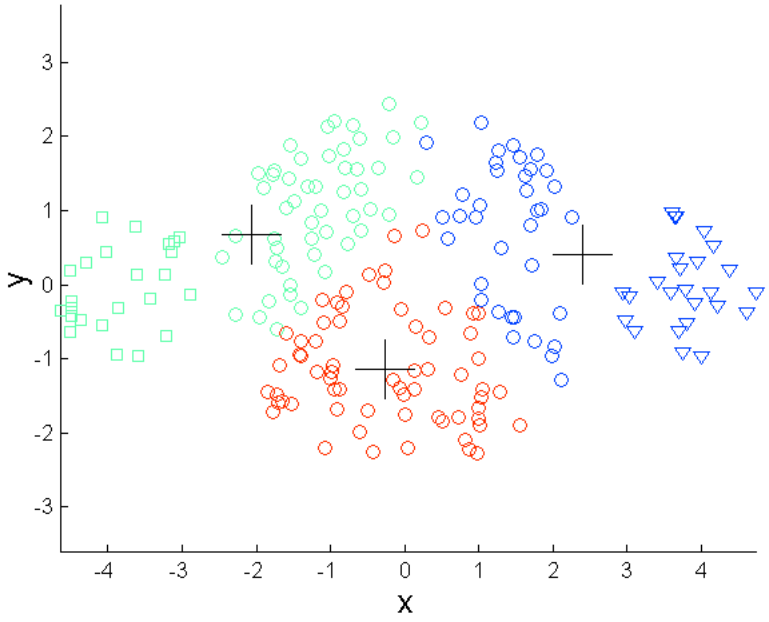
- K-means has problems when clusters are of differing
  - Sizes
  - Densities
  - Non-globular shapes

- K-means has problems when the data contains outliers.

Limitations of K-means: Differing Sizes

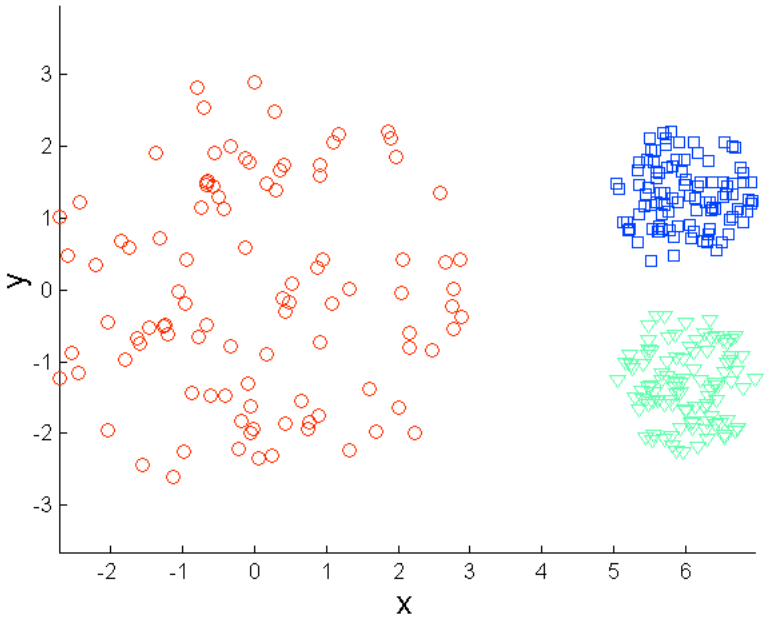


Original Points

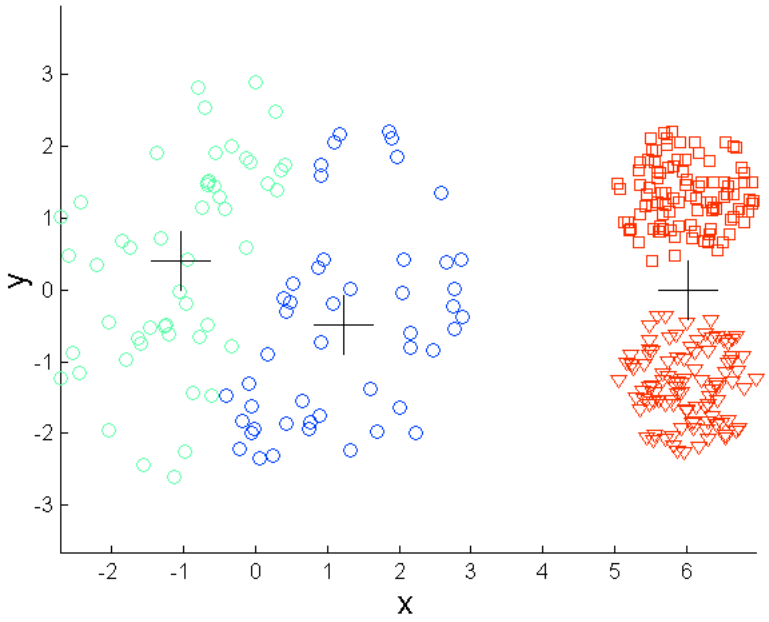


K-means (3 Clusters)

Limitations of K-means: Differing Density

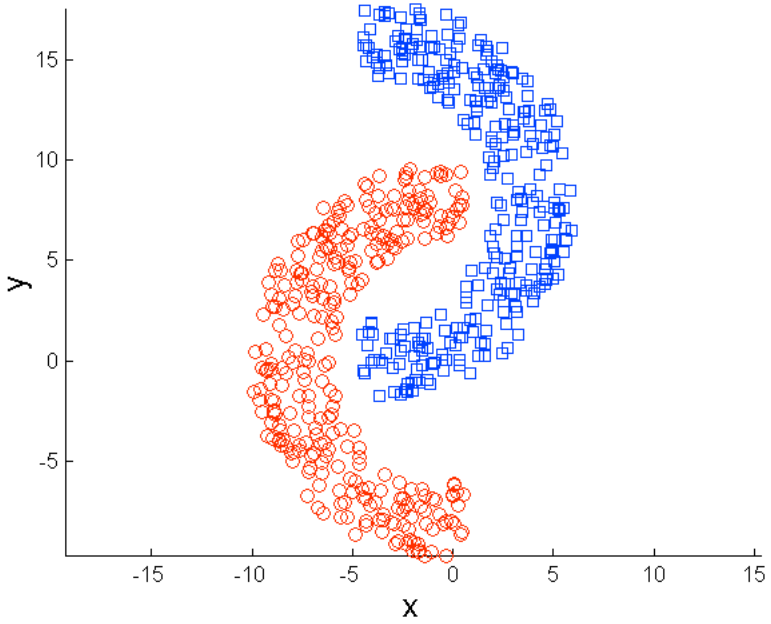


Original Points

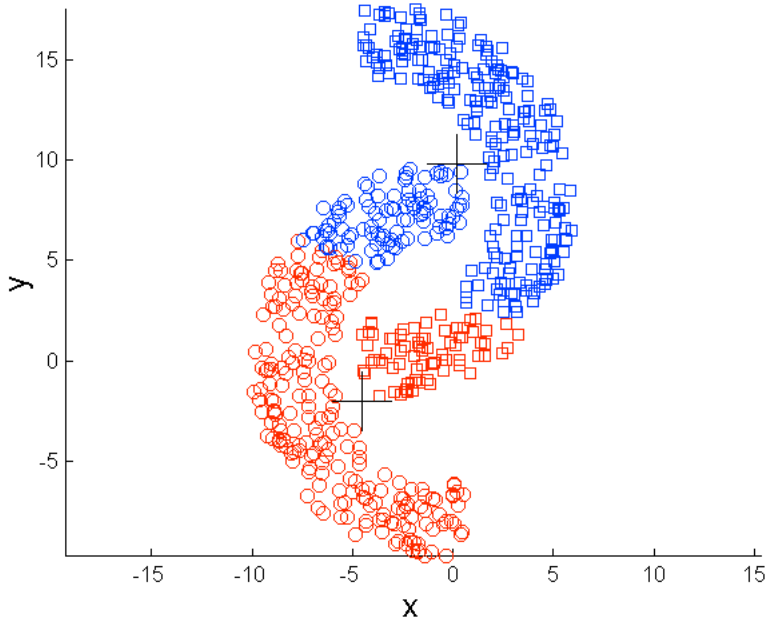


K-means (3 Clusters)

Limitations of K-means: Non-globular Shapes



Original Points



K-means (2 Clusters)

k-means and Its Variants

BITS Pilani
Pilani Campus

Swarna Chaudhary
Asst. Professor
WILP Division, BITS-Pilani

K-Means and its Variants

Objectives

- Elbow method
- kmeans++
- k-medoids
- Online version (with sequential update)
- Mini-Batch
- Outliers

Discussion on the K-means Method

- Efficiency: O(tkn) where n:#of objects, k:#of clusters, and t:#of iterations
  - Normally, k,t<<n;thus, an efficient method
- K-means clustering often terminates at a local optimal
  - Initialization can be important to find high-quality clusters
- Need to specify k, the number of clusters, in advance
  - In practise, one often runs a range of values and selected the “best” k value
- Sensitivity to noisy data and outliers
  - Variations: using k-medians, k-medoids, etc.
- K-means is applicable only to objects in a continuous n-dimensional space
  - Using k-modes for categorical data
- Not suitable to discover clusters with non-convex shapes
  - Using density based clustering, kernel k-means etc

Variations of k-Means

- There are many variants of the k-Means method, varying in different aspects
  - Choosing better initial centroids
    - K-Means++
  - Choosing different representative prototypes for the clusters
    - K-Medoids, K-median, K-modes
  - Applying feature transformation techniques
    - Kernel K-Means

K-Means++

- Different initializations may generate different clustering results
- For better initialisation of k-inital centroids
  - **K-Means++**
    - Randomly select the first centroid from the data points.
    - For each data point compute its distance from the nearest, previously chosen centroid.
    - Select the next centroid from the data points such that the probability of choosing a point as centroid is directly proportional to its distance from the nearest, previously chosen centroid. (i.e. the point having maximum distance from the nearest centroid is most likely to be selected next as a centroid)
    - Repeat steps 2 and 3 until k centroids have been sampled

K-Medians: Handling Outliers by Computing Medians

- Medians are less sensitive to outliers than means
- K-medians: instead of taking the mean value of the object in a cluster as a reference point, medians are used (L1-norm as the distance measure)
- The criterion function for the k-medians algorithm:

$$S = \sum_{k=1}^{K} \sum_{x_{ij} \in C_k} | x_{ij} - med_{kj} |$$

- The k-medians:
  - Select k points as the initial representative objects(i.e., as initial k medians)
  - Repeat
    - Assign every point to its nearest median
    - Re-compute the median using the median of each individual feature
  - Until convergence criterion is satisfied.

Handling Outliers: From K-Means to K-Medoids

- The k-means is sensitive to outliers-since an object with an extremely large value may substantially distort the distribution of the data
- K-Medoids: Instead of taking the mean value of the object in a cluster as a reference point, medoids can be used, which is the most centrally located object in a cluster
- **K-Medoids**

1. Initialize: select k random points out of the n data points as the medoids.

2. Associate each data point to the closest medoid by using any common distance metric methods.

3. While the cost decreases:

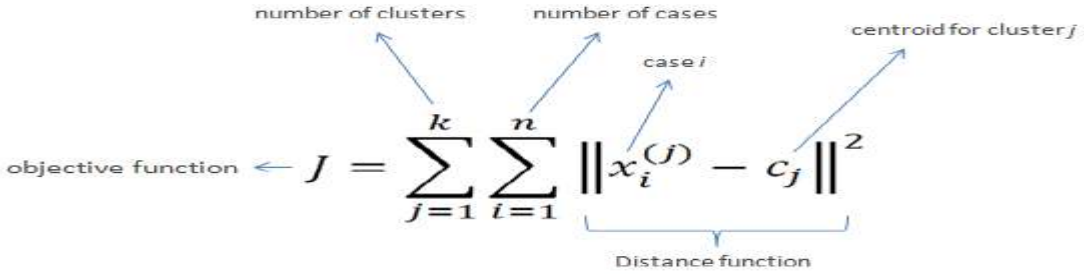
For each medoid m, for each data o point which is not a medoid:

1. Swap m and o, associate each data point to the closest medoid, recompute the cost.

2. If the total cost is more than that in the previous step, undo the swap.

How to choose a "good" k for k-means clustering?

- Choose the number of clusters by visually inspecting data points
- **Elbow method:**
  - Compute the sum of squared error (SSE) for some values of k (for example, 2,3,4,5,6…). The SSE is defined as the sum of the squared distance between each member of the cluster and its centroid.



  - If you plot k against SSE, you will see that the error decreases as k gets larger, this is because when the number of clusters increases, they should be smaller, so distortion is also smaller. The idea of the elbow method is to choose the k at which the SSE decreases abruptly.

Online Version (with Sequential update)

- Another way to modify the k-means procedure is to update the means one example at a time, rather than all at once.
- This is particularly attractive when we acquire the examples over a period of time, and we want to start clustering before we have seen all of the examples.
  - Make initial guesses for the means $\mathbf{m}_1$, $\mathbf{m}_2$, ..., $\mathbf{m}_k$
  - Set the counts n_1, n_2, ..., n_k to zero
  - Until interrupted
    - Acquire the next example, $\mathbf{x}$
    - If $\mathbf{m}_i$ is closest to $\mathbf{x}$
      - Increment n_i
      - Replace $\mathbf{m}_i$ by $\mathbf{m}_i + (1/n_i)*(\mathbf{x} - \mathbf{m}_i)$
    - end_if
  - end_until

https://www.cs.princeton.edu/courses/archive/fall08/cos436/Duda/C/sk_means.htm

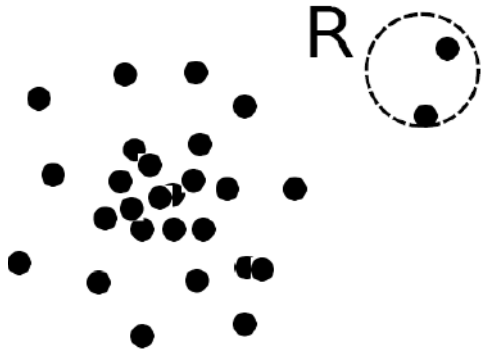
Mini Batch K-means

1. Initialize $\{\mu_k\}$ by randomly picking k instances from X
2. Initialize v [1...k] to 0's
3. For i = 1 … t do
    a. M ← pick b examples randomly from X [choosing the mini-batch]
    b. For each x ∈ M do
        i. d[x] ← f ($\{\mu_k\}$, x) [f returns μ_k closest to x]
    c. For each x ∈ M do
        i. μ_k ← d[x] [Get the μ_k closest to x]
        ii. v[μ_k] ← v[μ_k] + 1 [maintain the count of x's closer to μ_k]
        iii. η = 1 / v[μ_k] [per-center learning rate]
        iv. $\mu_k^{(new)} = \mu_k^{(old)} - \eta (\mu_k^{(old)} - x_n)$ [gradient step - same as prev. alg]
4. End For

What Are Outliers/Anomalies?

- **Outlier**: A data object that **deviates significantly** from the normal objects as if it were **generated by a different mechanism**
- Outliers are different from the noise data
  - Noise is random error or variance in a measured variable
  - Noise should be removed before outlier detection
- Outliers are interesting: It violates the mechanism that generates the normal data

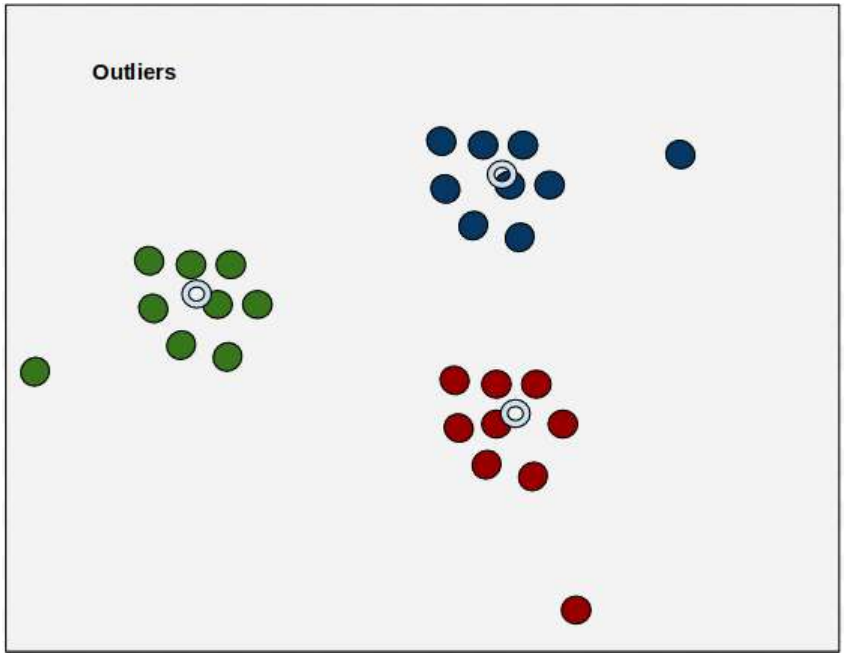
- Applications:
  - Credit card fraud detection
  - Intrusion detection
  - Medical analysis



(naive) K-Means for detecting outliers

- Let
  - dist(x, μ_k) be the distance of a point x, assigned to cluster k to its center μ_k.
  - Lμ_k be the the average distance of all the points assigned to cluster k with its center

- The ratio dist(x, μ_k) / Lμ_k for each point is the outlier score for each point.
  - Higher the ratio for a point x, more likely x is an outlier



Thank You!

Expectation Maximization and MoG

BITS Pilani
Pilani Campus

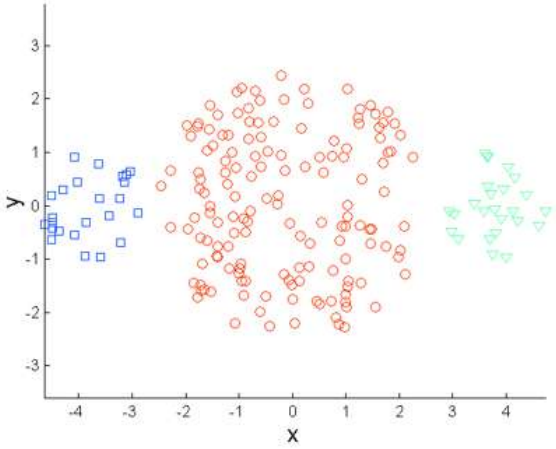
Swarna Chaudhary
Asst. Professor
WILP Division, BITS-Pilani

- *The slides presented here are obtained from the authors of the books and from various other contributors. I hereby acknowledge all the contributors for their material and inputs.*
- *I have added and modified a few slides to suit the requirements of the course.*

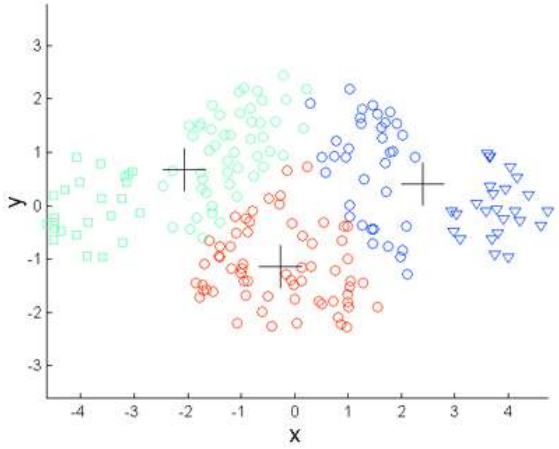
Expectation Maximization and MoG

Drawbacks of K-means Clustering

- Doesn't work on clusters of varying densities/sizes



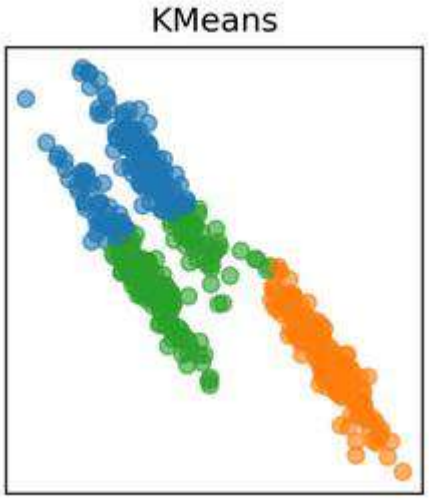
Original Points



K-means (3 Clusters)

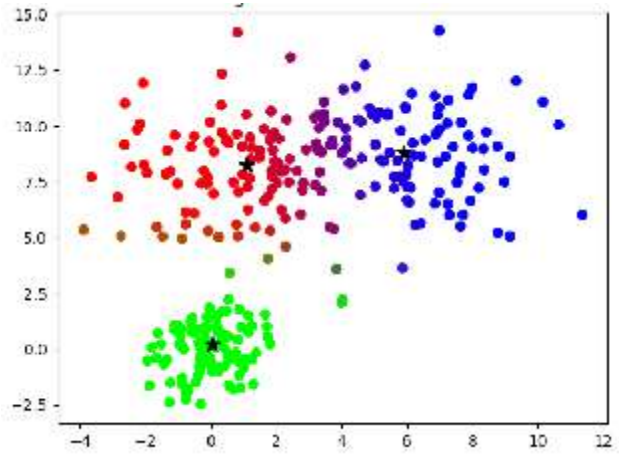
Drawbacks of K-means Clustering

- Doesn't work well when the distribution of points is *not* in a circular form.



Drawbacks of K-means Clustering

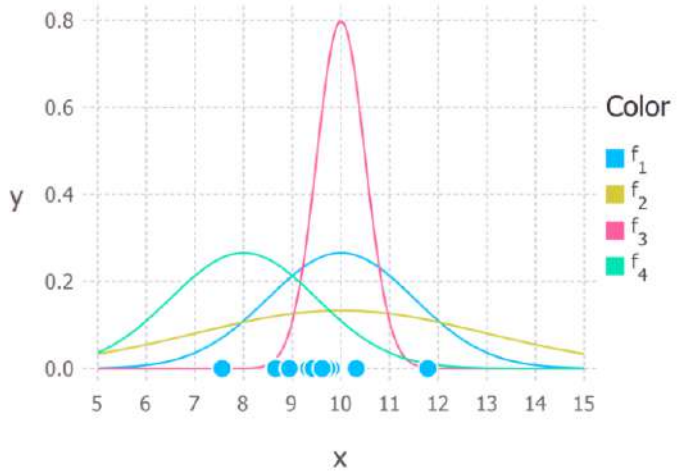
- Hard Clustering vs Soft Clustering



https://towardsdatascience.com/gaussian-mixture-models-d13a5e915c8e

Gaussian Distribution

- The normal curve is *bell-shaped* and has a single peak at the
- exact center of the distribution.
- The arithmetic mean, median, and mode of the distribution are equal and located at the peak.
- Half the area under the curve is above the peak, and the other half is below it.
- The normal distribution is symmetrical about its mean.



The 10 data points and possible Gaussian distributions from which the data were drawn. f1 is normally distributed with mean 10 and variance 2.25 (variance is equal to the square of the standard deviation), this is also denoted f1 ~ N (10, 2.25). f2 ~ N (10, 9), f3 ~ N (10, 0.25) and f4 ~ N (8, 2.25)

MATHEMATICAL FUNCTION (Pdf)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

This is a bell shaped curve with different centers and spreads depending on μ and σ

Note constants:

π=3.14159

e=2.71828

Multivariate Gaussian

- Gaussian for a single variable (Univariate) x:

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

- Given the x is D-dimensional vector & x is Gaussian distributed :

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

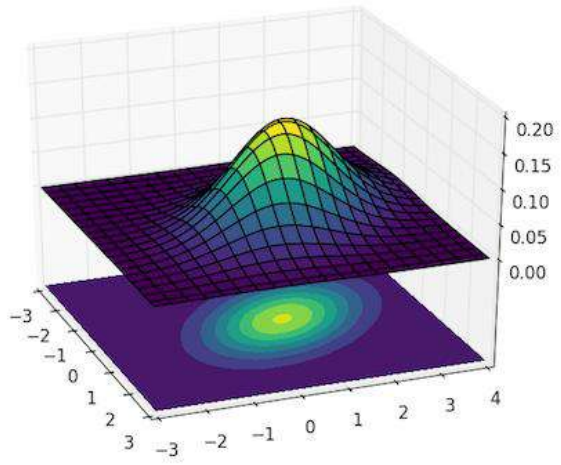
Where

x: input vector

μ : D-dimensional mean vector

Σ : D × D covariance matrix

|Σ| : Determinant of Σ



Univariate and Multivariate Gaussian density

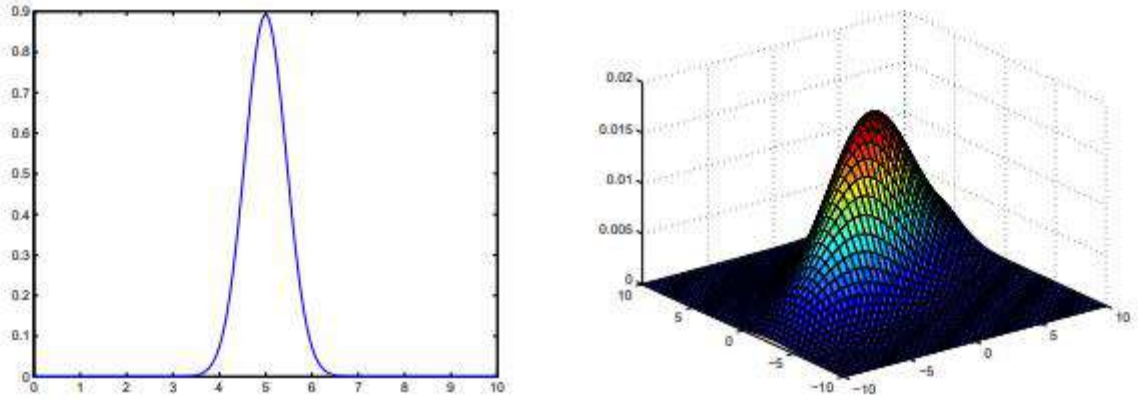
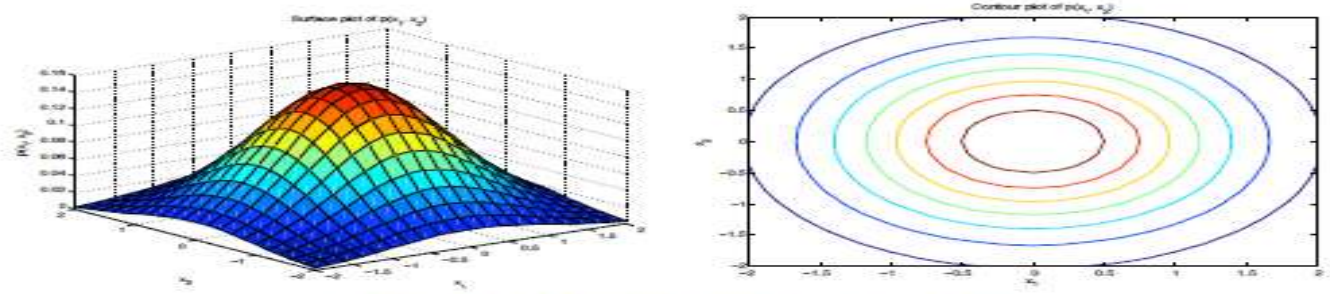
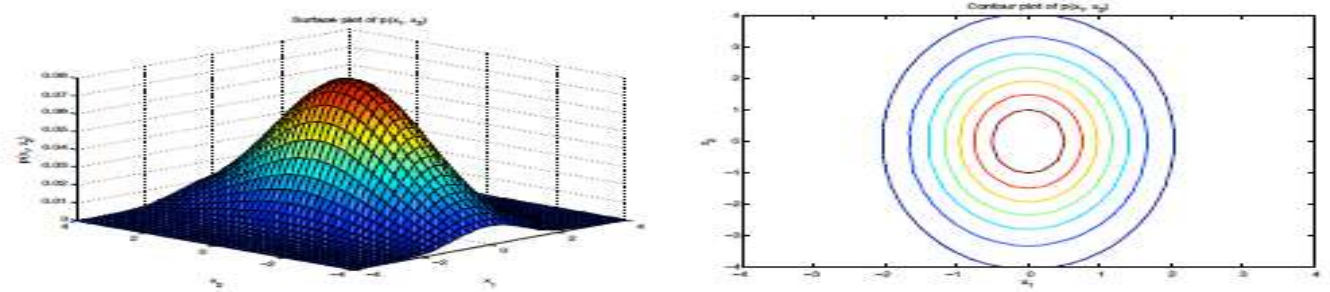


Figure 1: The figure on the left shows a univariate Gaussian density for a single variable X. The figure on the right shows a multivariate Gaussian density over two variables X_1 and X_2.

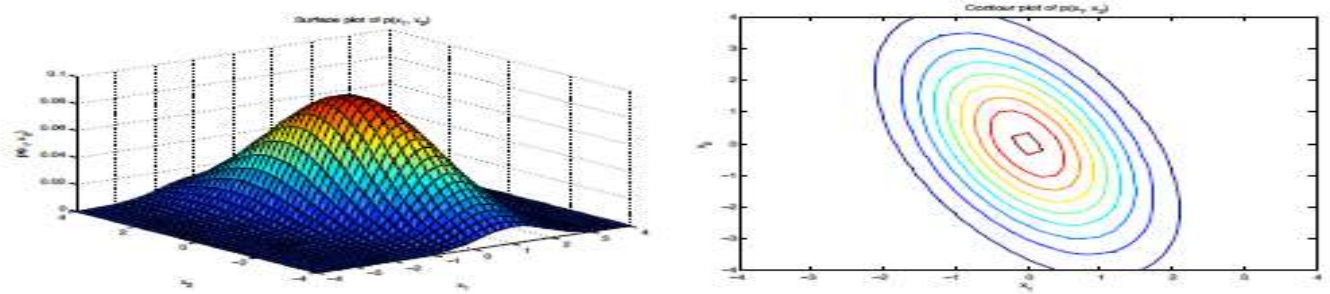
2-dimensional Gaussian distribution



(a) Spherical Gaussian (diagonal covariance, equal variances)

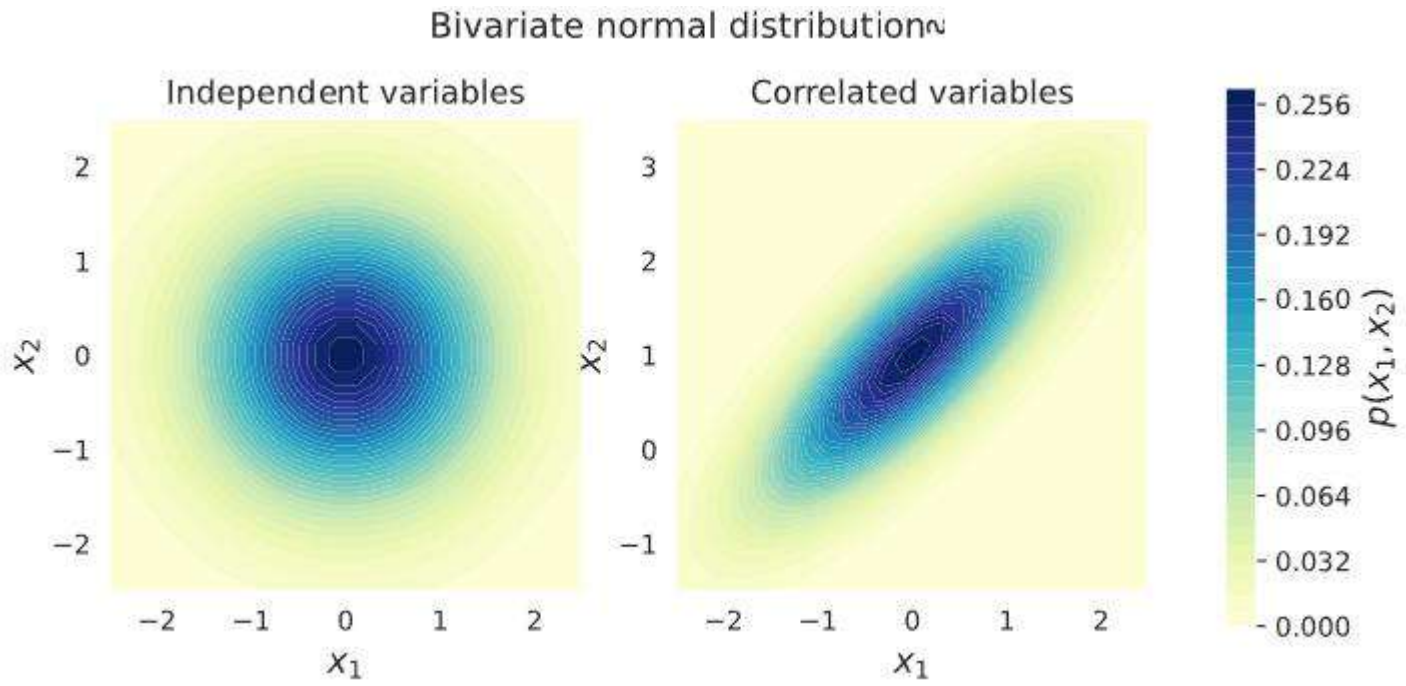


(b) Gaussian with diagonal covariance matrix



(c) Gaussian with full covariance matrix

Figure 7: Surface and contour plots of 2–dimensional Gaussian with different covariance structures



The figure on the left is a bivariate distribution with the covariance between x_1 and x_2 set to 0 so that these 2 variables are independent:

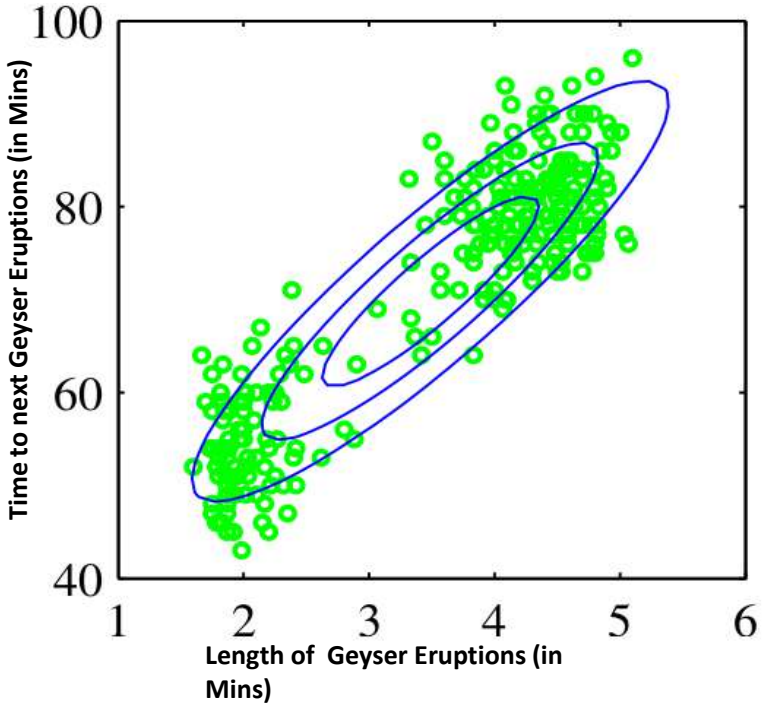
$$\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$$

The figure on the right is a bivariate distribution with the covariance between x_1 and x_2 set to be different than 0 so that both variables are correlated. Increasing x_1 will increase the probability that x_2 will also increase:

$$\mathcal{N}\left(\begin{bmatrix}0\\1\end{bmatrix}, \begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}\right)$$

Multimodal Data

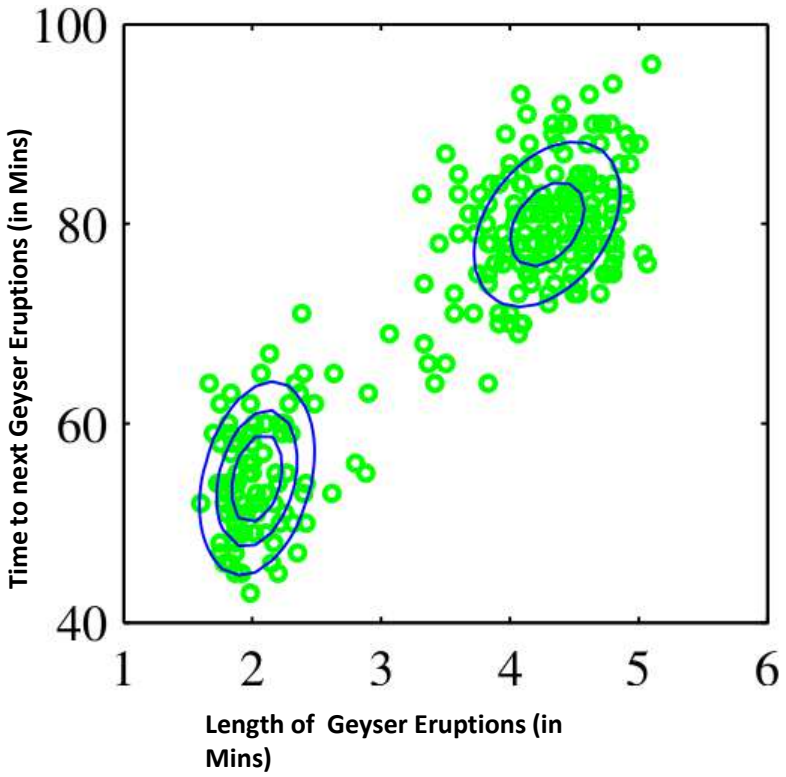
- Gaussian distribution is that it is intrinsically unimodal
- A single distribution is not sufficient to fit many real data sets which is multimodal



Demo of fitting a single gaussian to the data.

Multimodal Data

- A single distribution is not sufficient to fit many real data sets

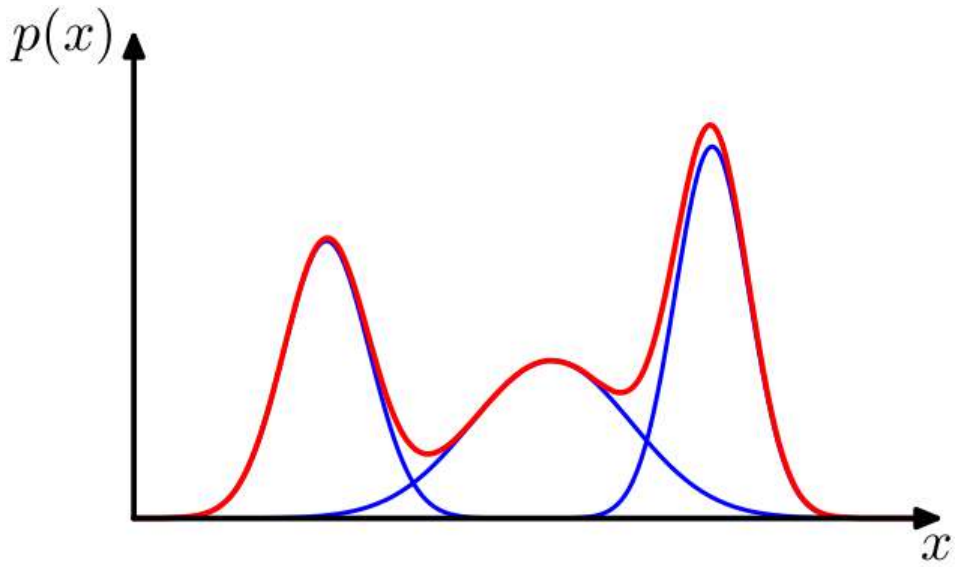


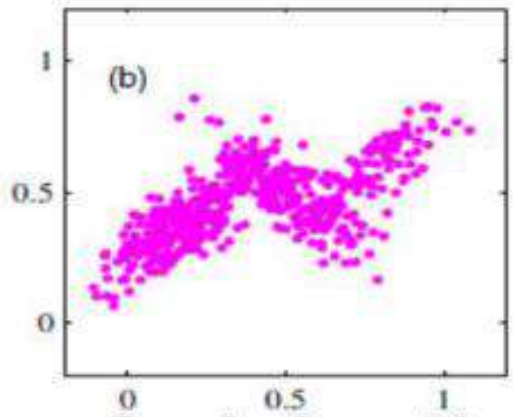
Demo of fitting a k gaussian to the data.

Mixture Model

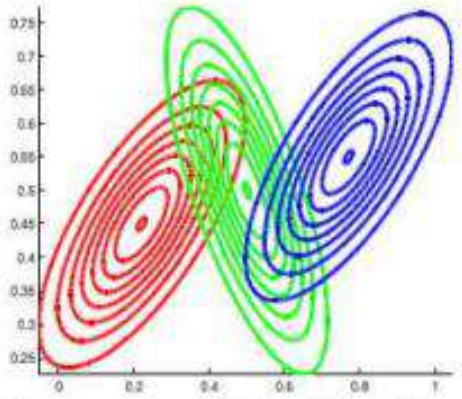
- A **mixture model** assumes that a set of observed objects is a mixture of instances from multiple probabilistic clusters, and conceptually each observed object is generated independently
  - *probabilistic cluster* is a distribution over the data space, which can be mathematically represented using a probability density function (or distribution function).

- **Out task**: infer a set of *k* probabilistic clusters that is mostly likely to generate *D*

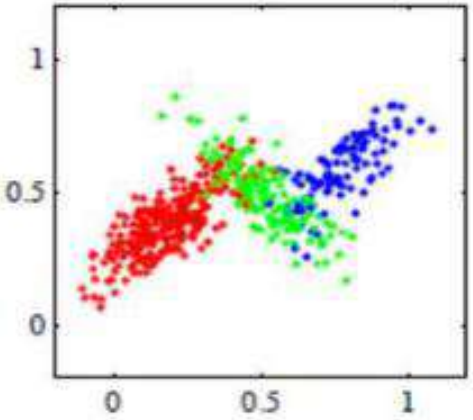




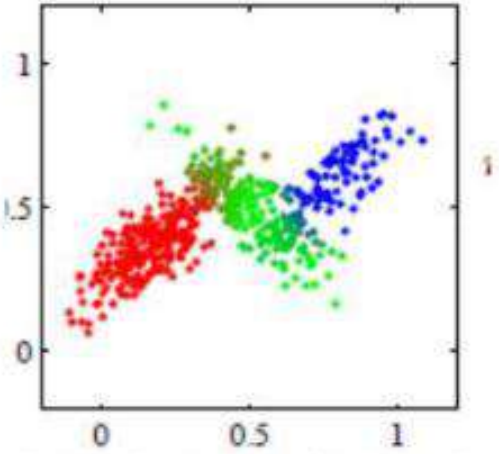
Samples from p(x)



Component Distributions



Samples labeled using their true component



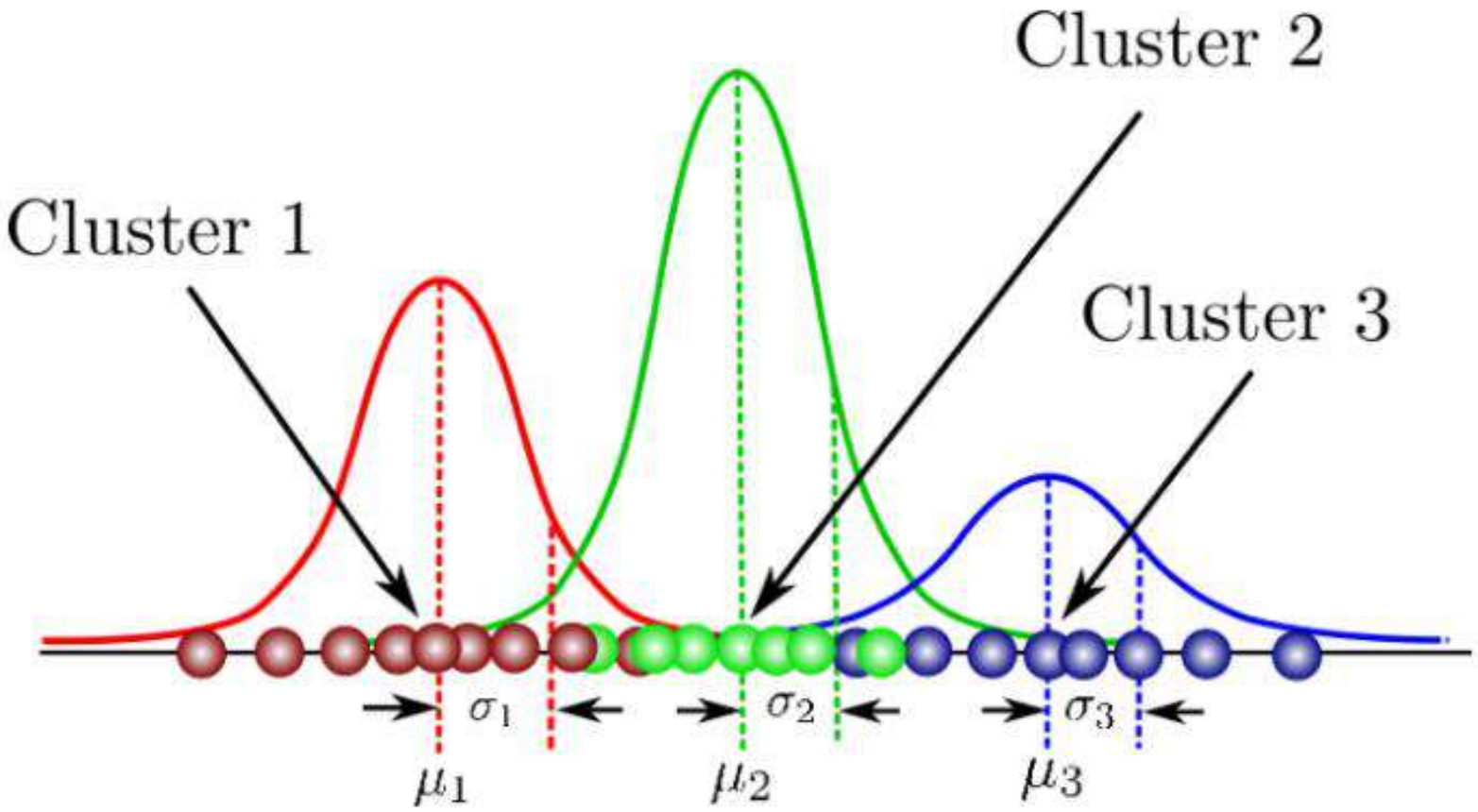
Soft clustering learned by a Gaussian mixture model

Gaussian Mixture Model

- A Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters.
  - A *Gaussian Mixture* is a function that is comprised of several Gaussians, each identified by $k \in \{1,\ldots, K\}$, where K is the number of clusters of our dataset. Each Gaussian k in the mixture is comprised of the following parameters:
  - A mean μ that defines its centre.
  - A covariance Σ that defines its width. This would be equivalent to the dimensions of an ellipsoid in a multivariate scenario.
  - A mixing probability π that defines how big or small the Gaussian function will be.

https://scikit-learn.org/stable/modules/mixture.html

Gaussian Mixture Model

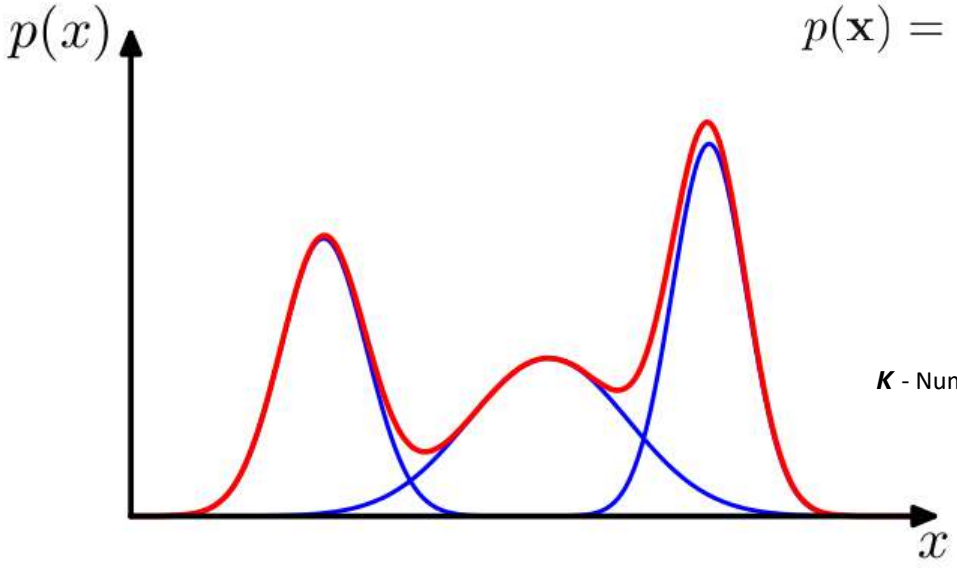


Mixture of Gaussians

- Mixture of Gaussians: Component densities are Gaussian.

Mixing Coefficients

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

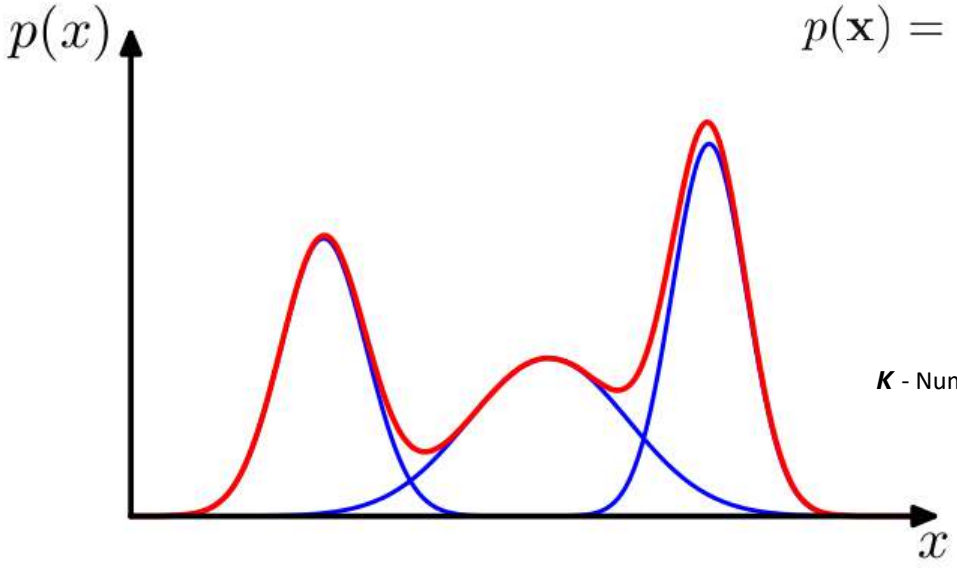


K - Number of Gaussians

Data Set: The data used this module came from 'Old Faithful Data' available from https://www.kaggle.com/janithwanni/old-faithful/data for download & used by the text book PRML.

Mixture of Gaussians

- Mixture of Gaussians: Component densities are Gaussian.



$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

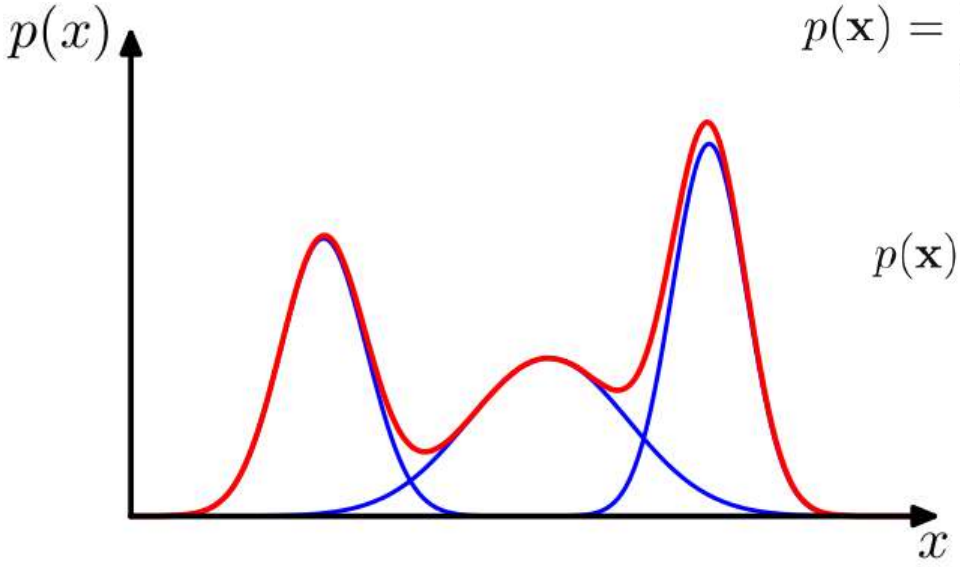
Component Densities

K - Number of Gaussians

Data Set: The data used this module came from 'Old Faithful Data' available from https://www.kaggle.com/janithwanni/old-faithful/data for download & used by the text book PRML.

Mixture of Gaussians

- Mixture of Gaussians: Component densities are Gaussian.



$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(k)p(\mathbf{x}|k)$$

Data Set: The data used this module came from 'Old Faithful Data' available from https://www.kaggle.com/janithwanni/old-faithful/data for download & used by the text book PRML.

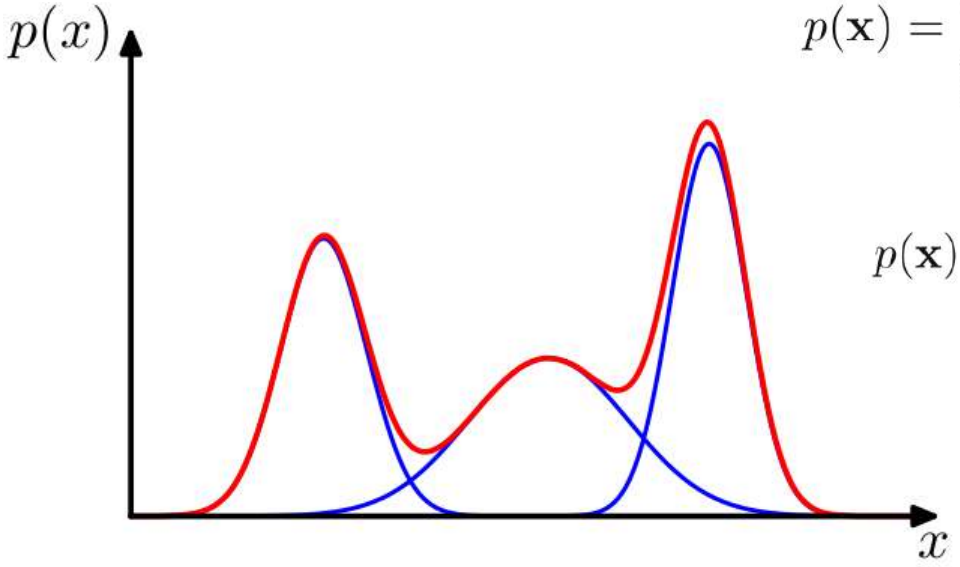
Mixture of Gaussians

- Mixture of Gaussians: Component densities are Gaussian.

Parameters of MoG:

$\pi : \{\pi_1, \ldots, \pi_K\}$,
$\mu : \{\mu_1, \ldots, \mu_K\}$
$\Sigma : \{\Sigma_1, \ldots \Sigma_K\}$

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$



$$p(\mathbf{x}) = \sum_{k=1}^{K} p(k)p(\mathbf{x}|k)$$

Data Set: The data used this module came from 'Old Faithful Data' available from https://www.kaggle.com/janithwanni/old-faithful/data for download & used by the text book PRML.

Log Likelihood

- Mixture of Gaussians: Component densities are Gaussian.

Parameters of MoG:

$\pi : \{\pi_1, \ldots, \pi_K\}$,
$\mu : \{\mu_1, \ldots, \mu_K\}$
$\Sigma : \{\Sigma_1, \ldots \Sigma_K\}$

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$p(\mathbf{X}) = \prod_{n=1}^{N} p(\mathbf{x}_n) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)$$

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(k)p(\mathbf{x}|k)$$

Log likelihood

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

Data Set: The data used this module came from 'Old Faithful Data' available from https://www.kaggle.com/janithwanni/old-faithful/data for download & used by the text book PRML.

γ(zk)

`

Parameters of MoG:

π : {π_1 , . . . , π_K },
μ : {μ_1 , . . . , μ_K }
Σ : {Σ_1, . . . Σ_K }

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x})$$
$$= \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)}$$
$$= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(k)p(\mathbf{x}|k)$$

Data Set: The data used this module came from 'Old Faithful Data' available from https://www.kaggle.com/janithwanni/old-faithful/data for download & used by the text book PRML.

Expectation-Maximization Algorithm

- Method for finding ML solutions for models with Latent Variables
  - Broad applicability for estimating parameters for various models
    - Estimating parameters of a MoG is one application
- Task : To Find ML parameters of MoG

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

- Set the derivative of ln p(X | $\boldsymbol{\pi}$, μ,Σ)
  - w.r.t μ_k to 0 & solve it for μ_k
  - w.r.t Σ_k to 0 & solve it for Σ_k
  - w.r.t $\boldsymbol{\pi}_k$ to 0 & solve it for $\boldsymbol{\pi}_k$ - [Constrained optimization using Lagrangian Multipliers since mixing coefficients sum should be 1]

Expectation-Maximization Algorithm

- We get μ_k as

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

Where

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

$$p(z_k = 1) = \pi_k$$

Expectation-Maximization Algorithm

- We get Σ_k as

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}$$

Where

$$N_k = \sum_{n=1}^{N}\gamma(z_{nk})$$

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

$$p(z_k = 1) = \pi_k$$

Expectation-Maximization Algorithm

- We get π_k as

$$\pi_k = \frac{N_k}{N}$$

Where

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

$$p(z_k = 1) = \pi_k$$

The EM (Expectation Maximization) Algorithm

- Powerful method for finding maximum likelihood solutions for models with latent variables
- **The (EM) algorithm:** A framework to approach maximum likelihood of parameters in statistical models.
  - **E-step** assigns objects to clusters according to the current parameters of probabilistic clusters
  - **M-step** finds the new clustering or parameters that maximize the expected likelihood

- The k-means algorithm has two steps at each iteration:
  - **Expectation Step** (E-step): Given the current cluster centers, each object is assigned to the cluster whose center is closest to the object: An object is *expected to belong to the closest cluster*
  - **Maximization Step** (M-step): Given the cluster assignment, for each cluster, the algorithm *adjusts the center* so that *the sum of distance* from the objects assigned to this cluster and the new center is minimized

EM Algorithm for MoG

1. Start by placing gaussians randomly.
2. Repeat until it converges.
    1. **E step**: With the current means and variances, find the probability of each data point xi coming from each gaussian.
    2. **M step**: Once it computed these probability assignments it will use these numbers to re-estimate the gaussians' mean and variance to better fit the data points.

EM for Gaussian Mixtures

Given a Gaussian mixture model, the goal is to maximize the likelihood function with respect to the parameters (comprising the means and covariances of the components and the mixing coefficients).

1. Initialize the means $\boldsymbol{\mu}k$, covariances $\boldsymbol{\Sigma}k$ and mixing coefficients πk, and evaluate the initial value of the log likelihood.

2. **E step**. Evaluate the responsibilities using the current parameter values

- For fixed values of μ_1, $\mu_{2, .. \mu k}$ and Σ_1, Σ_2, .. Σk and Π_1 ….$\Pi_{k,}$ Compute soft assignments per data point (allocating the probability of that data point belonging to each one of the clusters).

$$\gamma(z_k) = \frac{\pi_k N(x_i \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_i \mid \mu_j, \Sigma_j)}$$

  - For each observation i, vector γ (aka responsibility vector) is ($\gamma i1$, $\gamma i2$,… γiK), *wh*ere K is the total number of clusters, or often referred to as the number of components.
  - The cluster responsibilities for a single data point i should sum to 1.

EM for Gaussian Mixtures

3. M step. Re-estimate the parameters using the current responsibilities

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n \quad \text{, where} \quad N_k = \sum_{n=1}^{N}\gamma(z_{nk})$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}$$

$$\pi_k = \frac{N_k}{N}$$

4. Evaluate the log likelihood

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N}\ln\left\{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right\}$$

and check for convergence of either the parameters or the log likelihood. If the convergence criterion is not satisfied return to step 2.

Expectation-Maximization Algorithm

To Estimate:

M-Step

Initialize $\boldsymbol{\pi}$, μ,Σ *and also evaluate the log likelihood*

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\left(\mathbf{x}_n-\boldsymbol{\mu}_k^{\text{new}}\right)\left(\mathbf{x}_n-\boldsymbol{\mu}_k^{\text{new}}\right)^{\mathrm{T}}$$

$$\pi_k^{\text{new}} = \frac{N_k}{N}$$

$$N_k = \sum_{n=1}^{N}\gamma(z_{nk})$$

Perform E-Step Given $\gamma(z_k)$

E-Step

$$\gamma(z_k) \equiv p(z_k=1|\mathbf{x}) = \frac{p(z_k=1)p(\mathbf{x}|z_k=1)}{\sum_{j=1}^{K}p(z_j=1)p(\mathbf{x}|z_j=1)} = \frac{\pi_k\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K}\pi_j\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)}$$

Perform M-Step Given $\boldsymbol{\pi}$, μ,Σ

$$p(z_k=1)=\pi_k$$

Repeat Until Convergence - [Use log likelihood / parameters to decide this]

Complete vs incomplete data

- in general, the problem would be trivial if we had access to the complete data
- Gaussian mixture of *C* components
  - parameters $\theta = \{(\pi_1, \mu_1, \Sigma_1), \ldots, (\pi_C, \mu_C, \Sigma_C)\}$
- given the complete data D_c, we only need to split the training set according to the labels z_i
- $D^1 = \{x_i | z_i = 1\}$ $D^2 = \{x_i | z_i = 2\}, \ldots$, $D^C = \{x_i | z_i = C\}$
- and solve, for each c,

Latent Variables

The primary role of the latent variables is to allow a complicated distribution over the observed variables to be represented in terms of a model constructed from simpler (typically exponential family) conditional distributions.

Mixture Density

`

Parameters of MoG:

$\pi : \{\pi_1, \ldots, \pi_K\}$,
$\mu : \{\mu_1, \ldots, \mu_K\}$
$\Sigma : \{\Sigma_1, \ldots \Sigma_K\}$

X	Assuming Latent Variable z for 2 mixture components.	
	$z_{(1)}$	$z_{(2)}$
$x^{(1)}$	0	1
$x^{(2)}$	1	0
...		
$x^{(N)}$	1	0

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(k)p(\mathbf{x}|k)$$

$P(z_k=1) = \pi_k$

$0 \le \pi_k \le 1$ and $\Sigma_k \pi_k = 1$

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

Data Set: The data used this module came from 'Old Faithful Data' available from https://www.kaggle.com/janithwanni/old-faithful/data for download & used by the text book PRML.

Component Density

Joint Probability $p(x, z) = p(x|z)p(z)$

Parameters of MoG:

$\pi : \{\pi_1, \ldots, \pi_K\}$,
$\mu : \{\mu_1, \ldots, \mu_K\}$
$\Sigma : \{\Sigma_1, \ldots \Sigma_K\}$

Given a particular z_k :

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}$$

Data Set: The data used this module came from 'Old Faithful Data' available from https://www.kaggle.com/janithwanni/old-faithful/data for download & used by the text book PRML.

Marginal Distribution of x

Joint Probability $p(x, z) = p(x|z)p(z)$

Parameters of MoG:

$\pi : \{\pi_1, \ldots, \pi_K\}$,
$\mu : \{\mu_1, \ldots, \mu_K\}$
$\Sigma : \{\Sigma_1, \ldots \Sigma_K\}$

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}$$

Joint probability of all observations **X**

$$p(\mathbf{X}) = \prod_{n=1}^{N} p(\mathbf{x}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}$$

Data Set: The data used this module came from 'Old Faithful Data' available from https://www.kaggle.com/janithwanni/old-faithful/data for download & used by the text book PRML.

γ(zk)

`

Parameters of MoG:

π : {π_1 , . . . , π_K },
μ : {μ_1 , . . . , μ_K }
Σ : {Σ_1, . . . Σ_K }

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x})$$
$$= \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)}$$
$$= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(k)p(\mathbf{x}|k)$$

Data Set: The data used this module came from 'Old Faithful Data' available from https://www.kaggle.com/janithwanni/old-faithful/data for download & used by the text book PRML.

Alternative Formulation

Observed Data : X [$x^{(1)}$, $x^{(2)}$,..... $x^{(N)}$]

Latent Variables : Z [$z^{(1)}$, $z^{(2)}$,..... $z^{(N)}$]

Goal : Find ML solution to model parameters

	X			Z	
1	x_{11}	...	x_{1D}	z_{11}	z_{12}
2	x_{21}	...	x_{2D}	z_{21}	z_{22}
...				...	
N	x_{N1}	...	x_{ND}	z_{N1}	z_{N2}

Let us suppose Z is given :

The form of the **complete data log likelihood** becomes

ln p(X, Z|θ)

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln \left\{ \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \right\}$$

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

Alternative Formulation

Form of p(X,Z |θ) [For Mixture of Gaussians]:

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}$$

	X			Z	
1	x_{11}	...	x_{1D}	z_{11}	z_{12}
2	x_{21}	...	x_{2D}	z_{21}	z_{22}
...				...	
N	x_{N1}	...	x_{ND}	z_{N1}	z_{N2}

ln p(X, Z|θ)

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln \left\{ \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \right\}$$

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

Alternative Formulation

Form of p(X,Z |θ) [For Mixture of Gaussians]:

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}$$

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}$$

	X			Z	
1	x_{11}	...	x_{1D}	z_{11}	z_{12}
2	x_{21}	...	x_{2D}	z_{21}	z_{22}
...				...	
N	x_{N1}	...	x_{ND}	z_{N1}	z_{N2}

ln p(X, Z|θ)

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln \left\{ \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \right\}$$

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

Alternative Formulation

Form of p(X,Z |θ) [For Mixture of Gaussians]:

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}$$

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}$$

	X			Z	
1	x_{11}	...	x_{1D}	z_{11}	z_{12}
2	x_{21}	...	x_{2D}	z_{21}	z_{22}
...				...	
N	x_{N1}	...	x_{ND}	z_{N1}	z_{N2}

We know that, z_{ik} is 1 only for one of the components for ith data.

Alternative Formulation

Form of p(X,Z |θ) [For Mixture of Gaussians]:

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}$$

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}$$

	X			Z	
1	x_{11}	...	x_{1D}	0	1
2	x_{21}	...	x_{2D}	**1**	0
...				...	
N	x_{N1}	...	x_{ND}	**1**	0

We know that, z_{ik} is 1 only for one of the components for ith data.

Let us suppose we need the ML estimate for μ_k.

Alternative Formulation

Form of p(X,Z |θ) [For Mixture of Gaussians]:

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}$$

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}$$

	X			Z	
1	x_{11}	...	x_{1D}	0	1
2	x_{21}	...	x_{2D}	**1**	0
...				...	
N	x_{N1}	...	x_{ND}	**1**	0

We know that, z_{ik} is 1 only for one of the components for ith data.

Let us suppose we need the ML estimate for μ_k.

ML estimate for parameters is exactly for the parameters for a single gaussian for the fraction of examples with their z_{ik}=1

Alternative Formulation

Form of p(X,Z |θ) [For Mixture of Gaussians]:

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}$$

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}$$

	X			Z	
1	x_{11}	...	x_{1D}	0	1
2	x_{21}	...	x_{2D}	**1**	0
...				...	
N	x_{N1}	...	x_{ND}	**1**	0

The ML estimates takes the same form as earlier:

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \qquad \boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}$$

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}) \qquad \pi_k = \frac{N_k}{N}$$

Note: Here ***γ*** takes the same form as earlier
Since we have z, the solution is still closed form

Alternative Formulation

We have actually taken the discussion assuming, Z is given

- It is not given indeed !!!
  - Our knowledge of Z is only from P(Z|X,θ)

	X			Z	
1	x_{11}	...	x_{1D}	0	1
2	x_{21}	...	x_{2D}	**1**	0
...				...	
N	x_{N1}	...	x_{ND}	**1**	0

Alternative Formulation

We have actually taken the discussion assuming, Z is given

- It is not given indeed !!!
  - Our knowledge of Z is only from P(Z|X,θ)
  - Instead maximizing ln p(X,Z |θ), do the following
    - Maximize expectation of complete log likelihood under P(Z|X,θ) !!!
      - That is E_z [ln p(X,Z |θ)] denoted as $Q(\theta,\theta^{old})$

	X			Z	
1	x_{11}	...	x_{1D}	0	1
2	x_{21}	...	x_{2D}	1	0
...				...	
N	x_{N1}	...	x_{ND}	1	0

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

General EM Algorithm

1. Choose an initial setting for the parameters θ^{old} .
2. E-Step: Evaluate $p(Z|X,\theta^{old})$
3. M-Step:

   Evaluate θ^{new} given by

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$$

   Where

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

1. Check for convergence of either the log likelihood or the parameter values. If the convergence criterion is not satisfied, then let

$\theta^{old} \leftarrow \theta^{new}$

and return to step-2

General EM Algorithm - for MoG

1. Choose initial values for parameters μ^{old} , Σ^{old} and π^{old}
2. E-Step : Compute $\gamma(z_{nk})$
3. M-Step : Keep $\gamma(z_{nk})$ fixed, and maximize $E_z[\ln p(X,Z|\theta)]$ for μ_k , Σ_k and π_k, to get μ^{new} , Σ^{new} and π^{new} ,
4. Repeat E & M until convergence

Expectation-Maximization Algorithm

To Estimate:

Initialize $\boldsymbol{\pi}$, μ,Σ *and also evaluate the log likelihood*

M-Step

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\left(\mathbf{x}_n-\boldsymbol{\mu}_k^{\text{new}}\right)\left(\mathbf{x}_n-\boldsymbol{\mu}_k^{\text{new}}\right)^{\mathrm{T}}$$

$$\pi_k^{\text{new}} = \frac{N_k}{N}$$

$$N_k = \sum_{n=1}^{N}\gamma(z_{nk})$$

Perform E-Step Given $\gamma(z_k)$

E-Step

$$\gamma(z_k) \equiv p(z_k=1|\mathbf{x}) = \frac{p(z_k=1)p(\mathbf{x}|z_k=1)}{\sum_{j=1}^{K}p(z_j=1)p(\mathbf{x}|z_j=1)} = \frac{\pi_k\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K}\pi_j\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)}$$

Perform M-Step Given $\boldsymbol{\pi}$, μ,Σ

$$p(z_k=1)=\pi_k$$

Repeat Until Convergence

Thank You!

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Contact Session #4
Unsupervised Learning & Association Rule Mining

PCAM ZC221

Raja vadhana P
BITS - CSIS

Course Plan

M1	Introduction to Unsupervised Learning
M2	K-Means Algorithm
M3	EM Algorithm
M4	Hierarchical Clustering
M5	Density Based Clustering
M6	Assessing Quality of Clustering
M7	Association Rule Mining
M8	Time series Prediction and Markov Process

Module 4 : Hierarchical Clustering

A Introduction to hierarchical clustering

B Agglomerative Clustering Vs Divisive Clustering

C Distance Measures

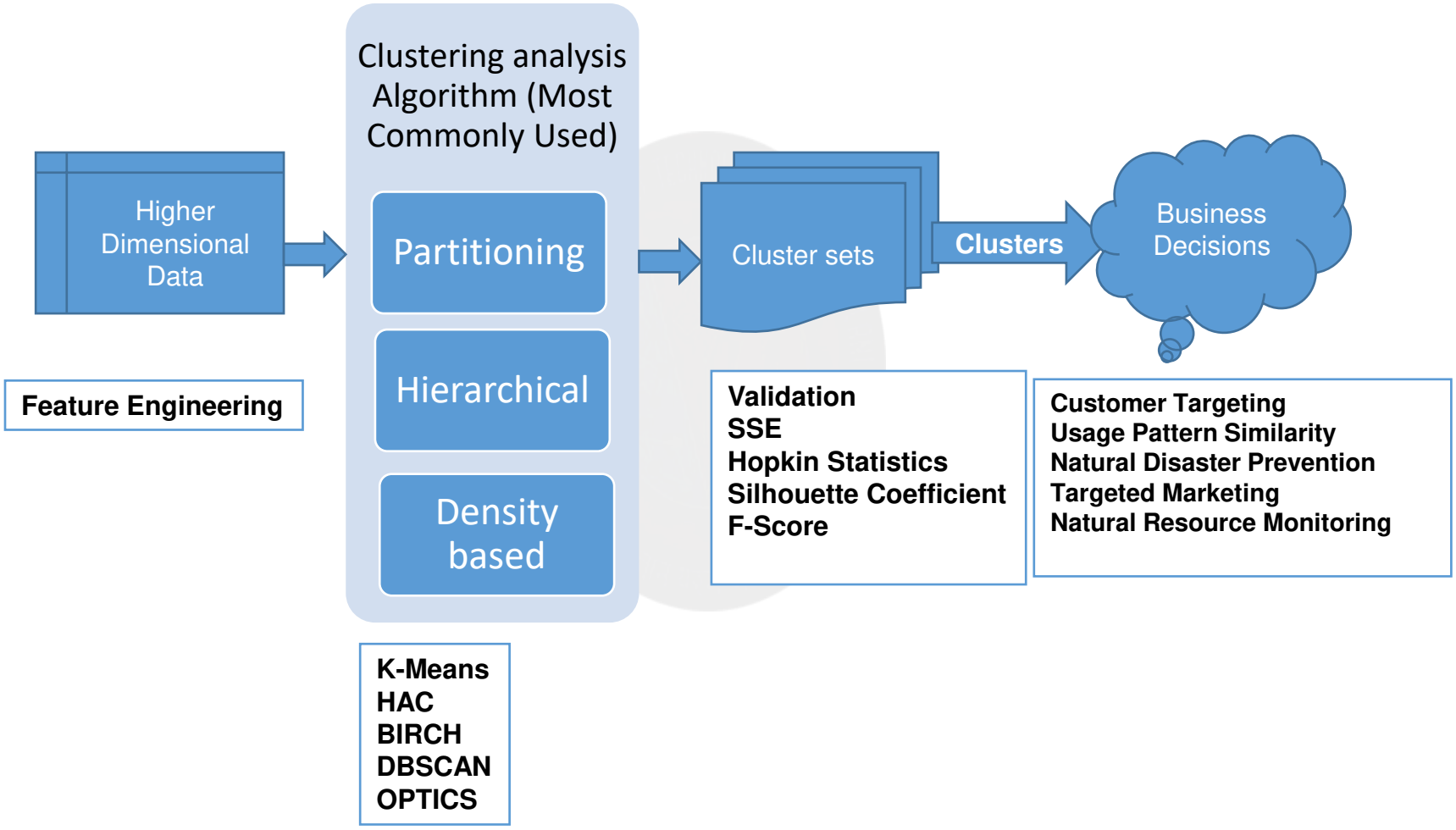
D Algorithms

Agenda

Learning Objective : Hierarchical clustering

- Identify the application specific need
- Choose the right inter-cluster dissimilarity measure
- Compare the working of agglomerative algorithms – Demo Session

Clustering Techniques



Identify the activity required?

Domain		
People Management	Divide the registered students by department name	Divide the students having similar learning pattern
Canvas LMS	Identify the most effective set of students who use the canvas in effective way	Segment the students based on their preferred mode of communication(Discussion forum/ Chat/Inbox) and find the popular one
Banking	Group the customers under Premium category and find the right age group to target	Segment the customers to target for marketing loan policies


INTERNET
Facebook News to forge business ties with media outlets, says Mark Zuckerberg

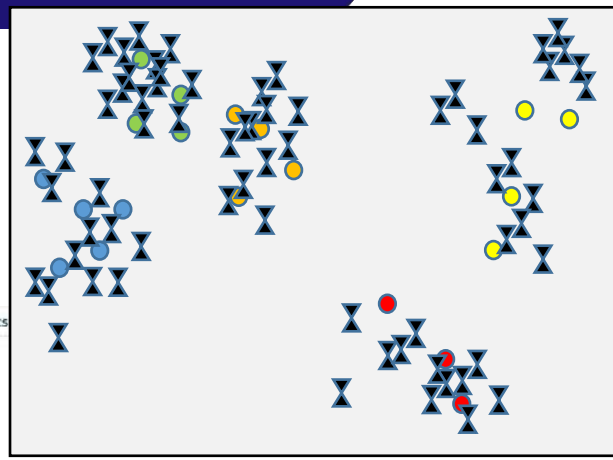

INTERNET
Indians binge watch for over 8 hours a week: report

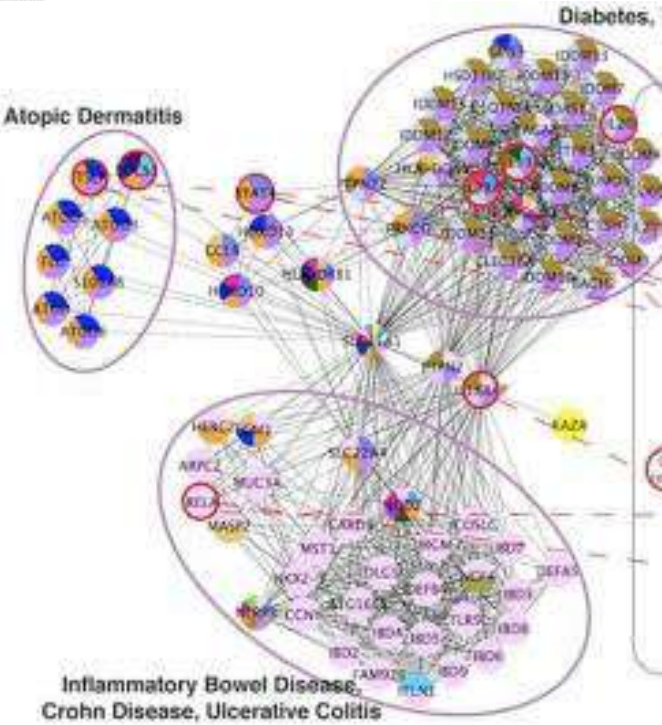

INTERNET
Facebook agrees to pay UK fine over Cambridge Analytica scandal

INTERNET
Facebook launches healthcare tool for check-up reminders


INDUSTRY
Amazon, Flipkart asked to disclose names of top 5 sellers, capital structure, inventory details


FOOD
Shameera PR has over a million subscribers on her culinary channel





Use Case – 1 : Customer Engagement

Customer Segmentation

Objective :

To segment the customer for targeted marketing

Use Case – 2 : Content Management

Document Clustering

Objective :

To group related news article for study

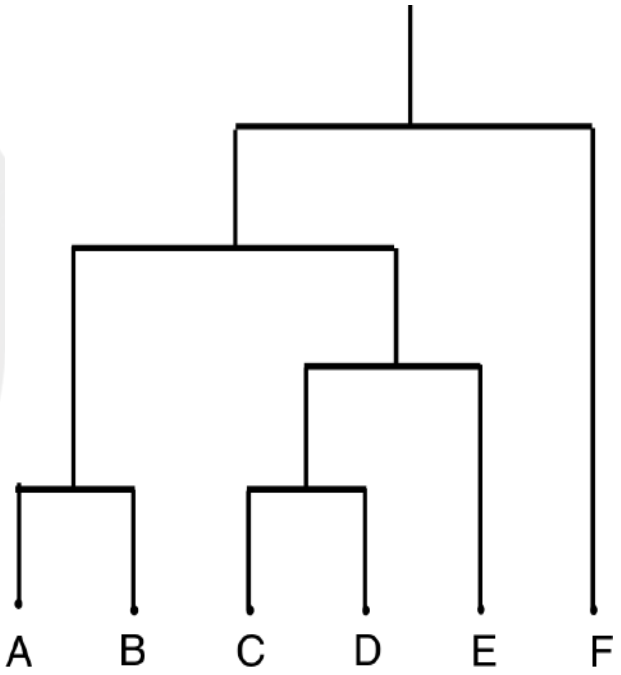
Hierarchical Clustering

A set of nested clusters organized as a hierarchical tree formed by the decomposition of the objects.

Cluster Objective function : Local

Bottom Up Approach : Agglomerative

Top Down Approach : Devisive



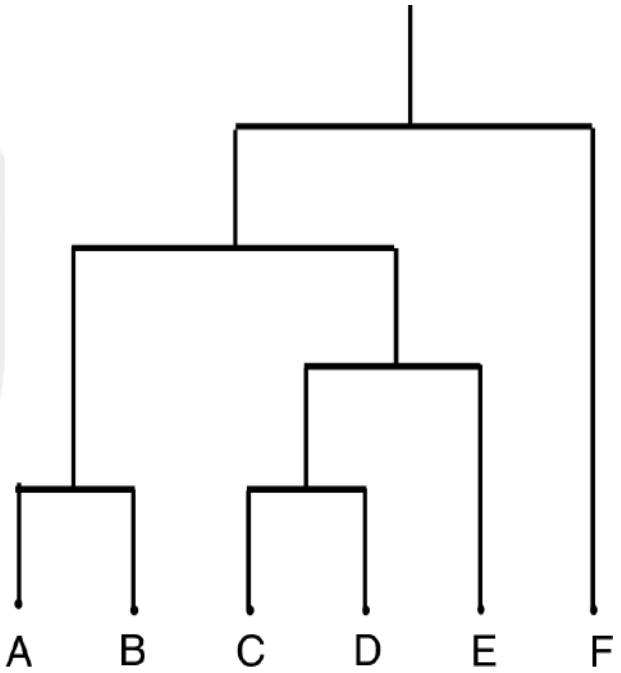
C: Hierarchical Clustering

Clusters
- Complete
- Singleton

Cost Evaluation
- Single Linkage
- Complete Linkage
- Average Linkage
- MST Single Linkage

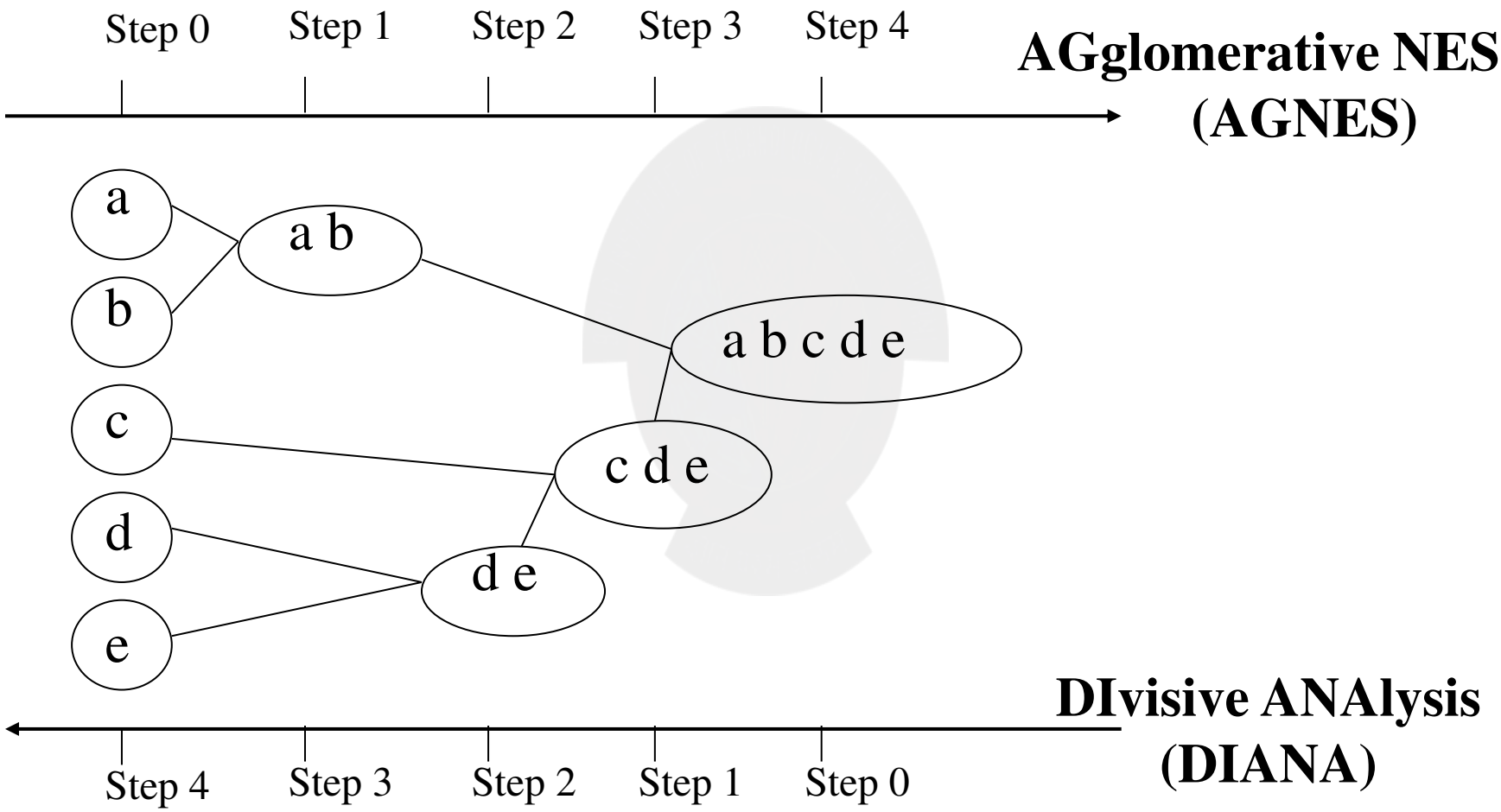
Cluster form & Update
- Partition
- Merge

Iterate till Termination Condition
- Singleton
- Complete Single Cluster



Hierarchical Clustering

Agglomerative Vs Divisive Approach

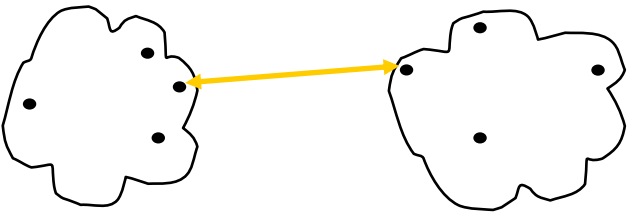


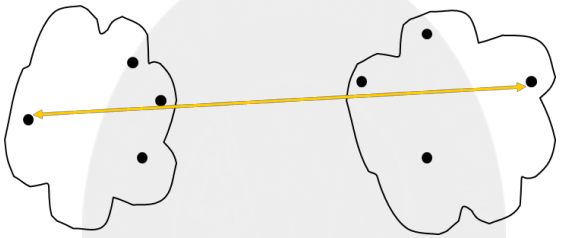
Hierarchical Clustering : Cost Function

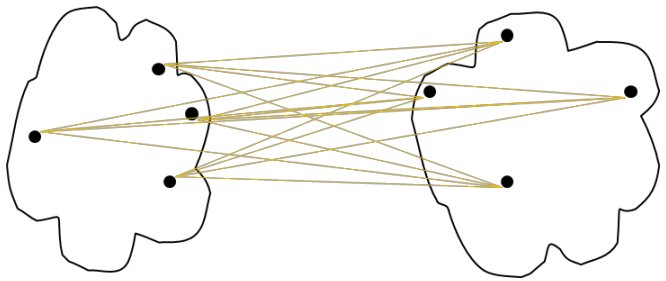
METHOD — **DISTANCE**

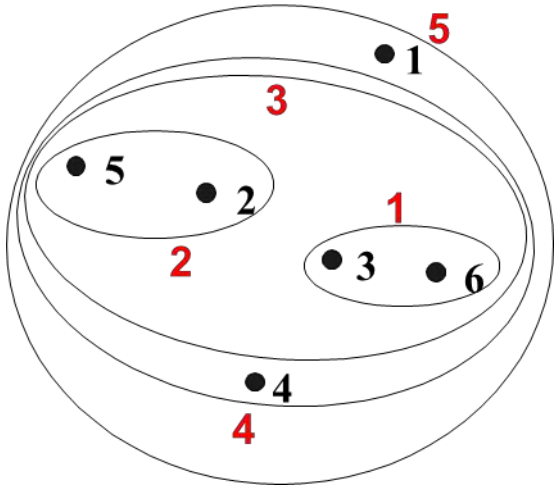
- Single / Connectedness : MIN — Nearest Neighbour Clustering Algorithm
- Complete / Diameter : MAX — Farthest Neighbour/ Clique
- Average Method : AVG

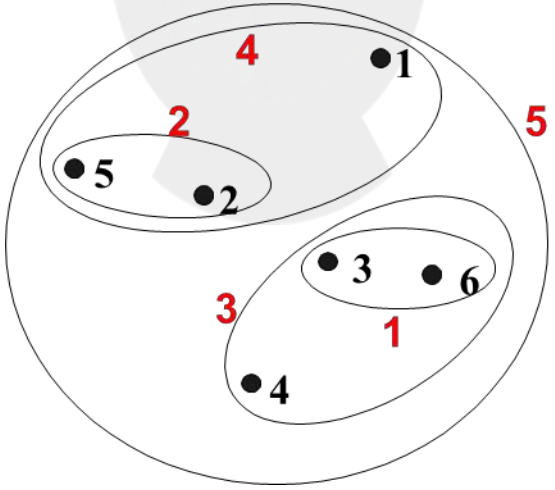
$$D_{avg}(C_i, C_j) = \frac{1}{|C_i| \times |C_j|} \sum_{x \in C_i, y \in C_j} d(x, y)$$

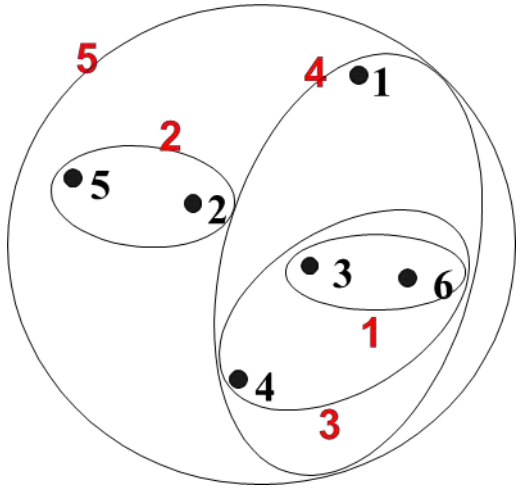












Nearest Neighbour Algorithm

Algorithm

1. Find the proximity Matrix
2. Consider each object as one cluster
3. Find the MIN distance and pair those objects as a single cluster
4. Delete the original objects which are merged and replace with the paired node
5. Re-compute the proximity matrix by **Linkage Rule**
6. Iterate over the steps till a single cluster is produced
7. Draw the Dendrogram and the graph representation

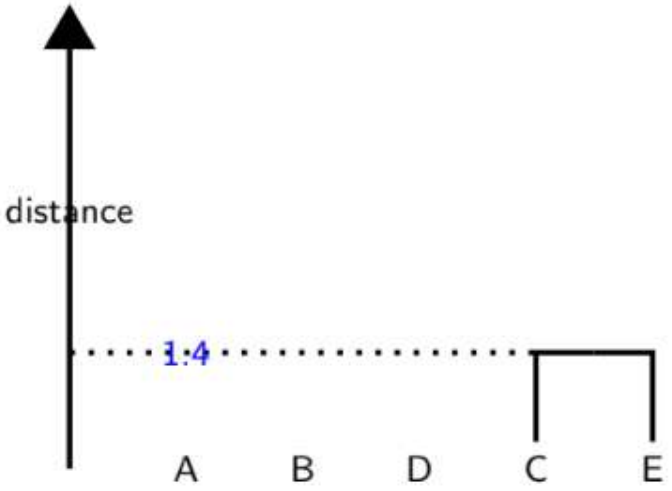
Nearest Neighbour Clustering

Single Linkage

The cluster proximity is the distance between closest pair of data points from different clusters.

$$Dist_{min}(C_i, C_j) = \min_{p_i \in C_i, p_j \in C_j} |p_i - p_j|$$

Movies	Imdb rating	Times of India (Scaled to 1-10)
A	2	2
B	5	8
C	2	4
D	4	3
E	3	5



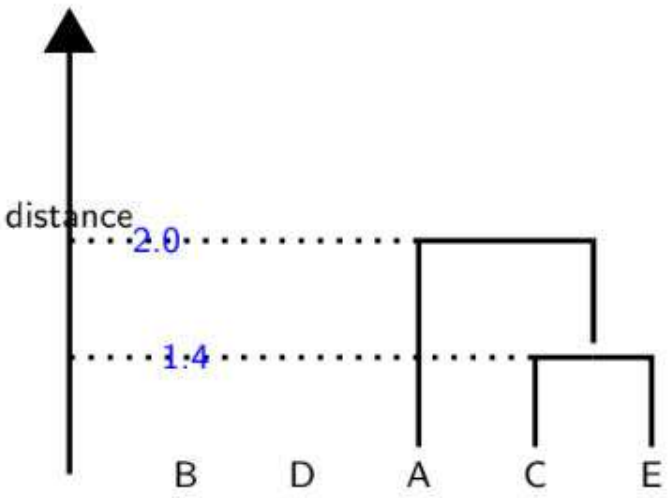
	A	B	C	D	E
A	0	6.7	2.0	2.2	3.2
B		0	5.0	5.1	3.6
C			0	2.2	1.4
D				0	2.2
E					0

Nearest Neighbour Clustering

$Dist_{(A, CE)} = \min\{AC, AE\} = \min\{2.0, 3.2\}$

$Dist_{(B, CE)} = \min\{BC, BE\} = \min\{5.0, 3.6\}$

$Dist_{(D, CE)} = \min\{DC, DE\} = \min\{2.2, 2.2\}$



Movies	Imdb rating	Times of India (Scaled to 1-10)
A	2	2
B	5	8
C	2	4
D	4	3
E	3	5

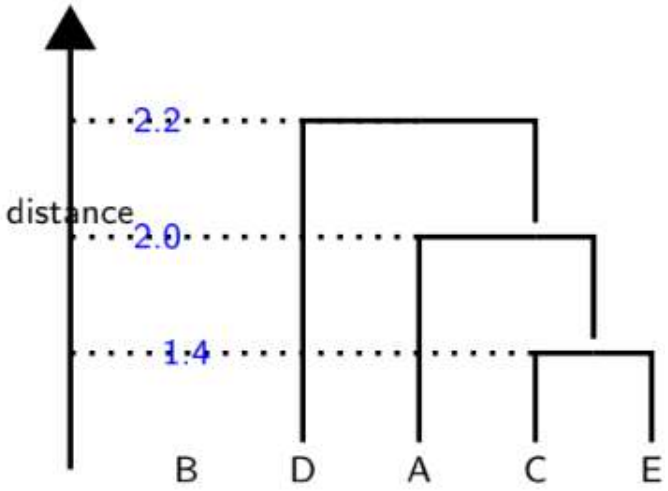
	A	B	C	D	E
A	0	6.7	2.0	2.2	3.2
B		0	5.0	5.1	3.6
C			0	2.2	1.4
D				0	2.2
E					0

	A	B	D	CE
A	0	6.7	2.2	2.0
B		0	5.1	3.6
D			0	2.2
CE				0

Nearest Neighbour Clustering

$Dist_{(B, ACE)} = \min\{AB, BC, BE\}$

$Dist_{(D, ACE)} = \min\{AD, CD, DE\}$



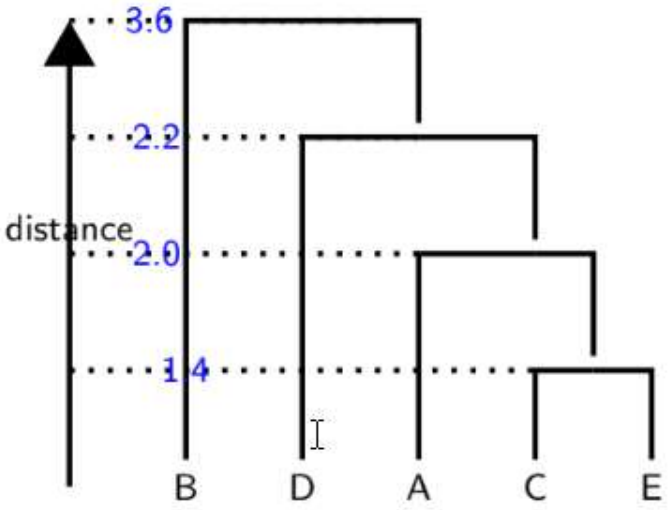
Movies	Imdb rating	Times of India (Scaled to 1-10)
A	2	2
B	5	8
C	2	4
D	4	3
E	3	5

	A	B	D	CE
A	0	6.7	2.2	2.0
B		0	5.1	3.6
D			0	2.2
CE				0

	B	D	ACE
B	0	5.1	3.6
D		0	2.2
ACE			0

Nearest Neighbour Clustering

Movies	Imdb rating	Times of India (Scaled to 1-10)
A	2	2
B	5	8
C	2	4
D	4	3
E	3	5



	B	D	ACE
B	0	5.1	3.6
D		0	2.2
ACE			0

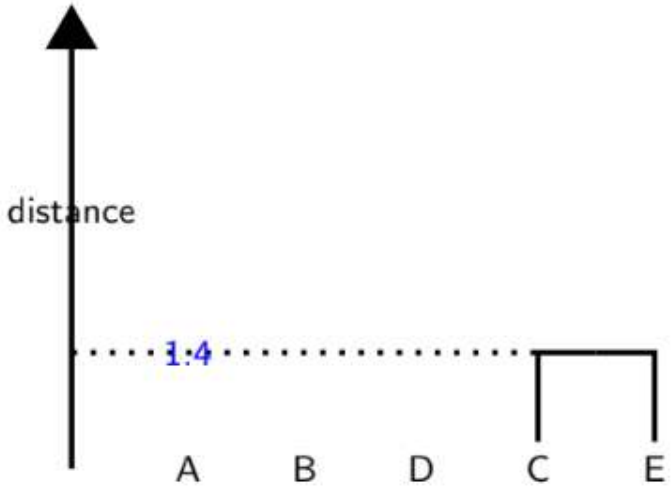
	B	DACE
B	0	3.6
DACE		0

Farthest Neighbour Clustering

- The cluster proximity is the distance between farthest pair of data points from different clusters.

$$Dist_{max}(C_i, C_j) = \max_{p_i \in C_i, p_j \in C_j} |p_i - p_j|$$

Movies	Imdb rating	Times of India (Scaled to 1-10)
A	2	2
B	5	8
C	2	4
D	4	3
E	3	5



	A	B	C	D	E
A	0	6.7	2.0	2.2	3.2
B		0	5.0	5.1	3.6
C			0	2.2	1.4
D				0	2.2
E					0

Complete Linkage

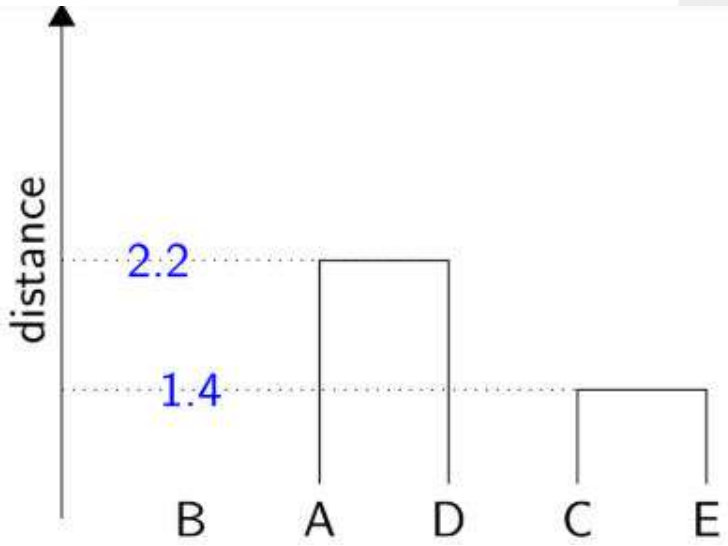
$Dist_{(A, CE)} = \max\{AC, AE\} = \max\{2.0, 3.2\}$

$Dist_{(B, CE)} = \max\{BC, BE\} = \max\{5.0, 3.6\}$

$Dist_{(D, CE)} = \max\{DC, DE\} = \max\{2.2, 2.2\}$

SSE -1 = {A} ,{B}, {CDE} = 0+0+4 = 4

SSE -2 = {A,D}, {B}, {CE} =2.5+0+1 = 3.5



Cohesion is a measure of compactness which determine how closely related the objects are within the cluster.

Movies	Imdb rating	Times of India (Scaled to 1-10)
A	2	2
B	5	8
C	2	4
D	4	3
E	3	5

	A	B	C	D	E
A	0	6.7	2.0	2.2	3.2
B		0	5.0	5.1	3.6
C			0	2.2	1.4
D				0	2.2
E					0

	A	B	D	CE
A	0	6.7	2.2	3.2
B		0	5.1	5.0
D			0	2.2
CE				0

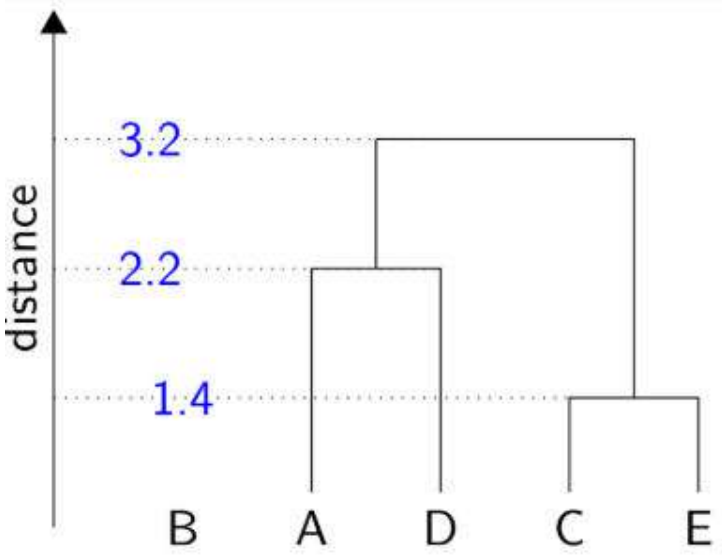
Complete Linkage

$Dist_{(B, AD)} = \max(AB, BD) = \max(6.7, 2.2) = 6.7$

$Dist_{(B, CE)} = \max(BC, BE) = \max(5.0, 3.6) = 5.0$

$Dist_{(AD, CE)} = \max(AC, AE, DC, DE)$

$= \max(2.0, 3.2, 2.2, 2.2) = 3.2$



Movies	Imdb rating	Times of India (Scaled to 1-10)
A	2	2
B	5	8
C	2	4
D	4	3
E	3	5

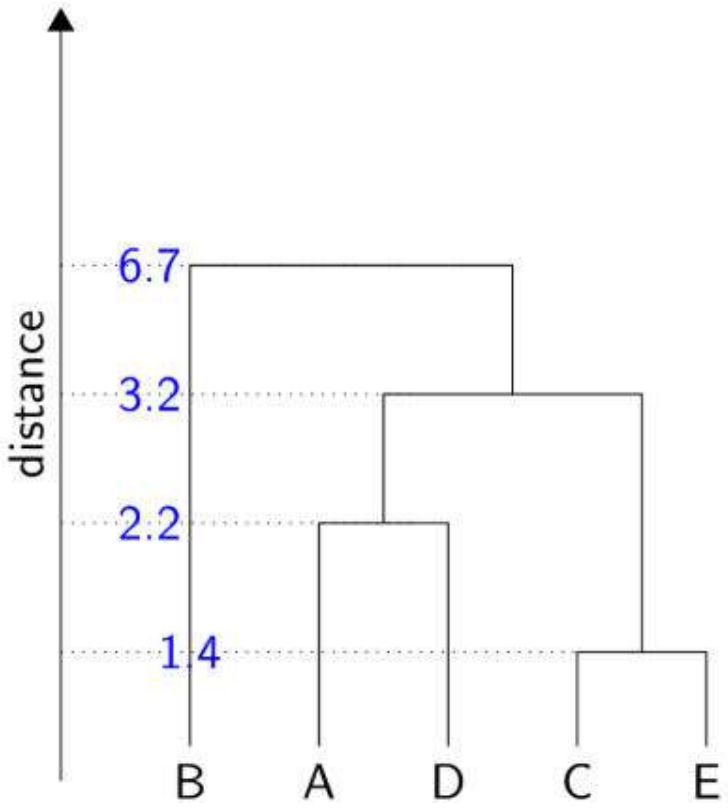
	A	B	D	CE
A	0	6.7	2.2	3.2
B		0	5.1	5.0
D			0	2.2
CE				0

	B	AD	CE
B	0	6.7	5.0
AD		0	3.2
CE			0

Complete Linkage

$Dist_{(B, ADCE)} = \max(AB, BC, BD, BE)$

$= \max(6.7, 5.0, 5.1, 3.6) = 6.7$



Movies	Imdb rating	Times of India (Scaled to 1-10)
A	2	2
B	5	8
C	2	4
D	4	3
E	3	5

	B	AD	CE
B	0	6.7	5.0
AD		0	3.2
CE			0

	B	ADCE
B	0	6.7
ADCE		0

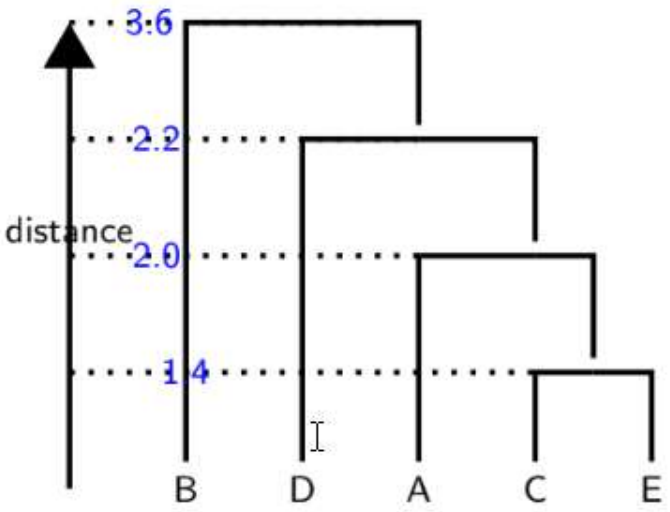
Nearest Neighbour Clustering

Single Linkage

The cluster proximity is the distance between closest pair of data points from different clusters.

$$Dist_{min}(C_i, C_j) = \min_{p_i \in C_i, p_j \in C_j} |p_i - p_j|$$

Movies	Imdb rating	Times of India (Scaled to 1-10)
A	2	2
B	5	8
C	2	4
D	4	3
E	3	5



	A	B	C	D	E
A	0	6.7	2.0	2.2	3.2
B		0	5.0	5.1	3.6
C			0	2.2	1.4
D				0	2.2
E					0

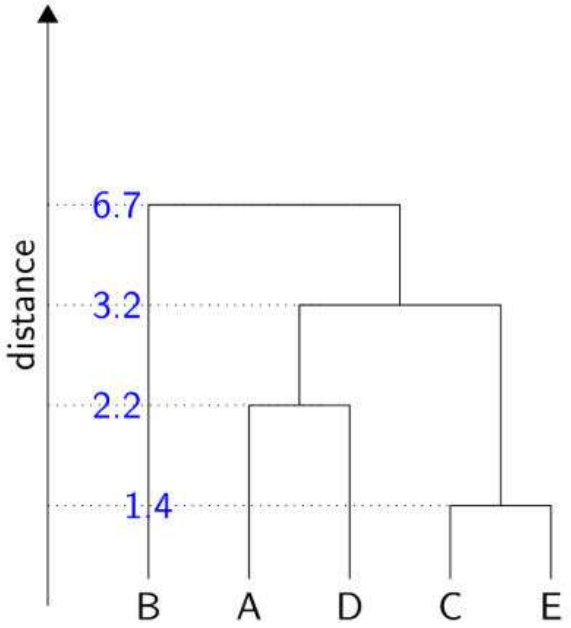
Farthest Neighbour Clustering

Complete Linkage

- The cluster proximity is the distance between farthest pair of data points from different clusters.

$$Dist_{max}(C_i, C_j) = \max_{p_i \in C_i, p_j \in C_j} |p_i - p_j|$$

Movies	Imdb rating	Times of India (Scaled to 1-10)
A	2	2
B	5	8
C	2	4
D	4	3
E	3	5



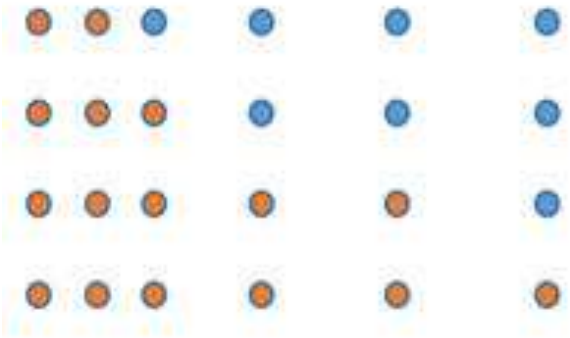
	A	B	C	D	E
A	0	6.7	2.0	2.2	3.2
B		0	5.0	5.1	3.6
C			0	2.2	1.4
D				0	2.2
E					0

Hierarchical Clustering

Challenges

+ Greedy & Deterministic
+ Devoid of initialization bias

- No Undo possible once clustered
- **Chaining Effect**
- Outlier Effect
- Difficulty in interpreting Dendrogram for larger size



Hierarchical Clustering

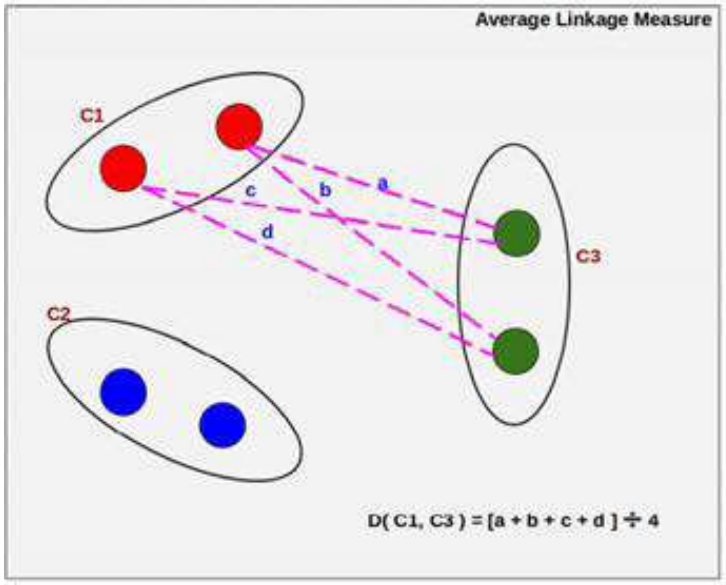
Challenges

+ Greedy & Deterministic
+ Devoid of initialization bias

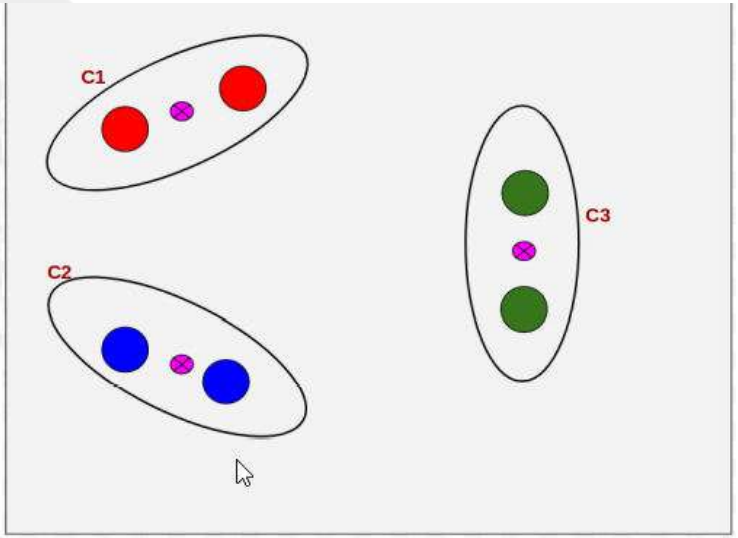
- No Undo possible once clustered
- Chaining Effect
- **Outlier Effect**
- Difficulty in interpreting Dendrogram for larger size

Hierarchical Clustering

Average Linkage

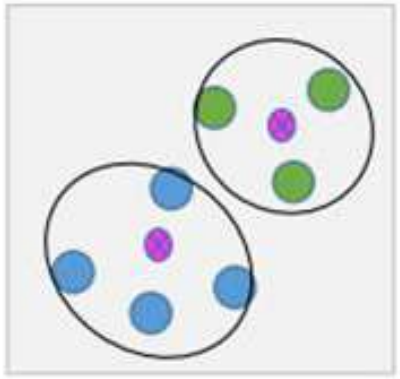


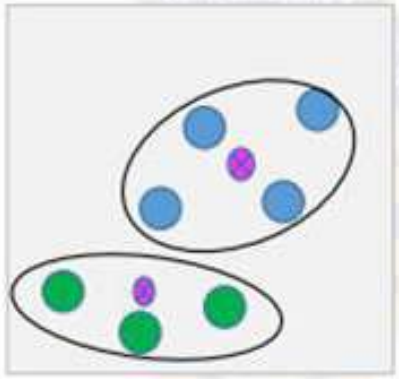
Mean Linkage - Centroid

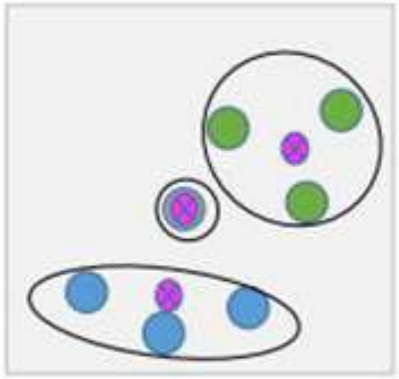


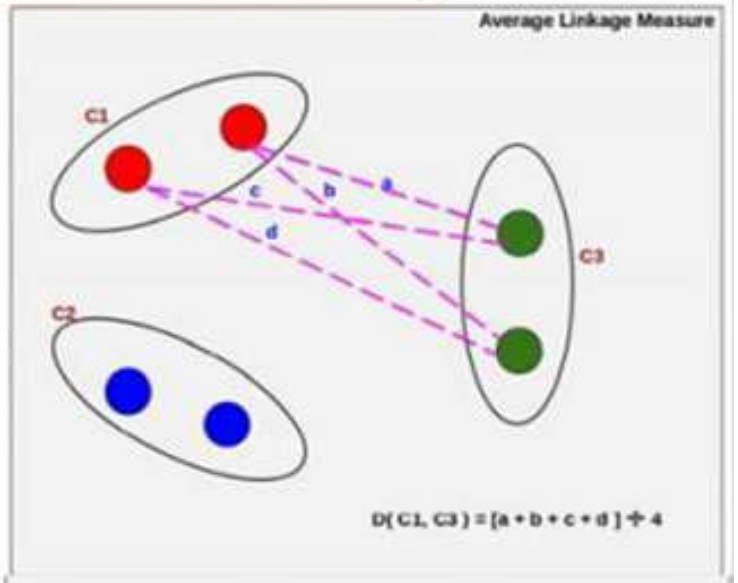
Hierarchical Clustering

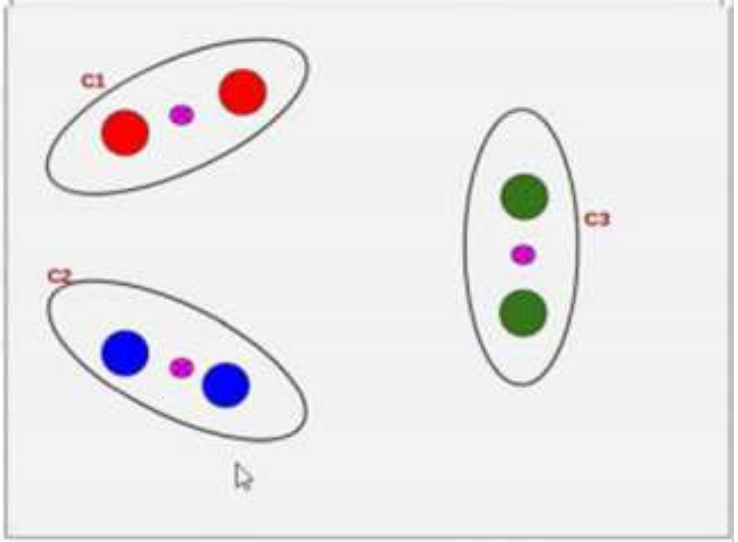
Wards Algorithm : Idea









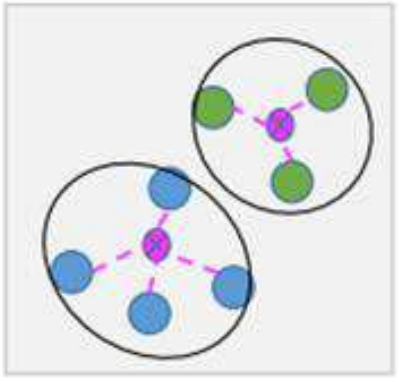


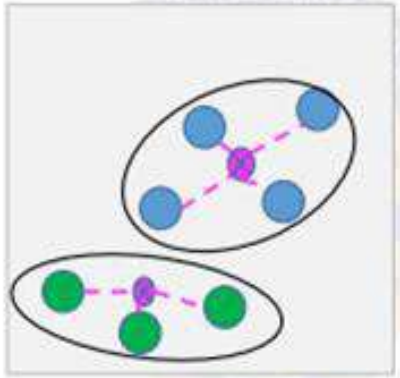
Hierarchical Clustering

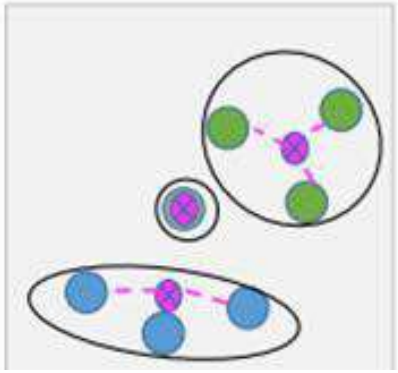
Wards Algorithm : Idea

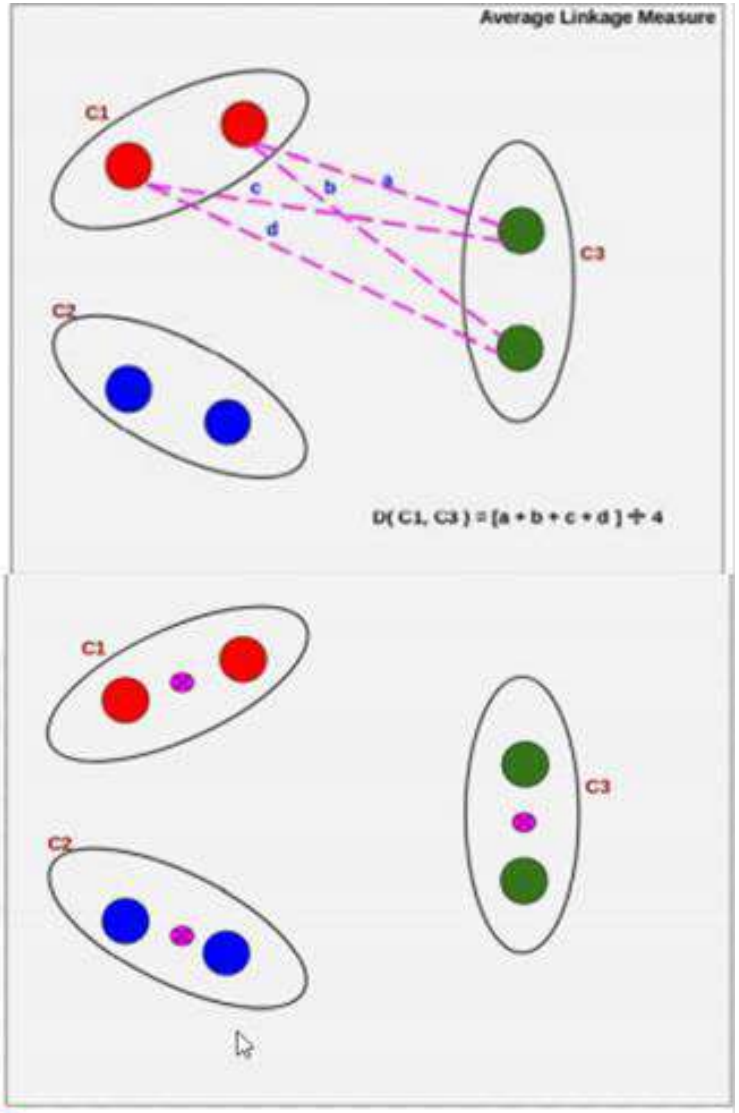
Decision Criteria: Minimum Total within cluster variance

Cluster Distance : Distance between the centroids









Clustering Tendency

The below is included for completion to support the lab illustrations as additional note. More on the clustering quality measures will be covered in your module 6

Need: Clustering analysis on a data set is meaningful only when there is a non-random structure in the data

Metric: Hopkins Statistic is a spatial statistic that tests the spatial randomness of a variable as distributed in a space

$X_i = \min_{v \in D} \{ dist(p_i, v) \}$

$Y_i = \min_{v \in D,\, v \neq q_i} \{ dist(q_i, v) \}$

$$H = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i}$$

Interpretation:

If D is uniformly distributed, H ≈ 0.5

If D is highly skewed, H ≈ 0

If H > 0.5 then D then significant clusters may be found.

Clustering Quality

Intrinsic Criterion

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)} & \text{if } a(i) < b(i) \\ 0 & \text{if } a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1 & \text{if } a(i) > b(i) \end{cases}$$

Metric: Silhouette Coefficient

$$S(p) = \frac{b(p) - a(p)}{\max(b(p), a(p))}$$

$$a(p) = \frac{\sum_{o` \in Ci,\, p \neq o`} dist(p, o`)}{||Ci|| - 1}$$

$$b(p) = \min_{Cj:0<j<=k,\, j \neq i} \frac{\sum_{o` \in Cj,\, p \neq o`} dist(p, o`)}{||Cj||}$$

a(p) reflects the compactness of the cluster to which an object belongs. LTB

It's the average distance between p and all other objects in the cluster to which p belongs

b(p) captures the degree to which an object is separated from other clusters. HTB

It's the minimum average distance from p and all clusters to which p does not belong

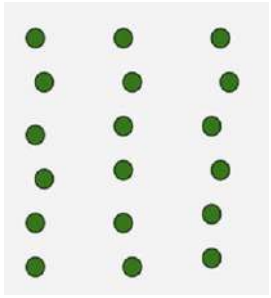
This slide is included for completion to support the lab illustrations as additional note. More on the clustering quality measures will be covered in your module 6

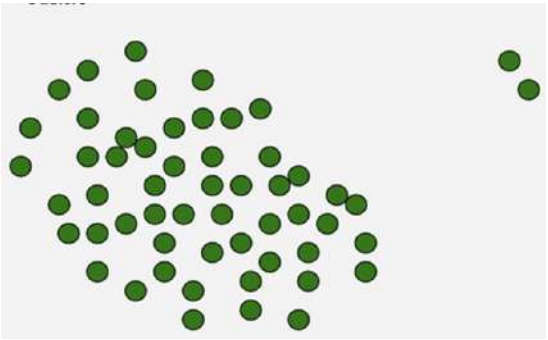
Interpretation:

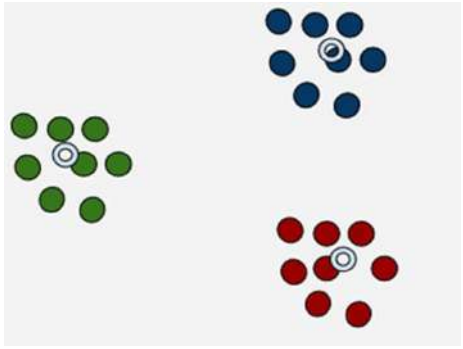
The value of the silhouette coefficient is between −1 and 1

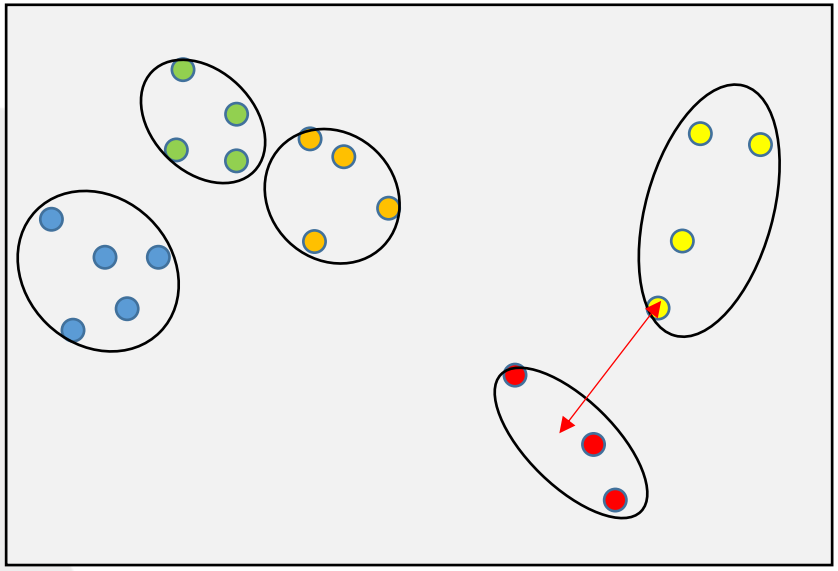
If S→1, Clusters are compact and well separated

Intuition









$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)} & \text{if } a(i) < b(i) \\ 0 & \text{if } a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1 & \text{if } a(i) > b(i) \end{cases}$$

Next Session

Density Based Clustering

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

on #5
ing &
lining

PCAM ZC221

Raja vadhana P
BITS - CSIS

Course Plan

M1 Introduction to Unsupervised Learning

M2 K-Means Algorithm

M3 EM Algorithm

M4 Hierarchical Clustering

M5 Density Based Clustering

M6 Assessing Quality of Clustering

M7 Association Rule Mining

M8 Time series Prediction and Markov Process

The Association Rule mining topic was recorded by me and the pace of delivery turned to be slower than expected. Request you to increase the speed to 1.5X while referring to the same. Thanks for your understanding

Module 5 : Density Based Clustering

A Introduction to Density based approach to clustering

B DBSCAN

C Performance & Scalability

D Clustering for Anomaly Detection

FYI There is an Industry Expert talk session available as pre-recorded content under the “Modules” page section titled “Industry Talk”. In this Anomaly Detection , Cluster Quality Metrics are available.

Agenda

Learning Objective : Density Based Clustering

- Identify the application specific need
- Understand the working of DBSCAN
- Understand the working of OPTICS and outlier detection mechanism using Density based clustering technique

D: Density Based Clustering

A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density.

Need : Clusters are irregular or intertwined, and when noise and outliers are present

Cluster Objective function : **Local**

DBSCAN: Ester, et al. (KDD'96)

OPTICS: Ankerst, et al (SIGMOD'99).

DENCLUE: Hinneburg & D. Keim (KDD'98)

CLIQUE: Agrawal, et al. (SIGMOD'98) (more grid-based)

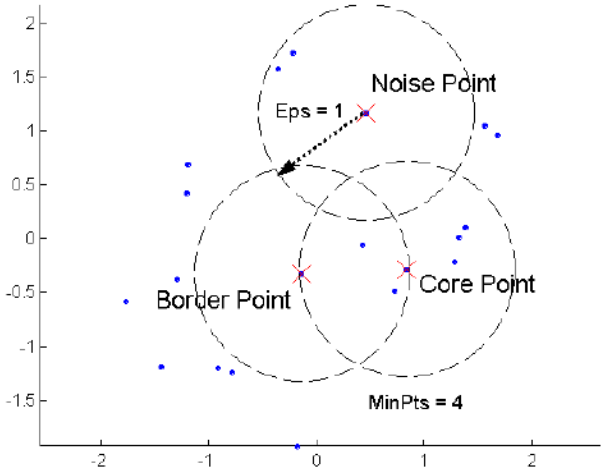
DBSCAN - Density-Based Spatial Clustering of Applications with Noise

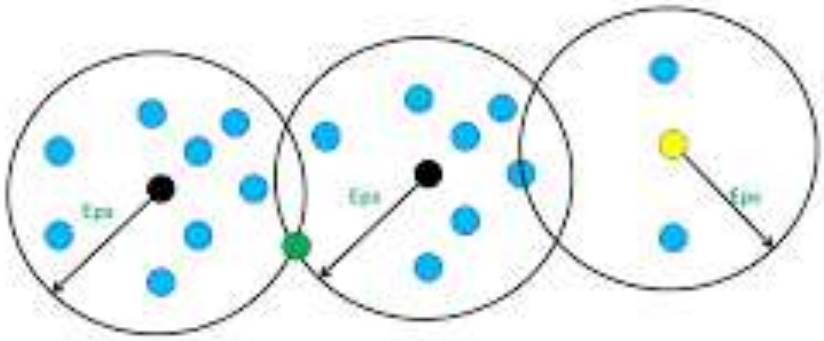
Notion of Density

The density of an object *O* can be measured by the number of objects close to *O*.

Є - neighbourhood is the space within a radius with *O* as centre.

MinPts – Threshold of dense region





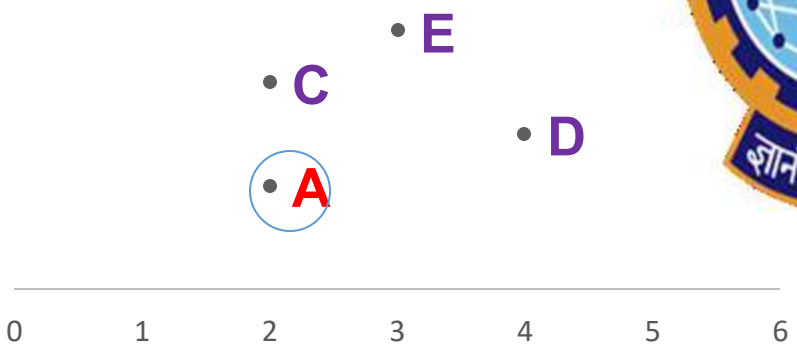
Example

Eps / ε = 2.5

MinPts = 4

Unvisited	{B,C,D,E}
Visited	{A}
Neighborhood(A)	{C,D}
Clusters	C1 = {A}
Core Points	
Noise Points	

Density	Neighbourhood	Movies	Imdb rating	Times of India (Scaled to 1-10)
3	{C, D}	A	2	2
1	{}	B	5	8
4	{A, D. E}	C	2	4
4	{A, C, E}	D	4	3
3	{C, D}	E	3	5



	A	B	C	D	E
A	0	6.7	2.0	2.2	3.2
B		0	5.0	5.1	3.6
C			0	2.2	1.4
D				0	2.2
E					0

*The circle denotes only the cluster formation not the exact size of the neighborhood

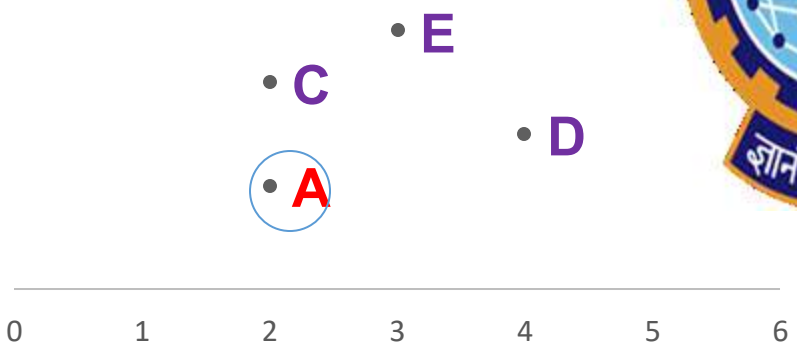
Example

Eps / ε = 2.5

MinPts = 4

Unvisited	{C,D,E}
Visited	{A, B}
Neighborhood(B)	{}
Clusters	C1={A}, C2={B}
Core Points	
Noise Points	{A, B}

Density	Neighbourhood	Movies	Imdb rating	Times of India (Scaled to 1-10)
3	{C, D}	A	2	2
1	{}	B	5	8
4	{A, D. E}	C	2	4
4	{A, C, E}	D	4	3
3	{C, D}	E	3	5



	A	B	C	D	E
A	0	6.7	2.0	2.2	3.2
B		0	5.0	5.1	3.6
C			0	2.2	1.4
D				0	2.2
E					0

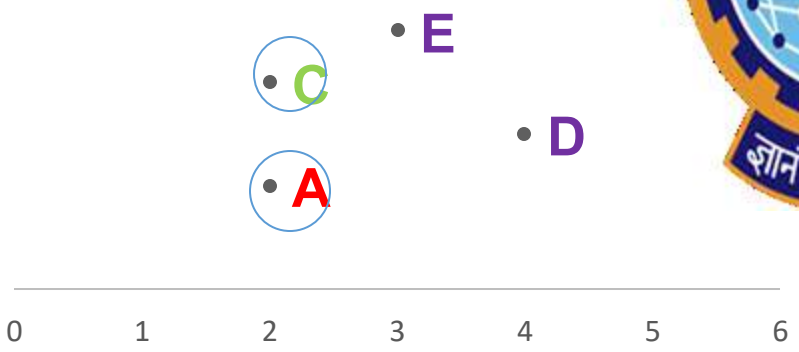
Example

Eps / ε = 2.5

MinPts = 4

Unvisited	{D,E}
Visited	{A, B, C}
Neighborhood(C)	{A,D,E}
Clusters	C1={A}, C2={B}, C3={C}
Core Points	{C}
Noise Points	{A, B}

Density	Neighbourhood	Movies	Imdb rating	Times of India (Scaled to 1-10)
3	{C, D}	A	2	2
1	{}	B	5	8
4	{A, D. E}	C	2	4
4	{A, C, E}	D	4	3
3	{C, D}	E	3	5



	A	B	C	D	E
A	0	6.7	2.0	2.2	3.2
B		0	5.0	5.1	3.6
C			0	2.2	1.4
D				0	2.2
E					0

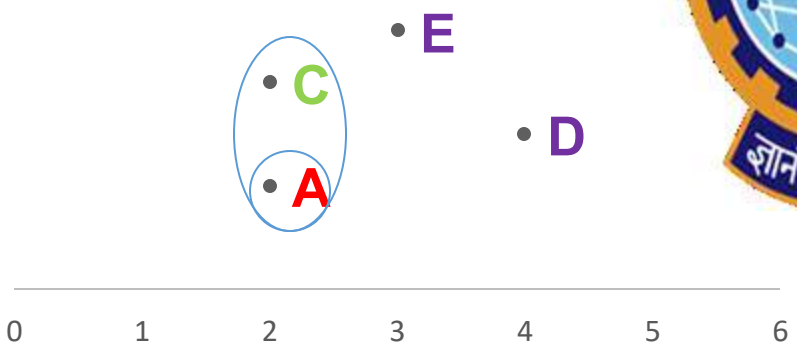
Example

Eps / ε = 2.5

MinPts = 4

Unvisited	{D,E}
Visited	{A, B, C}
Neighborhood(C)	{A,D,E}
Clusters	C2={B}, C3={C, A}
Core Points	{C}
Noise Points	{B}

Density	Neighbourhood	Movies	Imdb rating	Times of India (Scaled to 1-10)
3	{C, D}	A	2	2
1	{}	B	5	8
4	{A, D. E}	C	2	4
4	{A, C, E}	D	4	3
3	{C, D}	E	3	5



	A	B	C	D	E
A	0	6.7	2.0	2.2	3.2
B		0	5.0	5.1	3.6
C			0	2.2	1.4
D				0	2.2
E					0

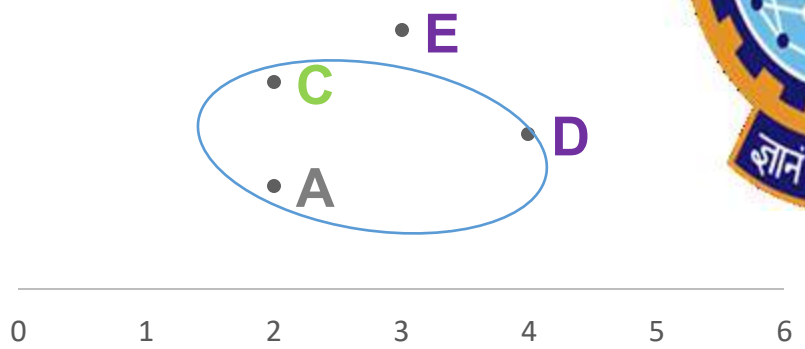
Example

Eps / ε = 2.5

MinPts = 4

Unvisited	{D,E}
Visited	{A, B, C}
Neighborhood(C)	{A,D,E}
Clusters	C2={B}, C3={C, A, D}
Core Points	{C}
Noise Points	{B}

Density	Neighbourhood	Movies	Imdb rating	Times of India (Scaled to 1-10)
3	{C, D}	A	2	2
1	{}	B	5	8
4	{A, D. E}	C	2	4
4	{A, C, E}	D	4	3
3	{C, D}	E	3	5



	A	B	C	D	E
A	0	6.7	2.0	2.2	3.2
B		0	5.0	5.1	3.6
C			0	2.2	1.4
D				0	2.2
E					0

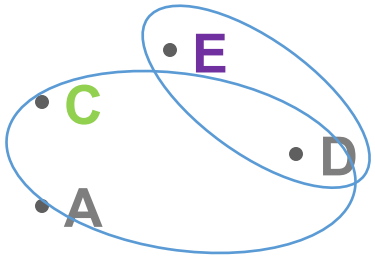
Example

Eps / ε = 2.5

MinPts = 4

Unvisited	{E}
Visited	{A, B, C, D}
Neighborhood(C)	{A,D,E} U {A,C,E}
Clusters	C2={B}, C3={C, A, D}
Core Points	{C, D}
Noise Points	{B}

Density	Neighbourhood	Movies	Imdb rating	Times of India (Scaled to 1-10)
3	{C, D}	A	2	2
1	{}	B	5	8
4	{A, D. E}	C	2	4
4	{A, C, E}	D	4	3
3	{C, D}	E	3	5



	A	B	C	D	E
A	0	6.7	2.0	2.2	3.2
B		0	5.0	5.1	3.6
C			0	2.2	1.4
D				0	2.2
E					0

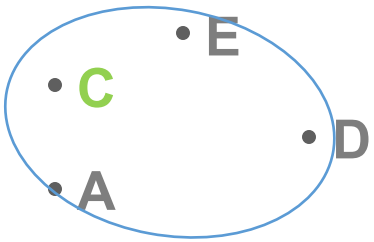
Example

Eps / ε = 2.5

MinPts = 4

Unvisited	{}
Visited	{A, B, C, D, E}
Neighborhood(C)	{A,D,E} U {C,D}
Clusters	C2={B}, C3={C, A, D, E}
Core Points	{C, D}
Noise Points	{B}

Density	Neighbourhood	Movies	Imdb rating	Times of India (Scaled to 1-10)
3	{C, D}	A	2	2
1	{}	B	5	8
4	{A, D. E}	C	2	4
4	{A, C, E}	D	4	3
3	{C, D}	E	3	5



	A	B	C	D	E
A	0	6.7	2.0	2.2	3.2
B		0	5.0	5.1	3.6
C			0	2.2	1.4
D				0	2.2
E					0

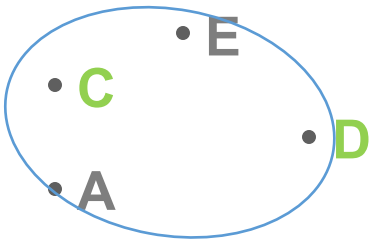
Example

Eps / ε = 2.5

MinPts = 4

Unvisited	{}
Visited	{A, B, C, D, E}
Neighborhood(C)	{A,D,E}
Clusters	C2={B}, C3={C, A, D, E}
Core Points	{C, D}
Noise Points	{B}

Density	Neighbourhood	Movies	Imdb rating	Times of India (Scaled to 1-10)
3	{C, D}	A	2	2
1	{}	B	5	8
4	{A, D. E}	C	2	4
4	{A, C, E}	D	4	3
3	{C, D}	E	3	5



	A	B	C	D	E
A	0	6.7	2.0	2.2	3.2
B		0	5.0	5.1	3.6
C			0	2.2	1.4
D				0	2.2
E					0

DBSCAN - Density-Based Spatial Clustering of Applications with Noise

Notion of Density

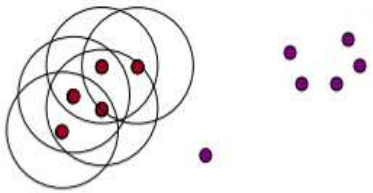
#1: For a core object q and an object p, p is **directly density-reachable** from q (with respect to ϵ and MinPts) if p is within the ϵ-neighborhood of q.

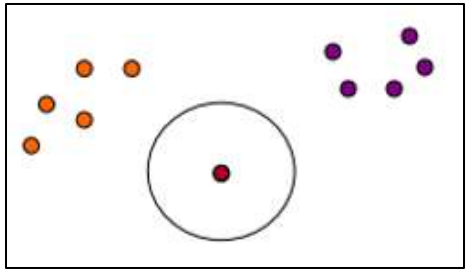
#2: Object p is **density-reachable** from q (with respect to ϵ and MinPts) if there is a chain of objects $p_1,\ldots,p_n$, such that $p_1 = q$ and $p_n = p$ and p_{i+1} is directly density-reachable from p_i.

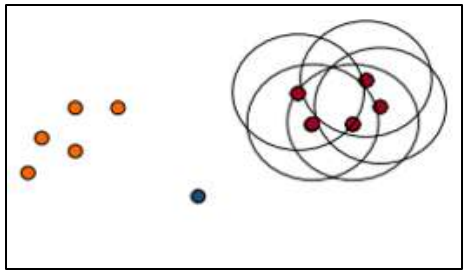
#3: Two objects $p_1, p_2 \in D$ are **density-connected** (with respect to ϵ and MinPts) if there is an object $q \in D$ such that both p_1 and p_2 are density reachable from q.

$\varepsilon = 2$ cm

MinPts = 3



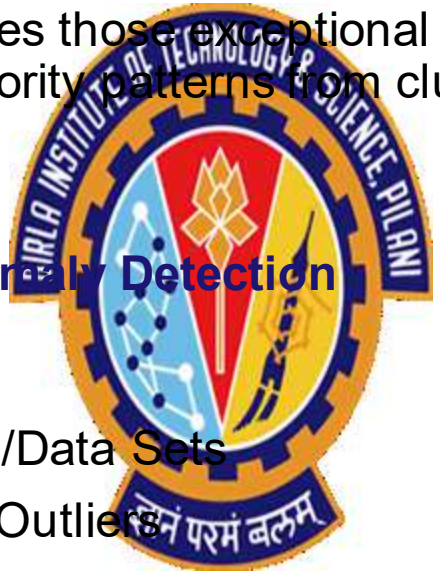




Concept of Outliers

Outliers

- Outliers are observations that deviate significantly from the general patterns & behaviour exhibited by the datasets
- Anomaly detection captures those exceptional cases that deviate substantially from the majority patterns from clustering



Goal of Outlier Analysis / Anomaly Detection

- Define Outlier w.r.t to APP/Data Sets
- Minimize the influence of Outliers
- Remove Outliers
- Extract Outliers

DBSCAN - Density-Based Spatial Clustering of Applications with Noise

Algorithm

Given : **Eps / ε, MinPts**

L1:

Until all the points are examined

Select any unexamined point **p**

Iterate to find all density reachable points from **p**

If **p** is a core point

Create a new/index a new

or existing cluster is extended.

Else if **p** is a border point

Go to L1

DBSCAN - Density-Based Spatial Clustering of Applications with Noise

Algorithm

current_cluster_label $\leftarrow 1$
for all core points **do**
  if the core point has no cluster label **then**
    current_cluster_label $\leftarrow$ *current_cluster_label* $+ 1$
    Label the current core point with cluster label *current_cluster_label*
  end if
  for all points in the *Eps*-neighborhood, except i^{th} the point itself **do**
    if the point does not have a cluster label **then**
      Label the point with cluster label *current_cluster_label*
    end if
  end for
end for

Density based Clustering

Limitations

Figure 8. DBScan results for DS1 with MinPts at 4 and Eps at (a) 0.5 and (b) 0.4.

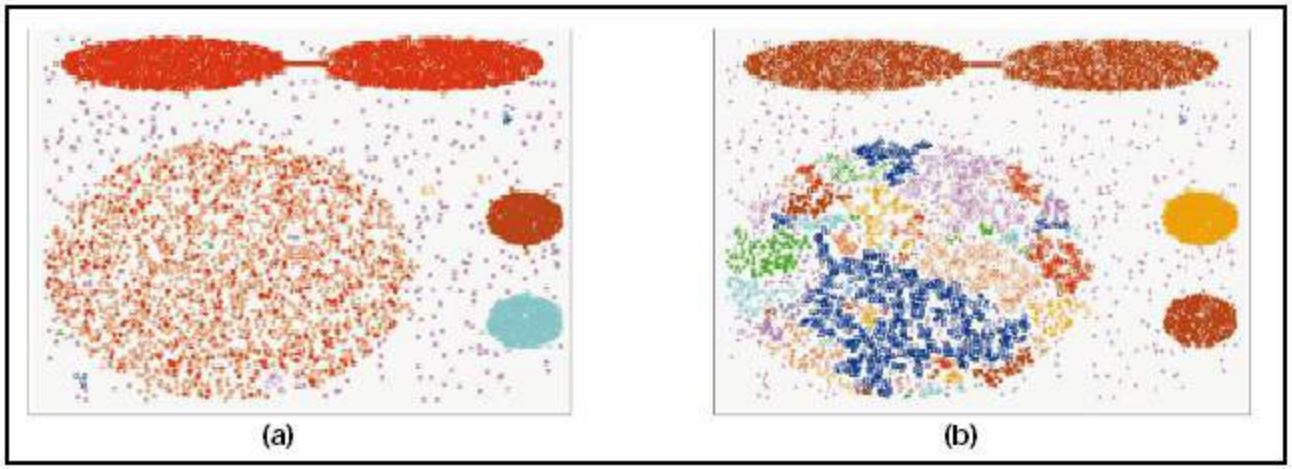
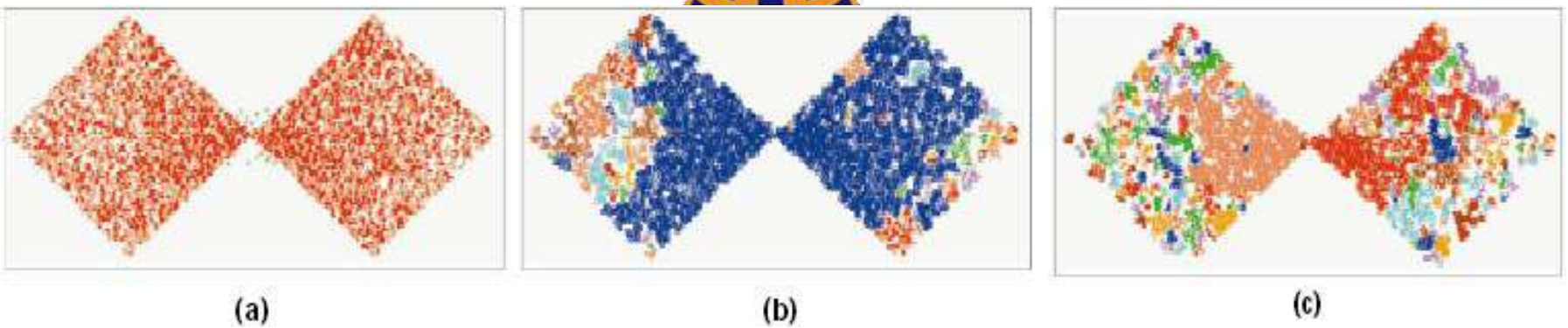


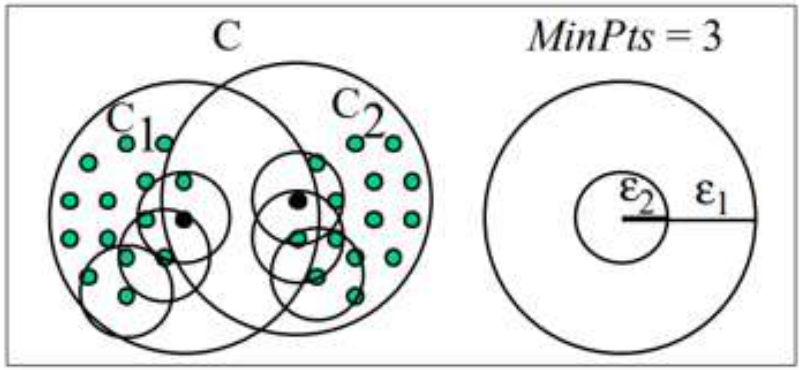
Figure 9. DBScan results for DS2 with MinPts at 4 and Eps at (a) 5.0, (b) 3.5, and (c) 3.0.



OPTICS

Ordering Points to Identify Cluster Structures

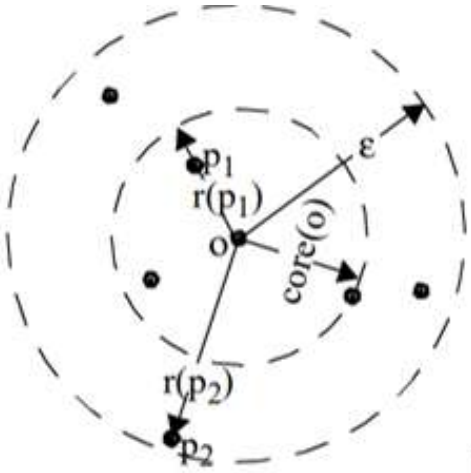
- Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, Jörg Sander (1999)
- Detects clusters of varying densities
- Outputs a cluster ordering, a linear list of all objects under analysis and represents the density-based clustering structure of the data.
- Objects in a denser cluster are listed closer to each other in the cluster ordering.



OPTICS

Ordering Points to Identify Cluster Structures

- core-distance $_{\varepsilon,MinPts}$ (o) – Minimum distance that makes o's neighbourhood dense
- reachability-distance $_{\varepsilon,MinPts}$ (p, o) = max{ core-distance(o), d(p, o) }



OPTICS

Ordering Points to Identify Cluster Structures

Init: Take random point & add to control list *(Point, reachability_distance) = (p,? Or undefined)*

Loop: Until the control List(points to be processed) is empty

Take the first point *p*

Find the ε – Neighbors of *P's* cluster denoted by O

If No.of.Neighbors >= MinPts

Find the Core-distance(*P*) = distance(P, K^{th} Neighbor) //For MinPts=3 K=2^{nd} neighbor

Dequeue P from control list & Add to the Core objects list & mark P as processed

Calculate the reach-dist(o ←p) from p to its neighbors = max{core-dist(p), dist(p,o)}

for every *O*

If o is already in control list update its reach-dist as min{Old-reach-dist, New-reach-dist}

else add (o , reach-dist(o←p))

end for

Reorder the entire control list is ascending order.

Else

Dequeue P from control list & Add to the Noise/border objects list & mark P as processed.

//Noise point if the reach-dist=undefined else it's a border point

OPTICS

Ordering Points to Identify Cluster Structures

-

	A	B	C	D	E
A	0				
B	3.16	0			
C	3.61	4.12	0		
D	2	1.414	3	0	
E	5	5	1.411	4.12	0

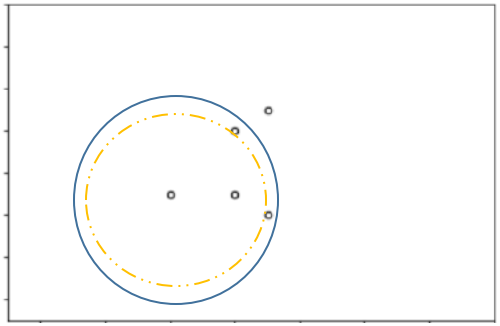
Priority Queue	Reachability Distance
A	? Or Undefined

OPTICS

Observe: The Neighborhood of P can contain more than MinPts if the distances are equal

In chart the reach-dist(A) = undefined not 5

- ε = 4, MinPts= 3

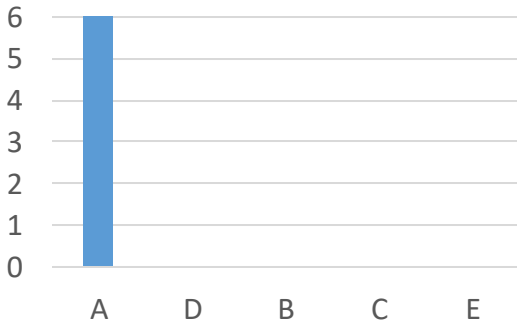


P	Neighbor	Core-dist	Reach-dist
A	D,B,C	3.16	Undef
D			

Observe: The heuristic used while updating the reach-dist in control list may impact the cluster structure formed. If (D,B) is ordered as (B,D) then we may have cluster/s of varying densities. This need to choose the order, is observed only when the value of B & D are same. One way to deal is **find the affinity rank** ie., if DA is lesser than DB then retain the same order else swap the order to (B,D)

	A	B	C	D	E
A	0				
B	3.16	0			
C	3.61	4.12	0		
D	2	1.414	3	0	
E	5	5	1.411	4.12	0

Priority Queue	Reachability Distance
D	**3.16**
B	3.16
C	**3.61**

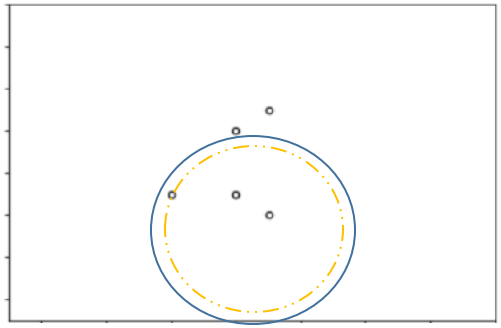


OPTICS

Observe: The Neighborhood of P can contain more than MinPts if the distances are equal

In chart the reach-dist(A) = undefined not 5

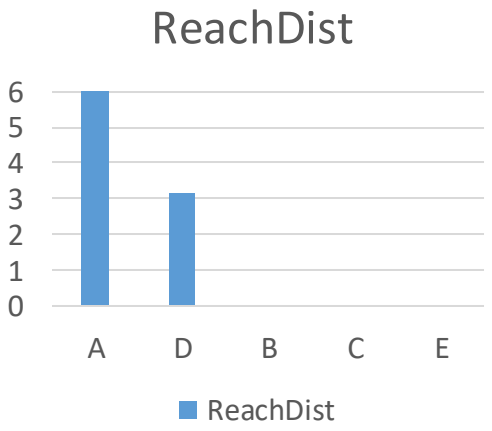
- ε = 4, MinPts= 3



P	Neighbor	Core-dist	Reach-dist
A	D,B,C	3.16	Undef
D	B,A,C	2	**3.16**

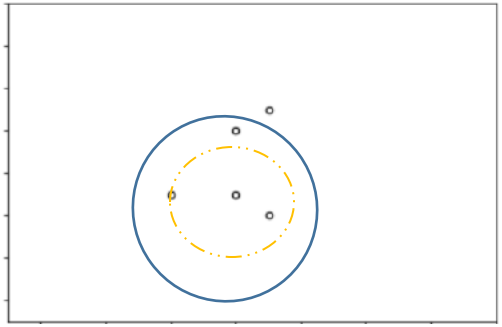
	A	B	C	D	E
A	0				
B	3.16	0			
C	3.61	4.12	0		
D	2	1.414	3	0	
E	5	5	1.411	4.12	0

Priority Queue	Reachability Distance
~~D~~	~~3.16~~
B	~~3.16~~ 2 (MIN update)
C	~~3.61~~ 3 (MIN update)



OPTICS

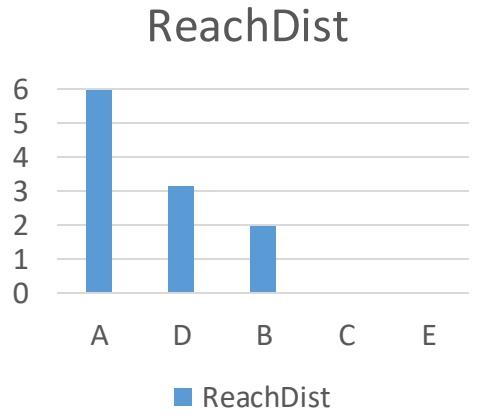
- ε = 4, MinPts= 3



P	Neighbor	Core-dist	Reach-dist
A	D,B,C	3.16	Undef
D	B, A,C	2	3.16
B	D,A	3.16	2
C			

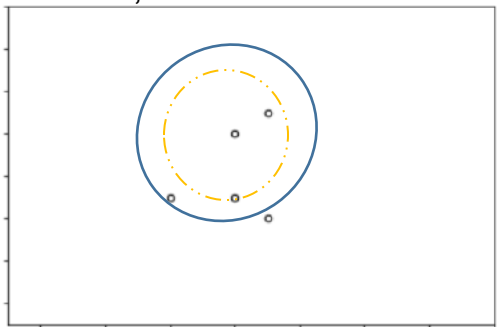
	A	B	C	D	E
A	0				
B	3.16	0			
C	**3.61**	4.12	0		
D	2	1.414	3	0	
E	5	5	1.411	4.12	0

Priority Queue	Reachability Distance
~~B~~	~~2~~
C	3



OPTICS

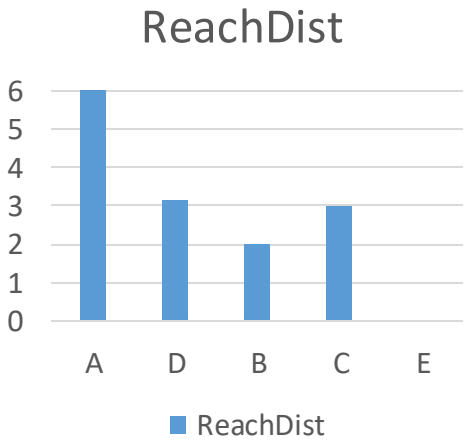
- ε = 4, MinPts= 3



P	Neighbor	Core-dist	Reach-dist
A	D,B,C	3.16	Undef
D	B, A,C	2	3.16
B	D,A	3.16	2
C	E,D,A	3	**3**

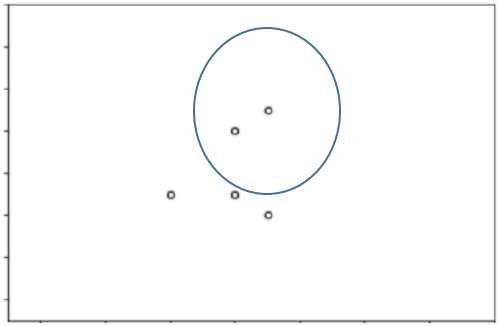
	A	B	C	D	E
A	0				
B	3.16	0			
C	**3.61**	4.12	0		
D	2	1.414	3	0	
E	5	5	1.411	4.12	0

Priority Queue	Reachability Distance
~~€~~	~~3~~
E	3



OPTICS

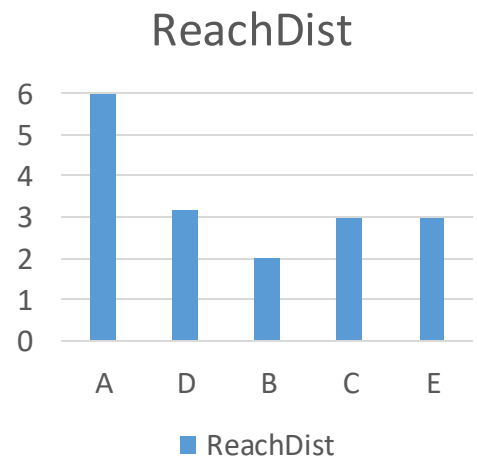
- ε = 4, MinPts= 3



P	Neighbor	Core-dist	Reach-dist
A	D,B,C	3.16	Undef
D	B, A,C	2	3.16
B	D,A	3.16	2
C	E,D,A	3	**3**
E	C	Undef	**3**

	A	B	C	D	E
A	0				
B	3.16	0			
C	**3.61**	4.12	0		
D	2	1.414	3	0	
E	5	5	1.411	4.12	0

Priority Queue	Reachability Distance
~~E~~	~~3~~



Observe:

No.of Clusters =1

E- Border Point

OPTICS – Scenario 2

- ε = 4, MinPts= 3

Observe: If the reachability plot had been like the one above then

No.of Clusters = 2 – Denoted by separate valleys formed

OPTICS

Ordering Points to Identify Cluster Structures

- ε = 4, MinPts= 3
- Objects are processed in a specific order. This order selects an object that is density-reachable with respect to the lowest ϵ - value so that clusters with higher density will be finished first

Use Case – 4 : Resource Monitoring

Performance Analysis

Training has significant influence in the rating

Objective :

To analyse the performance of customer service executives

Rating → Training
Rating → SpecificShift
…… …..

Call Center Executives Observation

EmployeeID	Rating	Training

Detect the Anomaly or Outlier Community

The lab file customized for density clustering with the same WineDataset is uploaded in the canvas

The Association Rule mining topic was recorded by me and the pace of delivery turned to be slower than expected. Request you to increase the speed to 1.5X while referring to the same. Thanks for your understanding

Next Session

Association Rule Mining- Module 6

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

on #6
ing &
lining

PCAM ZC221

Raja vadhana P
BITS - CSIS

Course Plan

M1 Introduction to Unsupervised Learning

M2 K-Means Algorithm

M3 EM Algorithm

M4 Hierarchical Clustering

M5 Density Based Clustering

M6 Assessing Quality of Clustering

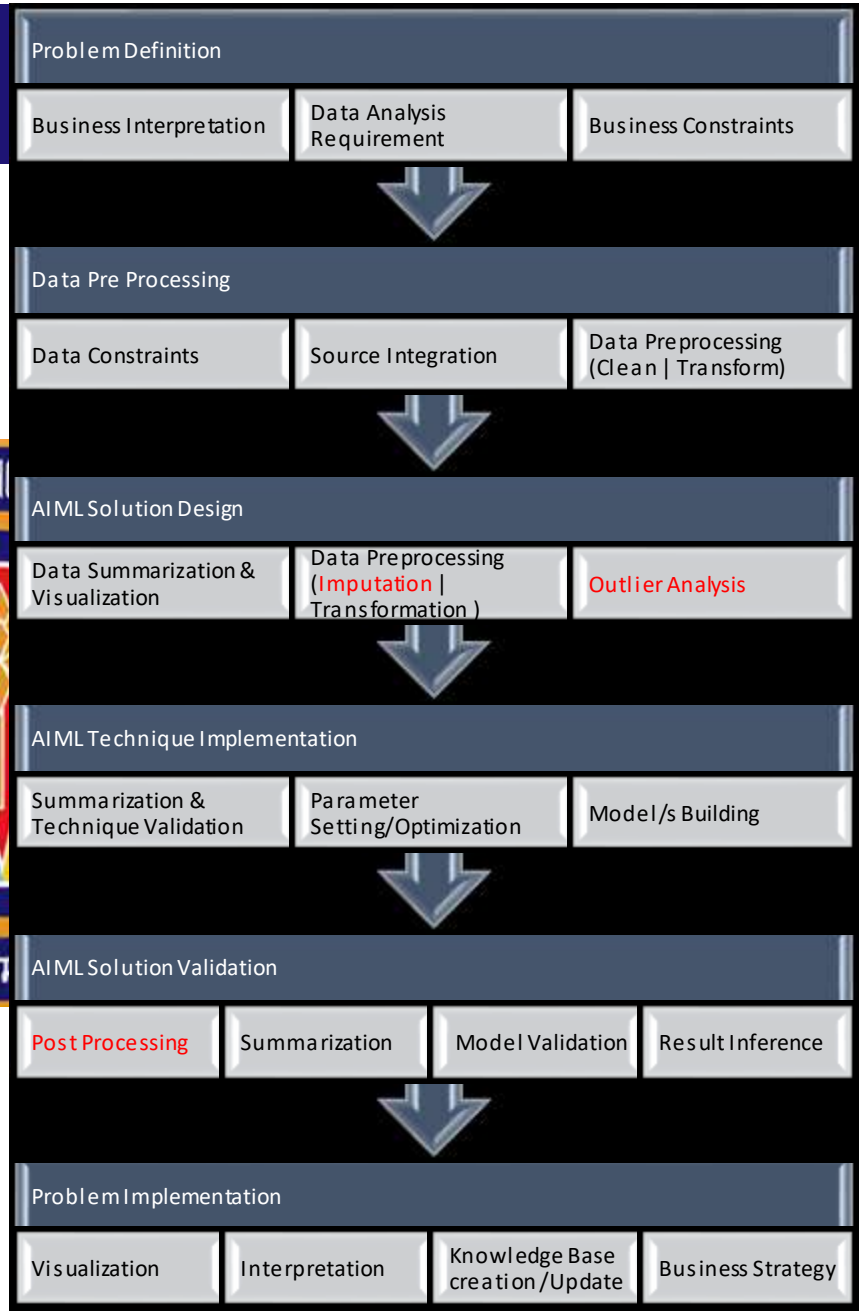
M7 Association Rule Mining

M8 Time series Prediction and Markov Process

Agenda

Learning Objective : Association Rule Mining

- Identify the application specific need
- Extension of introduced case studies for Post Processing
- Apriori Implementation & Metric Interpretations


Problem Definition
Business Interpretation
Data Analysis Requirement
Business Constraints
Data Pre Processing
Data Constraints
Source Integration
Data Preprocessing (Clean | Transform)
AIML Solution Design
Data Summarization & Visualization
Data Preprocessing (Imputation | Transformation)
Outlier Analysis
AIML Technique Implementation
Summarization & Technique Validation
Parameter Setting/Optimization
Model/s Building
AIML Solution Validation
Post Processing
Summarization
Model Validation
Result Inference
Problem Implementation
Visualization
Interpretation
Knowledge Base creation /Update
Business Strategy

Use Case – 1 : Customer Engagement

Customer Segmentation – Post Processing Case

Objective :

To segment the customer for targeted marketing

Use Case – 2 : Resource Monitoring

Performance Management | Optimization Strategies

Objective :

To monitor the performance / implementation efficiency of deployed resources

Use Cases:

1. Call Centre Representatives Monitoring
2. Product Support System by Log Monitoring

Association Rule Mining

Application in ML workflow

- **Pre-Processing**
- Post-Processing

	Symp-1	Symp-2	Symp-3	……	Symp-50	Med -1	Med -2	…..	……	…….	……..
P001											
P002											
P003											
P004											
P005											
P006											
…….											
P2000											

Association Rule Mining

Application in ML workflow

- Pre-Processing
- **Post-Processing**

	Symp-1	Symp-2	Med -1	Med -2	Age	Gender	BMI	……..
P001								
P002								
P003								
P004								
P005								
P006								
…….								
P2000								

Association Rule Mining

Application in ML workflow

- Pre-Processing
- **Post-Processing**

Cluster 1	Symp-1	Symp-2	Med -1	Med -2
P001				
P120				
P225				

Cluster 2	Symp-1	Symp-2	Med -1	Med -2
P002				
P003				
P004				
P005				
……				
P479				
…….				
P2000				

Cluster 3	Symp-1	Symp-2	Med -1	Med -2
P007				
P006				
P525				
P1045				
…….				
P1500				

Association Rule Mining

Application in ML workflow

- Pre-Processing
- **Post-Processing**

Class AGE = Y	Symp-1	Symp-2	Med -1	Med -2	BMI	BP……
P001						
P120						
P225						

CLASS AGE=S	Symp-1	Symp-2	Med -1	Med -2	BMI	BP…….
P002						
P003						
P004						
P005						
……						
P479						
…….						
P2000						

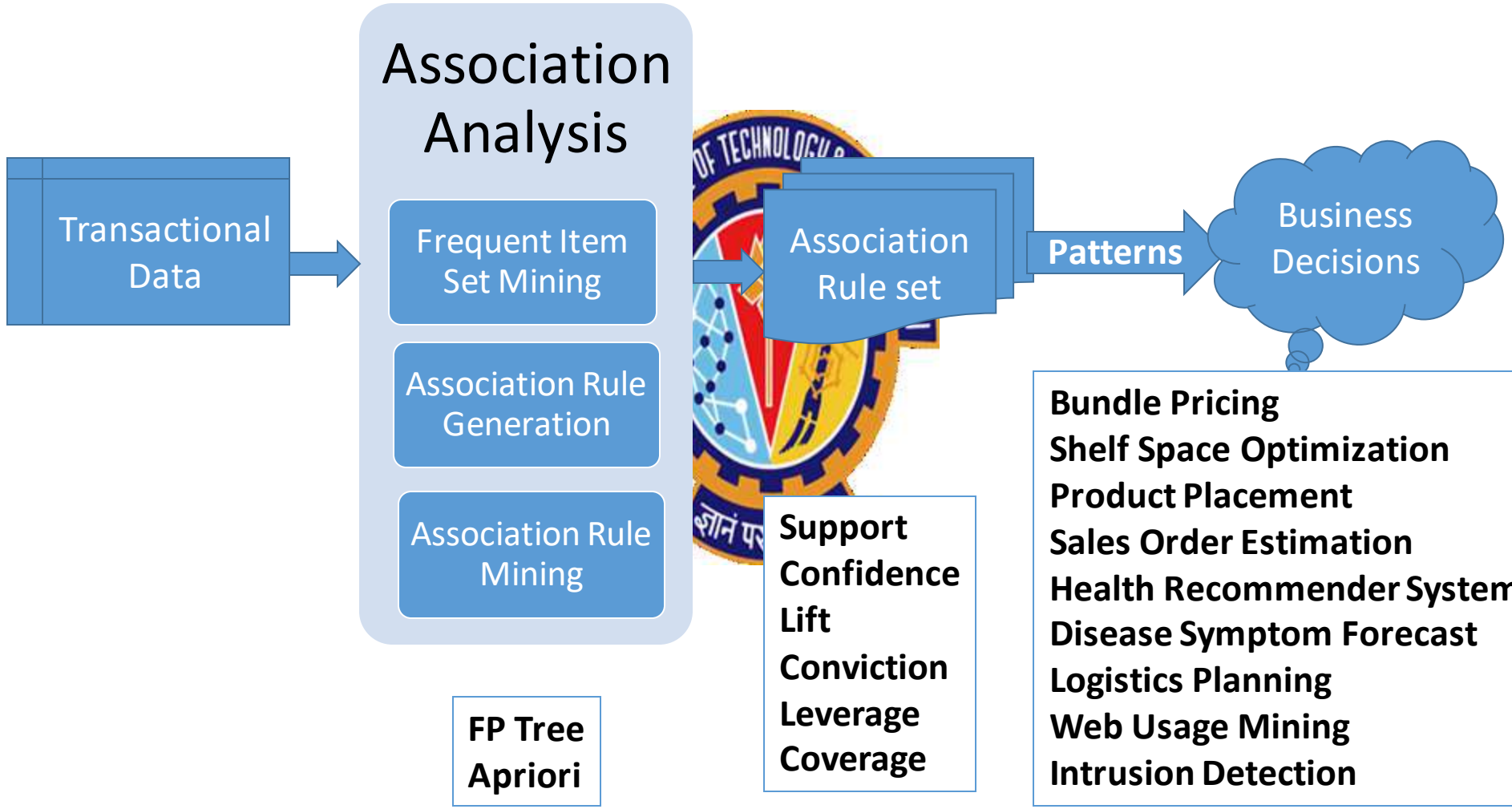
CLASS AGE =	Symp-1	Symp-2	Med -1	Med -2	BMI	BP …….
P007						
P006						
P525						
P1045						
…….						
P1500						

Association Analysis

1. Collect the data from a given collection
2. Filter the frequently occurring items
3. Derive co-occurrence dependency rules
4. Filter the best predictive rules using the measures of interest

GOAL : Discovery of co-occurrence pattern of higher confidence

Association Analysis



Common Pre-processing Techniques

1.Clean Spurious transaction/Data/fact

2.Data Anonymization

3.Binary Transformation

4.Data store Layout Optimization

Set the right objective

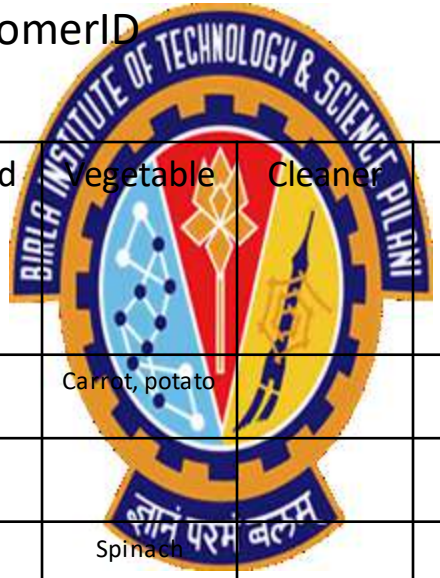
Type of associations:

Potato → Toiletries

GreenVegetables → LiquidCleaner

Vegetable : Potato → Cleaner : Shampoo

2018 : CustomerID → 2019 : CustomerID



Order ID	Customer Age	Customer PhoneNo	Milk	Bread	Vegetable	Cleaner	Medicine	Kitchen Staples	Pay Mode	Purchase Amount
101	25	9876478692	2	1	Carrot, potato				CC	250
102	6	9536478692					para		C	20
103	15	9876678692		1	Spinach			Stationary	C	300
104	29	9876410692	1		Carrot, potato	Toiletries	aa	Oil, Salt, Stationary	DC	570
105	32	9876489692	1	2	Beans, Onion, Potato, Tomato	Toiletries, Mob	bb	Utensils, Cereals, Pulses	DC	2000
106	60	9876400692	3		Potato, Tomato, Spinach	Towels, Phenol, Shampoo	aa		C	1500

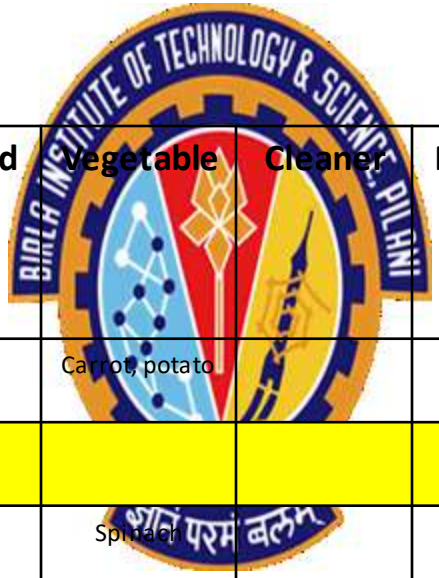
Association Analysis

1.Clean Spurious transaction/Data/fact

2.Data Anonymization

3.Binary Transformation

4.Data store Layout Optimization



Order ID	Cust omer Age	Customer PhoneNo	Milk	Bread	Vegetable	Cleaner	Medicine	Kitchen Staples	Pay Mode	Purchase Amount
101	25	9876478692	2	1	Carrot, potato				CC	250
102	**6**	**9536478692**					**para**		**C**	**20**
103	15	9876678692		1	Spinach			Stationary	C	300
104	29	9876410692	1		Carrot, potato	Toiletries	aa	Oil, Salt, Stationary	DC	570
105	32	9876489692	1	2	Beans, Onion, Potato, Tomato	Toiletries, Mob	bb	Utensils, Cereals, Pulses	DC	2000
106	60	9876400692	3		Potato, Tomato, Spinach	Towels, Phenol, Shampoo	aa		C	1500

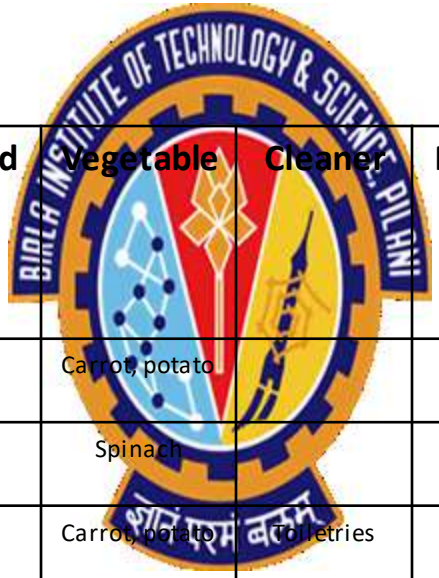
Association Analysis

1.Clean Spurious transaction/Data/fact

2.Data Anonymization

3.Binary Transformation

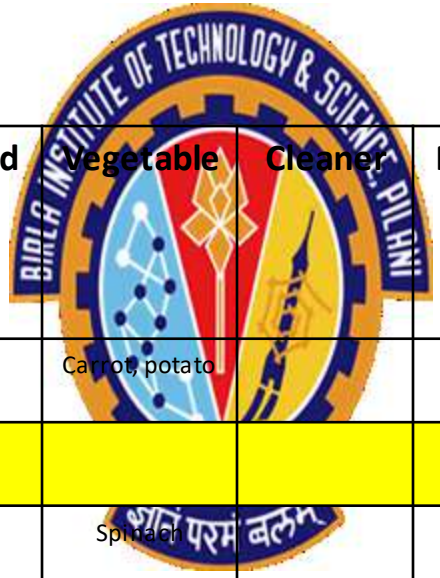
4.Data store Layout Optimization



Order ID	Customer Age	Customer PhoneNo	Milk	Bread	Vegetable	Cleaner	Medicine	Kitchen Staples	Pay Mode	Purchase Amount
101	25	**9876478692**	2	1	Carrot, potato				**CC**	250
103	15	**9876678692**		1	Spinach			Stationary	**C**	300
104	29	**9876410692**	1		Carrot, potato	Toiletries	aa	Oil, Salt, Stationary	**DC**	570
105	32	**9876489692**	1	2	Beans, Onion, Potato, Tomato	Toiletries, Mob	bb	Utensils, Cereals, Pulses	**DC**	2000
106	60	**9876400692**	3		Potato, Tomato, Spinach	Towels, Phenol, Shampoo	aa		**C**	1500

Association Analysis

1.Clean Spurious transaction/Data/fact
2.Data Anonymization
3.Binary Transformation
4.Data store Layout Optimization



Order ID	Customer Age	Customer PhoneNo	Milk	Bread	Vegetable	Cleaner	Medicine	Kitchen Staples	Pay Mode	Purchase Amount
101	25	9876478692	2	1	Carrot, potato				CC	250
102	**6**	**9536478692**					**para**		**C**	**20**
103	15	9876678692		1	Spinach			Stationary	C	300
104	29	9876410692	1		Carrot, potato	Toiletries	aa	Oil, Salt, Stationary	DC	570
105	32	9876489692	1	2	Beans, Onion, Potato, Tomato	Toiletries, Mob	bb	Utensils, Cereals, Pulses	DC	2000
106	60	9876400692	3		Potato, Tomato, Spinach	Towels, Phenol, Shampoo	aa		C	1500

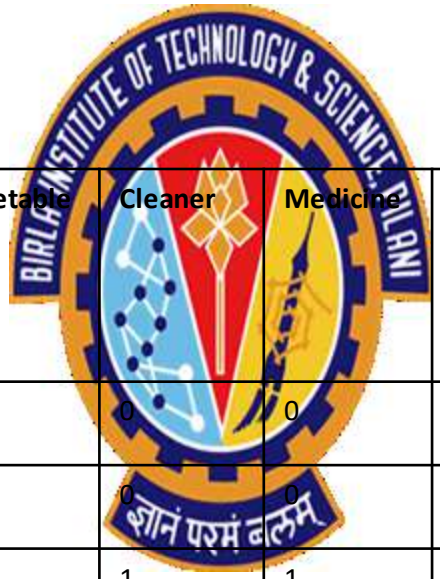
Association Analysis

1.Clean Spurious transaction/Data/fact
2.Data Anonymization
3.Binary Transformation
4.Data store Layout Optimization

Order ID	Customer Age	Customer PhoneNo	Milk	Bread	Vegetable	Cleaner	Medicine	Kitchen Staples	Pay Mode	Purchase Amount LEVEL
101	Adult	XX	2	1	Carrot, potato				XX	Low
103	Young	XX		1	Spinach			Stationary	XX	Low
104	Adult	XX	1		Carrot, potato	Toiletries	aa	Oil, Salt, Stationary	XX	Medium
105	Adult	XX	1	2	Beans, Onion, Potato, Tomato	Toiletries, Mob	bb	Utensils, Cereals, Pulses	XX	High
106	Senior	XX	3		Potato, Tomato, Spinach	Towels, Phenol, Shampoo	aa		XX	High

Association Analysis

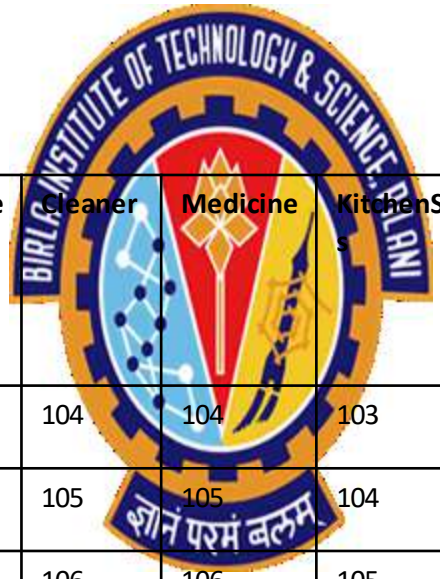
1.Clean Spurious transaction/Data/fact
2.Data Anonymization
3.Binary Transformation
4.Data store Layout Optimization



Order ID	Customer Age	Milk	Bread	Vegetable	Cleaner	Medicine	KitchenStaples	PurchaseAmount LEVEL
101	Adult	1	1	1	0	0	0	Low
103	Young	0	1	1	0	0	1	Low
104	Adult	1	0	1	1	1	1	Medium
105	Adult	1	1	1	1	1	1	High
106	Senior	1	0	1	1	1	0	High

Association Analysis

1.Clean Spurious transaction/Data/fact
2.Data Anonymization
3.Binary Transformation
4.Data store Layout Optimization



Age = Young	Age = Adult	Age = Senior	Milk	Bread	Vegetable	Cleaner	Medicine	KitchenStaples	LEVEL=Low	LEVEL=Medium	LEVEL=High
103	101	106	101	101	101	104	104	103	101	104	105
	104		104	103	103	105	105	104	103		106
	105		105	105	104	106	106	105			
			106		105						
					106						

Apriori Algorithm – Frequent Itemsets Mining

TID	Items
1	Bread, Milk
2	Bread, Diaper, Butter, Eggs
3	Milk, Diaper, Butter, Coke
4	Bread, Milk, Diaper, Butter
5	Bread, Milk, Diaper, Coke

Assume Minimum Support count = 3
or support = 0.6

Apriori Algorithm – Frequent Itemsets Mining

TID	Items
1	Bread, Milk
2	Bread, Diaper, Butter, Eggs
3	Milk, Diaper, Butter, Coke
4	Bread, Milk, Diaper, Butter
5	Bread, Milk, Diaper, Coke

1- Itemset

Item	Count
Bread	4
~~Coke~~	~~2~~
Milk	4
Butter	3
Diaper	4
~~Eggs~~	~~1~~

Assume Minimum Support count = 3
or support = 0.6

Apriori Algorithm – Frequent Itemsets Mining

TID	Items
1	Bread, Milk
2	Bread, Diaper, Butter, Eggs
3	Milk, Diaper, Butter, Coke
4	Bread, Milk, Diaper, Butter
5	Bread, Milk, Diaper, Coke

Assume Minimum Support count = 3
or support = 0.6

1- Itemset

Item	Count
Bread	4
~~Coke~~	~~2~~
Milk	4
Butter	3
Diaper	4
~~Eggs~~	~~1~~

Frequent Itemset = {Bread, Milk, Butter, Diaper}

2 - Itemset

Itemset	Count
{Bread,Milk}	3
~~{Bread,Butter}~~	~~2~~
{Bread,Diaper}	3
~~{Milk,Butter}~~	~~2~~
{Milk,Diaper}	3
{Butter,Diaper}	3

Apriori Algorithm – Frequent Itemsets Mining

TID	Items
1	Bread, Milk
2	Bread, Diaper, Butter, Eggs
3	Milk, Diaper, Butter, Coke
4	Bread, Milk, Diaper, Butter
5	Bread, Milk, Diaper, Coke

Assume Minimum Support count = 3
or support = 0.6

1- Itemset

Item	Count
Bread	4
~~Coke~~	~~2~~
Milk	4
Butter	3
Diaper	4
~~Eggs~~	~~1~~

Frequent Itemset = {Bread, Milk, Butter, Diaper}

2 - Itemset

Itemset	Count
{Bread,Milk}	3
~~{Bread,Butter}~~	~~2~~
{Bread,Diaper}	3
~~{Milk,Butter}~~	~~2~~
{Milk,Diaper}	3
{Butter,Diaper}	3

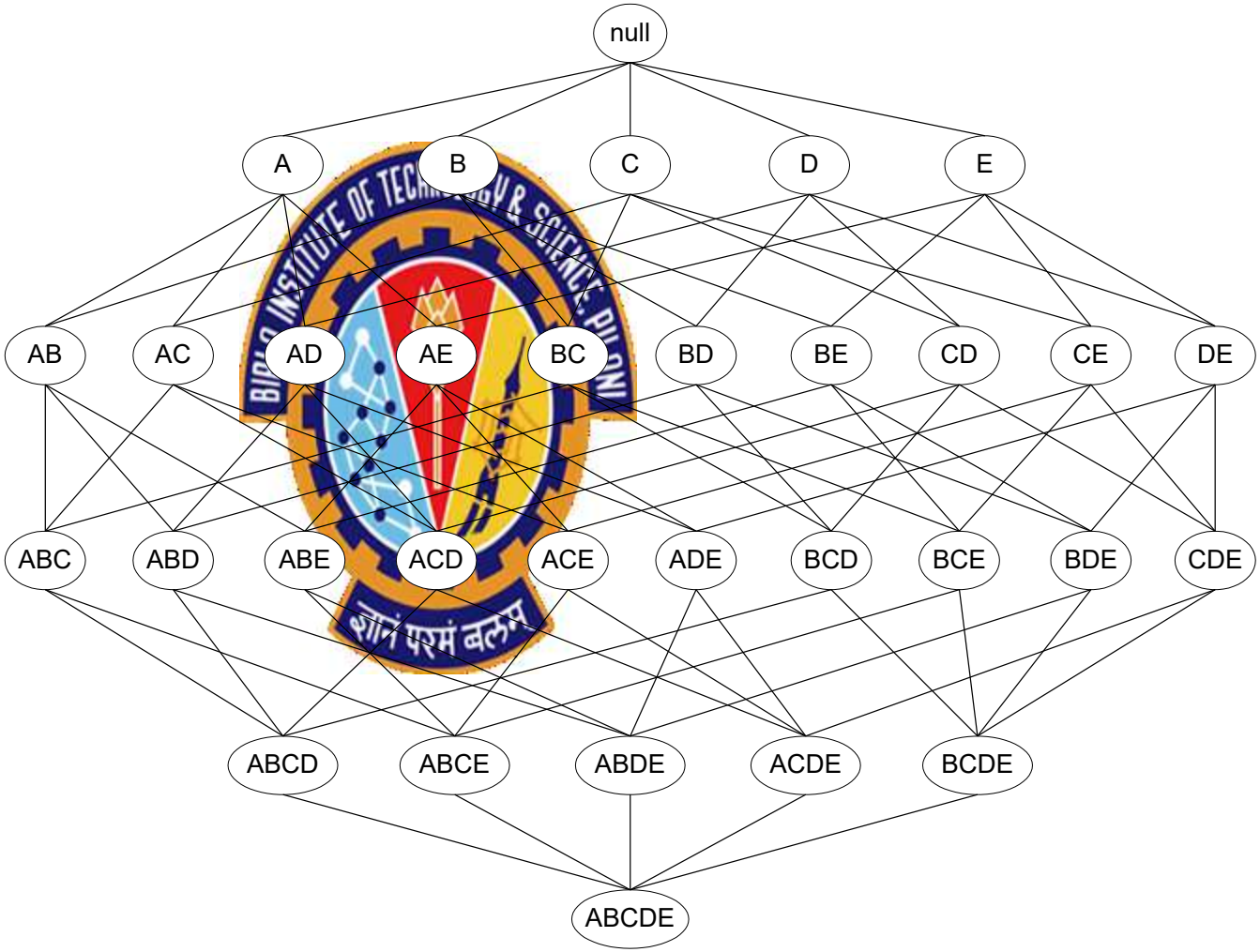
3 - Itemset

Itemset	Count
{Bread,Milk,Diaper}	~~2~~
{Bread,Butter,Diaper}	Pruned by Apriori principle and count is not even calculated
{Milk,Butter,Diaper}	

Association Analysis

Association Analysis

- Frequent Item Set Mining
- Association Rule Generation
- Association Rule Mining



Association Analysis

Association Analysis

- Frequent Item Set Mining
- Association Rule Generation
- Association Rule Mining

Apriori Principle :

All subsets of a frequent itemsets are also frequent.

All supersets of infrequent itemsets are also infrequent

Antimonotone Property: Support(Subset) >= Support(Itemset)

Apriori Pruning Principle : Supersets of infrequent itemsets need not be generated or tested

Apriori Algorithm – Strong Rules Mining

Association Analysis

- Frequent Item Set Mining
- Association Rule Generation
- Association Rule Mining

Apriori Algorithm – Strong Rules Mining

Assume Minimum Confidence = 60% or Confidence =0.6

1. Generate the rule by binary partition
2. Compute Confidence
3. Prune rules using Anti-monotone Property.

Assume the below itemset as frequent for this example.

Itemset	Count
{Bread,Milk,Diaper}	2

Candidate Rule	Confidence
{Bread, Milk} -> {Diaper}	
{Bread, Diaper} -> {Milk}	
{Diaper, Milk} -> {Bread}	

Apriori Algorithm – Strong Rules Mining

Itemset	Count
{Bread,Milk,Diaper}	

Candidate Rule	Confidence
{Bread, Milk} -> {Diaper}	0.66
{Bread, Diaper} -> {Milk}	0.66
{Diaper, Milk} -> {Bread}	0.66

Candidate Rule	Confidence
{Bread} -> {Diaper, Milk}	0.5
{Milk} -> {Diaper, Bread}	0.5

Item	Count
Bread	4
Coke	2
Milk	4
Butter	3
Diaper	4
Eggs	1

Itemset	Count
{Bread,Milk}	3
{Bread,Butter}	2
{Bread,Diaper}	3
{Milk,Butter}	2
{Milk,Diaper}	3
{Butter,Diaper}	3

Apriori Algorithm – Strong Rules Mining

Itemset	Count
{Bread,Milk,Diaper}	

Candidate Rule	Confidence
{Bread, Milk} -> {Diaper}	0.66
{Bread, Diaper} -> {Milk}	0.66
{Diaper, Milk} -> {Bread}	0.66

Candidate Rule	Confidence
{Diaper} -> {Bread, Milk}	0.5

Item	Count
Bread	4
Coke	2
Milk	4
Butter	3
Diaper	4
Eggs	1

Itemset	Count
{Bread,Milk}	3
{Bread,Butter}	2
{Bread,Diaper}	3
{Milk,Butter}	2
{Milk,Diaper}	3
{Butter,Diaper}	3

Apriori Algorithm – Strong Rules Mining

Itemset	Count
{Bread,Milk,Diaper}	

Itemset	Count
{Bread,Milk}	3
{Bread,Butter}	2
{Bread,Diaper}	3
{Milk,Butter}	2
{Milk,Diaper}	3
{Butter,Diaper}	3

Item	Count
Bread	4
Coke	2
Milk	4
Butter	3
Diaper	4
Eggs	1

Candidate Rule	Confidence
{Bread, Milk} -> {Diaper}	0.66
{Bread, Diaper} -> {Milk}	0.66
{Diaper, Milk} -> {Bread}	0.66

Association Rule Generation:

{ {Bread, Milk} -> {Diaper} ,
{Bread, Diaper} -> {Milk} ,
{Diaper, Milk} -> {Bread}

} Continue the process for another iteration where the antecedents are 1-itemsets moving one item from above generated rules to consequent of the rule to continue.

Please note that as per the problem inputs the taken 3-itemset is not frequent and hence the rules generated are not Frequent strong rules as per input transaction data. Kindly use this as practice to work on all eight frequent itemsets obtained in previous result

Important Note :
FP Growth algorithm PPT discussed in class in already available under the courseware recording section

Next Session

Time Series Prediction and Markov Process

Unsupervised Learning
&
Association Rule Mining

Contact Session 7

Raja vadhana P

Assistant Professor,

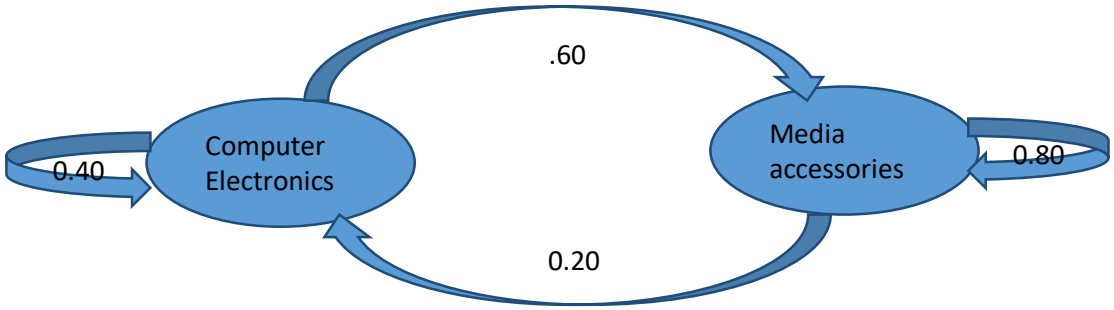
BITS - CSIS

BITS Pilani

Pilani Campus

Time Series Predication & Hidden Morkov Model

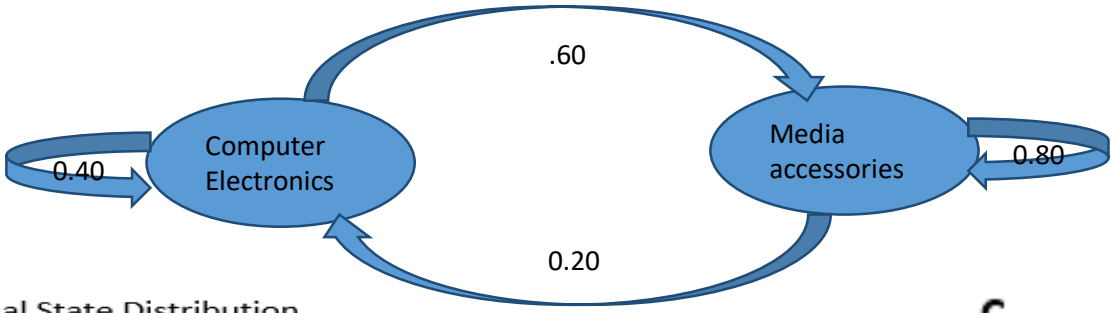
Morkov Model



Transition Model

C	M	
0.40	0.20	C
0.60	0.80	M

Morkov Model



Current State: Initial State Distribution

1	C
0	M

C	M	
0.40	0.20	C
0.60	0.80	M

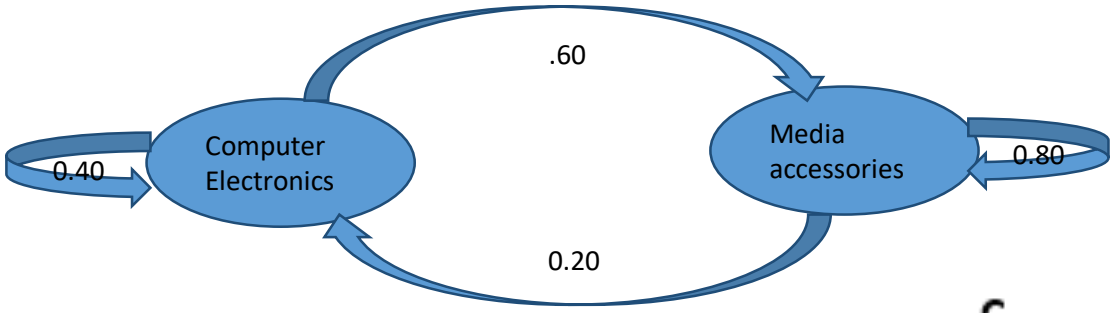
Next State : Likely to buy Media accessories on next visit

0.40	C
0.60	M

Next State : Likely to buy Media accessories on next visit

0.28	C
0.72	M

Morkov Model

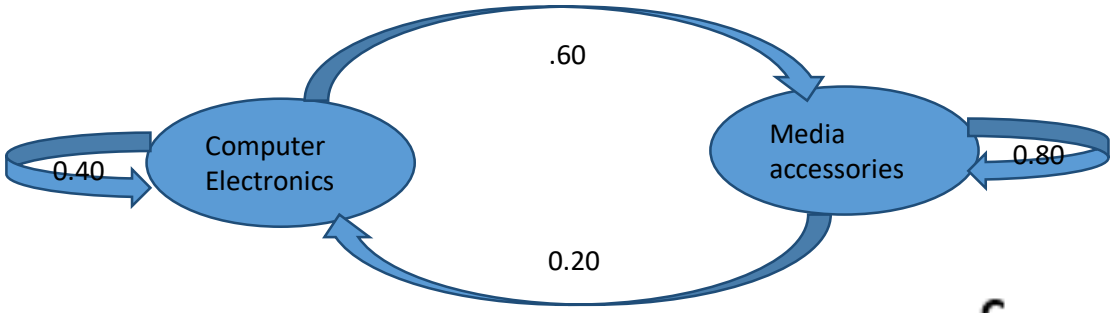


	C	M	
	0.40	0.20	C
	0.60	0.80	M

What is the probability that the purchasing behavior of the customer is in the below order sequentially observed?

(Computer , Media, Media, Computer)

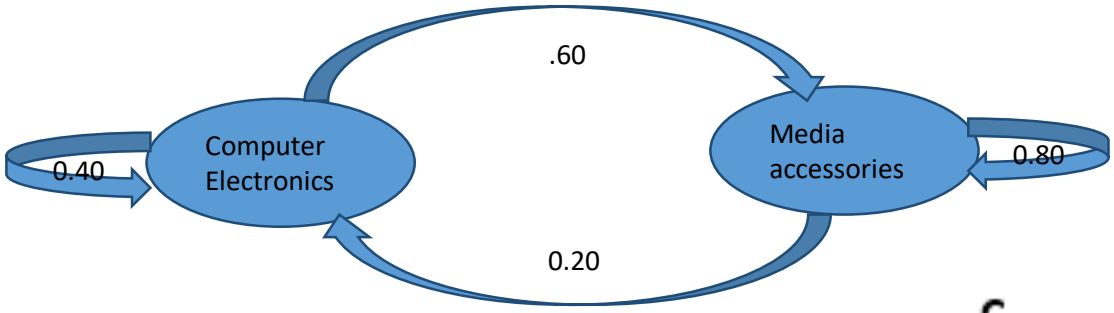
Morkov Model



	C	M	
	0.40	0.20	C
	0.60	0.80	M

What is the probability that a customer who purchased Media accessories will return back and keep purchasing Media accessories for only 2 consecutive visits?

Morkov Model



	C	M	
	0.40	0.20	C
	0.60	0.80	M

Given that a customer walked into a store and bought a computer electronics, find the expected purchase pattern in his next 3 visits.

Markov Process

States | Observations | Assumptions

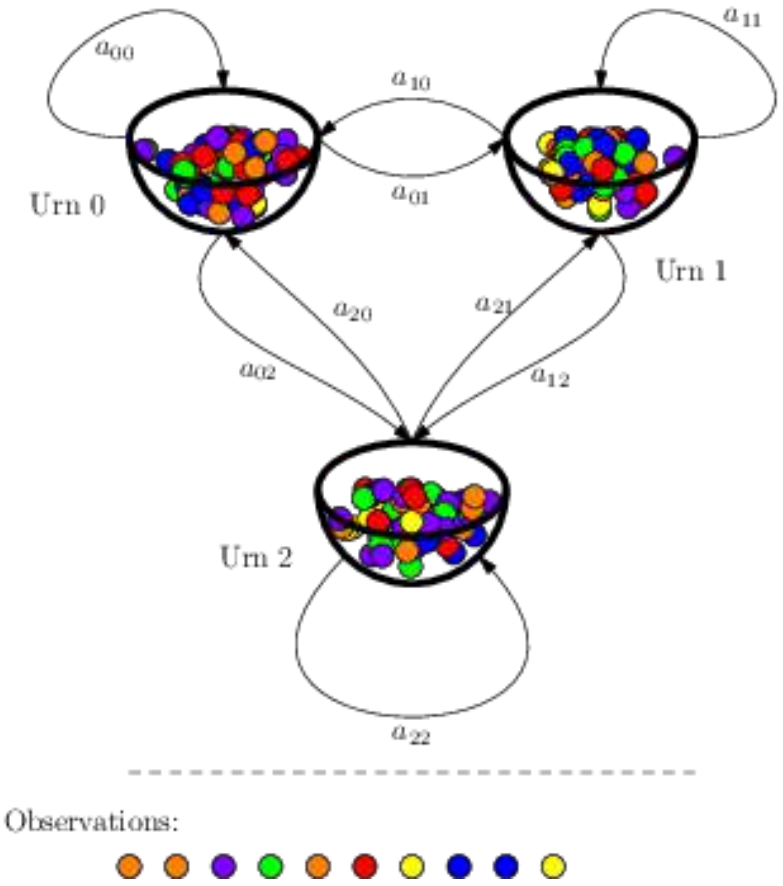
Modelling sequences of random events and transitions between states over time is known as Morkov chain

Transition Model / Probability Matrix :
Current state depends only finite number of previous states. :

Markov Process

States | Observations | Assumptions

Standard Mathematical Example:
Urn & Ball Model



Hidden Markov Process

States | Observations | Assumptions

Modelling sequences of random events and transitions between states over time is known as Morkov chain

Hidden Markov Process models events as the state sequences that are not directly observable but only be approximated from the sequence of observations produced by the system

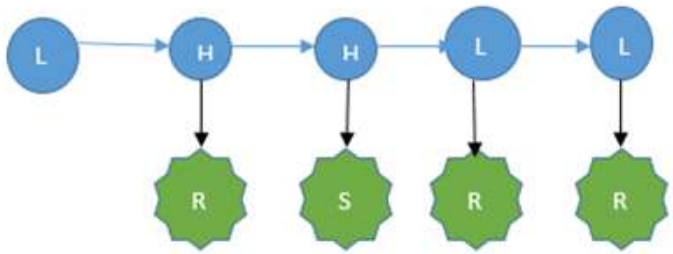
Transition Model / Probability Matrix :
Current state depends only finite number of previous states. :

Evidence / Sensor Model/ Emission Probability Matrix :
Current Evidence or Observation depends Current State of the world. Given the Current State Knowledge of the world, observation doesn't depend on history:

Hidden Morkov Model

States | Observations | Assumptions

Time Slice (t)	0	1	2	3	4	.....	$P(O_t \mid O_{t-1}, O_{t-2})$
Observed Evidence (O_t)	-	Rainy	Sunny	Rainy	Rainy		
Unobserved State(U_t)	Low Pressure	High Pressure	High Pressure	Low Pressure	Low Pressure		



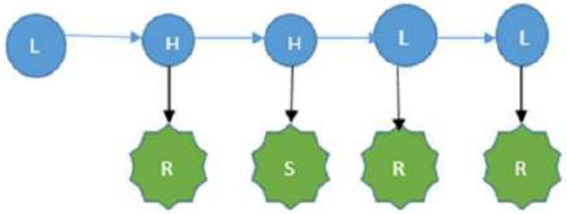
Transition Model / Probability Matrix

$P(U_{t-2} = LP, U_{t-1} = HP)$	$P(U_{t-2} = HP, U_{t-1} = HP)$	$P(U_{t-2} = HP, U_{t-1} = LP)$	$P(U_{t-2} = LP, U_{t-1} = LP)$	←Previous
0.2	0.40	0.85	0.5	$P(U_t = LP)$
0.8	0.60	0.15	0.5	$P(U_t = HP)$

Evidence / Sensor Model/ Emission Probability Matrix

$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t = Rainy)$
0.2	0.6	$P(E_t = Sunny)$

Hidden Morkov Model



innovate achieve lead

Filtering	Prediction	Smoothing	Most Likely Explanation (Viterbi Algorithm)
$P(L_3 \mid R\text{-}S\text{-}R\text{-}R)$	$P(L_3 \mid R\text{-}S)$	$P(H_2 \mid R\text{-}S\text{-}R\text{-}R)$	$P(H\text{-}H\text{-}L\text{-}L \mid R\text{-}S\text{-}R\text{-}R)$
$P(X_t \mid E_{1\ldots t})$	$P(X_{t+k} \mid E_{1\ldots t})$	$P(X_{k,\ o>k>t} \mid E_{1\ldots t})$	$\text{argmax}\ X_{1\ldots t} : P(X_{1\ldots t} \mid E_{1\ldots t})$

Find the probability of occurrence of this weather sequence observation: **S-S-R**

Find the Current Pressure if sequence of weather observations recorded till now are: **S-S-R**

Find the Pressure in past instance of time if sequence of following future weather observations recorded are: **S-S-R**

Find the pattern in pressure that might have caused this observation: **S-S-R**

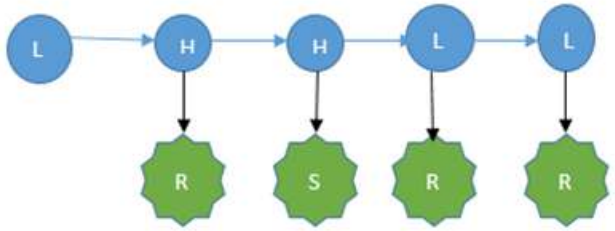
Hidden Morkov Model

Inference: Type -1

Sequence Evaluation : Likely hood Computation : Forward Algorithm

Find the probability of occurrence of this Pressure sequence observation: **S-S-R**

Intuition: $P(E_{1...t}) = \sum_{i=1}^{N} P(E_{1...t} \mid X_{1...t}) * P(X_{1...t}) = \sum_{i=1}^{N} \prod_{j=1}^{t} P(Ej \mid Xj) * P(Xj \mid X_{j-1})$



Transition Model / Probability Matrix

$P(U_{t-1} = HP)$	$P(U_{t-1} = LP)$	←Previous
0.2	0.5	$P(U_t = LP)$
0.8	0.5	$P(U_t = HP)$

Evidence / Sensor Model/ Emission Probability Matrix

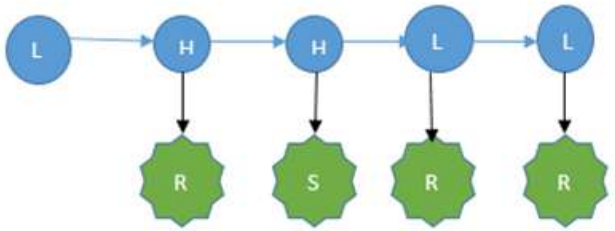
$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t = Rainy)$
0.2	0.6	$P(E_t = Sunny)$

Hidden Morkov Model

Inference: Type -1

Sequence Evaluation : Likely hood Computation : Forward Algorithm

Find the probability of occurrence of this Pressure sequence observation: **S-S-R**



Intuition: $P(E_{1...t}) = \sum_{i=1}^{N} P(E_{1...t} \mid X_{1...t}) * P(X_{1...t}) =$

$= \sum_{i=1}^{N} \prod_{j=1}^{t} P(Ej \mid Xj) * P(Xj \mid Xj_{-1})$

Transition Model / Probability Matrix

$P(U_{t-1} = HP)$	$P(U_{t-1} = LP)$	←Previous
0.2	0.5	$P(U_t = LP)$
0.8	0.5	$P(U_t = HP)$

Evidence / Sensor Model/ Emission Probability Matrix

$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t = Rainy)$
0.2	0.6	$P(E_t = Sunny)$

$P(SSR)$

$$= \sum_{X} P(SSR, X) = \sum_{X} P(SSR, X_1X_2X_3)$$

$$= \sum_{X} P(R, X_3, S, X_2, S, X_1) = \sum_{X} P(R \mid X_3) * P(X_3 \mid X_2) * P(S \mid X_2) * P(X_2 \mid X_1) * P(S \mid X_1) * P(X_1 \mid X_0)$$

$$= \sum_{X} P(R \mid X_3) * P(S \mid X_2) * P(S \mid X_1) * P(X_3 \mid X_2) * P(X_2 \mid X_1) * P(X_1 \mid X_0)$$

$$= \sum_{X} \prod_{j=1}^{t} P(Ej \mid X_j) * P(Xj \mid X_{j-1})$$

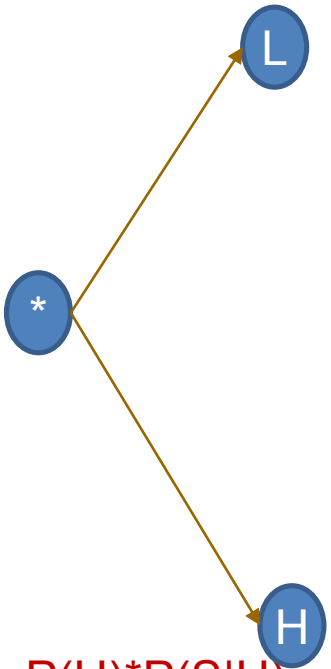
Hidden Morkov Model

Forward Propagation Algorithm

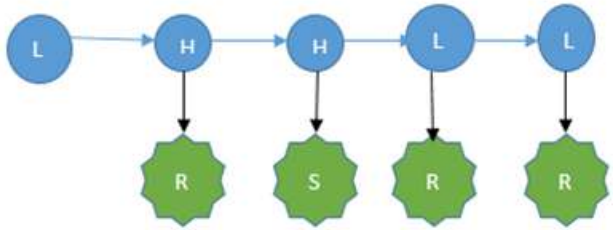
Find the probability of occurrence of this Pressure sequence observation: **S-S-R**

Initialization Phase:

P(L)*P(S|L) = 0.5*0.2 = 0.1 → 0.25



P(H)*P(S|H) = 0.5*0.6 = 0.3 → 0.75



Transition Model / Probability Matrix

$P(U_{t-1} = HP)$	$P(U_{t-1} = LP)$	←Previous
0.2	0.5	$P(U_t = LP)$
0.8	0.5	$P(U_t = HP)$

Evidence / Sensor Model/ Emission Probability Matrix

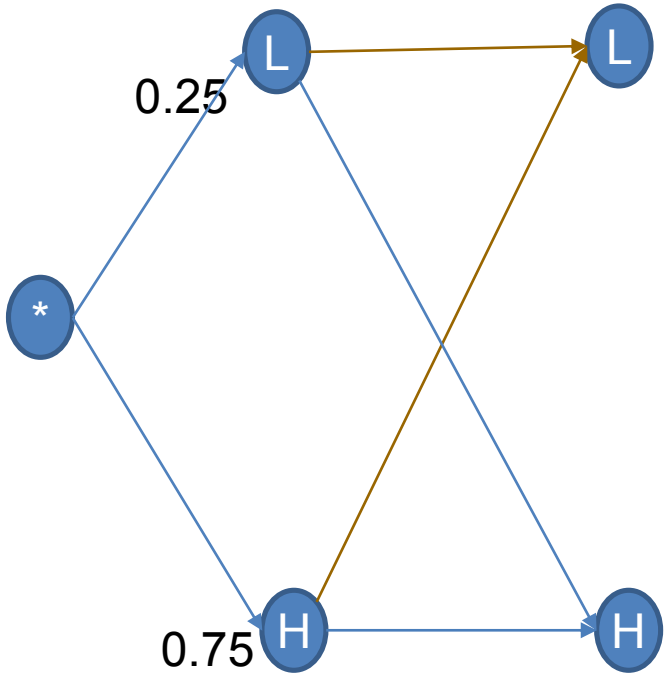
$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t = Rainy)$
0.2	0.6	$P(E_t = Sunny)$

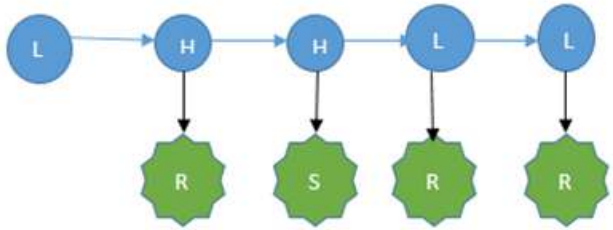
Hidden Morkov Model

Forward Propagation Algorithm : S-S-R

$P(L)*P(L|L)*P(S|L) = 0.25*0.5*0.2 = \mathbf{0.025}$

$P(H)*P(L|H)*P(S|L) = 0.75*0.2*0.2 = \mathbf{0.03}$





Transition Model / Probability Matrix

$P(U_{t-1} = HP)$	$P(U_{t-1} = LP)$	←Previous
0.2	0.5	$P(U_t = LP)$
0.8	0.5	$P(U_t = HP)$

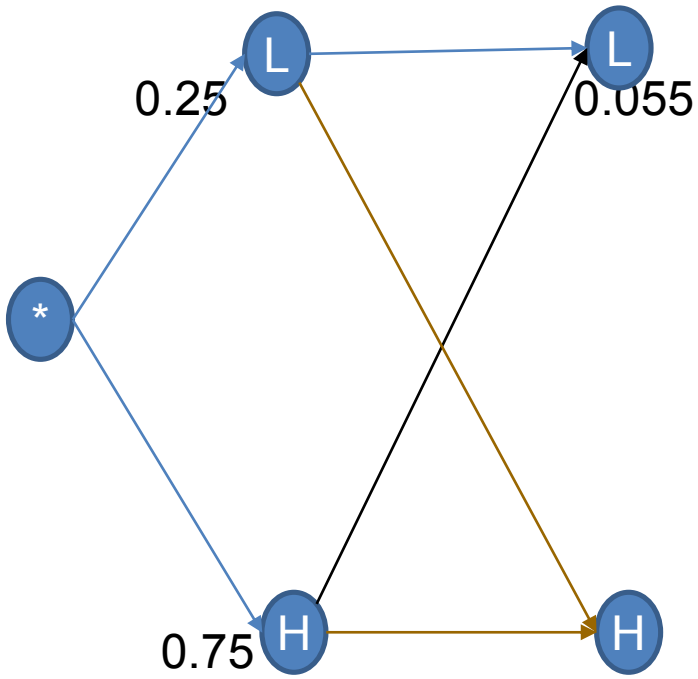
Evidence / Sensor Model/ Emission Probability Matrix

$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t = Rainy)$
0.2	0.6	$P(E_t = Sunny)$

Recursion Phase:

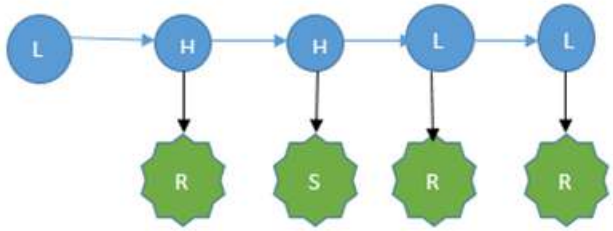
Hidden Morkov Model

Forward Propagation Algorithm : S-S-R



P(L)*P(H|L)*P(S|H) = 0.25*0.5*0.6 = **0.075**

P(H)*P(H|H)*P(S|H) = 0.75*0.8*0.6 = **0.36**



Transition Model / Probability Matrix

$P(U_{t-1} = HP)$	$P(U_{t-1} = LP)$	←Previous
0.2	0.5	$P(U_t = LP)$
0.8	0.5	$P(U_t = HP)$

Evidence / Sensor Model/ Emission Probability Matrix

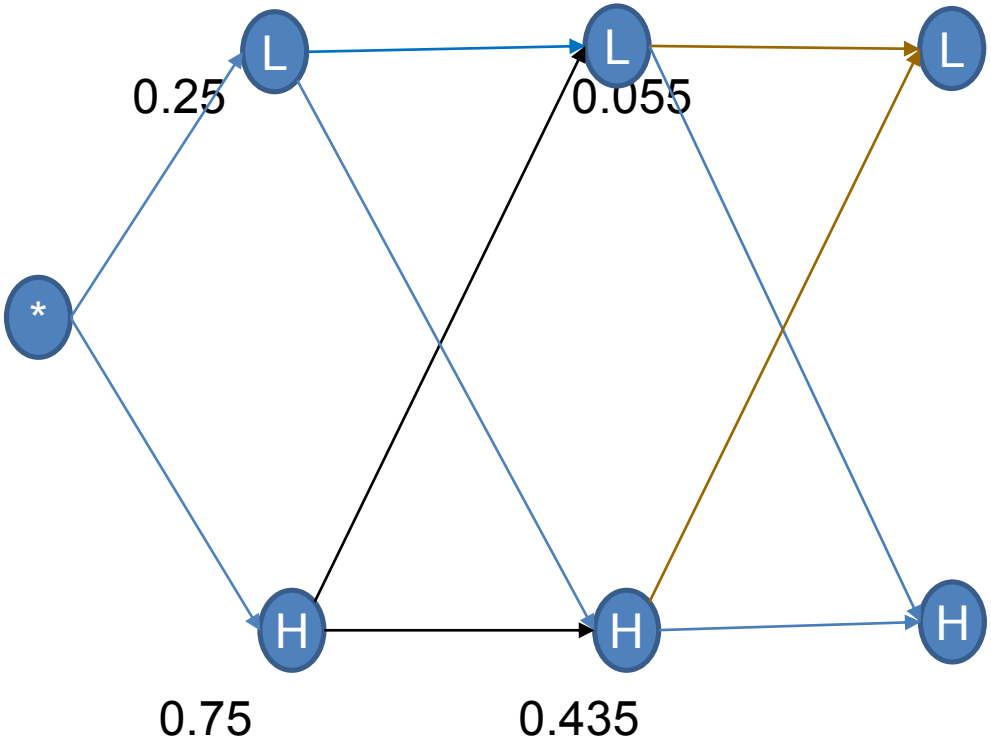
$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t = Rainy)$
0.2	0.6	$P(E_t = Sunny)$

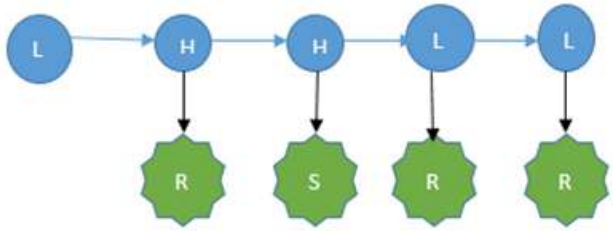
Hidden Morkov Model

Forward Propagation Algorithm : S-S-R

P(L)*P(L|L)*P(R|L) = 0.055*0.5*0.8 = **0.022**

P(H)*P(L|H)*P(R|L) = 0.435*0.2*0.8 = **0.0696**





Transition Model / Probability Matrix

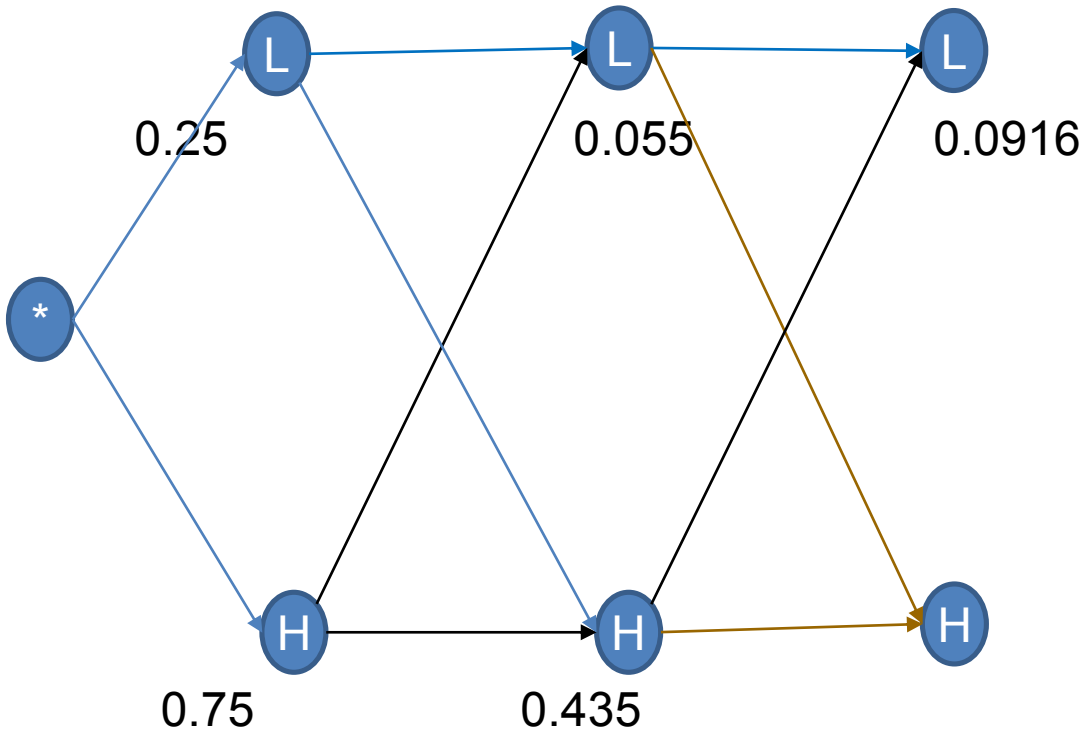
$P(U_{t-1} = HP)$	$P(U_{t-1} = LP)$	←Previous
0.2	0.5	$P(U_t = LP)$
0.8	0.5	$P(U_t = HP)$

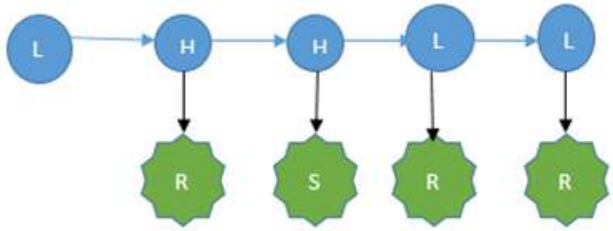
Evidence / Sensor Model/ Emission Probability Matrix

$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t = Rainy)$
0.2	0.6	$P(E_t = Sunny)$

Hidden Morkov Model

Forward Propagation Algorithm : S-S-R





Transition Model / Probability Matrix

$P(U_{t-1} = HP)$	$P(U_{t-1} = LP)$	←Previous
0.2	0.5	$P(U_t = LP)$
0.8	0.5	$P(U_t = HP)$

Evidence / Sensor Model/ Emission Probability Matrix

$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t = Rainy)$
0.2	0.6	$P(E_t = Sunny)$

$P(L)*P(H|L)*P(R|H) = 0.055*0.5*0.4 = \mathbf{0.011}$

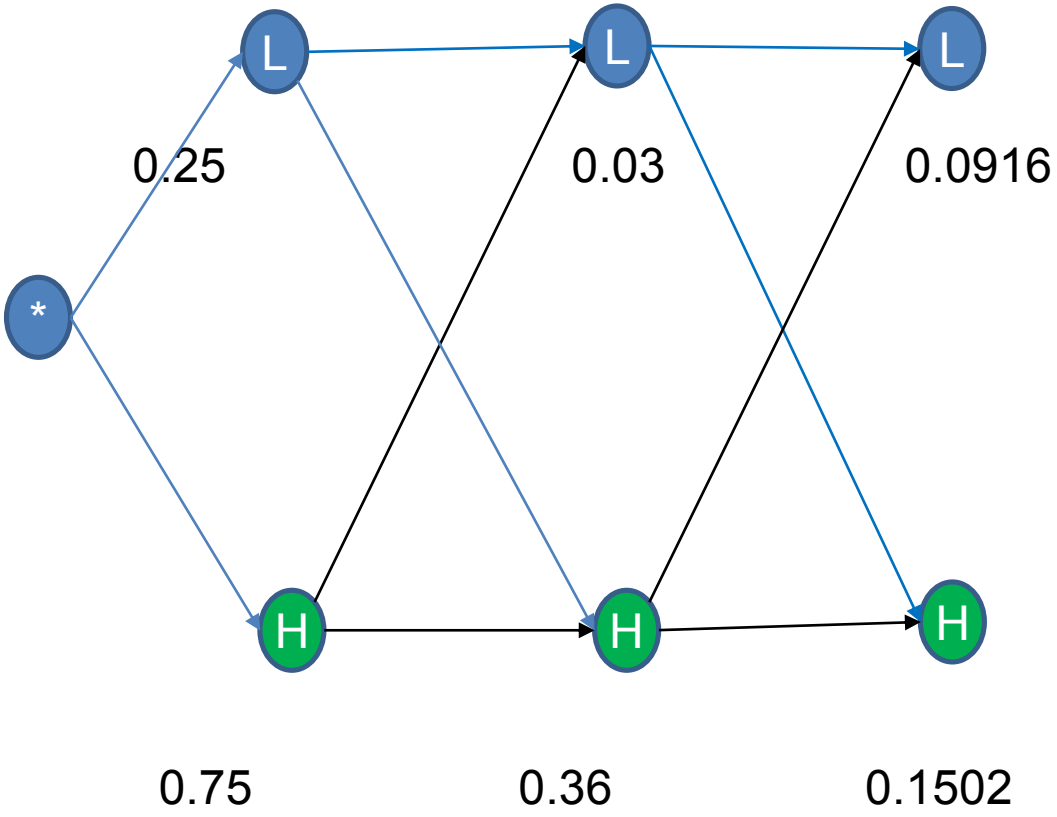
$P(H)*P(H|H)*P(R|H) = 0.435*0.8*0.4 = \mathbf{0.1392}$

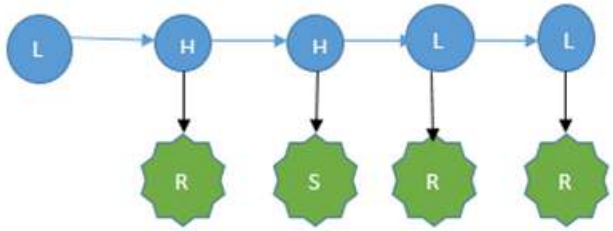
Hidden Morkov Model

Forward Propagation Algorithm : S-S-R

Termination Phase:

P(S-S-R) = 0.2418





Transition Model / Probability Matrix

$P(U_{t-1} = HP)$	$P(U_{t-1} = LP)$	←Previous
0.2	0.5	$P(U_t = LP)$
0.8	0.5	$P(U_t = HP)$

Evidence / Sensor Model/ Emission Probability Matrix

$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t = Rainy)$
0.2	0.6	$P(E_t = Sunny)$

Hidden Morkov Model

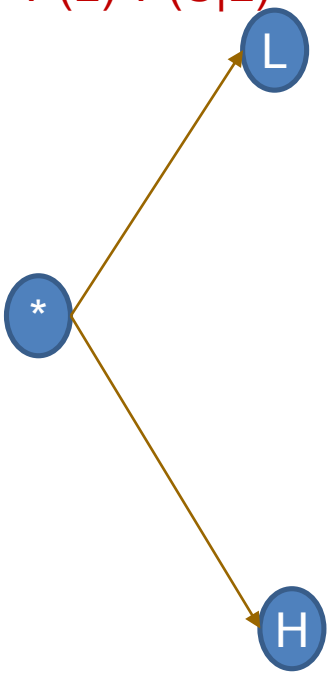
Inference: Type -2

Most Likely Explanation : Veterbi Algorithm

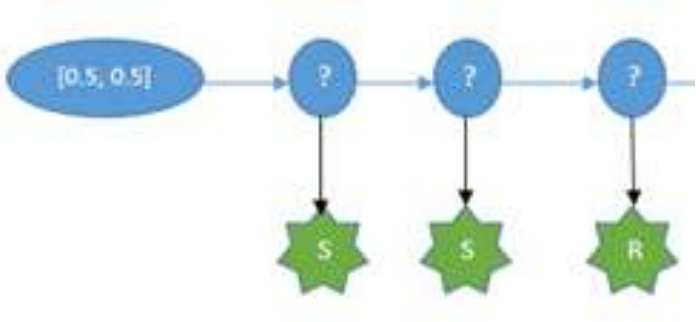
Find the pattern in pressure that might have caused this observation: **S-S-R**

argmax $X_{1...t}$: $P(X_{1...t} \mid E_{1...t})$

P(L)*P(S|L) = 0.5*0.2 = 0.1 → 0.25



P(H)*P(S|H) = 0.5*0.6 = 0.3 → 0.75



Transition Model / Probability Matrix

$P(U_{t-1} = HP)$	$P(U_{t-1} = LP)$	←Previous
0.2	0.5	$P(U_t = LP)$
0.8	0.5	$P(U_t = HP)$

Evidence / Sensor Model/ Emission Probability Matrix

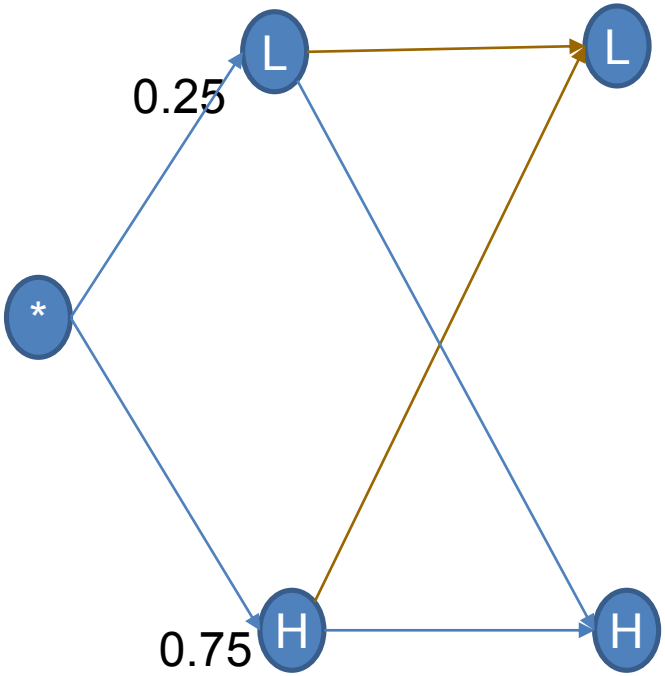
$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t = Rainy)$
0.2	0.6	$P(E_t = Sunny)$

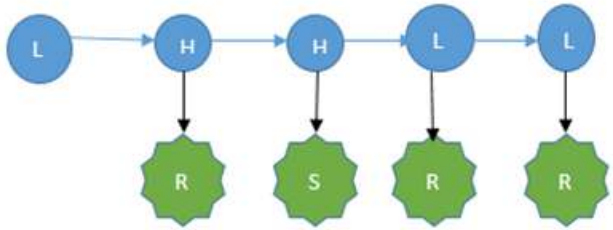
Hidden Morkov Model

Veterbi Algorithm : S-S-R

P(L)*P(L|L)*P(S|L) = 0.25*0.5*0.2 = 0.025

P(H)*P(L|H)*P(S|L) = 0.75*0.2*0.2 = **0.03**





Transition Model / Probability Matrix

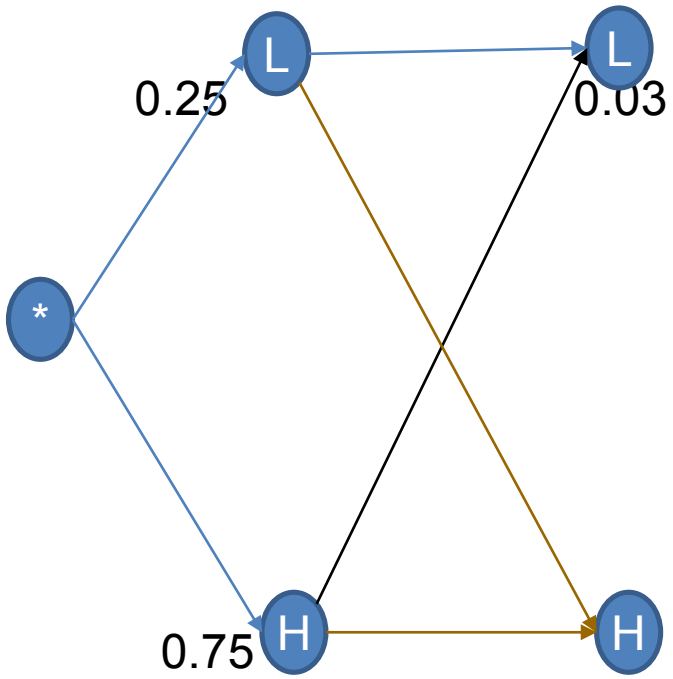
$P(U_{t-1} = HP)$	$P(U_{t-1} = LP)$	←Previous
0.2	0.5	$P(U_t = LP)$
0.8	0.5	$P(U_t = HP)$

Evidence / Sensor Model/ Emission Probability Matrix

$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t = Rainy)$
0.2	0.6	$P(E_t = Sunny)$

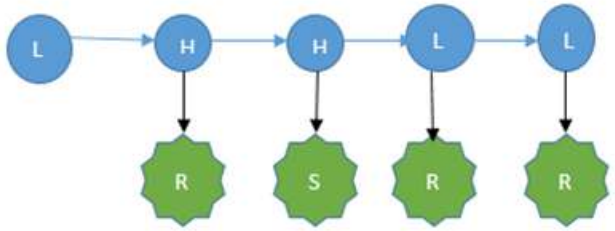
Hidden Morkov Model

Veterbi Algorithm : S-S-R



P(L)*P(H|L)*P(S|H) = 0.25*0.5*0.6 = 0.075

P(H)*P(H|H)*P(S|H) = 0.75*0.8*0.6 = **0.36**



Transition Model / Probability Matrix

$P(U_{t-1} = HP)$	$P(U_{t-1} = LP)$	←Previous
0.2	0.5	$P(U_t = LP)$
0.8	0.5	$P(U_t = HP)$

Evidence / Sensor Model/ Emission Probability Matrix

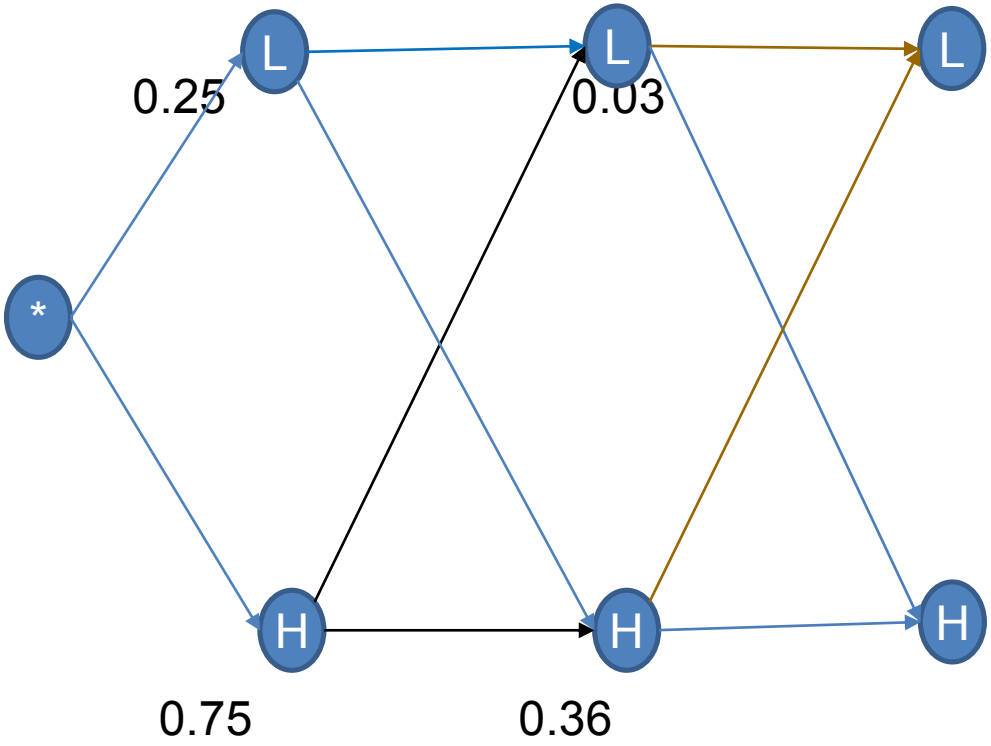
$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t = Rainy)$
0.2	0.6	$P(E_t = Sunny)$

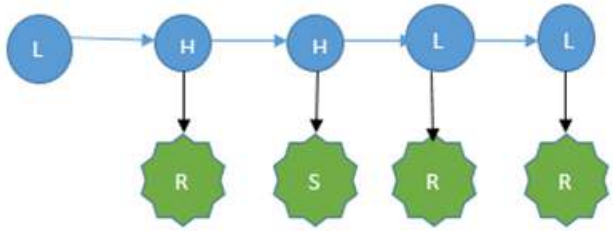
Hidden Morkov Model

Veterbi Algorithm : S-S-R

P(L)*P(L|L)*P(R|L) = 0.03*0.5*0.8 = 0.012

P(H)*P(L|H)*P(R|L) = 0.36*0.2*0.8 = **0.0576**





Transition Model / Probability Matrix

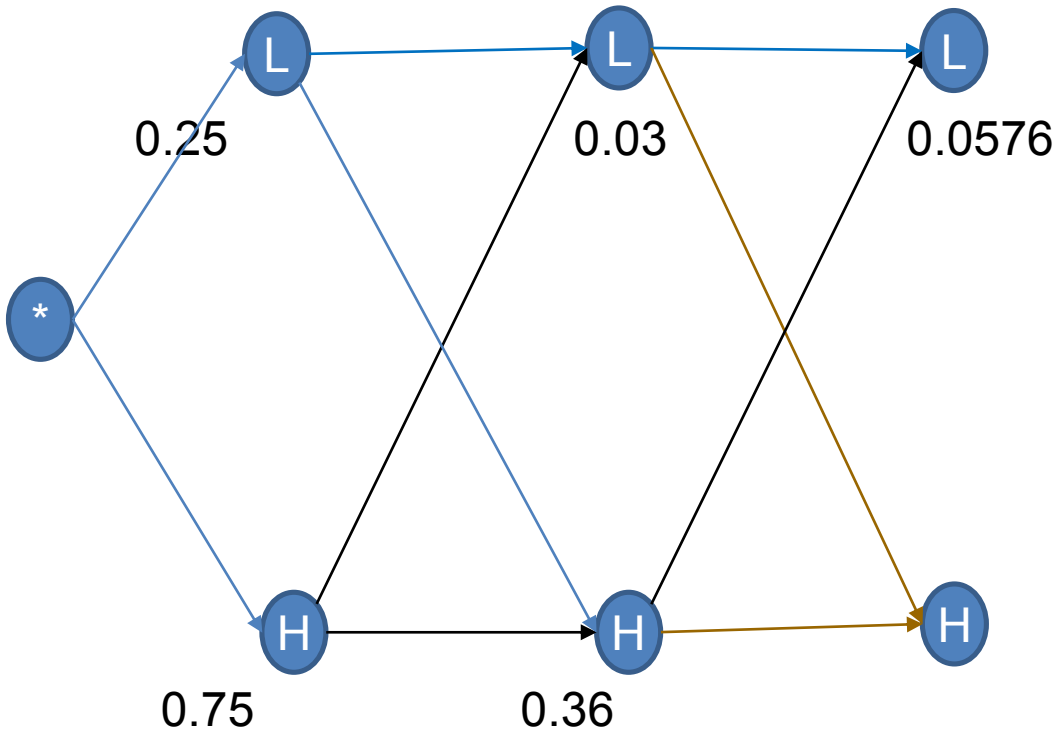
$P(U_{t-1} = HP)$	$P(U_{t-1} = LP)$	←Previous
0.2	0.5	$P(U_t = LP)$
0.8	0.5	$P(U_t = HP)$

Evidence / Sensor Model/ Emission Probability Matrix

$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t = Rainy)$
0.2	0.6	$P(E_t = Sunny)$

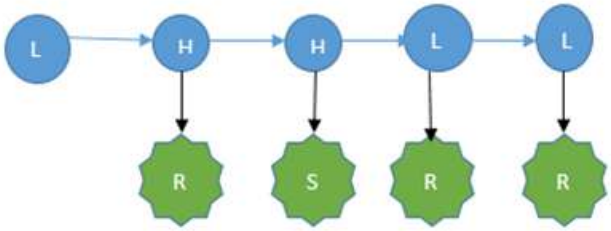
Hidden Morkov Model

Veterbi Algorithm : S-S-R



P(L)*P(H|L)*P(R|H) = 0.03*0.5*0.4 = 0.006

P(H)*P(H|H)*P(R|H) = 0.36*0.8*0.4 = **0.1152**



Transition Model / Probability Matrix

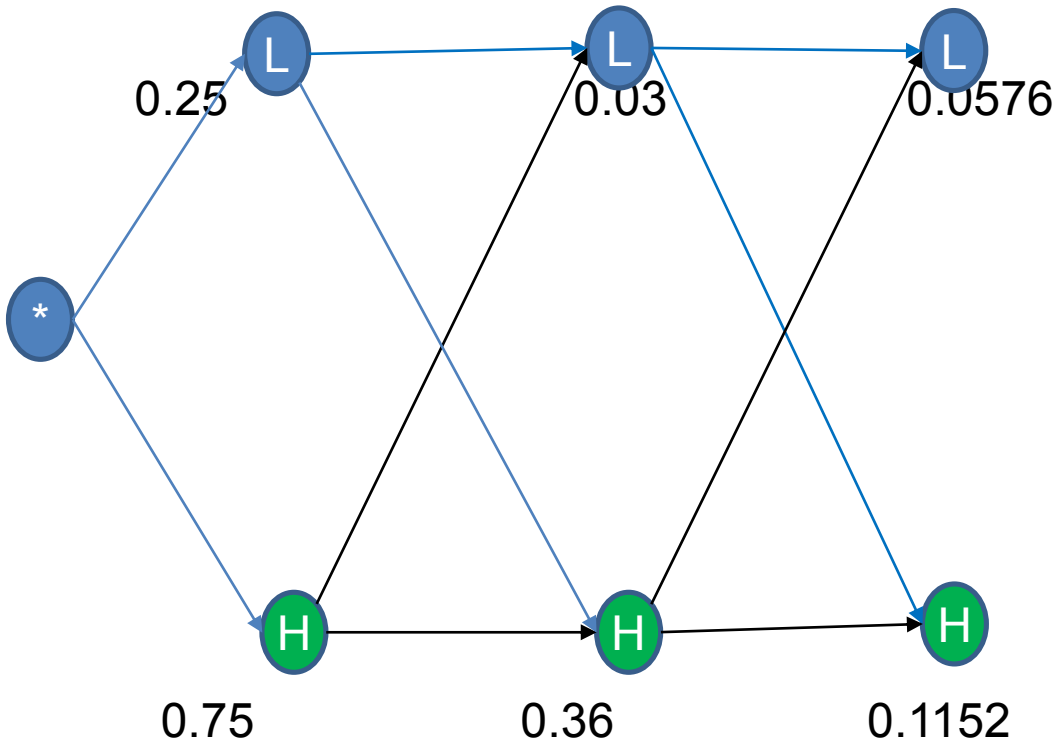
$P(U_{t-1} = HP)$	$P(U_{t-1} = LP)$	←Previous
0.2	0.5	$P(U_t = LP)$
0.8	0.5	$P(U_t = HP)$

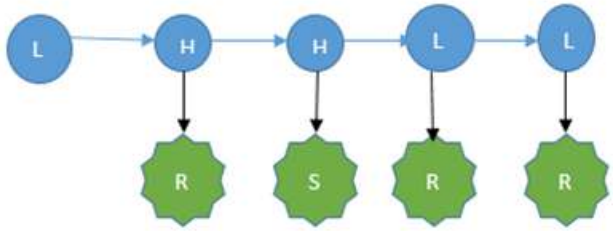
Evidence / Sensor Model/ Emission Probability Matrix

$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t = Rainy)$
0.2	0.6	$P(E_t = Sunny)$

Hidden Morkov Model

Veterbi Algorithm : S-S-R





Transition Model / Probability Matrix

$P(U_{t-1} = HP)$	$P(U_{t-1} = LP)$	←Previous
0.2	0.5	$P(U_t = LP)$
0.8	0.5	$P(U_t = HP)$

Evidence / Sensor Model/ Emission Probability Matrix

$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t = Rainy)$
0.2	0.6	$P(E_t = Sunny)$

Hidden Morkov Model

Inference: Type -3

Filtering : Forward Propagation Algorithm

Find the Current Pressure if sequence of weather observations recorded are: **S-S-R**

Intuition: $P(E_{1...t}) = \sum_{i=1}^{N} P(E_{1...t} \mid X_{1...t}) * P(X_{1...t}) = \sum_{i=1}^{N} \prod_{j=1}^{t} P(Ej \mid Xj) * P(Xj \mid Xj_{-1})$

Transition Model / Probability Matrix

$P(U_{t-1} = HP)$	$P(U_{t-1} = LP)$	←Previous
0.2	0.5	$P(U_t = LP)$
0.8	0.5	$P(U_t = HP)$

Evidence / Sensor Model/ Emission Probability Matrix

$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t = Rainy)$
0.2	0.6	$P(E_t = Sunny)$

Hidden Morkov Model

Filtering : Forward Propagation Algorithm

Find the Current Pressure if sequence of weather observations recorded are: **S-S-R**

Intuition: $P(Xt_{+1}|E_{1\ldots t+1}) = \boldsymbol{\alpha}\, P(et_{+1}|Xt_{+1}) * \sum_{X_t} P(X_{t+1}|X_t) * \boldsymbol{P}(\mathbf{X_t}|\mathbf{E_{1..t}})$

$P(X_3 \mid SSR) = P(X_3 \mid S,S,R)$

$$= \frac{\boldsymbol{P}(\boldsymbol{R}\,|\mathbf{X_3}\,\boldsymbol{S},\boldsymbol{S}) * P(X_3|\,S,S)}{P(R)}$$

$$= \frac{P(R\,|X_3) * \boldsymbol{P}(\mathbf{X_3}|\,\boldsymbol{S},\boldsymbol{S})}{P(R)}$$

$$= \frac{P(R\,|X_3) * \{\sum_{X_2} P(X_3|X_2) * \boldsymbol{P}(\mathbf{X_2}|\,\boldsymbol{S},\boldsymbol{S})\}}{P(R)}$$

$$= \frac{P(R\,|X_3) * \{\sum_{X_2} P(X_3|X_2) * P(S\,|X_2) * \{\sum_{X_1} P(X_2|X_1) * \boldsymbol{P}(\mathbf{X_1}|\,\boldsymbol{S})\}\}}{P(R) * P(S)}$$

Transition Model / Probability Matrix

$P(U_{t-1} = HP)$	$P(U_{t-1} = LP)$	←Previous
0.2	0.5	$P(U_t = LP)$
0.8	0.5	$P(U_t = HP)$

$$P(Xt_{+1}|E_{1\ldots t+1}) = \boldsymbol{\alpha}\, P(e_{t+1}|Xt_{+1}) * \sum_{X_t} P(X_{t+1}|X_t) * \boldsymbol{P}(\mathbf{X_t}|\,\mathbf{E_{1..t}})$$

Evidence / Sensor Model/ Emission Probability Matrix

$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t = Rainy)$
0.2	0.6	$P(E_t = Sunny)$

Hidden Morkov Model

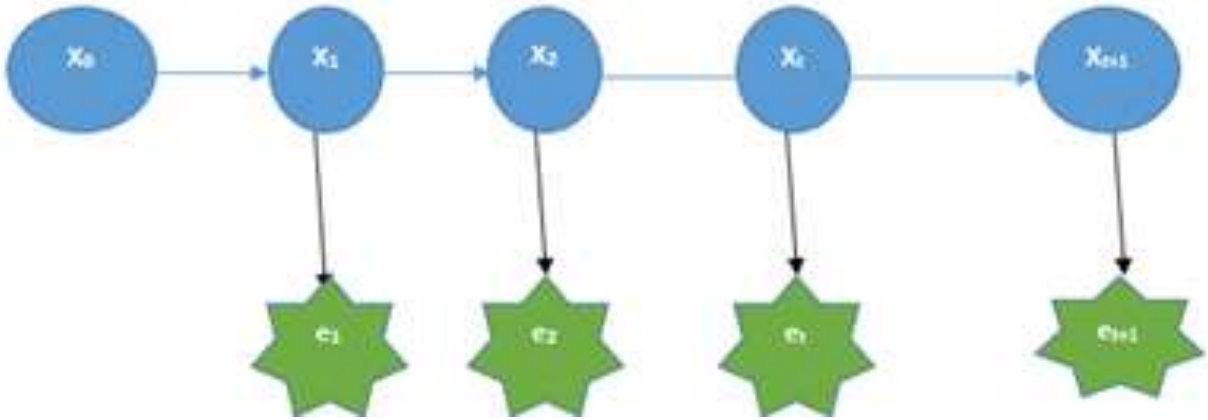
Inference: Type -3

Filtering : Forward Propagation Algorithm

Find the Current Pressure if sequence of weather observations recorded are: **S-S-R**

Intuition: $P(Xt_{+1}|E_{1...t+1}) = \boldsymbol{\alpha}\, P(et_{+1}|Xt_{+1}) * \sum_{X_t} P(X_{t+1}|X_t) * \boldsymbol{P}(\mathbf{X_t}|\mathbf{E_{1..t}})$

$$P(Xt_{+1}|E_{1...t+1}) = \boldsymbol{\alpha}\, P(e_{t+1}|Xt_{+1}) * \sum_{X_t} P(X_{t+1}|X_t) * \boldsymbol{P}(\mathbf{X_t}|\mathbf{E_{1..t}})$$



Transition Model / Probability Matrix

$P(U_{t-1} = HP)$	$P(U_{t-1} = LP)$	←Previous
0.2	0.5	$P(U_t = LP)$
0.8	0.5	$P(U_t = HP)$

Evidence / Sensor Model/ Emission Probability Matrix

$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t = Rainy)$
0.2	0.6	$P(E_t = Sunny)$

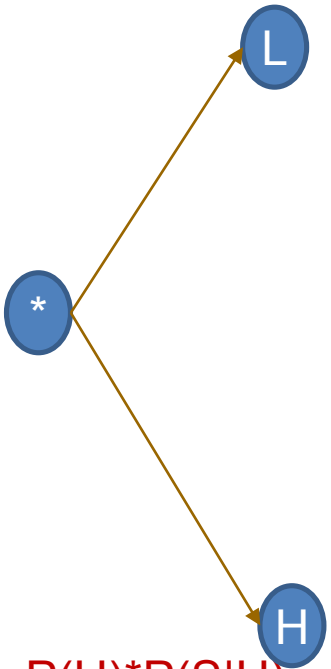
Hidden Morkov Model

Forward Propagation Algorithm

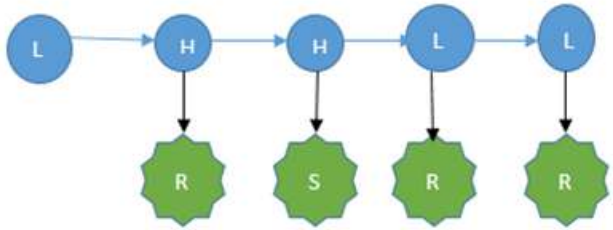
Pressure sequence observation: **S-S-R**

Initialization Phase:

P(L)*P(S|L) = 0.5*0.2 = 0.1 → 0.25



P(H)*P(S|H) = 0.5*0.6 = 0.3 → 0.75



Transition Model / Probability Matrix

$P(U_{t-1} = HP)$	$P(U_{t-1} = LP)$	←Previous
0.2	0.5	$P(U_t = LP)$
0.8	0.5	$P(U_t = HP)$

Evidence / Sensor Model/ Emission Probability Matrix

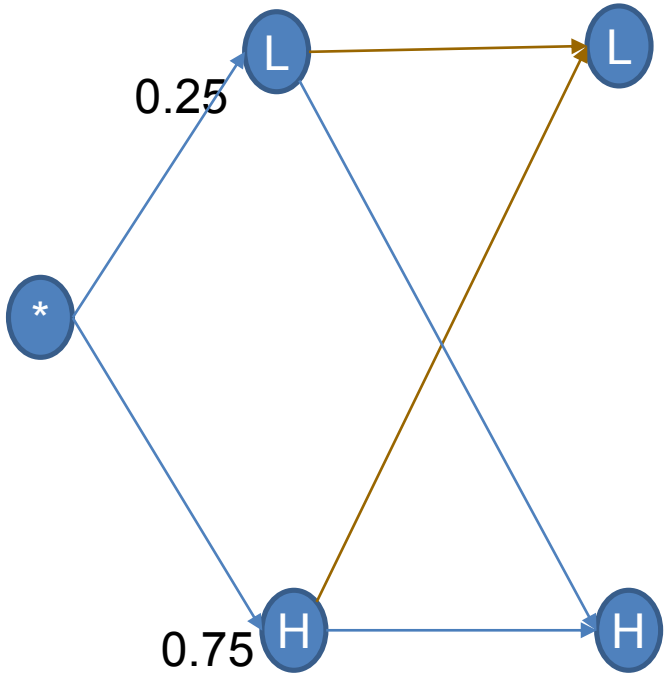
$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t = Rainy)$
0.2	0.6	$P(E_t = Sunny)$

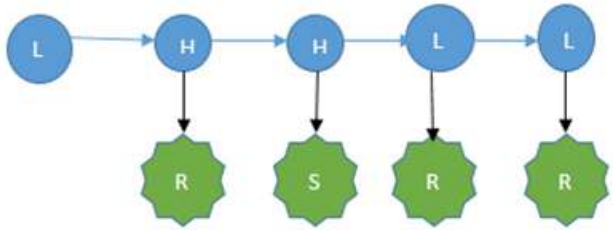
Hidden Morkov Model

Forward Propagation Algorithm : S-S-R

$P(L)*P(L|L)*P(S|L) = 0.25*0.5*0.2 = \mathbf{0.025}$

$P(H)*P(L|H)*P(S|L) = 0.75*0.2*0.2 = \mathbf{0.03}$





Transition Model / Probability Matrix

$P(U_{t-1} = HP)$	$P(U_{t-1} = LP)$	←Previous
0.2	0.5	$P(U_t = LP)$
0.8	0.5	$P(U_t = HP)$

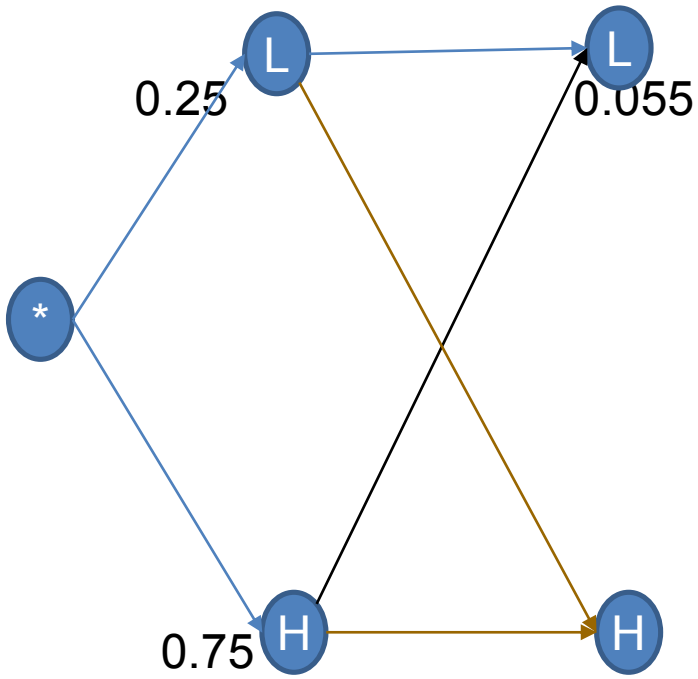
Evidence / Sensor Model/ Emission Probability Matrix

$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t = Rainy)$
0.2	0.6	$P(E_t = Sunny)$

Recursion Phase:

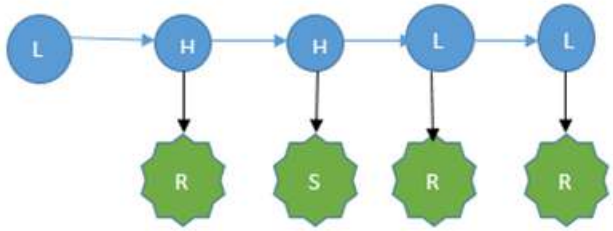
Hidden Morkov Model

Forward Propagation Algorithm : S-S-R



P(L)*P(H|L)*P(S|H) = 0.25*0.5*0.6 = **0.075**

P(H)*P(H|H)*P(S|H) = 0.75*0.8*0.6 = **0.36**



Transition Model / Probability Matrix

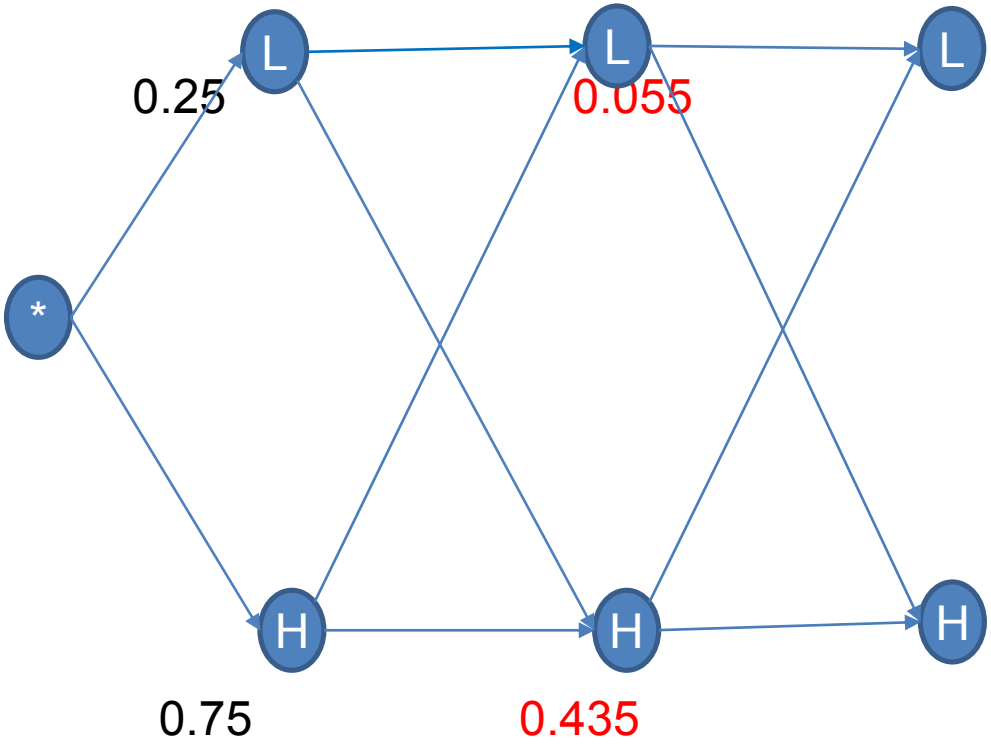
$P(U_{t-1} = HP)$	$P(U_{t-1} = LP)$	←Previous
0.2	0.5	$P(U_t = LP)$
0.8	0.5	$P(U_t = HP)$

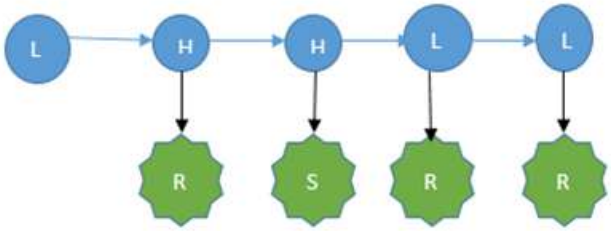
Evidence / Sensor Model/ Emission Probability Matrix

$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t = Rainy)$
0.2	0.6	$P(E_t = Sunny)$

Hidden Morkov Model

Forward Propagation Algorithm : S-S-R





Transition Model / Probability Matrix

$P(U_{t-1} = HP)$	$P(U_{t-1} = LP)$	←Previous
0.2	0.5	$P(U_t = LP)$
0.8	0.5	$P(U_t = HP)$

Evidence / Sensor Model/ Emission Probability Matrix

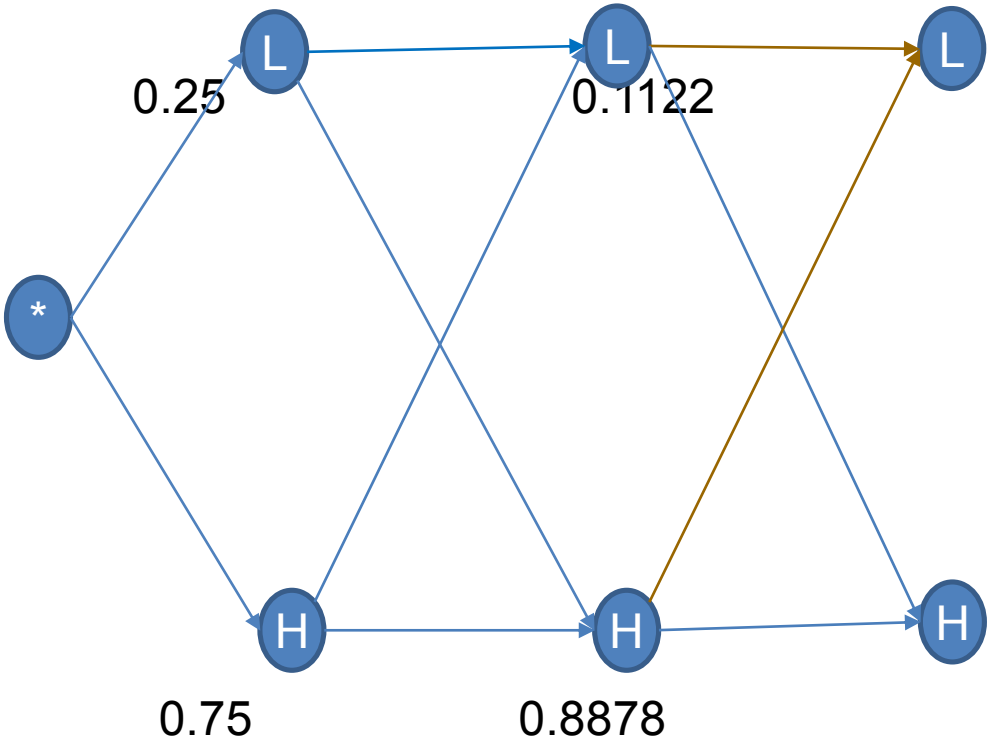
$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t = Rainy)$
0.2	0.6	$P(E_t = Sunny)$

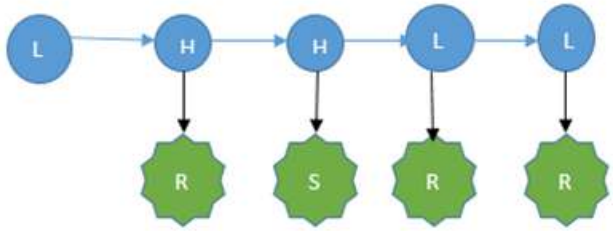
Hidden Morkov Model

Forward Propagation Algorithm : S-S-R

P(L)*P(L|L)*P(R|L) = 0.1122*0.5*0.8 = **0.04488**

P(H)*P(L|H)*P(R|L) = 0.8878*0.2*0.8 = **0.142048**





Transition Model / Probability Matrix

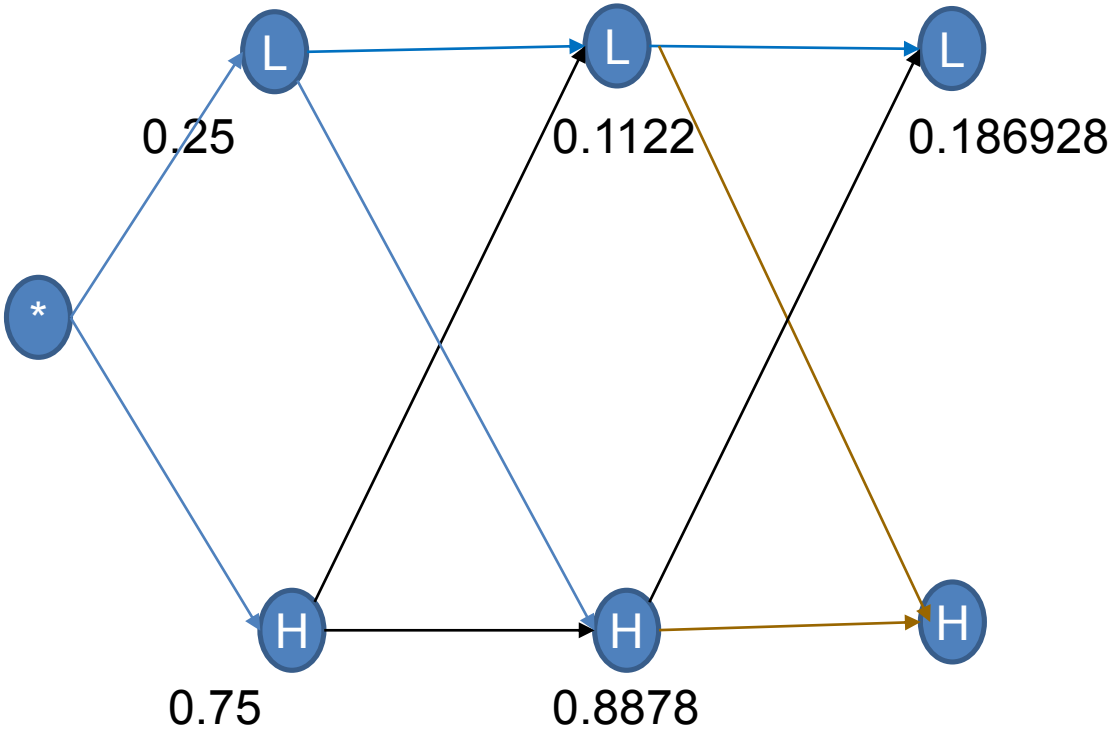
$P(U_{t-1} = HP)$	$P(U_{t-1} = LP)$	←Previous
0.2	0.5	$P(U_t = LP)$
0.8	0.5	$P(U_t = HP)$

Evidence / Sensor Model/ Emission Probability Matrix

$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t = Rainy)$
0.2	0.6	$P(E_t = Sunny)$

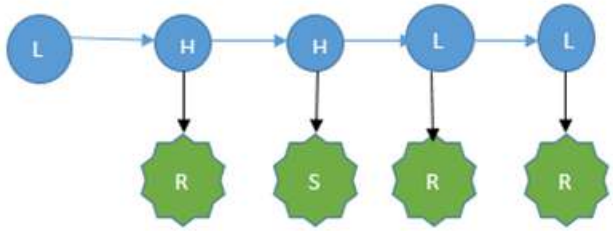
Hidden Morkov Model

Forward Propagation Algorithm : S-S-R



P(L)*P(H|L)*P(R|H) = 0.1122*0.5*0.4 = **0.02244**

P(H)*P(H|H)*P(R|H) = 0.8878*0.8*0.4 = **0.284096**



Transition Model / Probability Matrix

$P(U_{t-1} = HP)$	$P(U_{t-1} = LP)$	←Previous
0.2	0.5	$P(U_t = LP)$
0.8	0.5	$P(U_t = HP)$

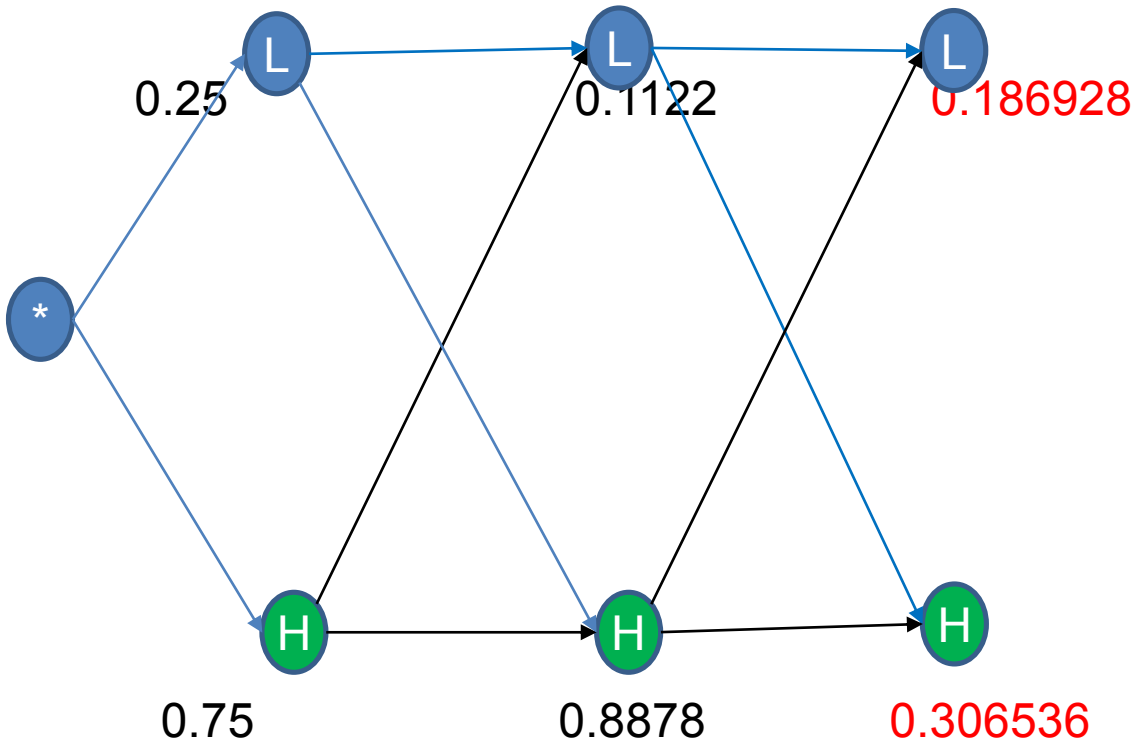
Evidence / Sensor Model/ Emission Probability Matrix

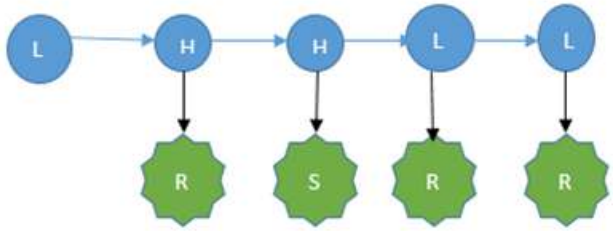
$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t = Rainy)$
0.2	0.6	$P(E_t = Sunny)$

Hidden Morkov Model

Forward Propagation Algorithm : S-S-R

Termination Phase:





Transition Model / Probability Matrix

$P(U_{t-1} = HP)$	$P(U_{t-1} = LP)$	←Previous
0.2	0.5	$P(U_t = LP)$
0.8	0.5	$P(U_t = HP)$

Evidence / Sensor Model/ Emission Probability Matrix

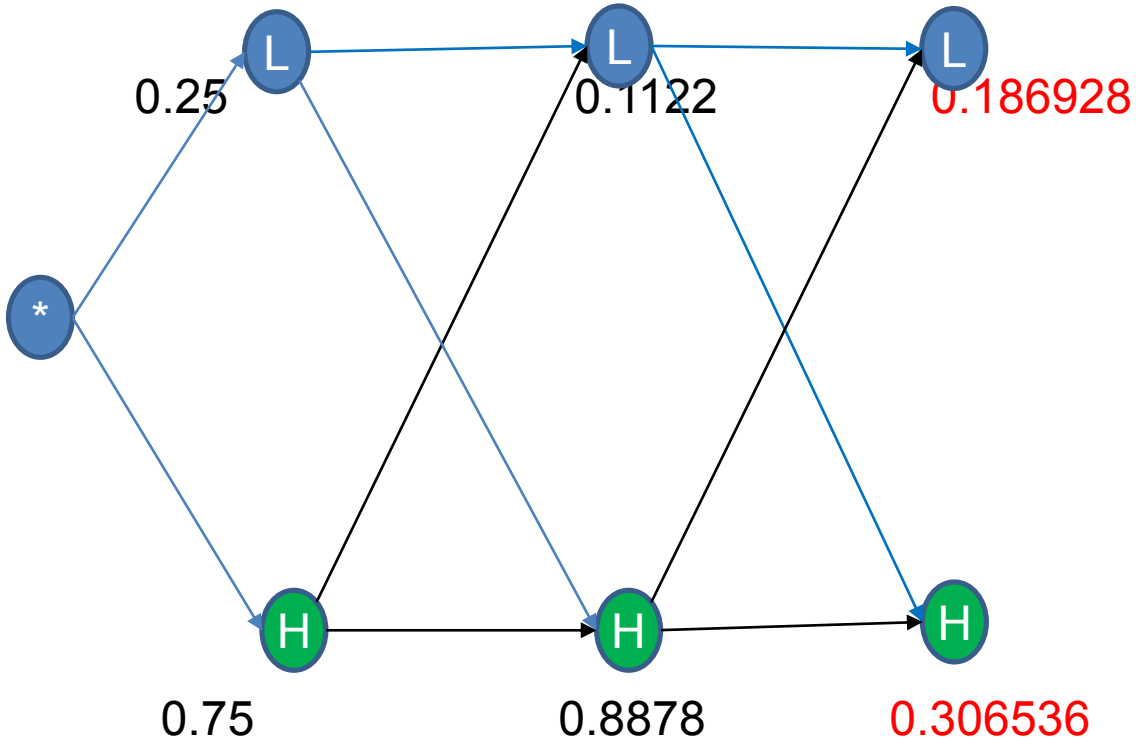
$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t = Rainy)$
0.2	0.6	$P(E_t = Sunny)$

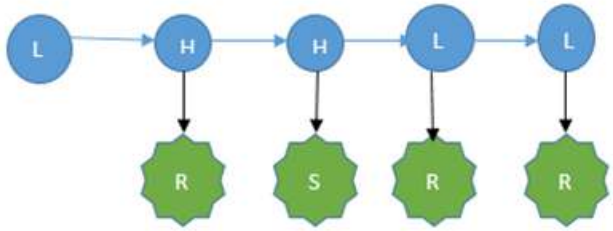
Hidden Morkov Model

Forward Propagation Algorithm : S-S-R

Termination Phase:

(0.37881,**0.62119**)





Transition Model / Probability Matrix

$P(U_{t-1} = HP)$	$P(U_{t-1} = LP)$	←Previous
0.2	0.5	$P(U_t = LP)$
0.8	0.5	$P(U_t = HP)$

Evidence / Sensor Model/ Emission Probability Matrix

$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t$ = **Rainy**)
0.2	0.6	$P(E_t$ = **Sunny**)

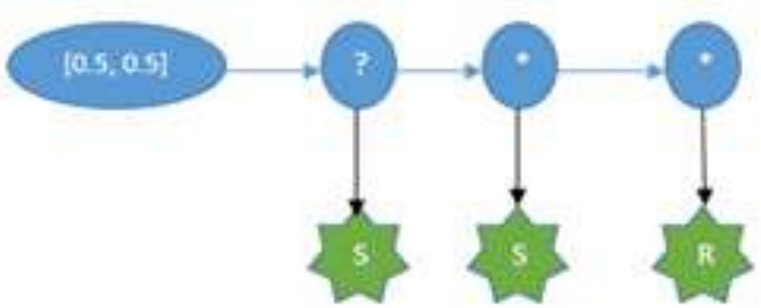
Hidden Morkov Model

Inference: Type -4

Smoothing : Backward Propagation Algorithm (Most Likely State Estimation)

Find the Pressure in past instance of time if sequence of following future weather observations recorded are: **S-S-R**

Intuition: $P(E_{1...t}) = \sum_{i=1}^{N} P(E_{1...t} \mid X_{1...t}) * P(X_{1...t}) = \sum_{i=1}^{N} \prod_{j=1}^{t} P(Ej \mid Xj) * P(Xj \mid Xj_{-1})$



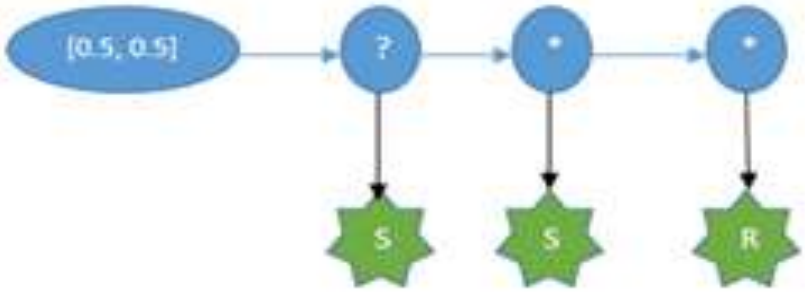
Transition Model / Probability Matrix

$P(U_{t-1} = HP)$	$P(U_{t-1} = LP)$	←Previous
0.2	0.5	$P(U_t = LP)$
0.8	0.5	$P(U_t = HP)$

Evidence / Sensor Model/ Emission Probability Matrix

$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t = Rainy)$
0.2	0.6	$P(E_t = Sunny)$

Hidden Morkov Model



Smoothing : Backward Propagation Algorithm

Find the Pressure in past instance of time if sequence of following future weather observations recorded are: **S-S-R**

Intuition: $P(Xt_{+1}|E_{1\ldots t+1}) = \boldsymbol{\alpha}\, P(et_{+1}|Xt_{+1}) * \sum_{X_t} P(X_{t+1}|X_t) * \boldsymbol{P}(\mathbf{X_t}|\mathbf{E_{1..t}})$

$P(X_1 \mid SSR) = P(X_1 \mid S,S,R)$

$$= \frac{\boldsymbol{P}(\boldsymbol{SR}\,|\mathbf{X_1}\,\boldsymbol{S}) * P(X_1|\,S)}{P(SR)}$$

$$= \frac{P(X_1|\,S) * \{\sum_{X_2} P(X_2|X_1) * \boldsymbol{P}(\mathbf{SR}\,|\,\mathbf{X_2}\,\mathbf{X_1})\}}{P(SR)}$$

$$= \frac{P(X_1|\,S) * \{\sum_{X_2} P(X_2|X_1) * \boldsymbol{P}(\mathbf{SR}\,|\,\mathbf{X_2})\}}{P(SR)}$$

$$= \frac{P(X_1|\,S) * \{\sum_{X_2} P(X_2|X_1) * P(S\,|\,X_2) * \boldsymbol{P}(\mathbf{R}\,|\,\mathbf{X_2})\}}{P(SR)}$$

$$= \frac{P(X_1|\,S) * \{\sum_{X_2} P(X_2|X_1) * P(S\,|\,X_2) * \{\sum_{X_3} P(X_3|X_2) * P(R\,|\,X_3) * \boldsymbol{P}(.\,|\,\mathbf{X_3})\}\}}{P(SR)}$$

Transition Model / Probability Matrix

$P(U_{t-1} = HP)$	$P(U_{t-1} = LP)$	←Previous
0.2	0.5	$P(U_t = LP)$
0.8	0.5	$P(U_t = HP)$

Evidence / Sensor Model/ Emission Probab

$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t = Rainy)$
0.2	0.6	$P(E_t = Sunny)$

$$P(Xt\,|E_{t+1,\,t+2\,\ldots\,z}) = \boldsymbol{\alpha} * \text{fwd msg} * \sum_{X_{t+1}} P(X_{t+1}|X_t) * P(e_{t+1}|X_{t+1}) * \boldsymbol{P}(\mathbf{E_{t+2..z}}|\,\mathbf{X_{t+1}})$$

Hidden Morkov Model

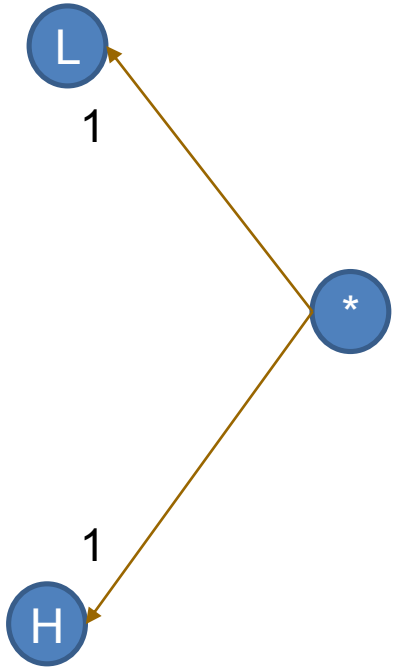
Backward Propagation Algorithm

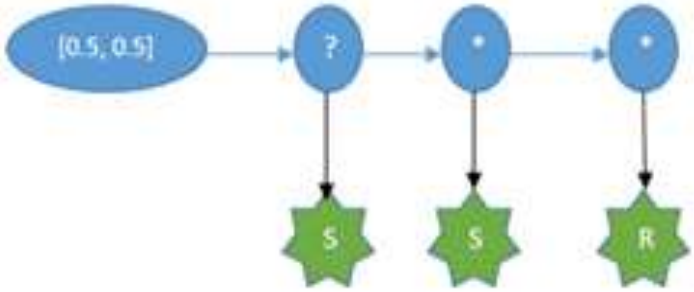
Pressure sequence observation: **S-S-R**

Initialization Phase: Set value 1 for the terminal state

P(L|L)*P(R|L)*P(.|L) = 0.5*0.8 * 1= 0.40

P(L|H)*P(R|L)*P(.|L) = 0.2*0.8 *1= 0.16





Transition Model / Probability Matrix

$P(U_{t-1} = HP)$	$P(U_{t-1} = LP)$	←Previous
0.2	0.5	$P(U_t = LP)$
0.8	0.5	$P(U_t = HP)$

Evidence / Sensor Model/ Emission Probability Matrix

$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t = Rainy)$
0.2	0.6	$P(E_t = Sunny)$

P(H|L)*P(R|H)*P(.|H) = 0.5*0.4 * 1= 0.20

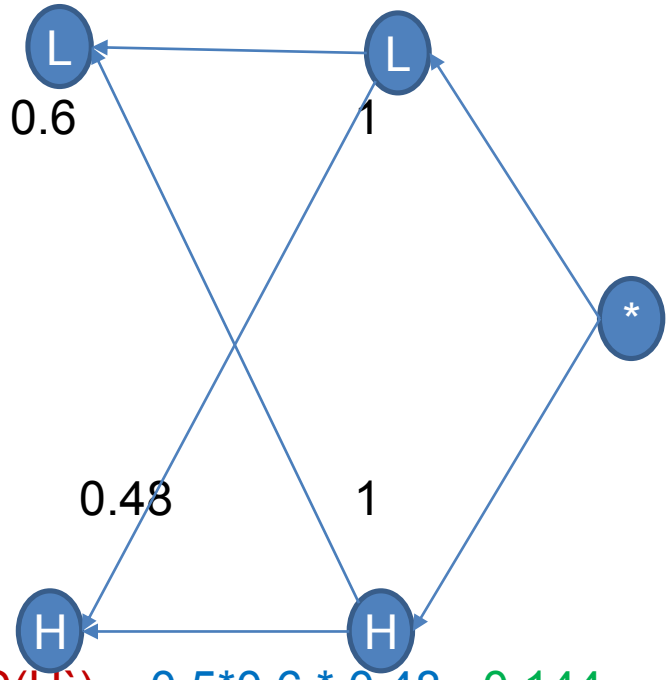
P(H|H)*P(R|H)*P(.|H) = 0.8*0.4 *1= 0.32

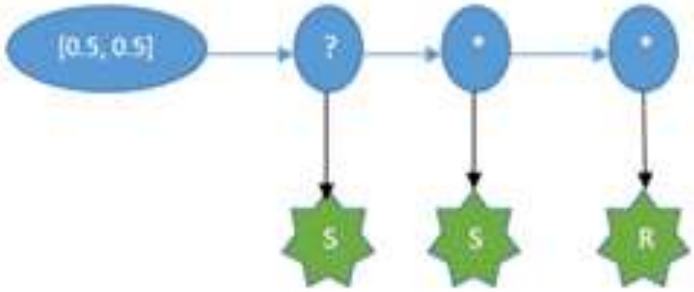
Hidden Morkov Model

Backward Propagation Algorithm : S-S-R

P(L|L)*P(S|L)*MSG(L`) = 0.5*0.2 * 0.6= 0.06

P(L|H)*P(S|L)*MSG(L`) = 0.2*0.2 *0.6= 0.024





Transition Model / Probability Matrix

$P(U_{t-1} = HP)$	$P(U_{t-1} = LP)$	←Previous
0.2	0.5	$P(U_t = LP)$
0.8	0.5	$P(U_t = HP)$

Evidence / Sensor Model/ Emission Probability Matrix

$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t$ = Rainy)
0.2	0.6	$P(E_t$ = Sunny)

P(H|L)*P(S|H)*MSG(H`) = 0.5*0.6 * 0.48= 0.144

P(H|H)*P(S|H)*MSG(H`) = 0.8*0.6 *0.48= 0.2304

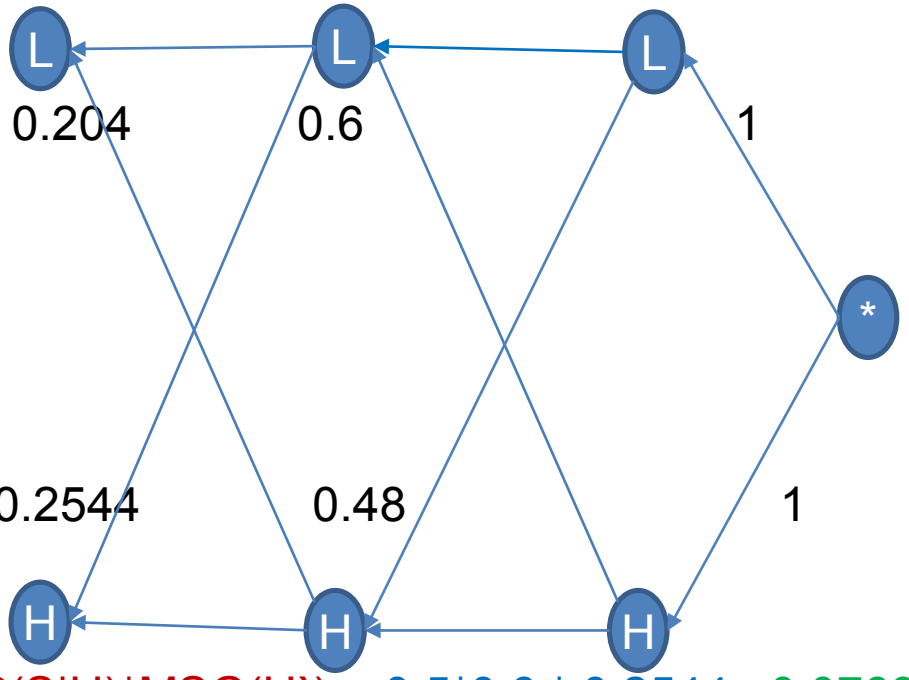
Recursion Phase:

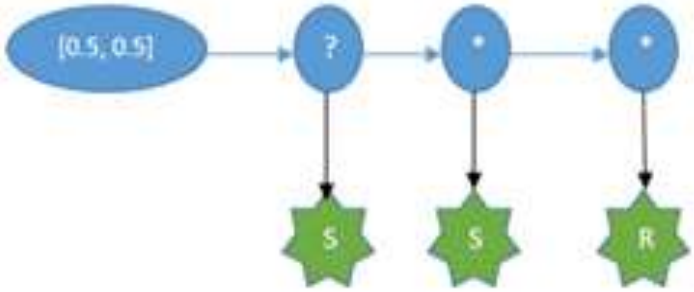
Hidden Morkov Model

Backward Propagation Algorithm : S-S-R

P(L|L)*P(S|L)*MSG(L`) = 0.5*0.2 *0.204= 0.0204

P(L|H)*P(S|L)*MSG(L`) = 0.2*0.2 *0.204= 0.00816





Transition Model / Probability Matrix

$P(U_{t-1} = HP)$	$P(U_{t-1} = LP)$	←Previous
0.2	0.5	$P(U_t = LP)$
0.8	0.5	$P(U_t = HP)$

Evidence / Sensor Model/ Emission Probability Matrix

$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t = Rainy)$
0.2	0.6	$P(E_t = Sunny)$

P(H|L)*P(S|H)*MSG(H`) = 0.5*0.6 * 0.2544= 0.07632

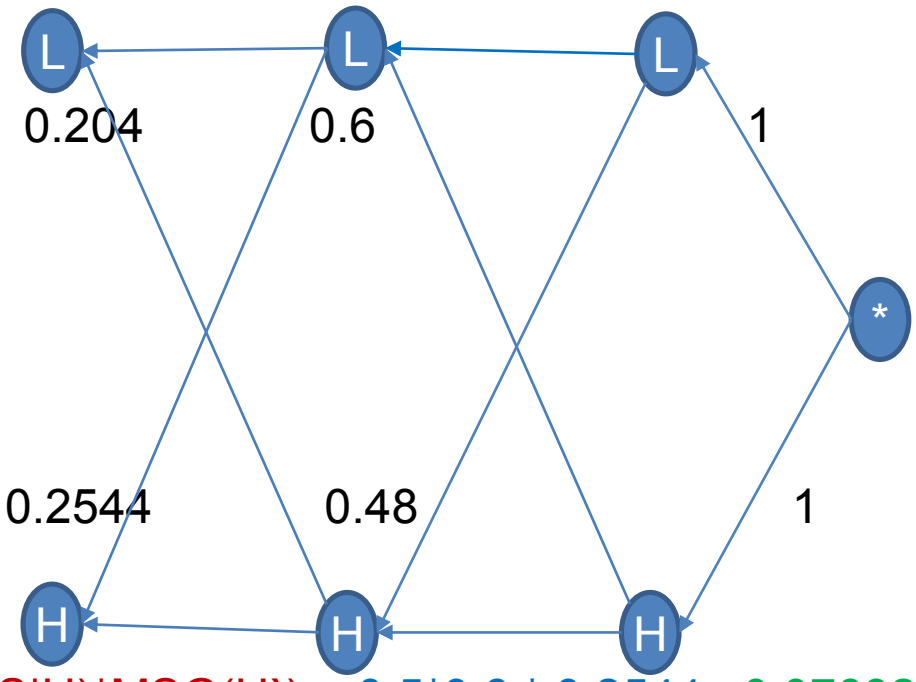
P(H|H)*P(S|H)*MSG(H`) = 0.8*0.6 * 0.2544= 0.122112

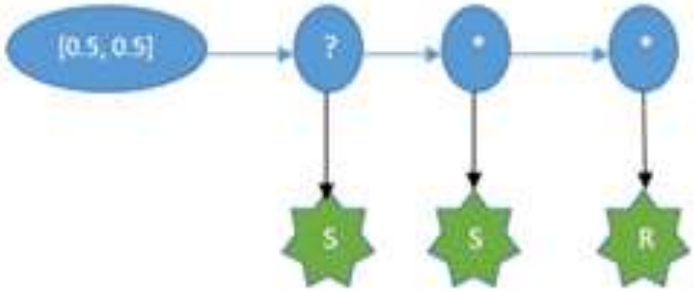
Recursion Phase: If it continues !!!!

Hidden Morkov Model—L17

Backward Propagation Algorithm : S-S-R

P(L)*P(S|L)*MSG(L`) = 0.5*0.2 * 0.204= 0.0204





Transition Model / Probability Matrix

$P(U_{t-1} = HP)$	$P(U_{t-1} = LP)$	←Previous
0.2	0.5	$P(U_t = LP)$
0.8	0.5	$P(U_t = HP)$

Evidence / Sensor Model/ Emission Probability Matrix

$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t = Rainy)$
0.2	0.6	$P(E_t = Sunny)$

P(H)*P(S|H)*MSG(H`) = 0.5*0.6 * 0.2544= 0.07632

Termination Phase: (0.2109,0.7890)

Normalize :Initial value * Emission at start* backMsg

Hidden Morkov Model

Inference: Type -5

Smoothing : Backward Propagation Algorithm

Find the Pressure in past instance of time if sequence of following future weather observations recorded are: **S-S-R**

Intuition: $P(Xt_{+1}|E_{1\ldots t+1}) = \boldsymbol{\alpha}\, P(et_{+1}|Xt_{+1}) * \sum_{X_t} P(X_{t+1}|X_t) * \boldsymbol{P}(\mathbf{X_t}|\mathbf{E_{1..t}})$

$P(X_1 \mid SSR) = P(X_1 \mid S,S,R)$

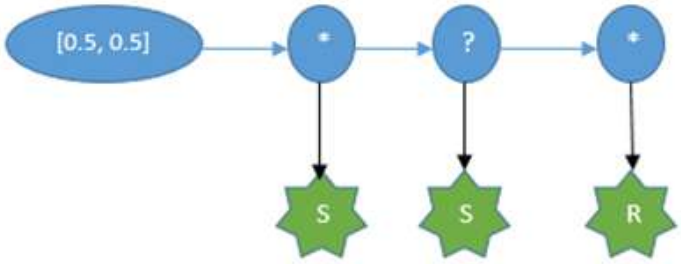
$$= \frac{\boldsymbol{P}(\boldsymbol{SR}\,|\mathbf{X_1}\,\boldsymbol{S}) * P(X_1|\,S)}{P(SR)}$$

$$= \frac{P(X_1|\,S) * \{\sum_{X_2} P(X_2|X_1) * \boldsymbol{P}(\mathbf{SR}\,|\,\mathbf{X_2}\,\mathbf{X_1})\}}{P(SR)}$$

$$= \frac{P(X_1|\,S) * \{\sum_{X_2} P(X_2|X_1) * \boldsymbol{P}(\mathbf{SR}\,|\,\mathbf{X_2})\}}{P(SR)}$$

$$= \frac{P(X_1|\,S) * \{\sum_{X_2} P(X_2|X_1) * P(S\,|\,X_2) * \boldsymbol{P}(\mathbf{R}\,|\,\mathbf{X_2})\}}{P(SR)}$$

$$= \frac{P(X_1|\,S) * \{\sum_{X_2} P(X_2|X_1) * P(S\,|\,X_2) * \{\sum_{X_3} P(X_3|X_2) * P(R\,|\,X_3) * \boldsymbol{P}(.\,|\,\mathbf{X_3})\}\}}{P(SR)}$$



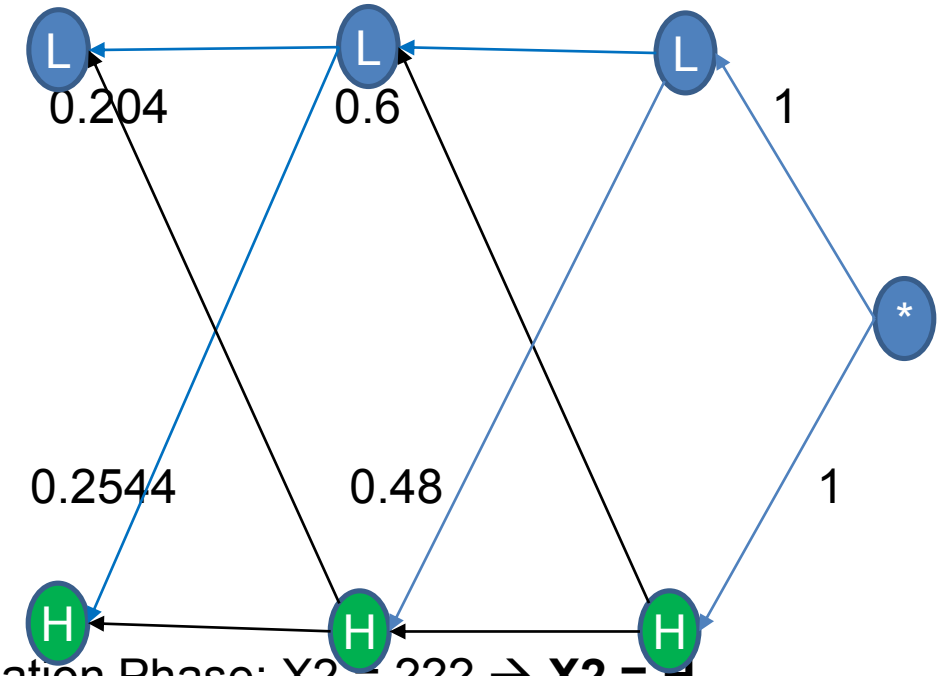
$$P(Xt\,|E_{t+1,\,t+2\,\ldots\,z}) = \boldsymbol{\alpha} * \text{fwd msg} * \sum_{X_{t+1}} P(X_{t+1}|X_t) * P(e_{t+1}|X_{t+1}) * \boldsymbol{P}(\mathbf{E_{t+2..z}}|\,\mathbf{X_{t+1}})$$

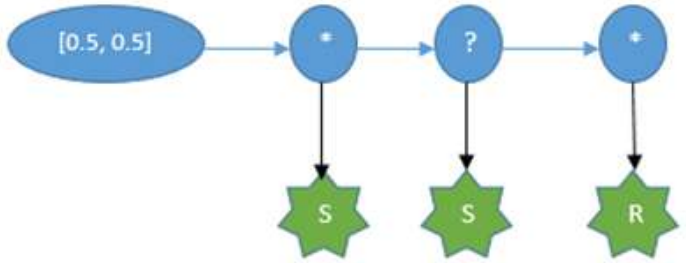
Hidden Morkov Model

Forward Backward Propagation Algorithm : S-S-R

P(X2 | SSR) = α * P(X2|SS) *P(R|X2)

P(X2 | SSR) = α * (0.1122 , 0.8878)* (0.6, 0.48) = (0.06732, 0.426144) = (0.1364,0.8636)





Transition Model / Probability Matrix

$P(U_{t-1} = HP)$	$P(U_{t-1} = LP)$	←Previous
0.2	0.5	$P(U_t = LP)$
0.8	0.5	$P(U_t = HP)$

Evidence / Sensor Model/ Emission Probability Matrix

$P(X_t = LP)$	$P(X_t = HP)$	←Unobserved Evidence v
0.8	0.4	$P(E_t$ = Rainy)
0.2	0.6	$P(E_t$ = Sunny)

Termination Phase: X2 = ??? → **X2 = H**

Initial Value * Emission at the start * backMsg

Text & Natural Language Processing

Initial	Prob
N	0.67
D	0.33
V	0
J	0
A	0
P	0

N	D	V	J	A	P	
	01675		0.67		1	N
		0.571				D
0.63	0.1675	0.143				V
	0.33	0.143				J
						A
	0.1675					P
0.37	0.1675	0.143	0.33			E

Given the corpus with tags to build training data:

1. Create initial probability matrix.
2. Transition probability matrix
3. Emission probability matrix
4. Use HMM Veterbi algorithm to predict the sequence of PoS Tags for given test data / sentence.

In the HMM model , the PoS tags act as the hidden states and the word in the given test sentence as the observed states.

- Boys are taller.
  N V J
- This is the tree.
  D V D N
- She is a tall girl.
  N V D J N
- Trees are more.
  N V D
- Girls are more than boys.
  N V D P N
- The tall tree is falling.
  D J N V V

Initial	Prob
N	0.67
D	0.33
V	0
J	0
A	0
P	0

N	D	V	J	A	P	
	01675		0.67		1	N
		0.571				
0.63	0.1675	0.143				
	0.33	0.143				
	0.1675					
0.37	0.1675	0.143	0.33			

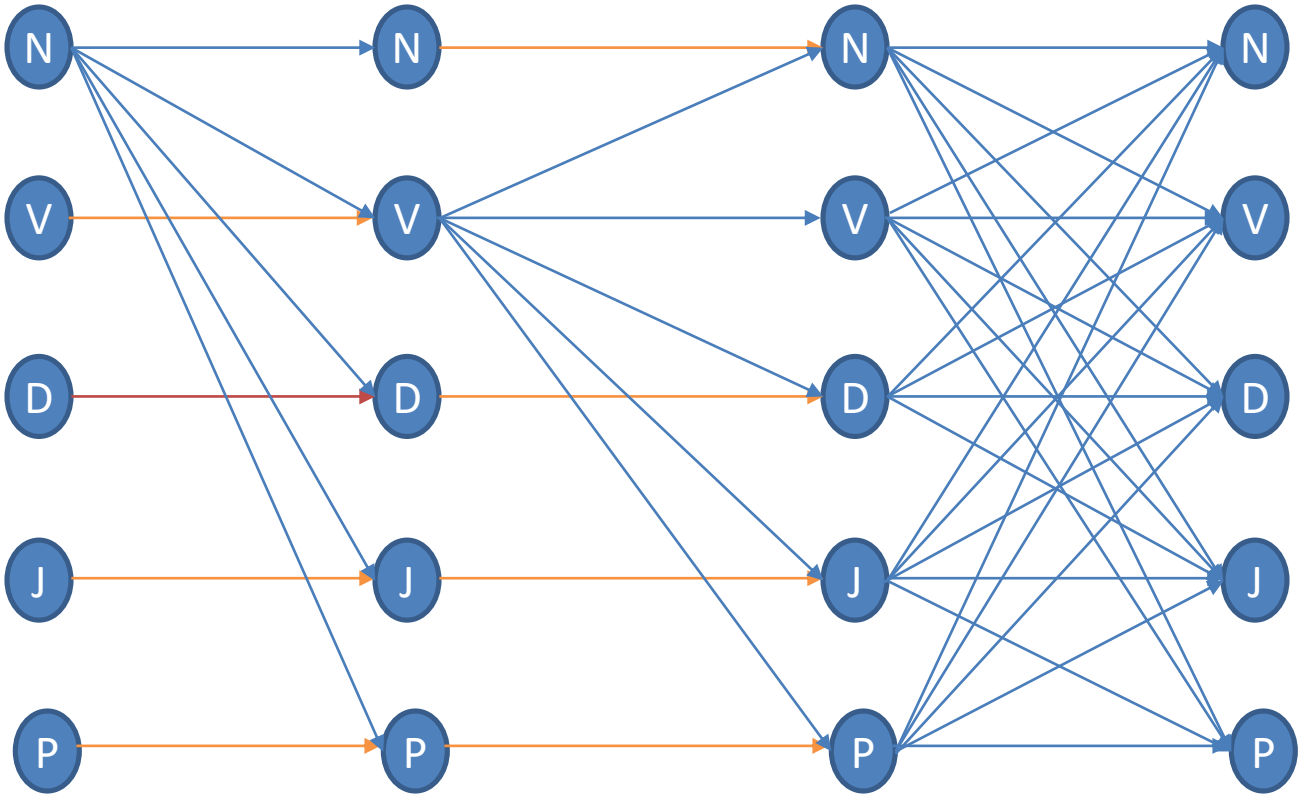
N	D	V	J	A	P	
0.25						Boys
		0.43				Are
			1			Tall
	0.17					This
		0.43				Is
	0.33					The
0.375						Tree
0.125						She
	0.17					A
0.25						Girl
	0.33					More
					1	Than
		0.14				fall

Exercise :

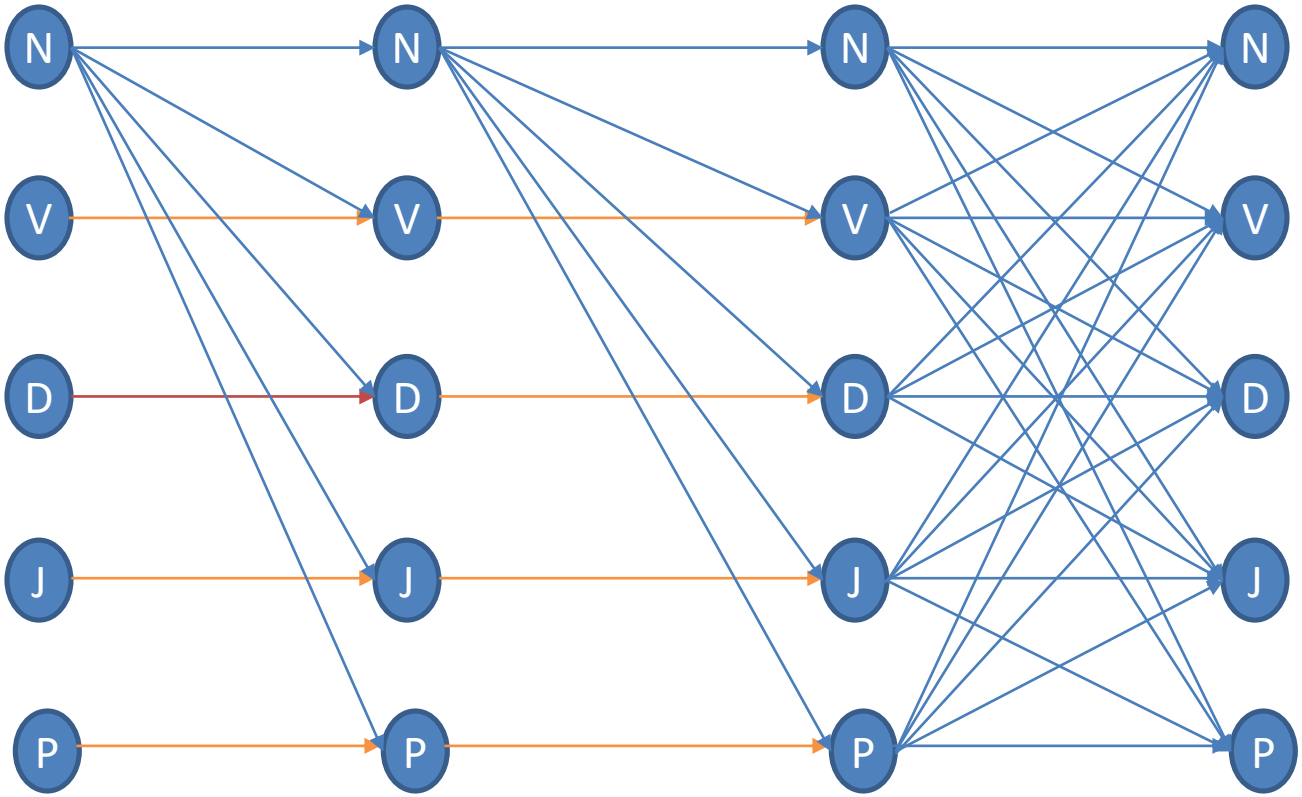
For the below test data/sentence, using the tables constructed using training data, predict the PoS tags.

"Girls are falling"

Sample Sequence under Test: Start → Noun → Verb → …….
Assume Noun→ Verb is the maximum Value



Sample Sequence under Test: Start → Noun → Noun→ …….
Assume Noun→ Noun is the maximum Value



Required Reading: AIMA

Thank You for all your Attention

Note : Some of the slides are adopted from AIMA TB materials