VIT BHOPAL
www.vitbhopal.ac.in

Expectation-Maximization Algorithm

Dr. Virendra Singh Kushwah

Assistant Professor Grade-II

School of Computing Science and Engineering

Virendra.Kushwah@vitbhopal.ac.in

7415869616

- In the real-world applications of machine learning, it is very common that there are many relevant features available for learning but only a small subset of them are observable. So, for the variables which are sometimes observable and sometimes not, then we can use the instances when that variable is visible is observed for the purpose of learning and then predict its value in the instances when it is not observable.

G5

	A	B	C	D	E
1	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
2	1	Male	19	15	39
3	2	Male	21	15	81
4	3	Female	20	16	6
5	4	Female	23	16	77
6	5	Female	31	17	40
7	6	Female	22	17	76
8	7	Female	35	18	6
9	8	Female	23	18	94
10	9	Male	64	19	3
11	10	Female	30	19	72
12	11	Male	67	19	14
13	12	Female	35	19	99
14	13	Female	58	20	15
15	14	Female	24	20	77
16	15	Male	37	20	13
17	16	Male	22	20	79
18	17	Female	35	21	35
19	18	Male	20	21	66
20	19	Male	52	23	29
21	20	Female	35	23	98
22	21	Male	35	24	35

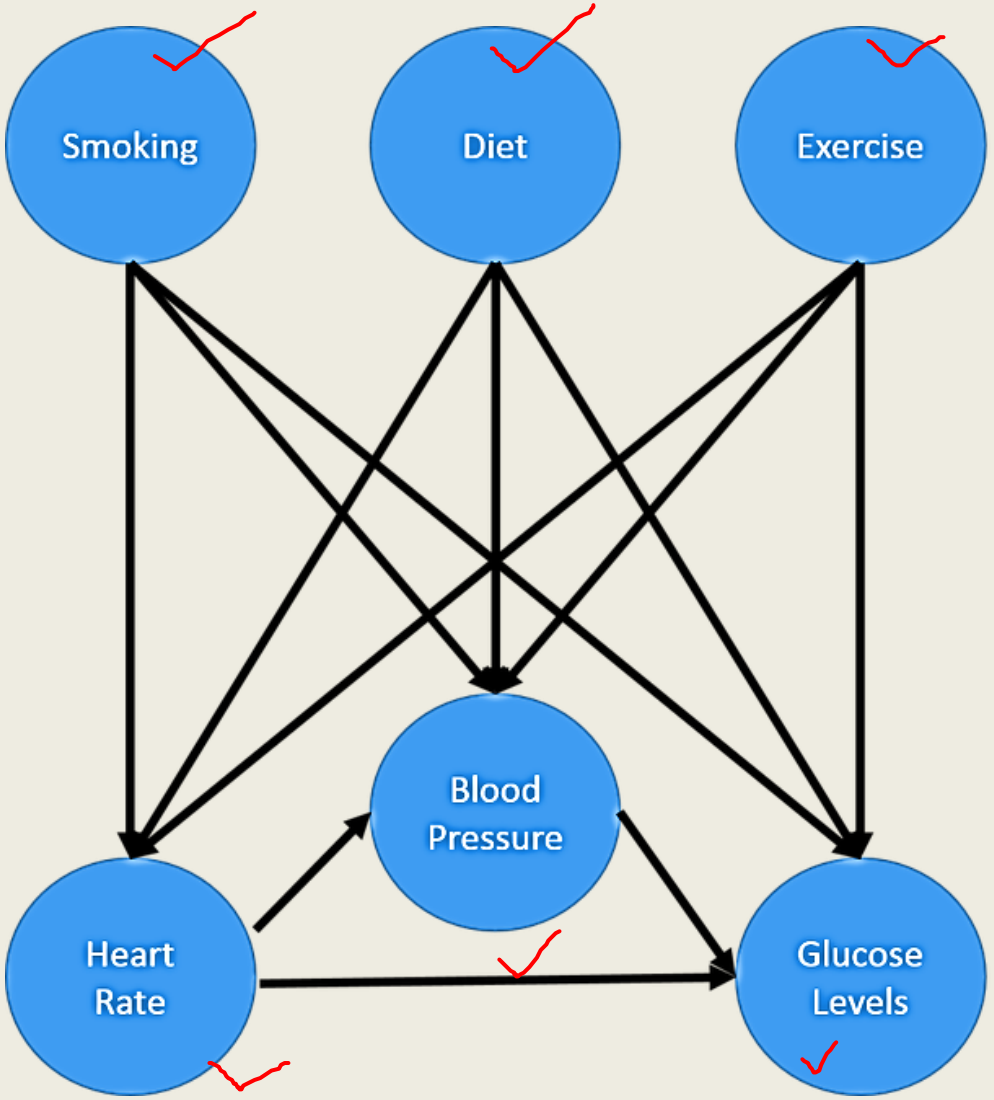
k-means

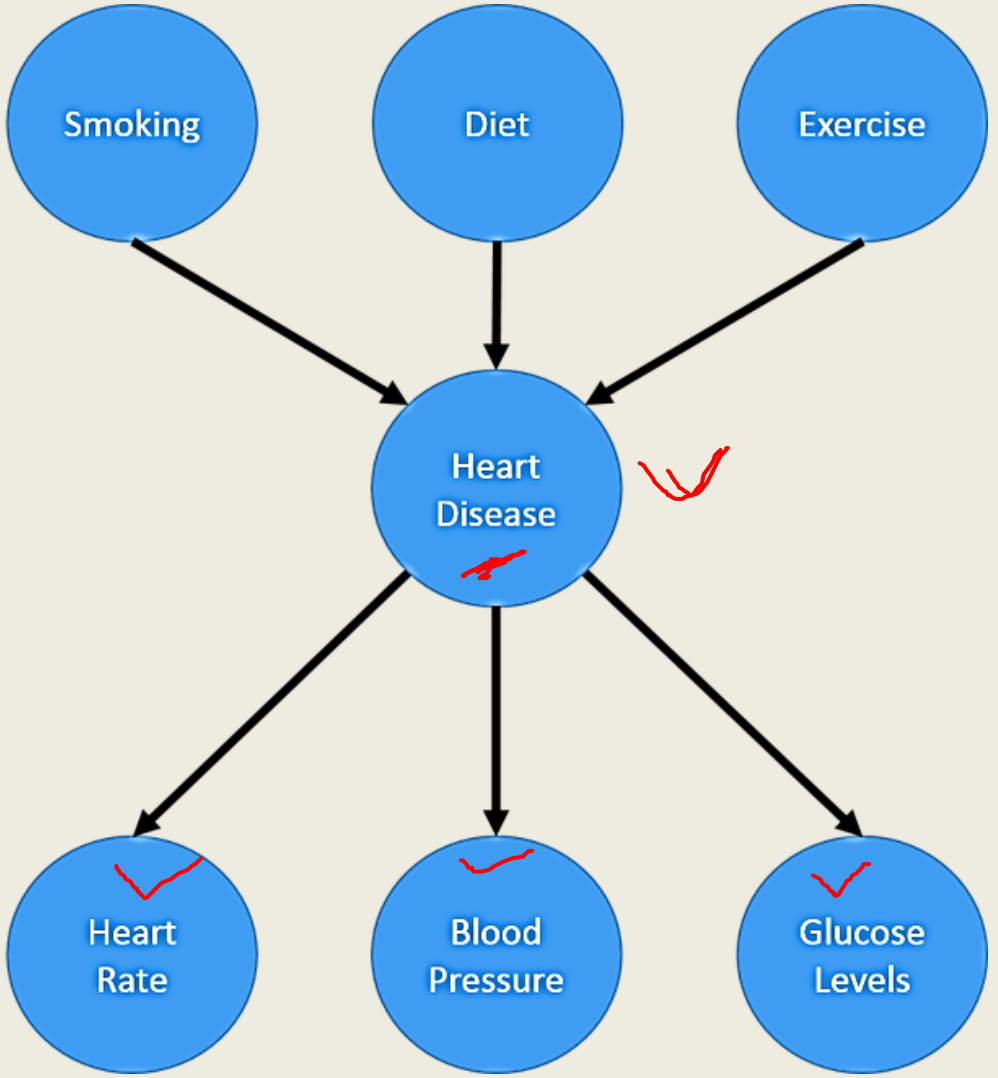
100%

- On the other hand, Expectation-Maximization algorithm can be used for the latent variables (variables that are not directly observable and are actually inferred from the values of the other observed variables) too in order to predict their values with the condition that the general form of probability distribution governing those latent variables is known to us. This algorithm is actually at the base of many unsupervised clustering algorithms in the field of machine learning.

- A latent variable model is a type of statistical model that contains two types of variables: observed variables and latent variables. Observed variables are ones that we can measure or record, while latent (sometimes called hidden) variables are ones that we cannot directly observe but rather inferred from the observed variables.

- Consider the problem of modelling medical symptoms such as blood pressure, heart rate and glucose levels (observed outcomes) and mediating factors such as smoking, diet and exercise (observed "inputs"). We could model all the possible relationships between the mediating factors and observed outcomes but the number of connections grows very quickly. Instead, we can model this problem as having mediating factors causing a non-observable hidden variable such as heart disease, which in turn causes our medical symptoms.


Smoking
Diet
Exercise
Blood Pressure
Heart Rate
Glucose Levels

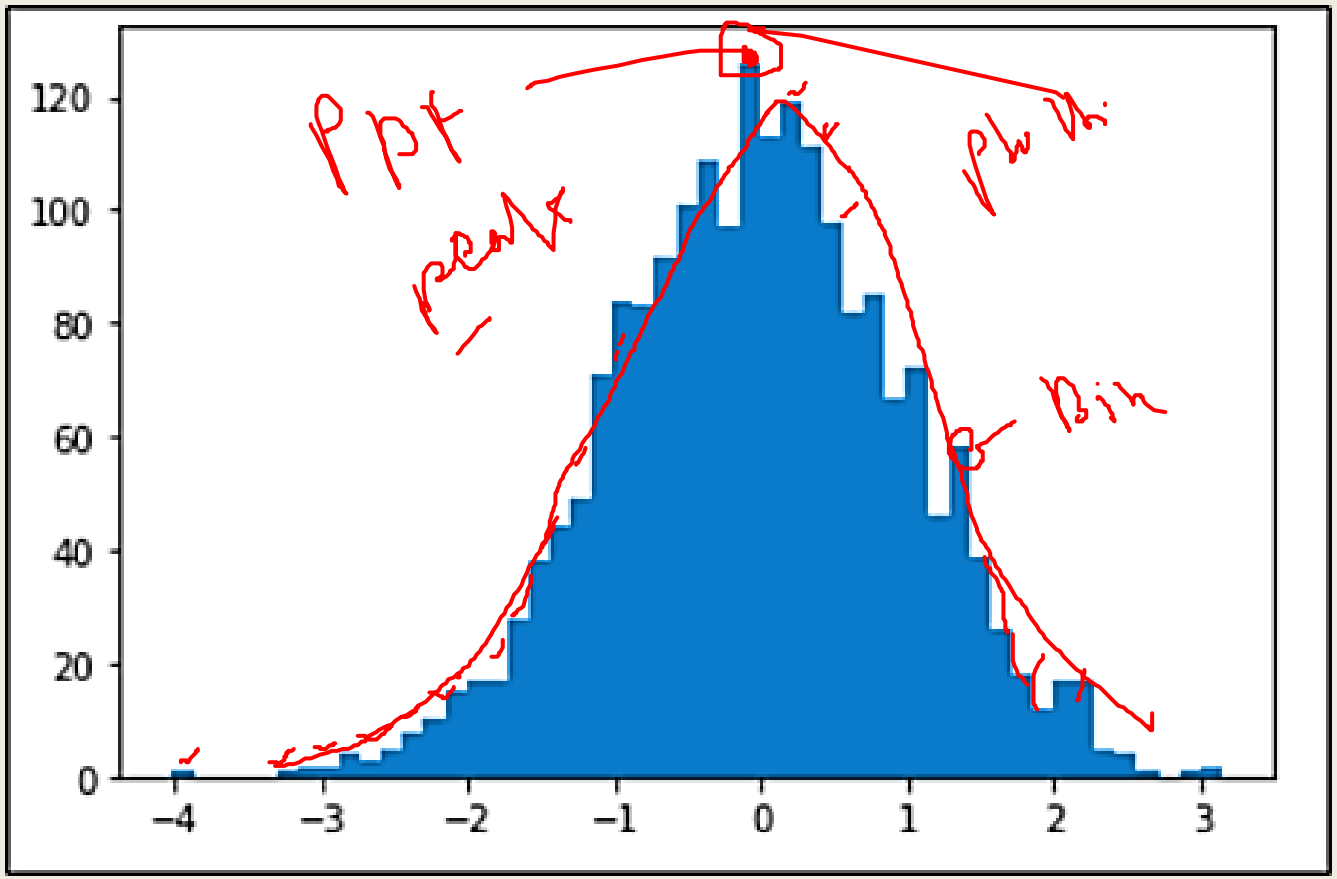

Smoking
Diet
Exercise
Heart Disease
Heart Rate
Blood Pressure
Glucose Levels

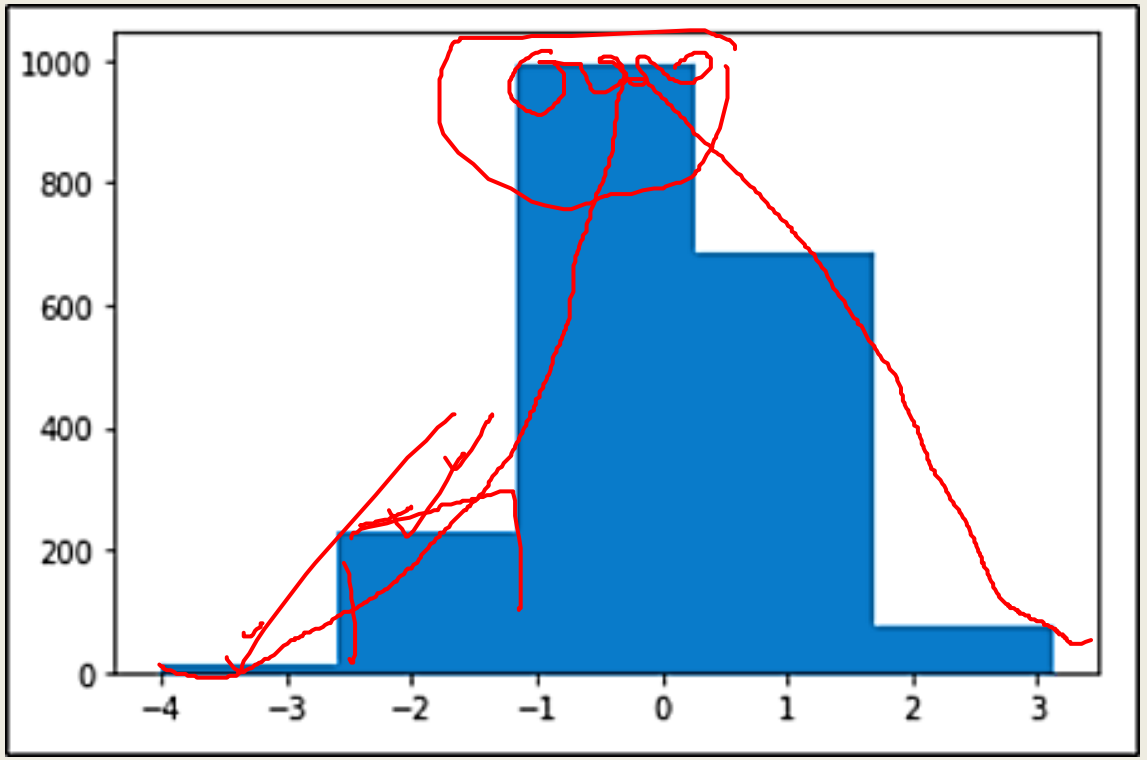
- Notice that the number of connections now grows linearly (in this case) instead of multiplicatively as you add more latent factors, this greatly reduces the number of parameters you have to estimate. In general, you can have an arbitrary number of connections between variables with as many latent variables as you wish. These models are more generally known as Probabilistic graphical models (PGMs).

Problem Of Latent Variables For Maximum Likelihood

- In statistic modeling, a common problem arises as to how can we try to estimate the joint probability distribution for a data set.

- Probability Density estimation is basically the construction of an estimate based on observed data. It involves selecting a probability distribution function and the parameters of that function that best explains the joint probability of the observed data.





The choice of number of bins plays an important role here in terms of the number of bars in the distribution and in terms of how well the density is plotted. If we change the bins to 5 in the above example, the distributions will be divided into 5 bins as shown in the image (side).

- Density estimation requires selecting a probability distribution function and the parameters of that distribution that best explain the joint probability distribution of the sample. The problem with the density estimation can be the following:

- How do you choose the probability distribution function?

- How do you choose the parameters for the probability distribution function?

- The most common technique to solve this problem is the Maximum Likelihood Estimation or simply “maximum likelihood”.

Maximum Likelihood Estimation

- In statistics, maximum likelihood estimation is the method of estimating the parameters of a probability distribution by maximizing the likelihood function in order to make the observed data most probable for the statistical model.

- But there lies a limitation with Maximum Likelihood, it assumes that the data is complete, fully observed, etc. It does not really mandate that the model will have access to all the data. Instead, it assumes that all the variables relevant to the model are already present. But in some cases, some relevant variables may remain hidden and cause inconsistencies.

- And these unobserved or hidden variables are known as Latent Variables.

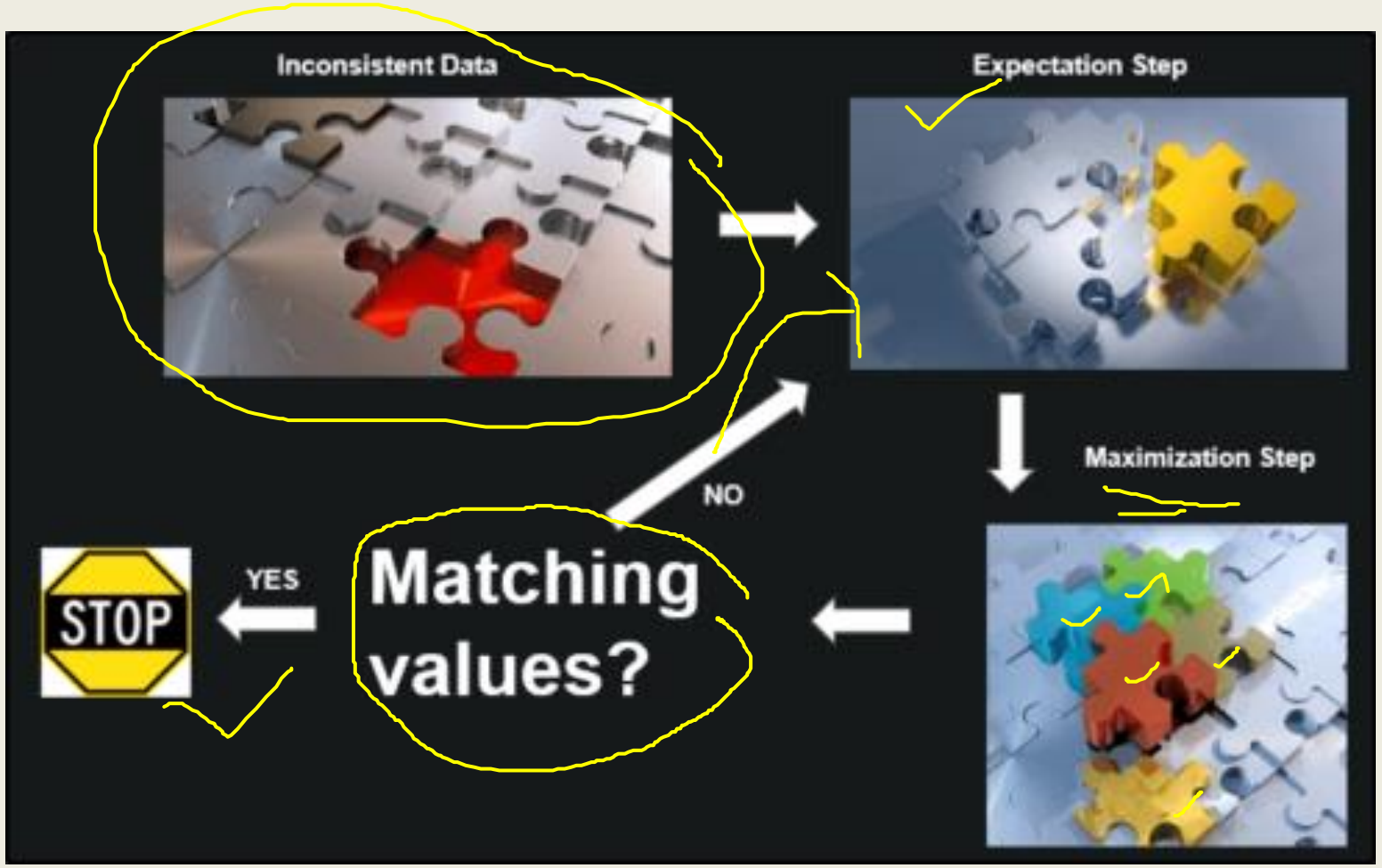
- The EM algorithm is an iterative approach that cycles between two modes. The first mode attempts to estimate the missing or latent variables, called the estimation-step or E-step. The second mode attempts to optimize the parameters of the model to best explain the data, called the maximization-step or M-step.

- E-Step. Estimate the missing variables in the dataset.
- M-Step. Maximize the parameters of the model in the presence of the data.

- The EM algorithm can be applied quite widely, although is perhaps most well known in machine learning for use in unsupervised learning problems, such as density estimation and clustering.

What is EM Algorithm In Machine Learning?

- EM algorithm was proposed in 1997 by Arthur Dempster, Nan Laird, and Donald Rubin. It is basically used to find the local maximum likelihood parameters of a statistical model in case the latent variables are present or the data is missing or incomplete.



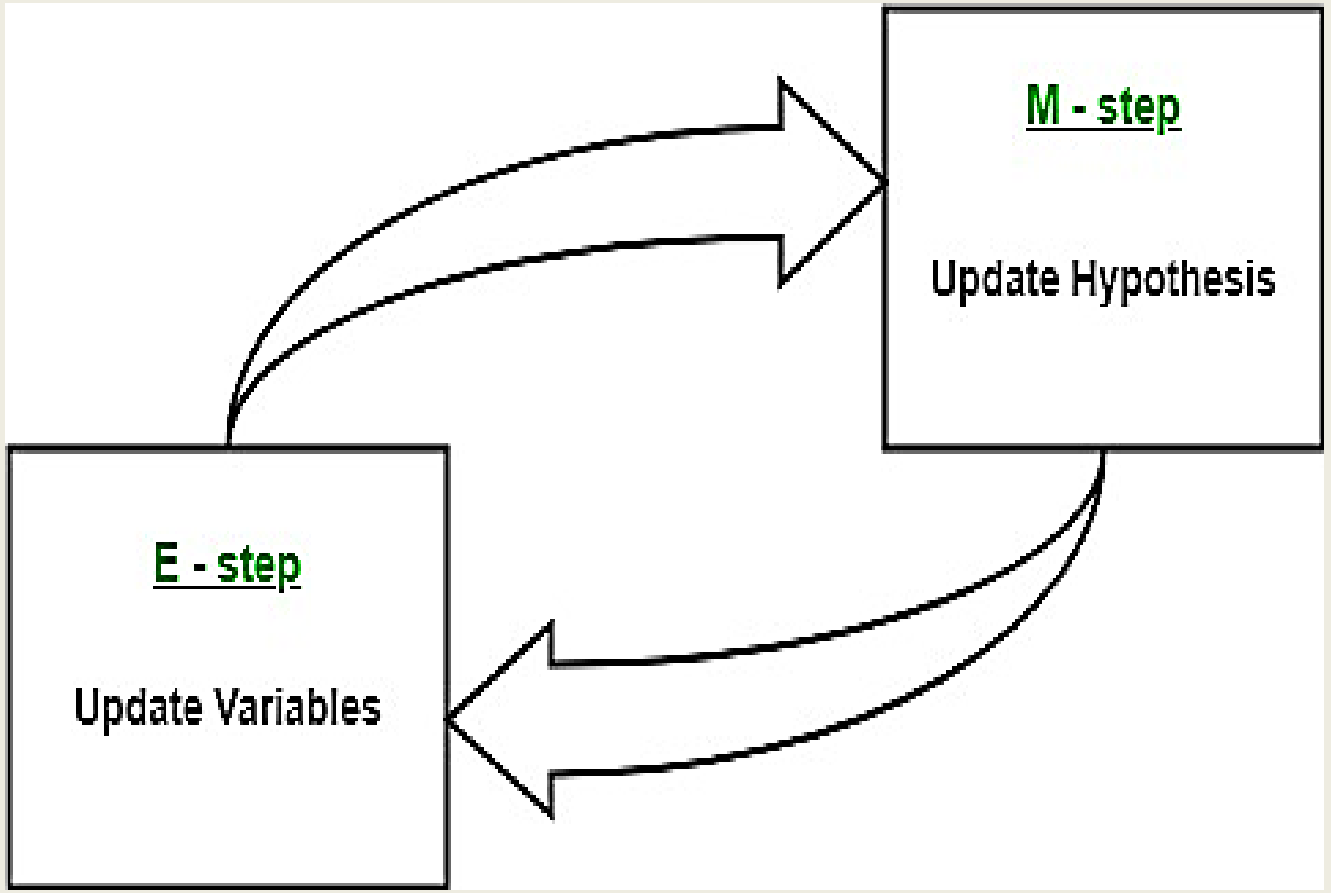
- The EM Algorithm follows the following steps in order to find the relevant model parameters in the presence of latent variables.

1. Consider a set of starting parameters in incomplete data.
2. Expectation Step – This step is used to estimate the values of the missing values in the data. It involves the observed data to basically guess the values in the missing data.
3. Maximization Step – This step generates complete data after the Expectation step updates the missing values in the data.
4. Execute the step 2 and 3 until the convergence is met.

- Convergence – The concept of convergence in probability is based on intuition. Let's say we have two random variables if the probability of their difference is very small, it is said to be converged. In this case, convergence means if the values match each other.

Algorithm

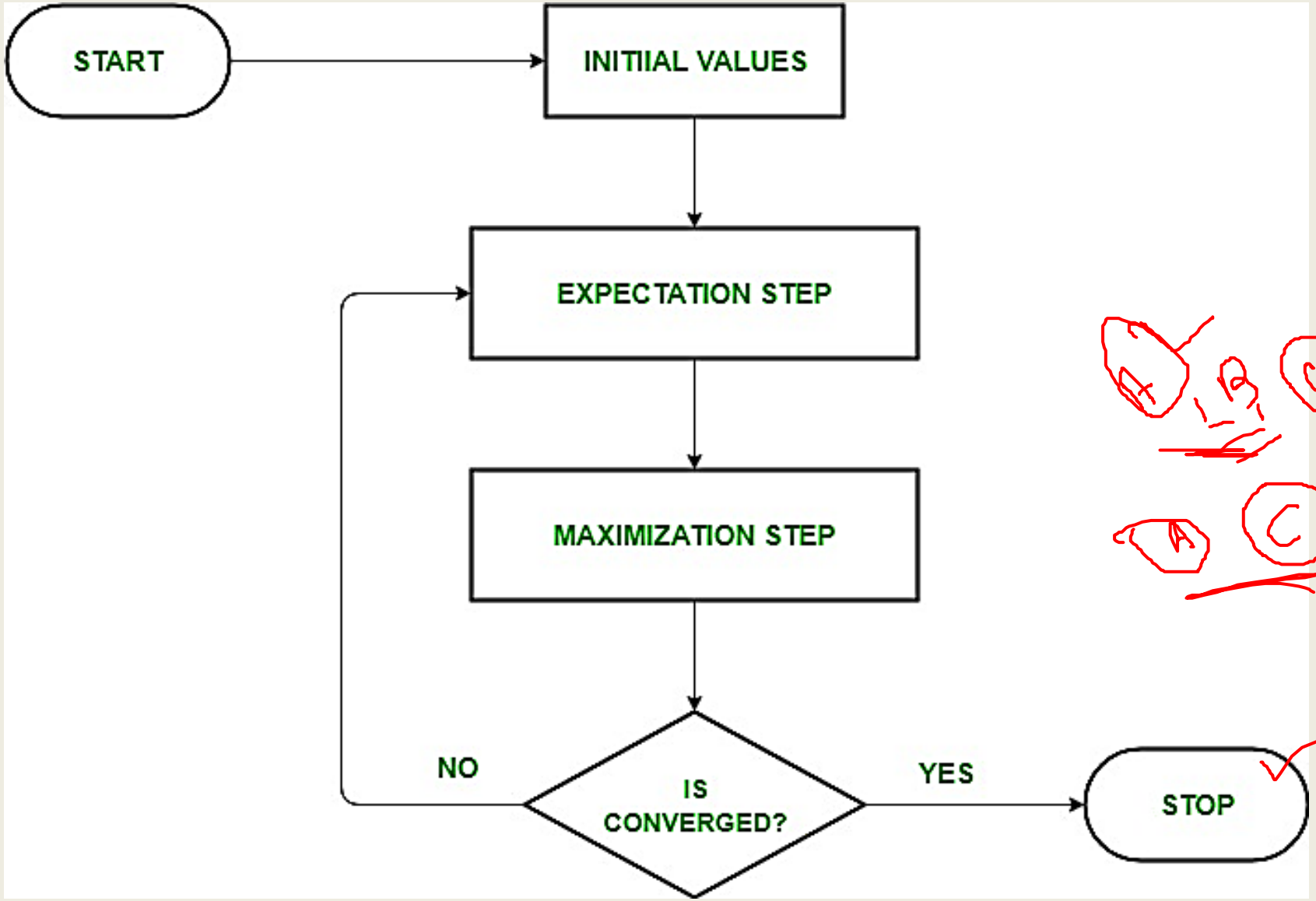
1. Given a set of incomplete data, consider a set of starting parameters.
2. Expectation step (E – step): Using the observed available data of the dataset, estimate (guess) the values of the missing data.
3. Maximization step (M – step): Complete data generated after the expectation (E) step is used in order to update the parameters.
4. Repeat step 2 and step 3 until convergence.

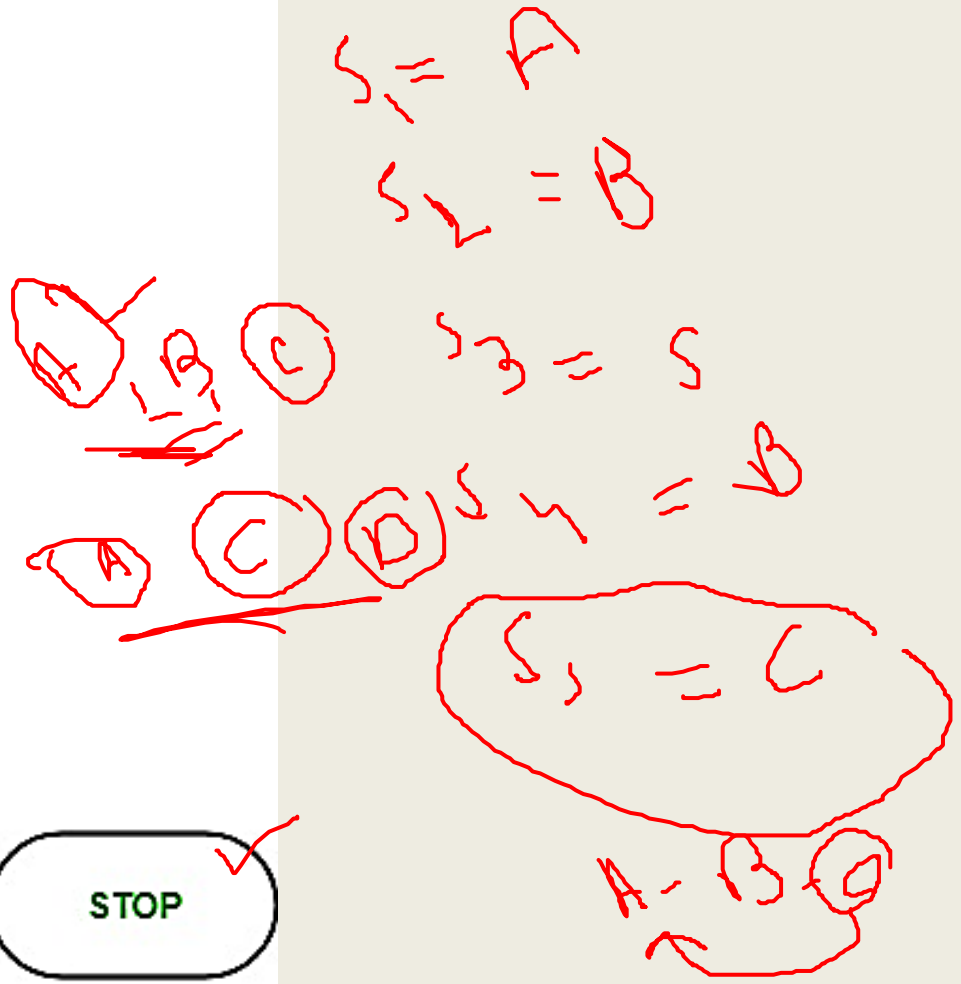


- The essence of Expectation-Maximization algorithm is to use the available observed data of the dataset to estimate the missing data and then using that data to update the values of the parameters. Let us understand the EM algorithm in detail.

1. Initially, a set of initial values of the parameters are considered. A set of incomplete observed data is given to the system with the assumption that the observed data comes from a specific model.
2. The next step is known as “Expectation” – step or E-step. In this step, we use the observed data in order to estimate or guess the values of the missing or incomplete data. It is basically used to update the variables.
3. The next step is known as “Maximization”-step or M-step. In this step, we use the complete data generated in the preceding “Expectation” – step in order to update the values of the parameters. It is basically used to update the hypothesis.
4. Now, in the fourth step, it is checked whether the values are converging or not, if yes, then stop otherwise repeat step-2 and step-3 i.e. “Expectation” – step and “Maximization” – step until the convergence occurs.

Flow chart for EM algorithm





Usage of EM algorithm

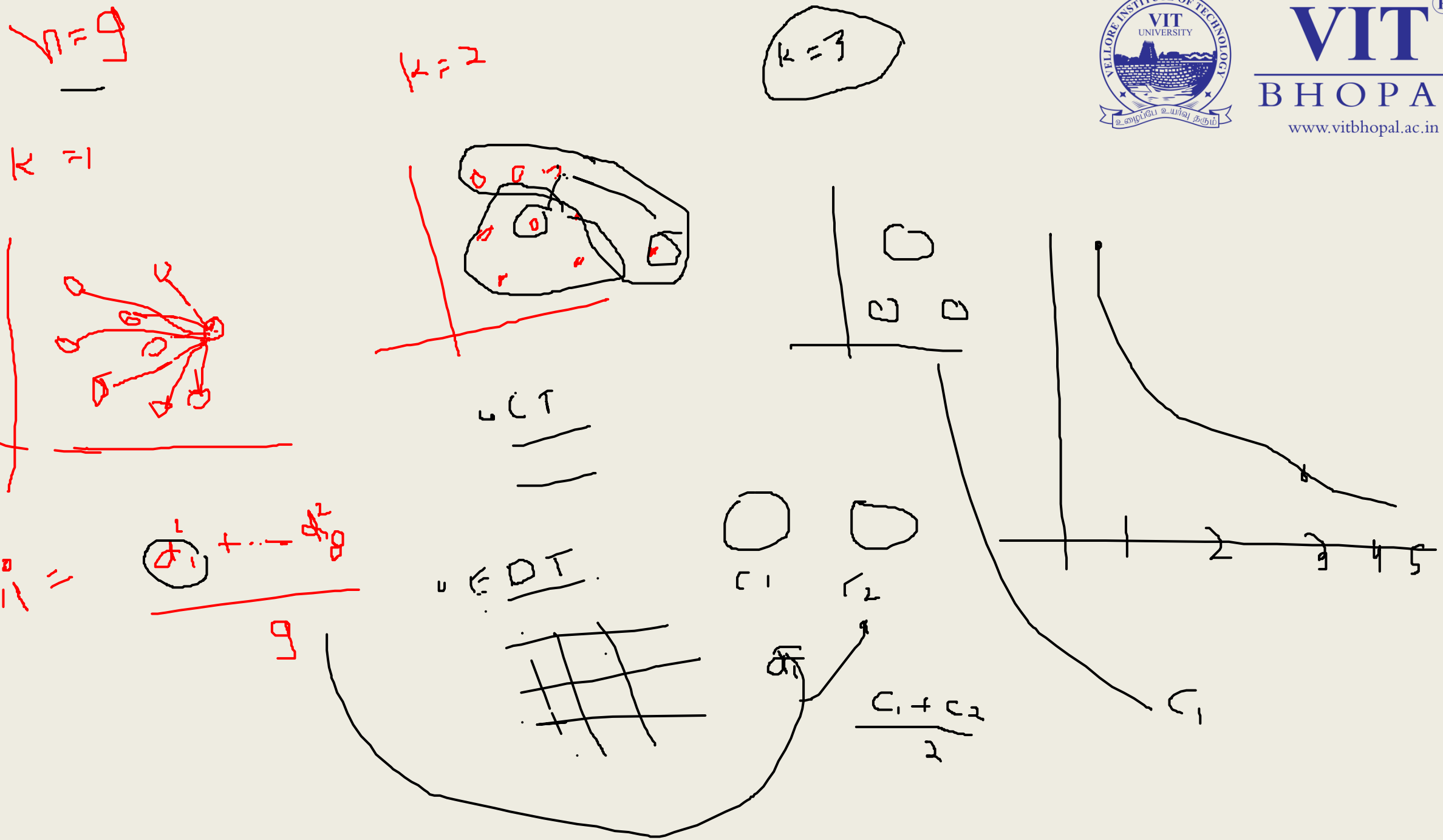
- It can be used to fill the missing data in a sample.
- It can be used as the basis of unsupervised learning of clusters.
- It can be used for the purpose of estimating the parameters of Hidden Markov Model (HMM).
- It can be used for discovering the values of latent variables.

Advantages of EM algorithm

- It is always guaranteed that likelihood will increase with each iteration.
- The E-step and M-step are often pretty easy for many problems in terms of implementation.
- Solutions to the M-steps often exist in the closed form.

Disadvantages of EM algorithm

- It has slow convergence.
- It makes convergence to the local optima only.
- It requires both the probabilities, forward and backward (numerical optimization requires only forward probability).



VIT®

BHOPAL

www.vitbhopal.ac.in