

Bayesian Learning

Bayes Theorem Application

Spam filtering

Google search

Weather forecasting

robotics

Bayes theorem

Bayes' Theorem is a way of finding a conditional probability.

Conditional probability is the probability of an event happening, given that it has some relationship to one or more other events.

For example, your probability of getting a parking space is connected to the time of day you park, where you park, and what conventions are going on at any time.

Conditional Probability

What is the probability of an event given knowledge of another event

- Example:

- $P(\text{raining} \mid \text{sunny})$
- $P(\text{raining} \mid \text{cloudy})$
- $P(\text{raining} \mid \text{cloudy, cold})$

$$p(\text{slept in movie}) = 0.5$$

$$p(\text{slept in movie} \mid \text{liked movie}) = 1/3$$

$$p(\text{didn't sleep in movie} \mid \text{liked movie}) = 2/3$$

Attribute				Classes
Gender	Car ownership	Travel Cost (\$)/km	Income level	Transportation mode
Male	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	1	Cheap	Medium	Train
Female	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Male	0	Standard	Medium	Train
Female	1	Standard	Medium	Train
Female	1	Expensive	High	Car
Male	2	Expensive	Medium	Car
Female	2	Expensive	High	Car

Bayes theorem

Bayes' Theorem is a way of finding a probability.

Also known as Bayes' rule, Bayes' law, Bayesian reasoning.

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

Which tells us:
written $P(A|B)$,

how often A happens given that B happens,

When we know:
written $P(B|A)$

how often B happens given that A happens,

and how likely A is on its own, written $P(A)$

and how likely B is on its own, written $P(B)$

Example : Picnic Day

You are planning a picnic today, but the morning is cloudy

50% of all rainy days start off cloudy!

But cloudy mornings are common (about 40% of days start cloudy)

And this is usually a dry month (only 3 of 30 days tend to be rainy, or 10%)

What is the chance of rain during the day?

$$P(\text{Rain}|\text{Cloud}) = \frac{P(\text{Rain}) P(\text{Cloud}|\text{Rain})}{P(\text{Cloud})}$$

P(Rain) is Probability of Rain = 10%

P(Cloud|Rain) is Probability of Cloud, given that Rain happens = 50%

P(Cloud) is Probability of Cloud = 40%

$$P(\text{Rain}|\text{Cloud}) = \frac{0.1 \times 0.5}{0.4} = .125$$

12.5% chance of rain. Not too bad, let's have a picnic!

Example 2 :

Find out a patient's probability of having liver disease if they are an alcoholic. "Being an alcoholic" is the **test** (kind of like a litmus test) for liver disease.

Past data tells you that 10% of patients entering your clinic have liver disease

Five percent of the clinic's patients are alcoholics.

You might also know that among those patients diagnosed with liver disease, 7% are alcoholics.

Solution:

A could mean the event “Patient has liver disease.” Past data tells you that 10% of patients entering your clinic have liver disease. $P(A) = 0.10$.

B could mean the litmus test that “Patient is an alcoholic.” Five percent of the clinic’s patients are alcoholics. $P(B) = 0.05$.

You might also know that among those patients diagnosed with liver disease, 7% are alcoholics. This is your **B|A**: the probability that a patient is alcoholic, given that they have liver disease is 7%

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

$$P(A|B) = (0.07 * 0.1)/0.05 = 0.14$$

Example : Drug Testing

Suppose that a test for using a particular drug is 99% sensitive and 99% specific.

Suppose that 0.5% of people are users of the drug. What is the probability that a randomly selected individual with a positive test is a drug user?

$$\begin{aligned}P(\text{User} \mid +) &= \frac{P(+ \mid \text{User})P(\text{User})}{P(+)} \\&= \frac{P(+ \mid \text{User})P(\text{User})}{P(+ \mid \text{User})P(\text{User}) + P(+ \mid \text{Non-user})P(\text{Non-user})} \\&= \frac{0.99 \times 0.005}{0.99 \times 0.005 + 0.01 \times 0.995} \\&\approx 33.2\%\end{aligned}$$

Advantages of probabilistic reasoning

Appropriate for complex, uncertain, environments

- Will it rain tomorrow?
- Applies naturally to many domains
- Robot predicting the direction of road, biology, Word paper clip
- Allows to generalize acquired knowledge and incorporate prior belief
- Medical diagnosis
- Easy to integrate different information sources
- Robot's sensors

Bayesian learning

Bayesian reasoning provides a probabilistic approach to inference.

Bayesian reasoning provides the basis for learning algorithms that directly manipulate probabilities, as well as a framework for analyzing the operation of other algorithms that do not explicitly manipulate probabilities.

Bayesian learning

In machine learning we are often interested in determining the best hypothesis from some space \mathbf{H} , given the observed training data \mathbf{D} .

best hypothesis is to say that we demand the **most probable** hypothesis, given the data \mathbf{D} plus any initial knowledge about the prior probabilities of the various hypotheses in H . Bayes theorem provides a direct method for calculating such probabilities. More precisely, Bayes theorem provides a way to calculate the probability of a hypothesis based on its prior probability, the probability of observing various data given the hypothesis, and the observed data itself.

Notations in Bayesian learning

Prior Probability

$P(h)$ to denote the initial probability that hypothesis **h** holds, before we have observed the training data. **$P(h)$** is called the ***prior probability*** of **h** and may reflect any background knowledge we have about the chance that **h** is a correct hypothesis.

$P(D/h)$: the probability of observing data **D** given some world in which hypothesis **h** holds.

Posterior probability

$P(h/D)$ is called the ***posterior probability*** of **h** , because it reflects our confidence that **h** holds after we have seen the training data **D** .

Bayesian learning

Bayes theorem is the cornerstone of Bayesian learning methods because it provides a way to calculate the posterior probability $P(h|D)$, from the prior probability $P(h)$, together with $P(D)$ and $P(D|h)$.

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Basic Probability formulas

Product rule: probability $P(A \wedge B)$ of a conjunction of two events A and B

$$P(A \wedge B) = P(A|B)P(B) = P(B|A)P(A)$$

Sum rule: probability of a disjunction of two events **A** and B

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

Naïve Bayes Algorithm

It is a [classification technique](#) based on Bayes' Theorem with an assumption of independence among predictors.

Naive Bayes model is easy to build and particularly useful for very large data sets.

How Naive Bayes classifier works?

Step 1: Calculate the prior probability for given class labels

Step 2: Find Likelihood probability with each attribute for each class

Step 3: Put these value in Bayes Formula and calculate posterior probability.

Step 4: See which class has a higher probability, given the input belongs to the higher probability class.

Problem: Players will play if weather is sunny. Is this statement is correct?

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	$\approx 4/14$	0.29
Rainy	3	2	$\approx 5/14$	0.36
Sunny	2	3	$\approx 5/14$	0.36
All	5	9		
	$\approx 5/14$	$\approx 9/14$		
	0.36	0.64		

Problem: Players will play if weather is sunny. Is this statement is correct?

$$P(\text{Yes} \mid \text{Sunny}) = P(\text{Sunny} \mid \text{Yes}) * P(\text{Yes}) / P(\text{Sunny})$$

Here we have $P(\text{Sunny} \mid \text{Yes}) = 3/9 = 0.33$,

$$P(\text{Sunny}) = 5/14 = 0.36$$

$$P(\text{Yes}) = 9/14 = 0.64$$

Now, $P(\text{Yes} \mid \text{Sunny}) = 0.33 * 0.64 / 0.36 = 0.60$, which has higher probability.

Problem: whether Players will play if weather is overcast?

$$P(\text{Yes} \mid \text{Overcast}) = P(\text{Overcast} \mid \text{Yes}) P(\text{Yes}) / P(\text{Overcast}) \dots\dots\dots(1)$$

Calculate Prior Probabilities:

$$P(\text{Overcast}) = 4/14 = 0.29$$

$$P(\text{Yes}) = 9/14 = 0.64$$

Calculate Posterior Probabilities:

$$P(\text{Overcast} \mid \text{Yes}) = 4/9 = 0.44$$

Put Prior and Posterior probabilities in equation (1)

$$P(\text{Yes} \mid \text{Overcast}) = 0.44 * 0.64 / 0.29 = 0.98(\text{Higher})$$

Problem: whether Players will play if weather is overcast?

$$P(\text{No} \mid \text{Overcast}) = P(\text{Overcast} \mid \text{No}) P(\text{No}) / P(\text{Overcast}) \dots\dots\dots(2)$$

Calculate Prior Probabilities:

$$P(\text{Overcast}) = 4/14 = 0.29$$

$$P(\text{No}) = 5/14 = 0.36$$

Calculate Posterior Probabilities:

$$P(\text{Overcast} \mid \text{No}) = 0/9 = 0$$

Put Prior and Posterior probabilities in equation (2)

$$P(\text{No} \mid \text{Overcast}) = 0 * 0.36 / 0.29 = 0$$

The probability of a 'Yes' class is higher. So you can determine here if the weather is overcast then players will play the sport.

Example of Naïve Bayes classifier

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

A: attributes

M: mammals

N: non-mammals

$$P(A|M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A|N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A|M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A|N)P(N) = 0.0042 \times \frac{13}{20} = 0.0027$$

$$P(A|M)P(M) >$$

$$P(A|N)P(N)$$

=> Mammals

Advantages of Naïve Bayes Classification

It is easy and fast to predict class of test data set. It also perform well in multi class prediction

When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you need less training data.

It perform well in case of categorical input variables compared to numerical variable(s). For numerical variable, normal distribution is assumed (bell curve, which is a strong assumption).

Limitations of naïve-Bayes classification

If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as “Zero Frequency”. To solve this, we can use the smoothing technique. One of the simplest smoothing techniques is called Laplace estimation.

On the other side naive Bayes is also known as a bad estimator, so the probability outputs from `predict_proba` are not to be taken too seriously.

Another limitation of Naive Bayes is the assumption of independent predictors. In real life, it is almost impossible that we get a set of predictors which are completely independent.

Applications of Naïve-Bayes theorem

Real time Prediction: Naive Bayes is an eager learning classifier and it is sure fast. Thus, it could be used for making predictions in real time.

Multi class Prediction: This algorithm is also well known for multi class prediction feature. Here we can predict the probability of multiple classes of target variable.

Text classification/ Spam Filtering/ Sentiment Analysis: Naive Bayes classifiers mostly used in text classification (due to better result in multi class problems and independence rule) have higher success rate as compared to other algorithms. As a result, it is widely used in Spam filtering (identify spam e-mail) and Sentiment Analysis (in social media analysis, to identify positive and negative customer sentiments)

Recommendation System: Naive Bayes Classifier and Collaborative Filtering together builds a Recommendation System that uses machine learning and data mining techniques to filter unseen information and predict whether a user would like a given resource or not.

Bayes Optimal classification

"what is the most probable ***hypothesis*** given the training data?"

"what is the most probable ***classification*** of the new instance given the training data? "

We can get answer by applying the MAP hypothesis.

MAP (Maximum a *posteriori*)

the learner considers some set of candidate hypotheses H and is interested in finding the most probable hypothesis $\mathbf{h} \in H$ given the observed data \mathbf{D} (or at least one of the maximally probable if there are several). Any such maximally probable hypothesis is called a ***maximum a posteriori*** (MAP) hypothesis

Choosing Hypotheses

Generally want the most probable hypothesis given the training data

Maximum a posteriori hypothesis h_{MAP} :

$$h_{\text{MAP}} = \arg \max_{h \in H} P(h|D)$$

$$\begin{aligned}h_{MAP} &= \arg \max_{h \in H} P(h|D) \\&= \arg \max_{h \in H} \frac{P(D|h)P(h)}{P(D)} \\&= \arg \max_{h \in H} P(D|h)P(h)\end{aligned}$$

If assume $P(h_i)=P(h_j)$ for all h_i and h_j , then can further simplify, and choose the

Maximum likelihood (ML) hypothesis

$$h_{ML} = \arg \max_{h_i \in H} P(D|h_i)$$

Example

Does patient have cancer or not?

A patient takes a lab test and the result comes back positive. The test returns a correct positive result (+) in only 98% of the cases in which the disease is actually present, and a correct negative result (-) in only 97% of the cases in which the disease is not present

Furthermore, 0.008 of the entire population have this cancer

Suppose a positive result (+) is returned...

$$P(cancer) = 0.008$$

$$P(\neg cancer) = 0.992$$

$$P(+|cancer) = 0.98$$

$$P(-|cancer) = 0.02$$

$$P(+|\neg cancer) = 0.03$$

$$P(-|\neg cancer) = 0.97$$

$$P(+|cancer) \cdot P(cancer) = 0.98 \cdot 0.008 = 0.0078$$

$$P(+|\neg cancer) \cdot P(\neg cancer) = 0.03 \cdot 0.992 = 0.0298$$

$$h_{MAP} = \neg cancer$$

Example

consider a hypothesis space containing three hypotheses, h_1 , h_2 , and h_3 .

Suppose that the posterior probabilities of these hypotheses given the training data are .4, **.3**, and **.3** respectively

MAP Hypothesis? $\rightarrow h_1$

Gibbs Algorithm

Although the Bayes classifier obtains the best performance that can be achieved from the given training data, it can be quite costly to apply.

The expense is due to the fact that it computes the posterior probability for every hypothesis in H and then combines the predictions of each hypothesis to classify each new instance.

An alternative, less optimal method is the Gibbs algorithm.

Algorithm

1. Choose a hypothesis **h** from H at random, according to the posterior probability distribution over H .
2. Use **h** to predict the classification of the next instance **x** .

Bayesian Belief Network

A Bayesian network is a probabilistic graphical model which represents a set of variables and their conditional dependencies using a directed acyclic graph.

It is also called a **Bayes network**, **belief network**, **decision network**, or **Bayesian model**.

A Bayesian Network can be used for building models from data and experts' opinions, and it consists of two parts:

Directed Acyclic Graph

Table of conditional probabilities.

A Bayesian network graph is made up of nodes and Arcs (directed links), where:

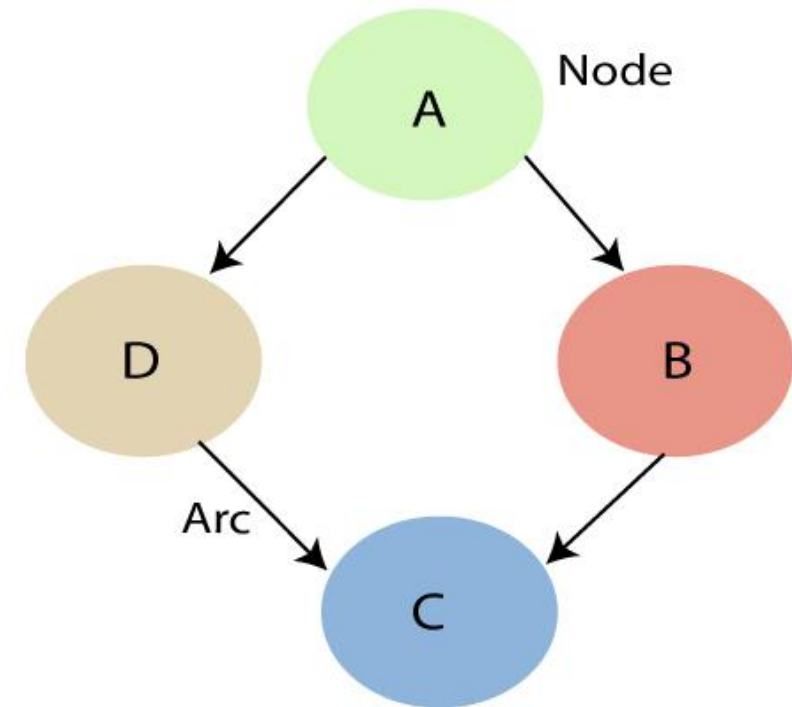
Each **node** corresponds to the random variables, and a variable can be **continuous** or **discrete**.

Arc or directed arrows represent the causal relationship or conditional probabilities between random variables. These directed links or arrows connect the pair of nodes in the graph.

These links represent that one node directly influence the other node, and if there is no directed link that means that nodes are independent with each other

- **In the above diagram, A, B, C, and D are random variables represented by the nodes of the network graph.**
- **If we are considering node B, which is connected with node A by a directed arrow, then node A is called the parent of Node B.**

- **Node C is independent of node A**



Text Classification and Naïve Bayes

THE TASK OF TEXT CLASSIFICATION

Is this spam?

From: googleteam **To:**
Subject: GOOGLE LOTTERY WINNER! CONTACT YOUR AGENT TO CLAIM YOUR PRIZE.

GOOGLE LOTTERY INTERNATIONAL
INTERNATIONAL PROMOTION / PRIZE AWARD .
(WE ENCOURAGE GLOBALIZATION)
FROM: THE LOTTERY COORDINATOR,
GOOGLE B.V. 44 9459 PE.
RESULTS FOR CATEGORY "A" DRAWS

Congratulations to you as we bring to your notice, the results of the First Ca
inform you that your email address have emerged a winner of One Million (1,
money of Two Million (2,000,000.00) Euro shared among the 2 winners in this
email addresses of individuals and companies from Africa, America, Asia, Au
CONGRATULATIONS!

Your fund is now deposited with the paying Bank. In your best interest to avo
award strictly from public notice until the process of transferring your claims
NOTE: to file for your claim, please contact the claim department below on e

Male or female author?

1. By 1925 present-day Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony of Cochinchina; the central area with its imperial capital at Hue was the protectorate of Annam...

2. Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of her greatest assets

Positive or negative movie review?

- unbelievably disappointing
- Full of zany characters and richly applied satire, and some great plot twists.
- this is the greatest screwball comedy ever filmed
- It was pathetic. The worst part about it was the boxing scenes.

What is the subject of this article?

Antagonists and Inhibitors

Blood Supply

Chemistry

Drug Therapy

Embryology

Epidemiology



Text classification

- Assigning subject categories, topics, or genres
- Spam detection
- Authorship identification
- Age/gender identification
- Language Identification
- Sentiment analysis

Text Classification: definition

Input:

- a document d
- a fixed set of classes $C = \{c_1, c_2, \dots, c_J\}$

Output: a predicted class $c \in C$

Classification Methods: Hand-coded rules

- Rules based on combinations of words or other features
 - spam: black-list-address OR (“dollars” AND “have been selected”)
- Accuracy can be high
 - If rules carefully refined by expert
- But building and maintaining these rules is expensive

Text Classification and Naïve Bayes

Simple (“naïve”) classification method based on Bayes rule

- Relies on very simple representation of document
- Bag of words

The Bag of Words Representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!





it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1

The Bag of Words Representation

Y
(

seen	2
sweet	1
whimsical	1
recommend	1
happy	1
...	...

) = C



Bayes' Rule Applied to Documents and Classes

- For a document d and a class c

$$P(c | d) = \frac{P(d | c)P(c)}{P(d)}$$

EM Algorithm (Expectation-Maximization Algorithm)

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.

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.

Algorithm

Given a set of incomplete data, consider a set of starting parameters.

Expectation step (E - step): Using the observed available data of the dataset, estimate (guess) the values of the missing data.

Maximization step (M - step): Complete data generated after the expectation (E) step is used in order to update the parameters.

Repeat step 2 and step 3 until convergence

Detailed EM Algorithm

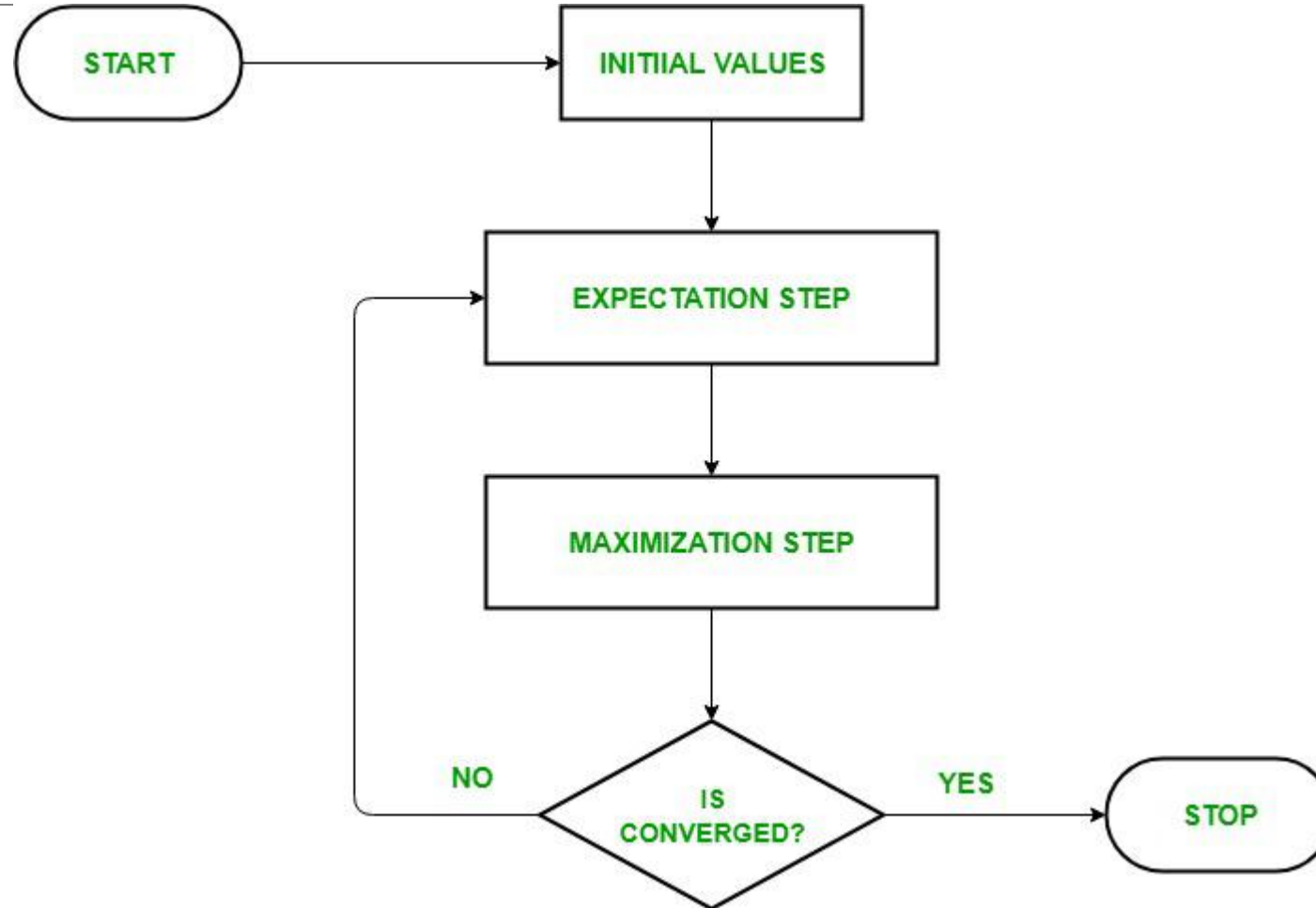
Initially, a set of initial values of the parameters are considered. A set of incomplete observed data is given to the system with the assumption that the observed data comes from a specific model.

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.

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.

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.

Flowchart 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).

References

1. https://web.stanford.edu/~jurafsky/slp3/slides/7_NB.pdf
2. <https://www.geeksforgeeks.org/ml-expectation-maximization-algorithm/>
3. <https://www.javatpoint.com/bayesian-belief-network-in-artificial-intelligence>