EM-ALGORITHM & APPLICATIONS

Presented By : Hemant Nishad

Indian Institute of Space Science and Technology | 2025

INTRODUCTION

The Expectation-Maximization (EM) algorithm is a widely used iterative method for finding maximum likelihood estimates of parameters in probabilistic models utilizing the observed data, especially when data has missing or latent variables.

Mathematical Framework

1. Observed Data (X)

•Represented as $X=\{x^1,x^2,\ldots,x^n\}$, where each x^i is a measurable quantity e.g., features in a dataset.

2. Latent Variables (Z)

•Unobserved variables $Z=\{z^1,z^2,\ldots,z^n\}$ that influence the distribution of X.

3. Parameters (θ)

•Unknown quantities θ to be estimated.

•The goal is to find θ that maximises the likelihood of X

Mathematical Framework

Likelihood Functions

- **Complete-Data Likelihood**

The joint probability of X and Z given θ

$$p(X, Z|\theta) = \prod_{i=1}^{n} p(x_i, z_i|\theta)$$

This incorporates both observed and latent variables but is intractable directly because Z is unobserved.

- **Observed-Data Likelihood**

The marginal likelihood of X, obtained by integrating out Z:

$$p(X|\theta) = \int p(X, Z|\theta)\, dZ$$

Maximizing log(p(X|θ)) directly is often difficult due to the integral.

HOW DOES EM ALGORITHM WORKS ?

Expectation Step (E – Step)

- Estimate the missing/ hidden variables using current parameter values.
- Computes the expected likelihood of the data.

Maximization Step (M-Step)

- Optimize parameters to maximize the expected likelihood found in the E-step.

EM Algorithm has two main Steps, which are repeated iteratively until convergence (which is guaranteed!).

THE CONNECTION

The EM algorithm bridges the complete and observed data likelihoods via:

$$\log p(X|\theta) \geq E_{Z|X,\theta'}[log\, p(X,Z|\theta)] - E_{Z|X,\theta'}[\log p(Z|X,\theta')]$$

Here, $E_{Z|X,\theta'}[log\, p(X,Z|\theta)]\ \big(= Q(\theta|\theta')\big)$ **say, represents expected complete –data log likelihood computed in E - step.**

Also, let $\mathrm{g}(\theta|\theta')$ **denote the right hand side of the inequality.**

THE CONNECTION

Now, $\log p(X|\theta)$ is always upper bounded and the inequality in previous slide lower bounds $\log p(X|\theta)$ by $g(\theta|\theta')$.

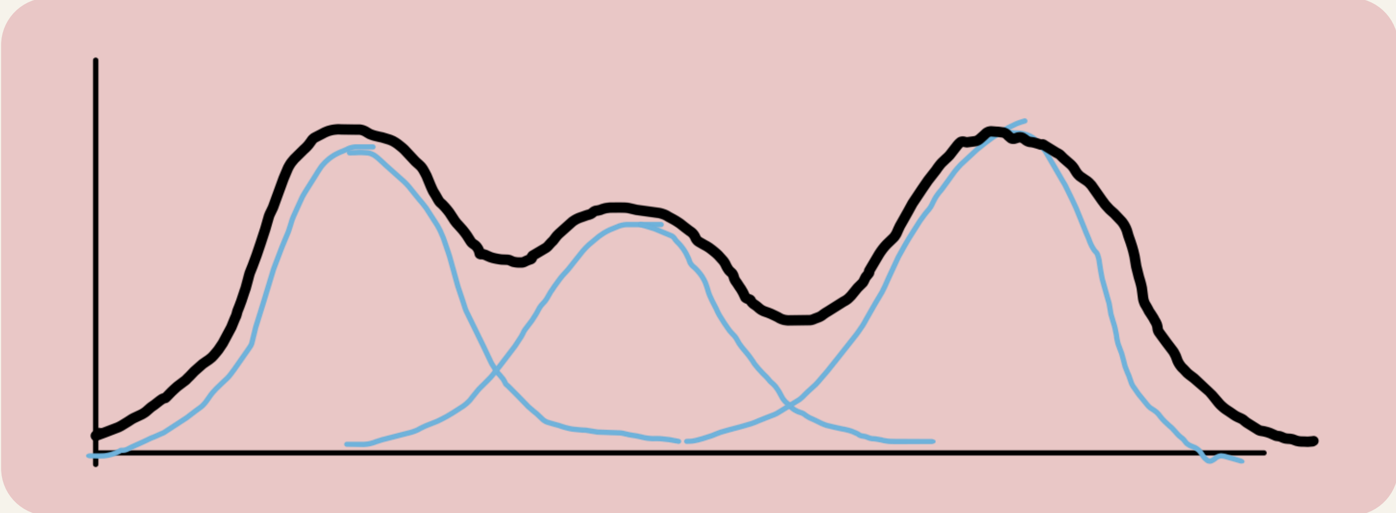
The M-step finds a new value of θ by maximizing $Q(\theta|\theta')$ over θ which is equivalent to maximizing the $g(\theta|\theta')$ over θ in previous equation.

This process in iteration ensures convergence to atleast a local maximum.

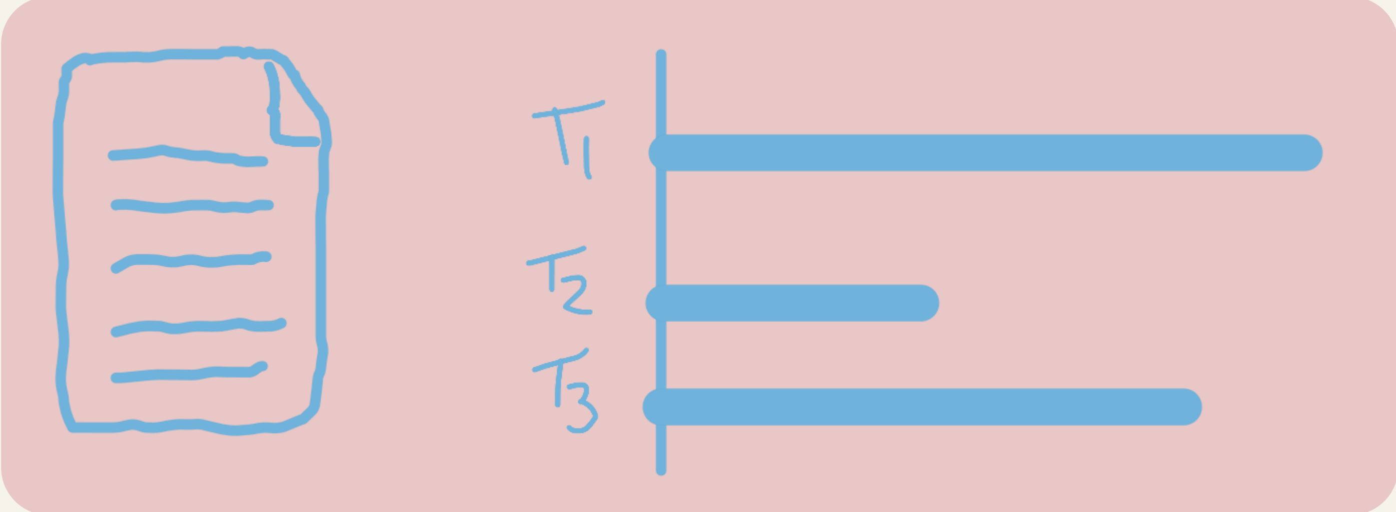
APPLICATIONS

The EM algorithm is a versatile tool that can be used in various problems. The feat of handling incomplete or missing data makes it a very sought-after algorithm in various ML applications.
Let us discuss some of them.

Gaussian Mixture Models



Latent Dirichlet Allocation



GAUSSIAN MIXTURE MODELS

The normal mixture model assumes that the observed data X=(X^{I},...,X^{n}) are independent and identically distributed (i.i.d.) random variables. The probability density function (PDF) is given by:

$$f_x(x) = \sum_{j=1}^{m} p_j \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left[-\frac{1}{2}\left(\frac{x_j - \mu_j}{\sigma_j}\right)^2\right]$$

Here:

- **m: Number of components in the mixture.**
- **pj: Mixing proportions ($p_j \geq 0$ and $\sum_{j=1}^{m} p_j = 1$).**
- μ_j,σ_j^2**: Mean and variance of the j-th normal distribution.**

The goal is to estimate the parameters $\theta = (p_1, \dots, p_m, \mu_1 \dots, \mu_m, \sigma_1, \dots, \sigma_m)$

GAUSSIAN MIXTURE MODELS

Instead of directly maximizing the likelihood $L(\theta; X) = p(X|\theta)$, lets introduce a latent variable Y_i, which indicates the component from which each observation X_i originates.

- **Then,** $p(Y = j) = p_j$ **for** $j = 1, \dots, m$.
- **And** $f_{x|y}(x_i|y_i = j; \theta) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left[-\frac{1}{2}\left(\frac{x_j - \mu_j}{\sigma_j}\right)^2\right]$
- **Subsequently, the likelihood function for complete data is given as**

$$L(\theta; X, Y) = P(\theta|X, Y) = \prod_{i=1}^{n} p(y = y_i) \frac{1}{\sqrt{2\pi\sigma_{y_i}^2}} \exp\left[-\frac{1}{2}\left(\frac{x_{y_i} - \mu_{y_i}}{\sigma_{y_i}}\right)^2\right]$$

GAUSSIAN MIXTURE MODELS

Now the EM-Algorithm starts with an initial guess θ_{old} and iterate between E-step and M-step as below.

E-Step : Calculate $Q(\theta|\theta_{old})$as

$$Q(\theta|\theta_{old}) = E_{Y|X,\theta_{old}}\left[log\, P(X,Y|\theta) \mid X, \theta_{old}\right] = \sum_{i=1}^{n}\sum_{j=1}^{m} P(Y_i = j|x_i,\theta)\, log\, P(X,Y|\theta)$$

M-Step : Updates μ_j, σ_j and p_j as

$$\mu_j = \frac{\Sigma_{i=1}^{n} x_i P(Y_i = j|x_i,\theta_{old})}{\Sigma_{i=1}^{n} P(Y_i = j|x_i,\theta_{old})};\ \sigma_j = \frac{\Sigma_{i=1}^{n}\left(x_i - \mu_j\right)^2 P(Y_i = j|x_i,\theta_{old})}{\Sigma_{i=1}^{n} P(Y_i = j|x_i,\theta_{old})}$$

$$p_j = \frac{1}{n}\Sigma_{i=1}^{n} P(Y_i = j|x_i,\theta_{old})$$

for each $j = 1, 2, \dots, m$

LATENT DIRICHLET ALLOCATION

In NLP, a topic model is a statistical model for discovering "topics" that occur in a collection of documents. The Latent Dirichlet Allocation (LDA) is the most common topic model currently being used.

Latent Dirichlet Allocation is a generative probabilistic model in which each item of a collection is modelled as a finite mixture over an underlying set of concepts. Also, each of these concepts is modelled as distributions over words.

LATENT DIRICHLET ALLOCATION

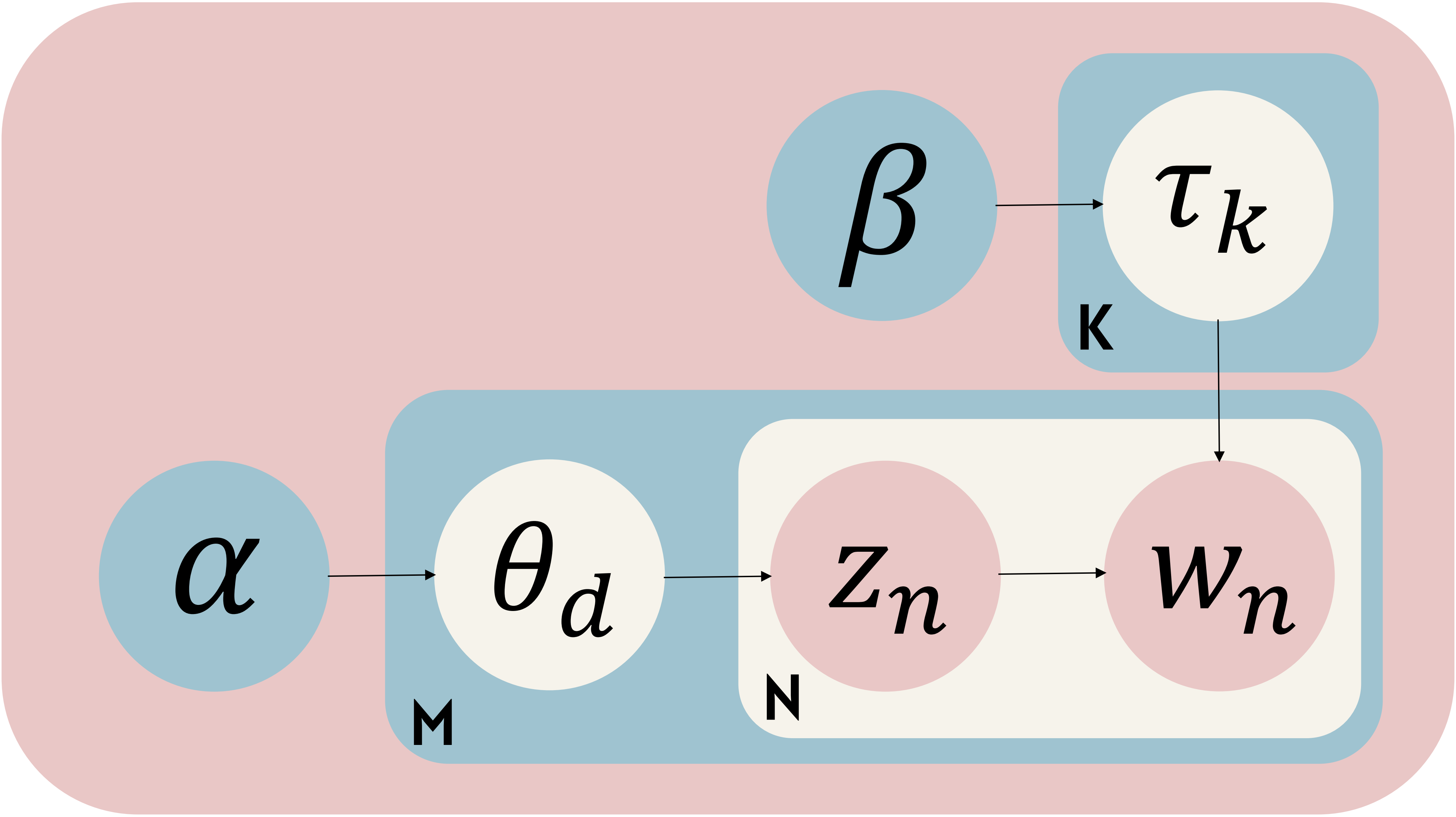
Topic Models, like LDA, work on the following assumptions :

1) Every document is a mix of topics.
2) Every topic is a mix of words.
3) Bag of Words: The order of words in a document, as well as the order of documents, has no importance.

LATENT DIRICHLET ALLOCATION

- A **word** is the basic unit of discrete text data. Each word is an item of a vocabulary indexed as $\{1,2,3, \dots, V\}$.
- A **document** is a sequence of N words denoted as $\overline{w} = (w_1, w_2, \dots, w_N)$
- A **corpus** is a collection of M documents denoted as $D = \{\overline{w}_1, \overline{w}_2, \dots, \overline{w}_M\}$

LATENT DIRICHLET ALLOCATION



- 'θ_d' denotes document topic distribution. It is a vector of dimension (1xk)
- 'τ_k' denotes topic word distribution matrix. It is a matrix of dimension (kxN).
- α and β denote Dirichlet priors $\mathrm{Dir}(\alpha)$ and $\mathrm{Dir}(\beta)$ from which θ_d and τ_k are sampled respectively.

LATENT DIRICHLET ALLOCATION

Total Probability of LDA model :

$$p(\overline{w}, \bar{z}, \bar{\theta}, \tau_k | \alpha, \beta\ )$$

$$= \prod_{i=1}^{k} p(\tau_i|\beta) \prod_{d=1}^{M} p(\theta_d|\alpha) \prod_{n=1}^{N} p\left(z_{d,n}|\theta_d\right) p\left(w_{d,n}|\tau_{z_{d,n}}\right)$$

Total Probability for observed corpus :

$$p(\overline{w}|\alpha, \beta\ ) = \int_{\tau_{z_{d,n}}} \int_{\theta_d} \sum_{z} p(\overline{w}, \bar{z}, \bar{\theta}, \tau_k|\alpha, \beta\ )\, d\, \theta_d d\tau_{z_{d,n}}$$

LATENT DIRICHLET ALLOCATION

Hence $Q(\theta|\theta')$ in current context becomes,

$$Q(\alpha,\beta|\alpha',\beta') = E_{\bar{z},\bar{\theta},\tau_k|\bar{w},\alpha',\beta'}\left[\log(p(\bar{w},\bar{z},\bar{\theta},\tau_k|\alpha,\beta))|\bar{w},\alpha',\beta'\right]$$

Now, similarly , the EM-Algorithm starts with an initial guess α', β' and iterate between E-step and M-step until convergence.

References

1 Haugh, M. (2015). Machine Learning for OR&FE [Lecture Notes] The EM Algorithm, Columbia University.

2 Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation." Journal of machine Learning research 3.Jan (2003): 993-1022.

IIST | 2025

THANK YOU

Presented By : Hemant Nishad