Gaussian Mixture Model with EM algorithm

GMM-EM

Kim Ying WONG

McMaster University
School of Computational Science and Engineering

January 30, 2024

Overview

1. Goal Statement
2. User Characteristics
3. Input and Output
4. Theorems and Figures
5. References

Goal Statement

Goal

Library on Gaussian Mixture Model. Given that dataset $\in R^{M \times N}$, we generate a predicted label for every data points corresponding to the cluster it belongs ($z \in R^{M}$ for $z_i \in 0, 1, 2, ..., K$)

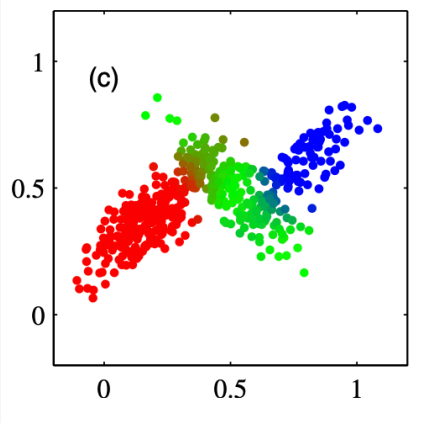

Figure: Figure caption

User Characteristics

Characteristics

1. Work in the fields related to Machine Learning or Data Science (employees, researchers, students etc.)
2. Basic knowledge in programming language and machine learning are needed
3. Basic skills in data pre-processing is needed

Input and Output

McMaster University

Input and Output

We illustrate our expected input and output here.

Input

Dataset $\in \mathrm{R}^{M \times N}$
where M = number of data-points,
N = number of predictor variables
and each data point x $\in R$,
Number of clusters in our dataset (k) (optional)

Output

$z \in \mathrm{R}^{M}$, *for*
$z_i \in 0, 1, 2, ..., K$, *where*
K is number of clusters needed

Table for mathematical notations

Notations	Meaning
M	number of data-points
N	number of predictor variables
K	number of clusters needed
x	a single data point
z	predicted label (latent variable in GMM)
p(x)	Gaussian mixture distribution
π	mixing coefficient
$N(x\|\mu_k, \Sigma_k)$	Gaussian distribution with mean μ_k and variance Σ_k

Table: Table 1

Theorems and Figures

McMaster University

Theory

We assume the data-points come from a mixture of Gaussian distributions. The training process maximizes the log-likelihood function, the parameters in the models (π_k, μ_k, Σ_k)will converge and give us the best model. This will be achieved by Expectation Maximization Algorithm (EM algorithm)

Definition (Gaussian Mixture)

$p(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k)$

Definition (log-Likelihood function)

$\ln p(X|\pi, \mu, \Sigma) = \sum_{n=1}^{N} ln \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k)$

EM algorith demostration

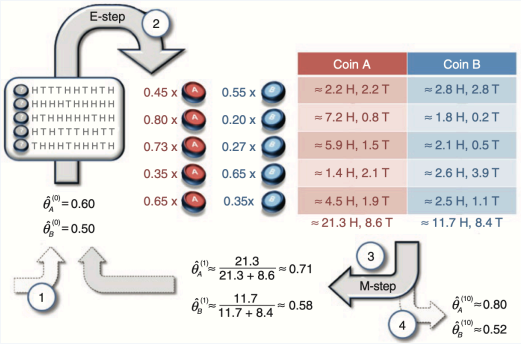

Figure: EM algorithm [Do,Batzoglou, 2008]

Constraint and Assumption

- Model assume that the data point follows Gaussian distribution.
- The problem assume and restrict the dataset to be well-processed (without missing value and infinity value).
- Convergence to optimal is not guaranteed based on the model nature.

References

References

What is the expectation maximization algorithm?

Chuong B Do , Serafim Batzoglou (2008)
What is the expectation maximization algorithm?
Nature biotechnology volume 26 number 8

Bishop, Christopher M.
Pattern Recognition and Machine Learning.
New York ,Springer, 2006.

Thank you for your attention

Kim Ying WONG

McMaster University
School of Computational Science and Engineering

January 30, 2024