Expectation Maximization

(Data Mining and Warehousing)

Outlines

- Expectation Maximization
- How EM works
- Applications of EM
- Example – Probabilistic Clustering
- Fuzzy Clustering Using EM
- K-means vs EM in terms of Clustering
- Gaussian Mixture Model with EM
- Optimal Number of Clusters
- Advantages and Limitations

Expectation Maximization

- An iterative approach to find local maximum likelihood or maximum a posteriori.
- Can handle latent variables.
- Gaussian mixture models are an approach to density estimation where the parameters of the distributions are fit using the expectation-maximization algorithm.

General MLE Process Vs. EM

- Both Maximum Likelihood Estimation (MLE) and EM can find the "best-fit" parameters, but with different methodologies.
- MLE accumulates all the data object to estimate the parameters; but EM takes a guess at the parameter first, and then tweaks the model to fit the guesses and the observed data.

- EM – The Chicken and Egg Problem

  - Need parameters (mean, covariance) to need the source of the points.

  - Need to know the source to estimate those parameters.

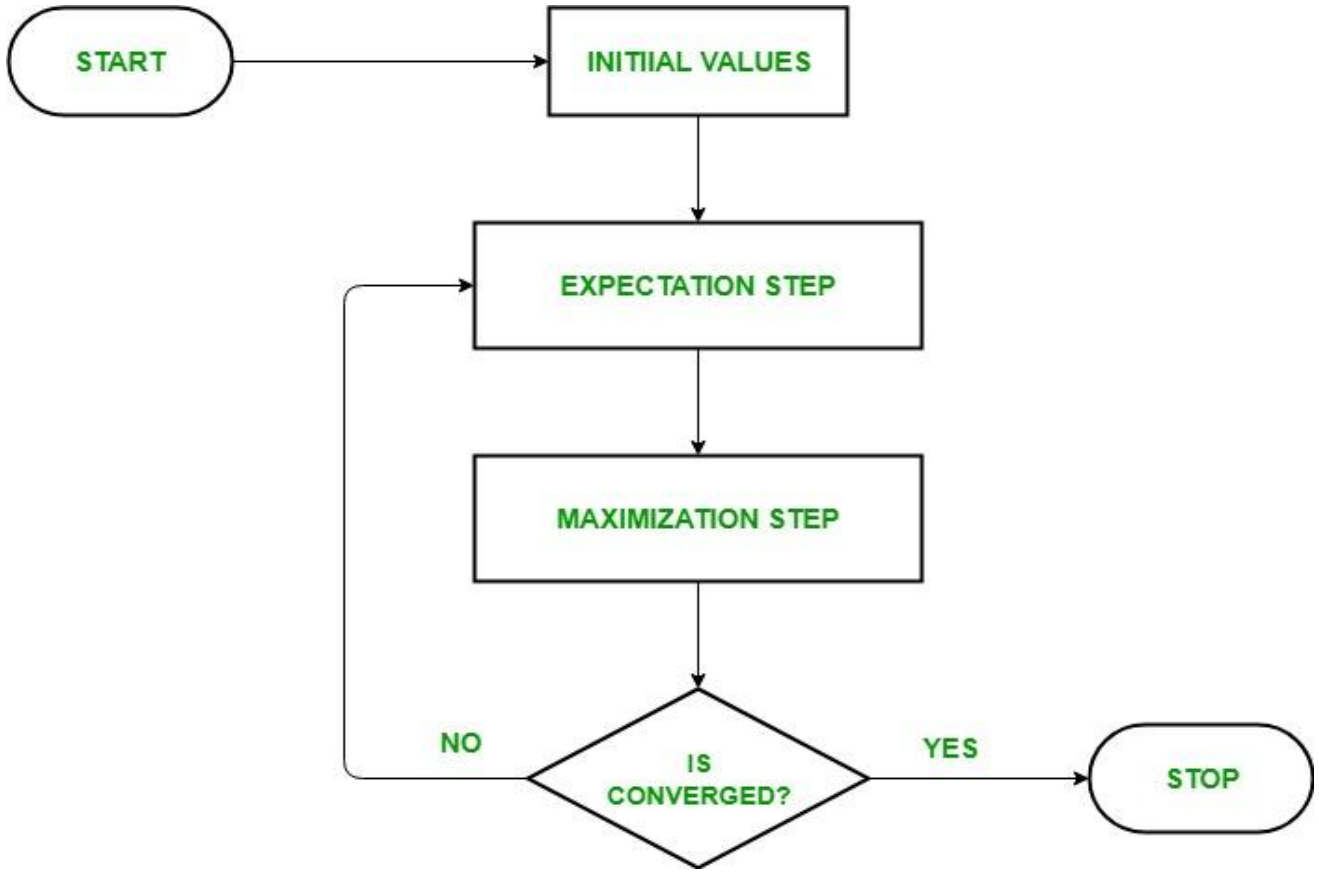
EM Algorithm – How It Works

Below are the steps of the Expectation Maximization Algorithm:

- The E-step: Estimate the missing variables (latent parameters) in the dataset.
- The M-step: Maximize the parameters of the model in the presence of the data.

These two steps are repeated until convergence.

EM Algorithm – How It Works (Contd.)



Applications of The EM Algorithm

- The EM algorithm is mostly used in probabilistic clustering methods (unsupervised), especially in Fuzzy Clustering and Probabilistic Model Based Clustering.
- Computer Visions and Machine Learning.
- Natural Language Processing (NLP).
- Estimation the parameters of Hidden Markov Model (HMM) Classifiers.
- Reconstruction of medical images.

Probabilistic Clustering

- A method for deriving cluster where each object is assigned a probability of belonging to a cluster.
- Data objects are assumed to be coming from some distribution function.
- If the probability is interpreted as degree of membership, then these are fuzzy clustering techniques.

Fuzzy Clustering

Given a set of objects, $X = \{x_1,\ldots,x_n\}$, a fuzzy set S is a subset of X that allows each object in X to have a membership degree between 0 and 1. Formally, a fuzzy set, S, can be modeled as a function, FS:X-> [0,1].

This concept can be applied on clustering.

- Fuzzy clustering allows an object to belong to more than one cluster.
- This clustering can be represented using a partition matrix M^T, where each object is assigned a membership degree, for each fuzzy cluster.
- Also called soft clustering.
- Used in text mining.

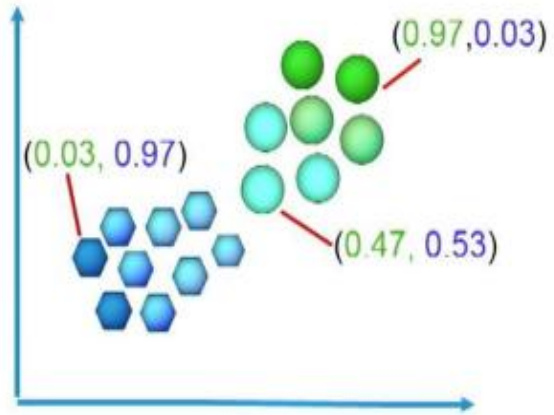
Fuzzy Clustering (Contd.)

Hard clustering

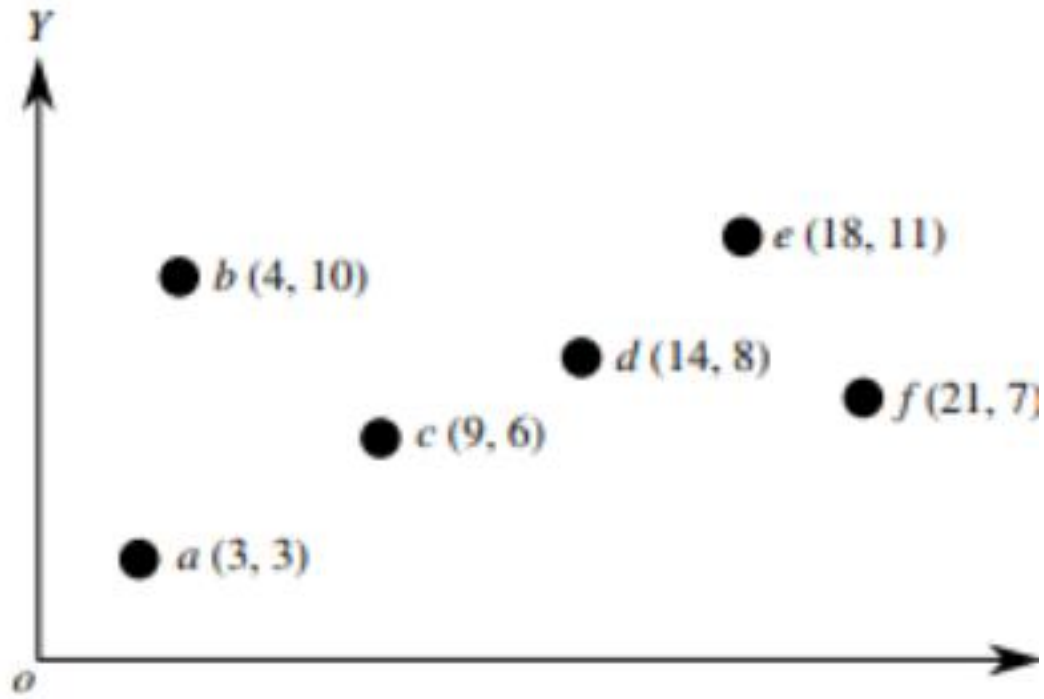
Each observation belongs to exactly one cluster

Soft clustering

An observation can belong to more than one cluster to a certain degree (e.g. likelihood of belonging to the cluster)



Fuzzy Clustering Using EM



Consider following six points:

a(3,3), b(4,10), c(9,6), d(14,8), e(18,11) and f(21,7)

We randomly select two points, say c_1 = a and c_2 = b, as the initial centers of the clusters.

Fuzzy Clustering Using EM (Contd.)

- 1st E step: Assign objects to the clusters.
- Calculate the weight (Membership Degree) of each object for each cluster: w_{ij} means the weight of object i in cluster j.
- For any data object o,

$$w_{ij} = \frac{\frac{1}{dist(o,c_1)^2}}{\frac{1}{dist(o,c_1)^2}+\frac{1}{dist(o,c_2)^2}} = \frac{dist(o,c_2)^2}{dist(o,c_1)^2+dist(o,c_2)^2} \text{ and } \frac{dist(o,c_1)^2}{dist(o,c_1)^2+dist(o,c_2)^2}$$

For data point, c(9,6), $w_{c,c1}= \frac{41}{41+45} = 0.48$ and $w_{c,c2}= \frac{45}{41+45} = 0.52$

Fuzzy Clustering Using EM (Contd.)

1st M Step: Update the previously assigned centroids.

$$c_j = \frac{\sum_{\text{each point } o} w_{o,c_j}^2 o}{\sum_{\text{each point } o} w_{o,c_j}^2},$$

where j=1,2.

According to this formula, after 1st iteration, we get the updated centers c1(8.47,5.12) and c2(10.42,8.99).

Fuzzy Clustering Using EM (Contd.)

In this example,

$$c_1 = \left(\frac{1^2 \times 3 + 0^2 \times 4 + 0.48^2 \times 9 + 0.42^2 \times 14 + 0.41^2 \times 18 + 0.47^2 \times 21}{1^2 + 0^2 + 0.48^2 + 0.42^2 + 0.41^2 + 0.47^2}, \right.$$
$$\left. \frac{1^2 \times 3 + 0^2 \times 10 + 0.48^2 \times 6 + 0.42^2 \times 8 + 0.41^2 \times 11 + 0.47^2 \times 7}{1^2 + 0^2 + 0.48^2 + 0.42^2 + 0.41^2 + 0.47^2} \right)$$
$$= (8.47, 5.12)$$

and

$$c_2 = \left(\frac{0^2 \times 3 + 1^2 \times 4 + 0.52^2 \times 9 + 0.58^2 \times 14 + 0.59^2 \times 18 + 0.53^2 \times 21}{0^2 + 1^2 + 0.52^2 + 0.58^2 + 0.59^2 + 0.53^2}, \right.$$
$$\left. \frac{0^2 \times 3 + 1^2 \times 10 + 0.52^2 \times 6 + 0.58^2 \times 8 + 0.59^2 \times 11 + 0.53^2 \times 7}{0^2 + 1^2 + 0.52^2 + 0.58^2 + 0.59^2 + 0.53^2} \right)$$
$$= (10.42, 8.99).$$

Fuzzy Clustering Using EM (Contd.)

Partition matrix and updated centroids after three iterations:

Iteration	*E-Step*	*M-Step*
1	$M^T = \begin{bmatrix} 1 & 0 & 0.48 & 0.42 & 0.41 & 0.47 \\ 0 & 1 & 0.52 & 0.58 & 0.59 & 0.53 \end{bmatrix}$	$c_1 = (8.47, 5.12)$ $c_2 = (10.42, 8.99)$
2	$M^T = \begin{bmatrix} 0.73 & 0.49 & 0.91 & 0.26 & 0.33 & 0.42 \\ 0.27 & 0.51 & 0.09 & 0.74 & 0.67 & 0.58 \end{bmatrix}$	$c_1 = (8.51, 6.11)$ $c_2 = (14.42, 8.69)$
3	$M^T = \begin{bmatrix} 0.80 & 0.76 & 0.99 & 0.02 & 0.14 & 0.23 \\ 0.20 & 0.24 & 0.01 & 0.98 & 0.86 & 0.77 \end{bmatrix}$	$c_1 = (6.40, 6.24)$ $c_2 = (16.55, 8.64)$

Comparison on Clustering Methods (k-means Vs. EM)

K-means	Expectation Maximization
1. Hard Clustering.	1. Soft clustering.
2. Based on Euclidean distance.	2. Based on density probability.
3. Works on numeric data only.	3. Works on both nominal and numeric data.
4. Inherently non-robust; sensitive to outliers.	4. Robust method.

Probabilistic Model Based Clustering and Mixture Models

Probabilistic Model Based Clustering

- Cluster analysis is to find hidden categories based on generative models.
- Each hidden category is a distribution over the data space.
- Each category represents a probabilistic cluster.

Mixture Models

- Probabilistically-grounded way of doing soft clustering.
- Each cluster is a generative model (Gaussian, Multinomial).
- The parameters are latent (Mean, covariance etc.).

Advantages and Limitations of EM

Advantages:

- 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.

Limitations:

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

EM – Dealing with The Local Maxima Problem

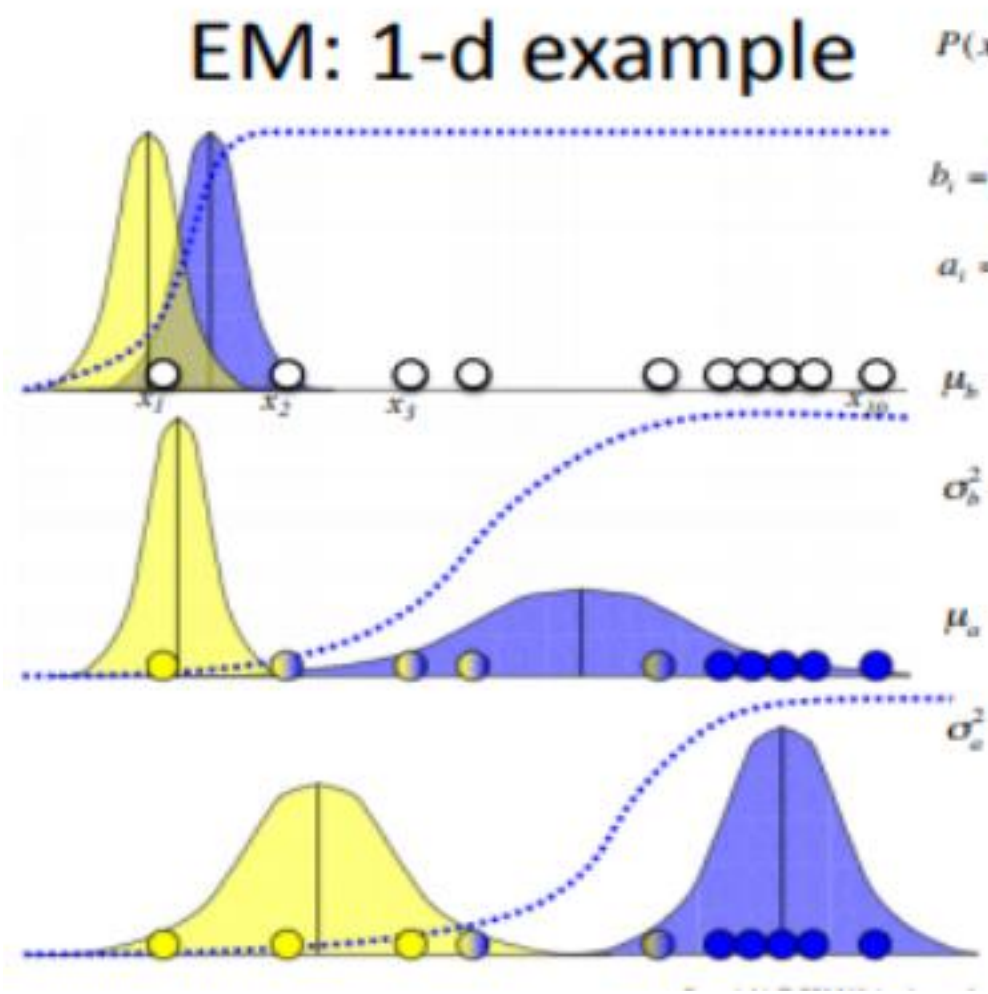
- The EM algorithm iterations always increases the likelihood, but it has a propensity to converge to local maxima.
- Restarting the algorithm from the initial "parameter guessing" step can be a solution.
- From all the possible (guessed) parameters, we can choose the one which yields the greatest maximum likelihood.

References

1. Jung YG, Kang MS, Heo J. Clustering performance comparison using *K*-means and expectation maximization algorithms. *Biotechnol Biotechnol Equip*. 2014;28(sup1):S44-S48. doi:10.1080/13102818.2014.949045
2. Gupta, Ujjwal Das, Vinay Menon, and Uday Babbar. "Detecting the number of clusters during expectation-maximization clustering using information criterion." In *2010 Second International Conference on Machine Learning and Computing*, pp. 169-173. IEEE, 2010.
3. https://towardsdatascience.com/a-comparison-between-k-means-clustering-and-expectation-maximization-estimation-for-clustering-8c75a1193eb7
4. https://www.geeksforgeeks.org/ml-expectation-maximization-algorithm/
5. http://www.inf.ed.ac.uk/teaching/courses/iaml/2011/slides/em.pdf
6. https://machinelearningmastery.com/expectation-maximization-em-algorithm/
7. https://www.statisticshowto.com/em-algorithm-expectation-maximization/

Thank you!!

Univariate Gaussian Mixture Model with EM



$$P(x_i|b) = \frac{1}{\sqrt{2\pi\sigma_b^2}} \exp\left\{-\frac{(x_i-\mu_b)^2}{2\sigma_b^2}\right\}$$

$$b_i = P(b|x_i) = \frac{P(x_i|b)P(b)}{P(x_i|b)P(b)+P(x_i|a)P(a)}$$

$$a_i = P(a|x_i) = 1 - b_i$$

$$\mu_b = \frac{b_1x_1 + b_2x_2 + \ldots + b_nx_n}{b_1 + b_2 + \ldots + b_n}$$

$$\sigma_b^2 = \frac{b_1(x_1-\mu_b)^2 + \ldots + b_n(x_n-\mu_b)^2}{b_1 + b_2 + \ldots + b_n}$$

$$\mu_a = \frac{a_1x_1 + a_2x_2 + \ldots + a_nx_n}{a_1 + a_2 + \ldots + a_n}$$

$$\sigma_a^2 = \frac{a_1(x_1-\mu_a)^2 + \ldots + a_n(x_n-\mu_a)^2}{a_1 + a_2 + \ldots + a_n}$$

could also estimate priors:

$P(b) = (b_1 + b_2 + \ldots b_n) / n$

$P(a) = 1 - P(b)$

Does it look more like a sample from yellow gaussian, or blue?

Bayesian Posterior

Each cluster follows 1-D Gaussian distribution.

Optimal Number of k for Gaussians

- Increasing the number of clusters?
  - The dimentionality of the model also increases.
  - Monotonous increase in likelihood.
- Focus on maximizing the likelihood with any number of clusters?
  - We may end up with k = n clusters for n data points.
- An Information Criteria parameter is used for selection among models with different number of parameters P.
  - Introduces a penalty term for each parameter.
  - Pick the simplest of the models.

  BIC : $\max_p \{ L - \frac{1}{2}*p*\log(n)\}$

  AIC : $\min_p \{2p - L\}$