Research and Implementation of EM Clustering Algorithm Based on Latent Variable Mining

Qiang Yue\(^1\),*\, Zhongyu Hu\(^2\), and Dongping Li\(^1\)

\(^1\)Kunming University, School of Information Engineering, 650214 Kunming, China
\(^2\)Kunming University, School of Mechanical and Electrical Engineering, 650214 Kunming, China

Abstract. Clustering analysis is one of the hot research fields in data mining. EM algorithm is an effective method to realize maximum likelihood estimation, which is mainly used for parameter estimation of incomplete data. It greatly simplifies the likelihood function equation by assuming the existence of latent variable, while maximum likelihood estimation is a commonly used parameter estimation method, and the EM algorithm makes its application more extensive. This paper takes clustering algorithm as the main research object, introduces the basic idea of maximum likelihood estimation, describes the basic theory of EM algorithm, and realizes EM algorithm. The experimental results show that compared with the traditional clustering algorithm, the EM algorithm has better convergence and clustering ability.

1 Introduction

With the era of big data coming, data begin to grow explosively. It is very difficult to obtain knowledge from these mass data. Data mining has attracted a great deal of attention in the information society, and has been widely used in many application fields. As one of the best important approaches of data mining, clustering analysis is a method that divides a dataset into groups of similar objects, thereby minimizing the similarities between different clusters and maximizing the similarities between objects in the same cluster [1]. Clustering is widely applied in data mining, such as in customer value analysis and AFC System in urban rail transit. Typical clustering methods are shown in Table 1.

Table 1. Clustering methods and algorithms.

| Clustering Method         | Clustering algorithm                              |
|--------------------------|--------------------------------------------------|
| Partitioning method      | K-means, K-Medoids, and K-Prototypes, etc.        |
| hierarchical method      | BIRCH, etc.                                       |
| density-based method     | DBSCAN, etc.                                      |
| grid-based method        | STRING, etc.                                      |
| model-based method       | EM, SOM, etc.                                     |

Latent variable model is an effective method for modelling, data analysis and data mining[2]. Assuming that \( x_1, x_2, \ldots, x_n \) are latent variables, which act on the observed
variable $y$. Assuming that the relationship between the latent variables and the observed variables is $y = f(x_1, x_2, \ldots, x_n)$, the distribution parameters and corresponding values of the latent variables can be estimated by the given values of the observed variables, as shown in Fig. 1. If the sample is regarded as the observed variable and the potential category as the latent variable, the clustering problem is also the parameter estimation problem, and the parameters are divided into the latent category variable and other parameters[3].

![Figure 1. Latent variable model.](image)

The rest of this paper is organized as follows. Section 2 introduces maximum likelihood estimation. Section 3 describes the EM algorithm. Section 4 presents and analyses the experimental results. Section 5 concludes this paper.

2 Maximum likelihood estimation

Maximum likelihood estimation is a classical estimation method, which treats the parameters to be estimated as an unknown constant. The conditional probability density function $f(x | \theta)$, as the observable measure of the $\theta$ function, is called likelihood function, which represents the conditional probability density function of $x$ under certain conditions of $\theta$.

Assuming that $x$ is a continuous random variable, and its distribution density function is $p(x | \theta)$, which is completely determined by the parameter $\theta$. Known N observations $x_1,\ldots,x_n$, assuming that they are extracted independently from the population whose distribution density is $p(x | \theta)$. Note $X = \{x_1,\ldots,x_n\}$, then there is Eq. (1):

$$p(X|\theta) = \prod_{i=1}^{N} p(x_i|\theta) \equiv L(\theta|X)$$  \hspace{1cm} (1)

The function $L(\theta|x)$ is called likelihood function. When $x$ is fixed, $L(\theta|x)$ is a function of $\theta$. The essence of maximum likelihood parameter estimation is to find out the value of $\theta$ when $L(\theta|x)$ reaches maximum, that is

$$\hat{\theta} = \arg \max_{\theta} L(\theta|X)$$  \hspace{1cm} (2)

where $\theta$ is the parameter space. In order to calculate the value of $\hat{\theta}$ that makes $L(\theta|x)$ reach the maximum, the logarithm is usually taken from both sides of Eq. (1), that is

$$\ln L(\theta|X) = \sum_{i=1}^{N} \ln p(x_i | \theta)$$  \hspace{1cm} (3)

In Eq. (3), the partial derivatives of $\theta_i$ are obtained respectively, so that the partial derivatives are equal to zero. The following equations are obtained

$$\frac{\partial}{\partial \theta_i} \ln L(\theta|X) = 0, \ i = 1,\ldots,M$$  \hspace{1cm} (4)

The maximum likelihood estimate $\hat{\theta}$ can be obtained by solving equations (4).
3 EM algorithm

First of all, a brief explanation of the symbols in the discussion of EM algorithm is given: \( x \) are actually observed data, but they are usually incomplete, \( y \) are latent variables, because they cannot be observed directly, so they are also called “unobserved data” or “missing data”.

3.1 Jensen’s inequality

Lemma 1 for \( X \in I \), if \( f'(x) \) is differentiable twice and \( f''(x) \geq 0 \), then \( f(x) \) is a convex function.

Lemma 2 for a convex function \( f(x) \) on \( I = [a, b] \), the following inequality holds.

\[
f\left(\sum_{i=1}^{n} \lambda_i x_i\right) \leq \sum_{i=1}^{n} \lambda_i f(x_i)
\]  

(5)

3.2 EM algorithm

EM algorithm is an effective method for maximum likelihood estimation, which is mainly applied to the following two kinds of incomplete data parameter estimation: first, the observed data is incomplete, which is caused by the limitations of the observation process, second, the likelihood function is not analytical, or the expression of the likelihood function is too complex, which leads to the failure of the traditional estimation method of the maximum likelihood function[4].

Assuming that there is a complete data set \( Z = (X, Y) \), \( X \) is an incomplete observed data set subject to a certain distribution, then the density function of \( Z \) is:

\[
p(z|\theta) = p(x, y|\theta) = p(y|x, \theta)p(x|\theta)
\]  

(6)

It can be seen from Eq. (6) that the density function \( p(z|\theta) \) is determined by the edge density function \( p(x|\theta) \), the hypothesis of the latent variable \( y \), the initial estimate of parameter \( \theta \) and the relationship between the latent variable and the observed variable.

Then a new likelihood function is defined:

\[
L(\theta|Z) = L(\theta|X,Y) \equiv p(X,Y|\theta)
\]  

(7)

\( L(\theta|Z) \) is called the full data likelihood function. Because the latent variable \( Y \) is unknown, the likelihood function \( L(\theta|Z) \) is random and determined by the latent variable \( Y \).

3.2.1 E-step

E-step is the first step of EM algorithm. Given the observed data \( X \) and the current parameter estimates, the expectation of \( \log (p(X, Y | \theta)) \) about the unknown data \( Y \) is calculated. Therefore, the expectation of log likelihood function is defined:

\[
Q(\theta, \theta^{i-1}) = E(\log p(X,Y|\theta)|X,\theta^{i-1})
\]  

(8)

where \( \theta^{i-1} \) is the known current parameter estimate. \( X \) and \( \theta^{i-1} \) are constants, \( \theta \) is the parameter to be optimized, \( Y \) is a random variable, which is assumed to obey a certain distribution \( f(.) \),

\[
y \sim f(y|X, \theta^{i-1})
\]  

(9)

Therefore, Eq. (8) can be written as:

\[
Q(\theta, \theta^{i-1}) = E(\log p(X,Y|\theta)|X,\theta^{i-1}) = \int_{y \in D} \log p(X,y|\theta)| f(y|X, \theta^{i-1}) dy
\]  

(10)
where \( f(y|X, \theta^{i-1}) \) is the edge distribution density function of the unobserved data \( Y \), and depends on the observed data \( X \) and the current parameter \( \theta^{i-1} \). In some special cases, \( f(y|X, \theta^{i-1}) \) is a simple analytic function of \( X \) and \( \theta^{i-1} \), but this function is usually difficult to get. From the multiplication formula, Eq. (11) can be derived:

\[
f(y,X,|\theta^{i-1}) = f(y|X, \theta^{i-1})f(X|\theta^{i-1})
\]  

Define binary functions:

\[
h(\theta, Y) \triangleq \log L(\theta|X,Y)
\]  

where \( y \) follows a certain distribution \( f_Y(y) \), then

\[
E_y[h(\theta, Y)] = \int_y h(\theta, Y)f_Y(y)dy \triangleq Q(\theta)
\]  

From Eq. (13), it can be seen that \( E_y[h(\theta, Y)] \) is a function of \( \theta \), and the estimated value \( \hat{\theta} \) of parameter \( \theta \) can be obtained by simple optimization method. The calculation of \( E_y[h(\theta, Y)] \) is just the E-step of EM algorithm.

### 3.2.1 M-step

M-step is the second step of EM algorithm. Maximize the expectation value \( Q(\theta, \theta^{i-1}) \), that is, find a \( \theta^i \), and meet Eq. (14).

\[
\theta^i = \arg \max_\theta Q(\theta|\theta^{i-1})
\]  

where \( \theta \) is the parameter space.

### 4 Experimental results and analysis

The experimental data come from the birth rate and death rate of various provinces and cities in a certain year. The data set is a two-dimensional data set, including the birth rate and death rate of 31 provinces and cities. The experimental result of the algorithm is shown in Fig. 2.

![Figure 2. Operation result of EM clustering algorithm.](image-url)
From the above output result, we can see that the optimal number of cluster is 5, and each cluster contains 3, 11, 8, 2 and 7 samples respectively. Different symbols represent different clusters, and the asterisk "*" represents the center of each cluster.

After estimating the density of each sample, we draw 2-dimensional and 3-dimensional density maps, as shown in Fig. 3 and Fig. 4.

5 Conclusions and future work

Maximum likelihood estimation is widely used in many fields such as parameter estimation, pattern recognition and signal processing. EM algorithm makes the maximum likelihood function easy to process by assuming the existence of latent variables, which makes the application of maximum likelihood estimation more simple. This paper first introduces the basic idea of maximum likelihood estimation in detail, and then studies and implements EM algorithm. When using this algorithm to realize clustering, the data set is regarded as a probability model with latent variables. The optimal solution is found by repeatedly estimating the parameters of the model, and the corresponding optimal cluster number is given to realize the optimization of the model. As future work, we will apply EM algorithm to AFC system of urban rail transit to solve the problem of knowledge discovery in large-scale operation data.

In this paper, the research was sponsored by the Science Research Foundation Project of the Education Department of Yunnan Province (Project No. 2016ZDX097).

References

1. Yue Qiang, Hu Zhongyu, Shen Shikai, et al. Research of Intrusion Detection Method Based on Ant Colony Clustering [C]. 2016 4th International Conference on Machinery, Materials and Computing Technology (ICMMCT 2016). (2016):34-39.
2. Tarpey T, Petkova E. Latent regression analysis[J]. Statistical Modelling, (2010): 133-158.
3. Han M E, Mohamed S, Marlin B M, et al. A stick-Breaking Likelihood for categorical data analysis with latent Gaussian models [C]. Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), (2012):610-618.
4. Ma Jinwen, Xu Lei. A symptotic convergence properties of the EM algorithm with respect to the overlap in the mixture[J]. Neurocomputing, (2005):105-129.