• Cover page with the project title and the names of the group members with their student IDs

anyone

Image Segmentation Using K-means Clustering, Gaussian Mixture model and Expectation Maximization

ENCS 6161

Group Members:

40232982 Zunao Hu

40258305 Qian Sun

• Abstract

Sun:

K-Means clustering and Gaussian Mixture Models (GMM) are two popular methods for data clustering. K-Means divides data into K clusters based on distance between data points, while GMM models the data as a combination of Gaussian distributions. Both methods use the Expectation-Maximization (EM) algorithm for parameter estimation. The EM algorithm alternates between an estimation step, which assigns data to clusters or distributions, and a maximization step, which updates the cluster/distribution parameters to better fit the data. Convergence is checked by evaluating the log likelihood, and the process is repeated until convergence is achieved. This paper explores the K-means algorithm and Gaussian Mixture Model (GMM) with Expectation Maximization (EM). We first implement the K-means algorithm for greyscale images, using intensity histograms in one scenario and without in another. Then we implement the K-means algorithm for color image without using intensity histograms. Additionally, we develop a custom function for fitting GMMs and use it to segment color images, comparing the results to those in a provided PDF. We also create our own function for fitting GMMs to data and compare the convergence plot to the PDF. Overall, this study provides a thorough examination of both K-means and GMM-EM, and offers new insights into their capabilities and limitations.

• Table of contents

Anyone:

I. Introduction A. Background information B. Statement of the problem C. Objectives of the study

II. Literature Review A. Previous studies on the topic B. Gaps in the literature

III. Methodology A. Research design B. Data collection methods C. Sampling procedure

IV. Results A. Presentation of findings B. Analysis of data V. Discussion A. Interpretation of results B. Implications of the findings

VI. Conclusion A. Summary of the study B. Recommendations for future research

• List of abbreviations in alphabetical order

Anyone:

Example 1: List of Abbreviations for a Medical Research Paper

AIDS - Acquired Immune Deficiency Syndrome BMI - Body Mass Index CVD - Cardiovascular Disease DM - Diabetes Mellitus HIV - Human Immunodeficien

• List of symbols

Anyone:

List of Symbols for a Physics Thesis

a - Acceleration c - Speed of Light E - Energy F - Force G - Gravitational Constant h - Planck's Constant m - Mass p - Momentum r - Radius v - Velocity

• List of Figures

anyone

a - Acceleration c - Speed of Light E - Energy F - Force G - Gravitational Constant h - Planck's Constant m - Mass p - Momentum r - Radius v - Velocity

• List of Tables

anyone

Table 1: Market Analysis by Product Type Table 2: Sales Forecasts by Region Table 3: Price Comparison of Competitors' Products Table 4: Operating Expenses by Quarter Table 5: Break-Even Analysis Table 6: Cash Flow Projections Table 7: Return on Investment Analysis Table 8: Key Performance Indicators

**1. Introduction**

**Sun**

K-means: The k-means algorithm is a commonly employed traditional clustering method that involves a division approach. Its similarity calculation entails determining the distance between a data object and the cluster center, and then dividing the distance from the cluster center into a cluster. This process is repeated until the criterion function converges. The algorithm's time complexity is O(nkt), where n is the total number of objects, k is the number of clusters, and t is the number of iterations. It is highly efficient; however, it has some limitations. Specifically, it can only handle numeric data, cannot process categorical data, is highly sensitive to exception data, and cannot handle clusters with non-convex shapes.

GMM: The Expectation-Maximization (EM) algorithm is a commonly used tool for estimating the parameters of a Gaussian Mixture Model (GMM). It is an iterative procedure that serves as a maximum-likelihood estimator. Despite its appealing properties, the EM algorithm has some well-documented drawbacks, such as the need for good initial values and the possibility of being trapped in local optima. However, despite these limitations, EM plays an important role in parameter estimation for mixture models.

**2. Scope and objectives of the project**

Zunao

The scope of this project is to perform image segmentation using concepts, ideas, and techniques covered in the course. The project objectives are as follows:

1) Learning the K-means algorithm: The K-means algorithm is a clustering technique used to partition a set of data points into K clusters based on similarity. The project aims to implement the K-means algorithm and its variations to gain proficiency in unsupervised learning techniques.

2) Learning Gaussian Mixture Model (GMM) and Expectation Maximization (EM): Gaussian Mixture Model (GMM) is a statistical model that uses a mixture of Gaussian distributions to represent a given set of data points. Expectation Maximization (EM) is an iterative algorithm used to estimate the parameters of a GMM. The project objective is to learn how to implement GMM and EM algorithms.

3) Implementing Image Segmentation: Image segmentation is the process of dividing an image into multiple segments or regions based on similar characteristics such as color, texture, or intensity. The project aims to apply the K-means algorithm and GMM/EM algorithms learned in objectives 1 and 2 for image segmentation.

4) Fitting 2-D Gaussian Mixture Data: The objective of this project is to learn how to fit 2-D Gaussian mixture data using the GMM and EM algorithms learned in objective 2.

By achieving these objectives, this project aims to provide a comprehensive understanding of the unsupervised learning techniques used in image segmentation and their practical applications, as taught in the course. It also aims to provide hands-on experience in implementing these techniques using Python and relevant libraries such as scikit-learn and OpenCV.

**3. Detailed methodology and implementation**

3.1 k-means methodology: Sun

The K-means algorithm belongs to the partition-based clustering algorithms. It defines the initial centroid value to determine the number of groups [12]. Determining the number of clusters (k) precisely is crucial for the K-Means algorithm as the initial cluster center may change, leading to unstable data grouping [13]. The output of the K-Means algorithm depends on the selected center values for clustering. The algorithm determines the clusters based on the initial value of the cluster's center point. Randomly assigning the initial cluster centroid can have an impact on the performance of the cluster [14]. K-Means is a distance-based clustering algorithm that partitions set data into K clusters. It works well for numerical attributes.

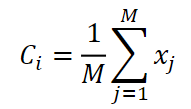
[12] Eltibi M F and Ashour W M 2011 Initializing K-Means Clustering Algorithm using Statistical Information. K-means clustering algorithm is one of the best known, XXIX 7 p51

[13] Aristidis L, Nikos V, Jacob J V 2011 The global k-means Clustering algorithm IAS technical report series, IAS-UVA-01-02

[14] Cosmin M P, Marian C M, Mihai M An Optimized Version of the K-Means Clustering Algorithm, Proceedings of the 2014 Federated Conference on Computer Science and Information Systems (ACSIS) 2 p695

1. Determine the number of clusters (K) and the maximum number of iterations.

2. Initialize the K midpoint clusters, followed by counting the feature centroids using the appropriate equation:



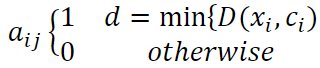
Equation 1 is done as much as p dimensions from i = 1 to i = p

3. Assign each observation data to the nearest cluster using the Euclidean distance metric, which can be calculated using equation 2.



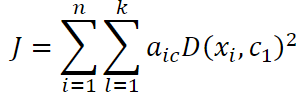
1. The algorithm reallocates the data to each group based on the comparison of the distance between the data and each group's centroid. [9].

[9] Joshi K D and Nalwade P S 2013 Modified K-Means for Better Initial Cluster Centres. International Journal of Computer Science and Mobile Computing II7 p2



5. Recalculate the cluster midpoint position.

The value of the membership of point xi to the centers of group c1 is denoted as aij, d represents the shortest distance from data point xi to group K after comparison, and c1 is the center of group 1. The method uses an objective function that combines the distance and the value of the data membership in the group. The objective function can be determined using the following equation.



In the equation, n represents the amount of data, k represents the number of groups, ai1 denotes the membership value of data point xi to group c1, and a can have a value of either 0 or 1. If the data point belongs to a group, the value of ai1 is 1; otherwise, it is 0.

6. If there is a change in the position of the cluster centroid or if the number of iterations is less than the maximum number of iterations, return to step 3. Otherwise, return the clustering result.

---------------------------------------------------------------------------------------------------

3.2 k-means implementation:

Zunao

1. Grayscale image segmentation with histogram

The first step is to load a grayscale image and compute its intensity histogram. The histogram is then reshaped into a 2D array of pixel intensities, and a user-defined number of clusters (K) is chosen.



Cluster centroids are initialized using the intensity histogram, and pixels are assigned to clusters based on their intensity values. The centroids are then updated by computing the mean of each cluster, and the process is repeated until convergence.

Finally, the segmented image is obtained by assigning each pixel to the closest centroid and visualized alongside the original image. The implementation uses OpenCV for image loading and manipulation, numpy for numerical computations, and matplotlib for data visualization.

3.3 GMM-EM methodology: Sun

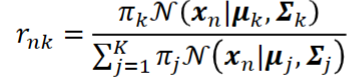
The Gaussian Mixture Model (GMM) is a probabilistic model that uses a combination of Gaussian (Normal) probability distributions to estimate the density of the data. In GMM, the parameters of the Gaussian distributions, such as mean and standard deviation, are estimated using the Expectation-Maximization (EM) algorithm. The EM algorithm consists of two steps: the estimation step and the maximization step. In the estimation step, the algorithm assigns a probability value to each data point to indicate the likelihood of it belonging to a particular cluster. In the maximization step, the algorithm optimizes the parameters of the probability distributions to best capture the density of the data. This iterative process continues until a good set of latent values and a maximum likelihood that fit the data are achieved.

1. Initializing the clusters

Foremost, we must select how many segments we'd like to partition our data into. After selecting k segments to partition the data into, we initialize random Gaussian models.

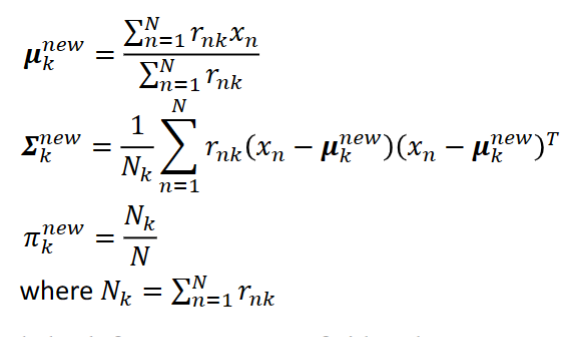
2. Probabalistic assignment to clusters (Expectation)

After initializing k random Gaussian models we can calculate our expectation of responsibilities rnk, a vector of probabilities that xi belongs to the kth clusters for k=1 to k=k.

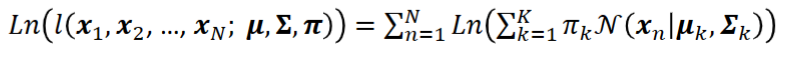


3. Reformulating the Gaussian models (Maximization)

We'll then recalculate our Gaussian models leveraging the weights we found in the expectation step. The expectation, rnk, represents the likelihood that the ith observation belongs to cluster kth.



4.Evaluate the log likelihood and check for convergence of either the parameters or the log likelihood. If the convergence criterion is not satisfied return to step 2.



--------------------------------------------------------------------------------------------------------------

3.4 GMM-EM implementation: Zunao

**4. Experimental results**

Zunao

**5. Conclusion**

Zunao

**References**

Sun

Zunao

**Appendices**

Zunao