Assignment Report

A Comparative Study of Gaussian Mixture Models and K-Means for Data Clustering

Aaron Modiyil Joseph
22018497

University of Hertfordshire
11 January 2024

Introduction

This report compares Gaussian Mixture Model clustering and K-means clustering on a dataset of apartments for rent in the United States. The objective is to perform and understand data preprocessing and use the two clustering approaches to uncover valuable insights into the underlying structures and trends inherent in data, as well as to compare and contrast the two approaches to clustering.

Prototype-Based Clustering

Prototype-based clustering involves grouping objects into clusters closer to the cluster's prototype than any other cluster's prototype. The concept of prototype-based clustering can be expanded in many ways. An approach is to treat a cluster as a statistical distribution, which means that objects are generated randomly from a statistical distribution defined by a set of statistical parameters such as mean and variance. This viewpoint generalizes the concept of a prototype and allows for the application of well-established statistical techniques. A basic approach is to use the cluster's centroid as its prototype(Tan et al., 2019).

Mixture Models - Gaussian Mixture Models (GMM)

In mixture model clustering, the clusters are modelled as a mixture of probability distributions, one for each cluster. Each distribution corresponds to a cluster; its parameters describe the cluster, usually in terms of its centre and spread(Tan et al., 2019).

The probability distributions can be anything, Gaussian mixture models assume they are Gaussian distributions. In the case of a single variable x, the Gaussian distribution can be written in the form

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{1}$$

A superposition of K Gaussian distributions can be written as

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k,\textstyle\sum_k) \tag{2}$$

where each multi-variant Gaussian distribution $\mathcal{N}(\mathbf{x}|\mu_k,\sum_k)$ has its own mean μ_k and covariance $\sum_k$ and π_k is the coefficient of the linear combination. Using a sufficient number of Gaussians and modifying their means and covariances as well as the coefficients in the linear combination, almost any continuous data can be approximated with arbitrary accuracy.

The form of the Gaussian mixture distribution is governed by the parameters π, μ and $\sum$, where we have used the notation $\pi = \{\pi_1, \ldots, \pi_K\}, \mu = \{\mu_1, \ldots, \mu_K\}$ and $\sum$=$\{\sum_1,\ldots,\sum_k\}$. One way to set the values of these parameters is to use maximum likelihood estimation. From (2) the log of the likelihood function is given by

$$\ln p(\mathbf{X}|\pi,\mu,\textstyle\sum) = \displaystyle\sum_{n=1}^{N} \ln\left\{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k,\textstyle\sum_k)\right\} \tag{3}$$

where $\mathbf{X} = \{x_1, \ldots, x_k\}$.

An elegant and powerful method to maximise the likelihood function is *expectation-maximization* algorithm or EM algorithm. Briefly, given a guess for the parameter values, the EM algorithm estimates the probabilities that each point belongs to each distribution. Afterwards, it utilizes these probabilities to create a more accurate estimate for the parameters. (These are the parameters that maximize the likelihood.) This iteration continues until the estimates of the parameters either do not change or change very little(Bishop, 2006).

K-means

In K-means clustering, the prototype is the centroid of the cluster. K-means algorithm randomly chooses k initial centres from the data. Then for each center, we identify the subset of data points (its cluster) that is closer to it than any other centre. The means of each cluster are computed, and this mean vector becomes the new centre for that cluster. This iteration continues until the centres do not change or change very little(Hastie et al., 2009).

Data Overview and Pre-processing

The data chosen is a dataset of apartments for rent in the USA. The dataset contains 10,000 rows and 22 columns. The dataset is taken from the UC Irvine Machine Learning Repository.

The data was pre-processed before applying the clustering algorithms of Gaussian mixture models and k-means. The pre-processing steps included checking for duplicate rows as they create redundancy and bias, selecting only the columns of price, square feet, latitude, and longitude as the clustering features, dropping missing values to ensure that each property had complete information for all four features, performing initial analysis and visualisation using summary statistics, box plots, and correlation heat maps, and detecting and removing outliers using the Z-score method, as they could distort the clustering results by creating artificial clusters or affecting the cluster centroids. In addition, the data was scaled using MinMaxScaler, which transformed the values into the range of 0 to 1. This was done to make the data more comparable and compatible, as the k-means algorithm is sensitive to the scale and range of the features.

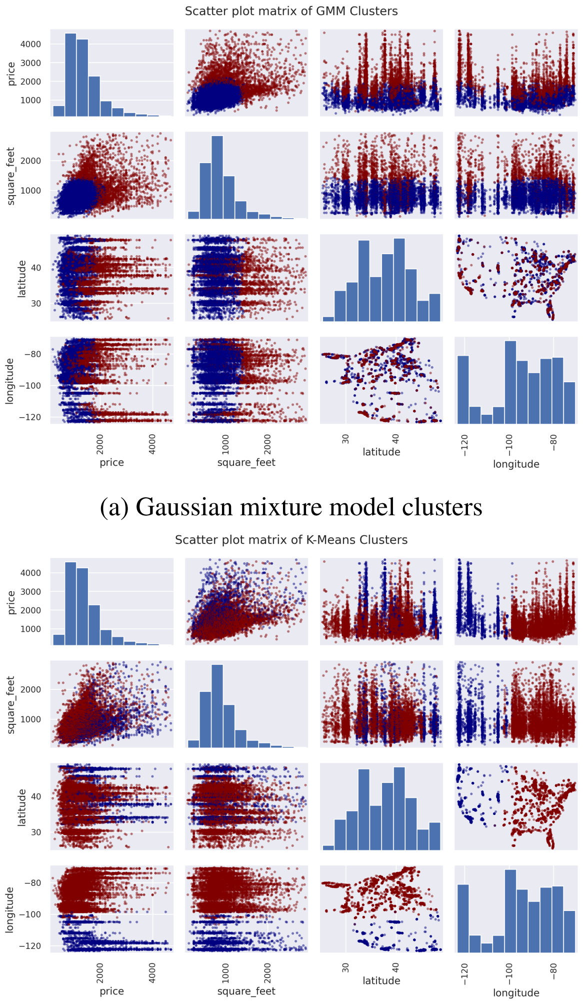
Results

The data were clustered with Gaussian mixture models (GMM) and k-means clustering. Figure 1 shows the scatter plot matrix of the clusters based on each method, with the color coding of the clusters colored blue and red, respectively.

The GMM clusters show a clear separation in the price vs square feet plots, indicating that the blue cluster consists of apartments with low price and small size, while the red cluster consists of apartments with high price and large size. The GMM clusters also have different shapes and densities, reflecting the heterogeneity of the data. The latitude vs longitude plots show that the GMM clusters are distributed across the country.

The k-means clusters, on the other hand, show a clear separation in the latitude vs longitude plots, indicating that the blue cluster corresponds to apartments in the west part of the USA and the red cluster corresponds to apartments in the east part of the USA. The k-means clusters do not show any meaningful separation in the price vs square feet plots, suggesting that the price and size of the apartments are not related to their geographical location.

We can conclude that the GMM clusters capture a more meaningful distinction in the data than the k-means clusters. The GMM clusters can be used to identify apartments with different price and size ranges, while the k-means clusters only provide geographical information. The performance of GMM is better than that of k-means on our data, as k-means is limited by the assumption of spherical clusters in four-dimensional space.



(a) Gaussian mixture model clusters

(b) K-means clusters

Figure 1: Scatter plot matrix of clusters

Conclusion

In this report, we compared Gaussian mixture models and k-means clustering on a dataset of apartments for rent in the USA. We found that GMM performed better than k-means in capturing the price and size differences among the apartments, while k-means only separated the apartments by their geographical location. We also discussed the advantages and limitations of each method. Our findings demonstrate the importance of choosing the appropriate clustering technique for the data and the research question.

References

[1] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer

[2] Hastie T., Tibshirani R. and Friedman J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition. Springer.

[3] Tan P. N., Steinbach M., Kumar V. and Karpatne A. (2019). *Introduction to Data Mining*, Global Edition. Pearson Education Limited.