San Jose State University

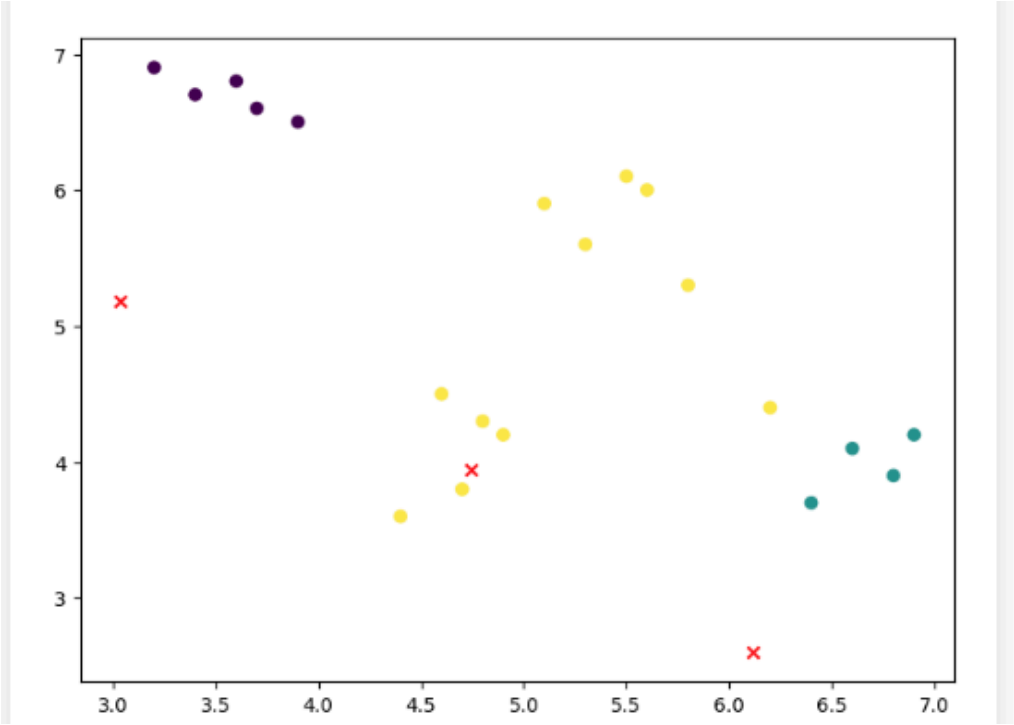

Figure 1

“ClusterMania”

Cluster Based Unsupervised Learning Application - K-Means and Expectation Maximization

---

With the current rapid advancements that are happening in the field of machine learning, an application to input personal data to train and test your own model is an essential step in understanding a given algorithm, including concepts of K-Means Clustering and Expectation Maximization.

By Shreyas Prabhudev, Virali Patel

Department of Computer Science, San Jose State University CS 174: Server-Side Web Programming

December 01, 2023

Table of Contents

Abstract

As there is a growing interest in machine learning, we discover up-to-date and advanced techniques to approach handling large amounts of ambiguous data and the concise conclusions one may be able to derive from analyzing the relationship between data points, ranging from small scale to large scale. While there are a plethora of resources and algorithms to train larger sets of models, understanding how to not only implement and develop a given machine learning algorithm with the implementation of a user-side interactive application was initially a challenge we were faced with, but we were able to combat it in a variety of different trial-and-error scenarios. Within this paper, we will be covering algorithms like K-Means Clustering and Expectation Maximization and their implementations within the project; creating an interactive user-side model of the application; the development of the backend for storing user data and returning the correct visual output to the user; and lastly, covering the next steps for the application.

Introduction

Simplicity and easy understanding within a certain criteria are essential to application development, not only for the user but also for the developer. When building an application, it is important to thoroughly consider concepts; therefore, if there is an obstacle that comes up during the implementation process, it can be avoided by planning for edge cases earlier on. Becoming more and more comfortable with working with the application and learning by failing is a long-driven process in which time, dedication, and perseverance are essential pillars to maintain progress and ultimately have a finished, working product. In terms of machine learning, testing

and learning from failing is the fundamental concept that supports processes within the field. Creating and implementing algorithms, analyzing the cause of failure to further develop methods for potential solutions, building models, and constantly testing how they may be improved for further improvement and efficiency of the project are all methods that machine learning closely follows. This varies from testing for feedback data to create a newer, more commendable iPhone, training human-like applications to provide us with up-to-date and accurate feedback where it will make conclusions based upon an entire swarm of user input such as ChatGPT, or even when you are shuffling a playlist on Spotify (Harvard 2018) when you realize it isn't playing in the order you added the music, but rather the order of the music is based upon the songs you more frequently choose to play, in which it will queue those before ones with less popularity within your unique, personal dashboard. This project aims to allow users to input their own data tested as "scores", in which the application will produce a trained output of a scatter plot graph to display the centroids and clusters within the produced data. Users are initially prompted to log in and will be able to choose from previously trained models to further test. Two algorithms that are implemented include the K-Means Clustering Algorithm and the Expectation Maximization Algorithm.

K-Means Clustering Algorithm

What is a K-means clustering algorithm? This algorithm is one of the more well-known algorithms in which there is cluster analysis that takes place but does so while keeping the euclidean distance to a minimum per data point for each cluster. The procedure behind performing K-means clustering includes assigning a number of centers that are dependent on a client, where these points may be sampled according to the available observations within the

nearest center. This is an iterative process where it will continuously aim to assign data points to an updated, nearest center through calculating the centroid mean of the clustering variables, according to the Columbia Mailman School of Public Health. Until the algorithm is converged and there are no more methods to support an iteration, there will be a continuous process of reclustering and assigning.

Expectation Maximization Algorithm

The Expectation Maximization Algorithm, also known as the EM Algorithm, is known to find the maximum likelihood estimation for a dataset. Typically, EM is applied to datasets that either have missing or hidden elements, which means that they cannot be directly observed, or are entirely incomplete datasets. Expectation maximization is an iterative process in which it will try through estimates of existing data to come up with solutions and inputs for the missing data until there is the best possible match. (Zuniga 2021) Therefore, this algorithm is excellent for working with datasets that are not complete. There are three main initialization steps that are needed to be taken in order to proceed with the procedure of performing expectation maximization derived from principles from the Gaussian Mixture Model (GMM). By setting the parameters for this model, including initializing the mean, covariances, and mixing coefficients, it promotes future convergence and performance for the expectation maximization model, which is far more efficient and accurate.

Training Data	Testing Data
4.1 5.2	5.5 6.1
3.3 4.4	4.7 3.8
2.2 3.6	6.2 4.4
6.7 2.1	3.6 6.8
4.5 3.8	5.1 5.9
5.9 4.2	4.9 4.2
2.1 6.3	6.4 3.7
3.4 5.5	3.9 6.5
6.1 2.8	5.3 5.6
4.2 3.7	4.6 4.5
5.4 2.9	6.6 4.1
2.5 5.8	3.2 6.9
3.2 4.6	5.8 5.3
6.3 3.0	4.4 3.6
4.8 4.1	6.8 3.9
5.7 2.3	3.4 6.7
2.8 6.0	5.6 6.0
3.7 5.2	4.8 4.3
6.5 2.5	6.9 4.2
4.3 3.9	3.7 6.6

Figure Two

Approach

"ClusterMania" was the desired end goal of a functional application that allowed users to create an account, input a text file, or enter data into a text box to test and train data through expectation maximization and the k-means clustering algorithm in which it will then plot the data. Figure 1 displays the plotted results from the tested,trained data displayed in Figure Two. In order to begin building the application, understanding the algorithms above was a crucial and essential step that needed to be taken prior to building the application. Django was used as the primary backend web framework and allows for efficient Python abstraction for web development. Form handling—the registration portion—was handled by Django. Django has

built-in features for log in and log out, in which a table stores, will hash and salt the passwords internally, and has an admin panel in order to observe and control the models created. The framework abstracts away the SQL and database querying and writing, making the development process far more efficient. This Python framework was chosen as it is highly beneficial to become familiar with future career aspects; therefore, implementing this into the project was a clear pathway to take. The application implements client-side validation using Javascript in order to ensure the correct type of values are being entered prior to sending them to the server, resulting in further increased effectiveness. In order to implement the algorithms of K-means and expectation maximization, the Python library scikit-learn was utilized. Default values of 300 maximum iterations, 8 clusters, selecting initial cluster centers to speed up convergence, 1e-4 tolerance to declare convergence, and allowing algorithms (scikit-learn 1.3.2 documentation) to be run with 10 different centroid seeds were chosen to start with.

Conclusion

This project was incredibly informative and interesting to build and watch come to life. From the planning portion at the beginning up to drawing out logistics and having a fully fledged application at the end, it has been a rewarding experience that has taught us a great deal regarding machine learning, web development, maintaining security, and database management. Future works for this project hope to include adding this to a cloud server for public external access and handling large data models with an increase in algorithm variety.

References

- Harvard. “How Spotify Beat Apple, Amazon, and Google Using Machine Learning.” Technology and Operations Management, 13 Nov. 2018, d3.harvard.edu/platform-rctom/submission/how-spotify-beat-apple-amazon-and-google-using-machine-learning/.
- Piech, Chris. “K Means.” CS221, Stanford CS221, 2021, stanford.edu/~cpiech/cs221/handouts/kmeans.html.
- Garett, Renee, et al. “A Literature Review: Website Design and User Engagement.” Online Journal of Communication and Media Technologies, U.S. National Library of Medicine, July 2016, www.ncbi.nlm.nih.gov/pmc/articles/PMC4974011/.
- “K-Means Cluster Analysis.” Columbia University Mailman School of Public Health, 13 Mar. 2023, www.publichealth.columbia.edu/research/population-health-methods/k-means-cluster-analysis.
- “14.4 - the Expectation-Maximization (EM) Algorithm: Stat 508.” PennState: Statistics Online Courses, online.stat.psu.edu/stat508/lesson/14/14.4. Accessed 6 Dec. 2023.
- Zuniga, Christian. “A Primer on the EM Algorithm.” Medium, Towards Data Science, 23 Apr. 2021, towardsdatascience.com/a-primer-on-the-em-algorithm-7bd60e9813e.
- Ravihara, Ransaka. “Gaussian Mixture Model Clearly Explained.” Medium, Towards Data Science, 11 Jan. 2023, towardsdatascience.com/gaussian-mixture-model-clearly-explained-115010f7d4cf.