CLUSTER ANALYSIS ON FASHION MNIST DATASET USING UNSUPERVISED LEARNING

Akshara Santharam
Department of Computer Science
University at Buffalo
Buffalo, NY 14214
aksharas@buffalo.edu

Abstract

Unsupervised learning is a type of machine learning algorithm which can be used to draw inferences from datasets consisting of input data without labeled responses. The most common unsupervised learning method is cluster analysis, which is used for exploratory data analysis to find hidden patterns or grouping in data. The aim of this project is to perform cluster analysis on fashion MNIST dataset using unsupervised learning algorithms like k-means clustering algorithm, Auto-Encoder based K-Means clustering algorithm and Auto-Encoder based GMM clustering algorithm. The main task will be that of clustering images and identify it as one of many clusters. Training of unsupervised model using Fashion-MNIST clothing images is required. The dataset in use is the Fashion-MNIST dataset. This dataset contains a training set of 60,000 examples and a test set of 10,000 examples. After performing cluster analysis, the performance can be tested on the test data by calculating accuracy, precision and recall.

1 Introduction

Unsupervised machine learning algorithms are the type of algorithms that can infer patterns from a dataset without referring to labeled outcomes. Unlike supervised learning, unsupervised learning methods cannot be directly applied to a classification or regression problem. The most common unsupervised learning method is cluster analysis, which is used for exploratory data analysis to find hidden patterns or grouping in data. The clusters are modeled using a measure of similarity which is defined upon metrics such as Euclidean or probabilistic distance.

Unsupervised learning problems can be grouped into clustering and association problems. Clustering is where we discover the inherent groupings in the data such as grouping customers by purchasing behavior. Association is where we discover the rules that describe the large portions of your data.

Different types of clustering algorithms include Hierarchical clustering, K-Means clustering, Gaussian mixture models, Self-organizing maps and Hidden Markov models. Hierarchical clustering creates a cluster tree to build a multilevel hierarchy of clusters. K-means clustering uses distance to the centroid of a cluster to partition data into k distinct clusters. Gaussian mixture model models clusters as a mixture of multivariate normal density components. Self-organizing maps uses neural networks that learn the topology and distribution of the data. Hidden Markov models recovers the sequence of states using the observed data.

Unsupervised learning methods are used in bioinformatics for sequence analysis and genetic clustering; in medical imaging for image segmentation; in data mining for sequence and pattern mining; and in computer vision for object recognition. Unsupervised learning methods can be used for predicting or understanding handwritten digits, clustering websites based on words count on each webpage to understand what those websites are talking about. This algorithm can also be used in market segmentation for targeting appropriate customers, image segmentation, fraud detection in banking sector, deriving climate indices based on clustering of earth science data, gene clustering for grouping gene with similar expression levels, document clustering based on content.

2 Dataset

In this project, we use the Fashion-MNIST) dataset. The Fashion-MNIST dataset will be used for training, and testing. It is a dataset of Zalando's article images, consisting of a training set of 60,000 examples and a test set of 10,000 examples. Each example is a 28x28 grayscale image, associated with a label from 10 classes.

Each image is 28 pixels in width and 28 pixels in height., for a total of 784 pixels in total. Each pixel has a single pixel-value associated with it, indicating the lightness or darkness of that pixel, with higher numbers meaning darker. The pixel-value is an integer between 0 and 255. The training and testing data sets have 785 columns. The first column consists of the class labels and represents the article of clothing. The rest of the columns contain the pixel-values of the associated image.

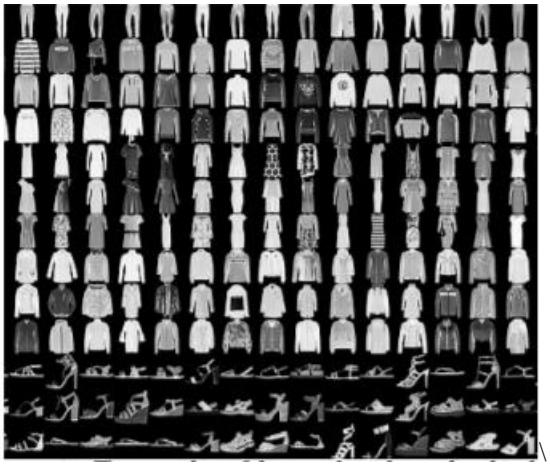

Fig 1: Fashion MNIST Dataset

The labels for the Fashion-MNIST dataset are:

- T-shirt/top
- Trouser
- Pullover
- Dress
- Coat
- Sandal
- Shirt
- Sneaker
- Bag
- Ankle Boot

3 Architecture

4.1 Part A – K-Means Clustering

K-Means clustering is one of the simplest and popular unsupervised machine learning algorithms, which is used when unlabeled data is used. A cluster refers to a collection of data points aggregated together because of certain similarities. A centroid is the location representing the center of cluster. Every data point is allocated to each of the clusters. The main task of the algorithm is to find the groups in data and it works recursively to assign data points to each cluster. Data points are clustered based on similarity feature. Euclidean Distance is used in K- Means clustering and can be calculated by using the following formula:

$$\sqrt{((x1-y1)^2 + (x2-y2)^2}$$

K-Means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. K-Means algorithm in data mining starts with a first group of randomly selected centroids and then performs iterative calculations to optimize the positions of the centroids. The result of K-Means clustering are the centroids of the K- Clusters and the labels of the training data. The K-Means algorithm chooses centroids that minimize the inertia, or within-cluster sum of squares criterion:

$$\sum_{i=0}^{n} \min_{\mu_j \in C} (||x_i - \mu_j||^2)$$

Inertia can be recognized as a measure of how internally coherent clusters are. Inertia is not a normalized metric. Euclidean distance tends to become inflated. So, running a dimensionality reduction algorithm such as Principal Component Analysis (PCA) or Auto-encoder prior to k-means clustering can alleviate the problem and speed up the computations. K-Means algorithms are very Fast, robust and easier to understand. It is very efficient and gives the best result when data set are distinct. But, K-Means cannot handle noisy data and outliers. It fails for non-linear data set. K -Means clustering can be used for customer segmentation, Document clustering, Image Segmentation, Recommendation Engines etc.

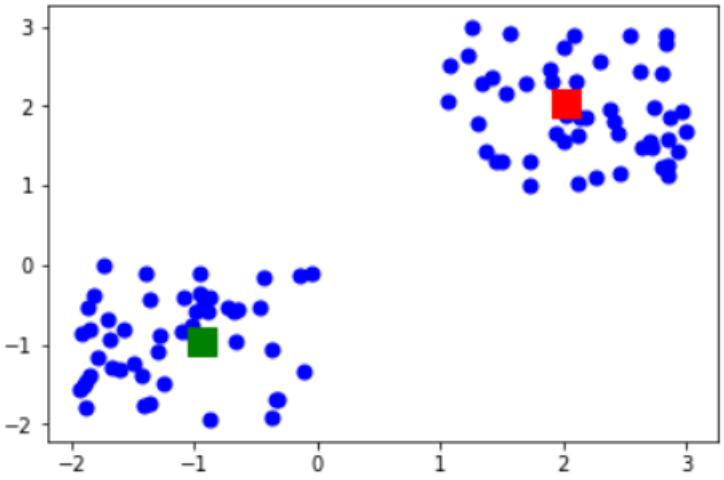

Fig 2: K-Means Clustering [1]

4.2 Part B – Train using Auto-Encoder based K-Means clustering

Autoencoder is an unsupervised machine learning and data compression algorithm that sets the target values to be equal to the inputs by using back propagation. It has two major parts, encoder and decoder. The job of the encoder is to compress the input data to lower dimensional features. For example, one sample of the 28x28 MNIST image has 784 pixels in total, the encoder we built can compress it to an array with only ten floating point numbers also known as the feature of an image.

The decoder part takes the compressed features as input and reconstruct an image as close to the original image. Autoencoder is unsupervised learning algorithm in nature because it takes only the images themselves and not the labels during the training period. The autoencoder built is one fully connected symmetric model. The sparsity parameter is as follows

$$\hat{\rho}_j = \frac{1}{m} \sum_{i=1}^{m} \left[a_j^{(2)} (x^{(i)}) \right]$$

Types of Autoencoders are Denoising autoencoder, Sparse autoencoder, Variational autoencoder (VAE) and Contractive autoencoder (CAE). Auto-Encoders are used in Dimensionality Reduction, Information Retrieval, Anomaly Detection, Image Processing and Drug Discovery.

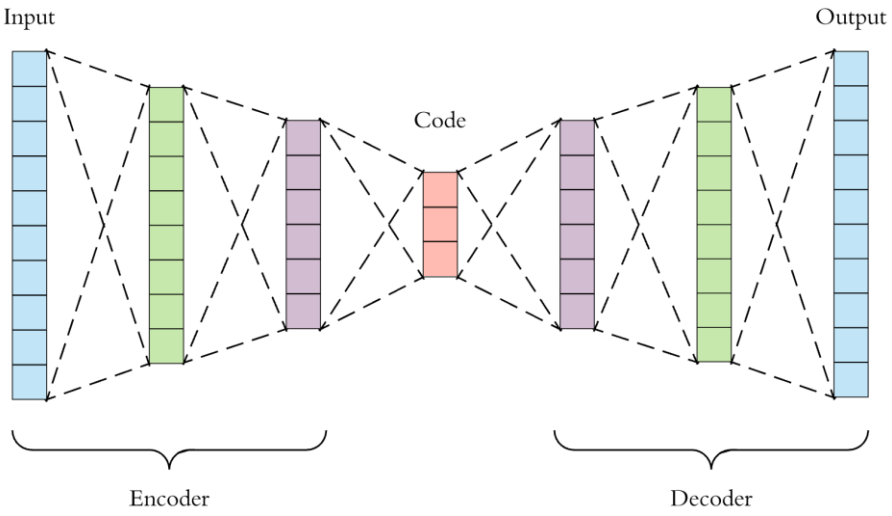

Fig 3: Architecture of auto-encoder [2]

The Auto-Encoder consists of three parts: Encoder, Code and Decoder.

- **Encoder**
  Encoder compresses the input into a latent space representation. This encodes the input image as a compressed representation in a reduced dimension. The compressed image is the distorted version of the original image.
- **Code**
  Code represents the compressed input which is fed to the decoder.
- **Decoder**
  Decoder decodes the encoded image back to the original dimension. The decoded image is a reconstruction of the original image which has been constructed before.

4.3 Part C – Train using Auto Encoder based Gaussian Mixture Model

Gaussian Mixture Models are probabilistic models that uses soft clustering approach for distributing the points in different clusters. Gaussian Mixture Models (GMM) assumes that there are certain number of Gaussian distributions, and each of these distributions represent a cluster. Hence, a Gaussian Mixture Model tends to group the data points belonging to a single distribution together. Gaussian Mixture Models use the soft clustering technique for assigning data points to Gaussian distributions.

A Gaussian mixture model (GMM) attempts to find a mixture of multi-dimensional Gaussian probability distributions. Gaussian mixture models are a probabilistic model for representing normally distributed subpopulations within an overall population. GMM have been used for feature extraction from speech data and have also been used extensively in object tracking of multiple objects, where the number of mixture components and their means predict object locations at each frame in a video sequence. A Gaussian mixture model is parameterized by two types of values, the mixture component weights and the component mean and variances/covariances.

In one dimension the probability density function of a Gaussian Distribution is given by:

$$G(X|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

For Multivariate Gaussian Distribution, the probability density function is given by

$$G(X|\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)|\mathbf{\Sigma}|}} \exp\left(-\frac{1}{2}(X-\mu)^T\mathbf{\Sigma}^{-1}(X-\mu)\right)$$

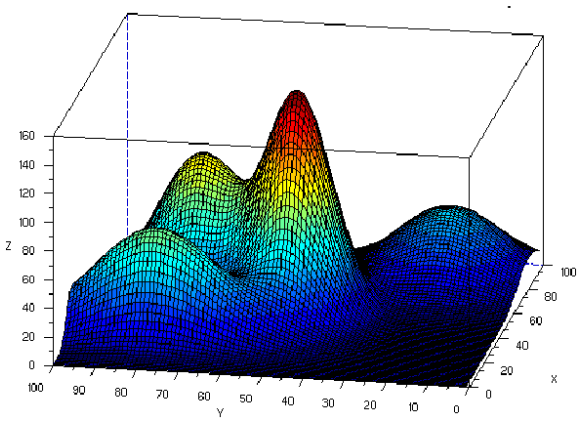

Figure 4: Graph of Gaussian Distribution [3]

5 Result

The task of the project is to perform cluster analysis on fashion MNIST dataset using unsupervised learning. Cluster analysis is one of the unsupervised machine learning technique which doesn't require labeled data. The task is to cluster images and identify it as one of many clusters. In this project, three tasks were performed:

- K-Means algorithm to cluster original data space of Fashion – MNIST dataset using Sklearns library.
- An Auto-Encoder based K-Means clustering model is built to cluster the condensed representation of the unlabeled fashion MNIST dataset using Keras and Sklearns library.
- An Auto-Encoder based Gaussian Mixture Model clustering model is built to cluster the condensed representation of the unlabeled fashion MNIST dataset using Keras and Sklearns library.

Evaluation Metrics: We have evaluated the model using clustering Accuracy and by deriving the confusion matrix. The clustering accuracy achieved in the K-Means clustering layer is 52.4%, Auto Encoder based K-Means clustering is 51.7% and Auto Encoder based GMM Clustering is 52.7%. Confusion matrix is constructed for K-Means clustering algorithm, Auto Encoder based K-Means clustering algorithm and Auto Encoder based GMM clustering algorithm.

- **Auto-Encoder based K-Means Clustering model**

```
[[  1 300  66  54   0   4 562   9   4   0]
 [  0  21 930  13   0   0  34   2   0   0]
 [  0 393   3 587   0   4  11   0   2   0]
 [  0 127 620  16   0   2 216  13   6   0]
 [  0 182  38 674   0   5  88   2  11   0]
 [489   4   0   0  48   0   0 436   0  23]
 [  1 454  35 325   0  16 160   6   3   0]
 [789   0   0   0 188   0   0  21   0   2]
 [ 27  61  16  57   4 382   3  51 398   1]
 [ 15   0   0   0 542   1   1  57   0 384]]
```

Figure 6: Confusion matrix for auto-encoder based K-Means Clustering model

- **Auto-Encoder based Gaussian Mixture Model**

```
[[202  48   2  80   1 592  42   3   0  30]
 [ 17   9   0  59   0   3  14   0   0 898]
 [ 37 652   1  39   0   8 258   2   0   3]
 [112  17   0 436   0  30  25   1   0 379]
 [ 28 658   1 190   0   5 107   5   0   6]
 [ 20   0  56   0 821   0   8   0  95   0]
 [147 368   3  95   1 131 243   3   0   9]
 [  0   0  78   0 919   0   0   0   3   0]
 [ 40  71 357   9  14   1  87 419   1   1]
 [  0   0 465   2  98   0   3   0 432   0]]
```

Figure 7: Confusion matrix for Auto-Encoder based Gaussian Mixture Model

The model is evaluated by plotting graph of training loss and validation vs number of epochs while training for auto-encoder.

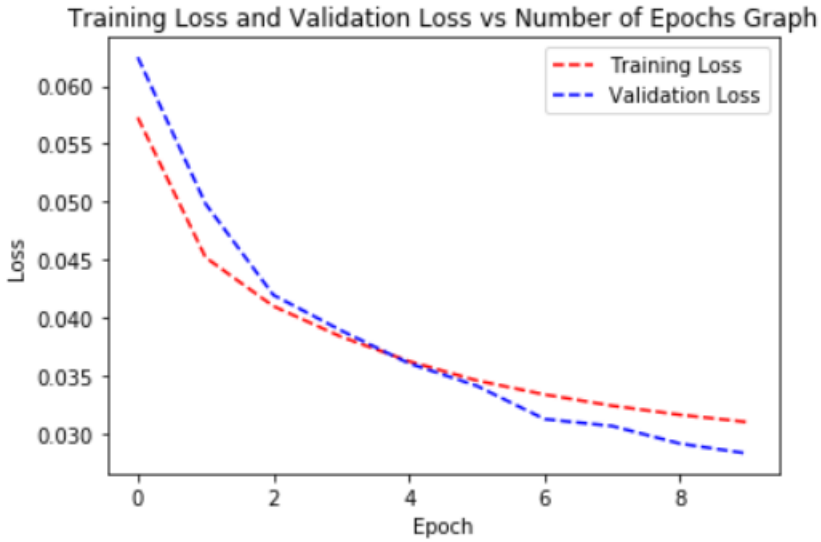

Figure 8: Training loss and validation loss vs Epochs for Auto-Encoder

6 Conclusion

To perform cluster analysis on Fashion MNIST dataset using unsupervised learning, K-Means clustering, and Gaussian Mixture Model clustering is used. The main task is to cluster images and identify it as one of many clusters and to perform cluster analysis on fashion MNIST dataset using unsupervised learning. The model's effectiveness is measured by testing the machine learning scheme on the testing set and the performance can be evaluated by its clustering accuracy. Three tasks performed are K-Means algorithm to cluster original data space of Fashion – MNIST dataset using Sklearns library, an Auto-Encoder based K-Means clustering model is built to cluster the condensed representation of the unlabeled fashion MNIST dataset using Keras and Sklearns library, an Auto-Encoder based Gaussian Mixture Model clustering model is built to cluster the condensed representation of the unlabeled fashion MNIST dataset using Keras and Sklearns library. The clustering accuracy achieved by K-Means clustering, Auto-Encoder based K-Means clustering and Auto-Encoder based GMM Clustering is 52.4%, 51.7% and 52.7%. Henceforth, Gaussian Mixture Model has the highest accuracy and is much better and efficient than the other models.

References

[1] https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

[2]https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders-1c083af4d798

[3]https://www.researchgate.net/figure/3D-view-of-a-4D-Gaussian-Mixture-Model-used-in-our-experiments_fig1_45872040