Exploring Unsupervised Learning on Fashion MNIST dataset

By Yagna Hari Muni

In this case, I'm attempting to train a K-means clustering model on Fashion MNIST data so that I can cluster images from the FMINST dataset with reasonable accuracy. The clustering results will be analyzed using plotly and matplotlib in comparison to the actuals to determine how k-means clustering performs on a Fashion MNIST image dataset.

About Dataset

Fashion-MNIST is a collection of images from Zalando articles. It consists of 60,000 training examples and 10,000 test examples. It is designed to be a direct drop-in replacement for the MNIST dataset in standard machine learning algorithms.

Labels: Each training and test example is assigned to one of the following labels:

0. T-shirt/top
1. Trouser
2. Pullover
3. Dress
4. Coat
5. Sandal
6. Shirt
7. Sneaker
8. Bag
9. Ankle boot

What we trying to achieve

Principal Component Analysis or **PCA** is a technique for reducing the dimensions of a given dataset while retaining most of its variance.

K-means clustering based on the number of clusters given, assigns a few centroids Each data point is assigned to the cluster with the closest centroid to it.

Clustering is an unsupervised machine learning algorithm that recognizes patterns without labels and clusters data based on features.

Unsupervised learning is based on the assumption that no true labels are available. As a result, two variables are created, the first holding labels and the second holding images. The same holds true for test data.

I divided the training and test data into two holdings, one with label data on one variable and the other with plot data on the other.

Visualize the training images

Let’s look at the pictures. We’re using plotly.express to draw the first 16 training images with labels Let's take a look at the images. The first 16 training images with labels mapped in the input data frame are drawn with plotly.express.

Here's an example of how the data appears. Each image is a greyscale 28 by 28 image.

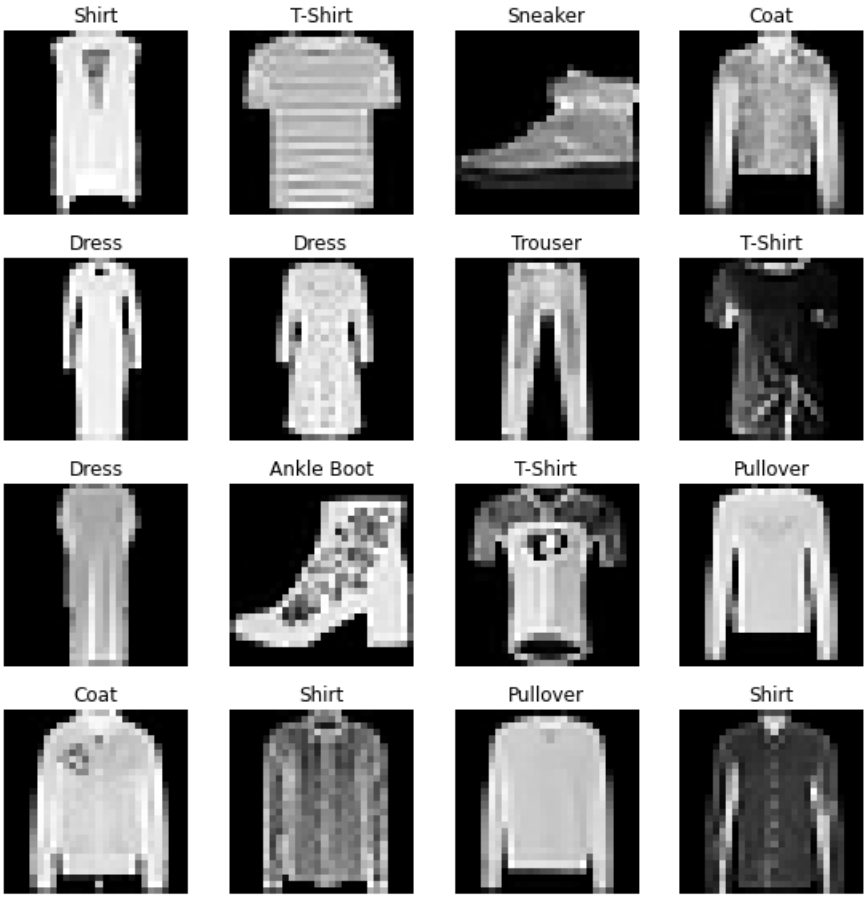

Figure 1: Plots of some random samples from our dataset with corresponding labels

Visualize using Principal Component Analysis

PCA is a technique for reducing the number of dimensions in a dataset while retaining the majority of the data. It is also known as a statistical procedure. It is most commonly used for dimension reduction, especially when dealing with large amounts of data.

By comparing or calculating the correlations between dimensions, this method provides the fewest variables while retaining the most variation (explanation) about how the original data is distributed. In other words, it skips the dimensions with less explained variance and keeps the ones that are more meaningful. We also attempted to standardize the data so that it ranged from 0 to 8000. However, this pre-processing step had no effect on the final result. We believe this is because we used batch normalization in our models, which already used standardizing data.

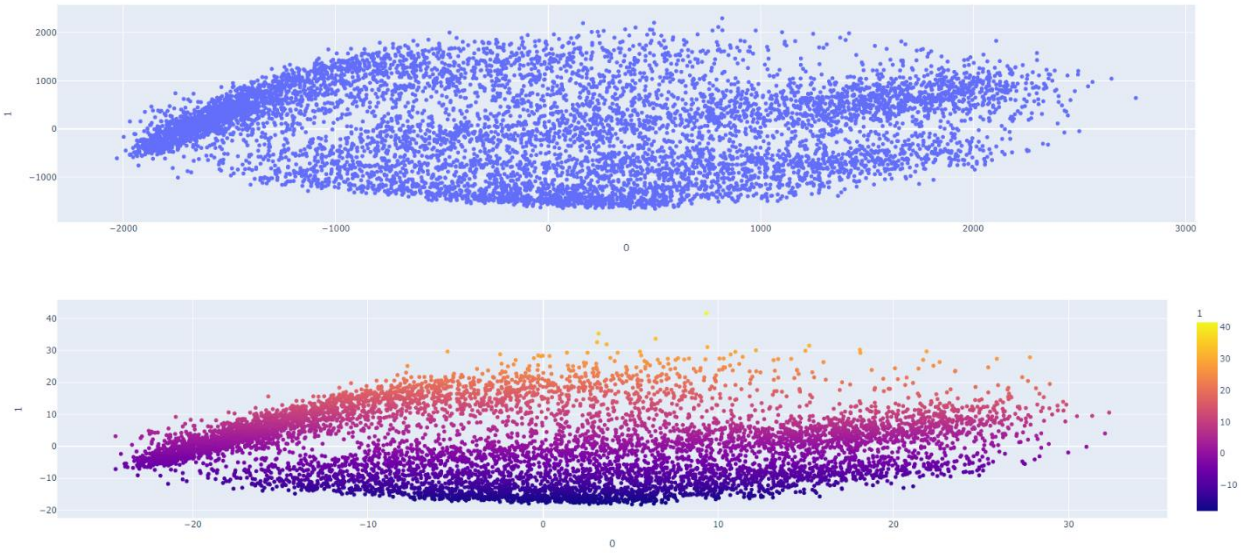

Figure 2: PCA instance on normalized data 2-D

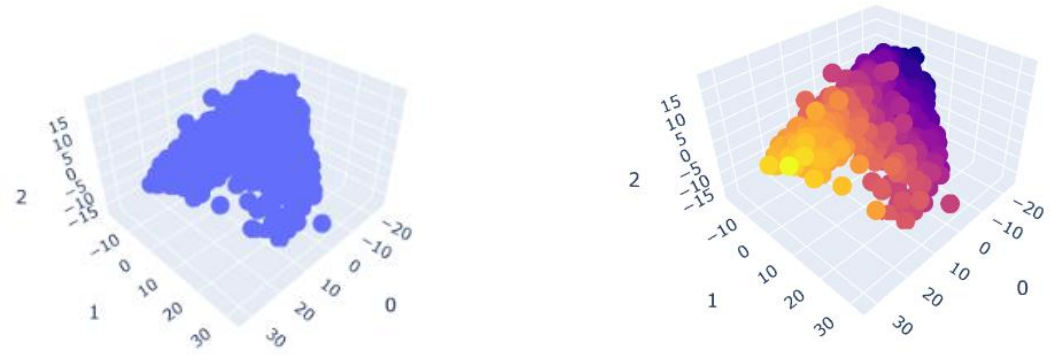

Figure 3: PCA instance on normalized data 3-D

The above figure shows that the principal components captured some variation because there is some structure in the points when projected along the two principal component axes. The points in the same class are close together, while the points or images that are very different semantically are further apart.

K-Means clustering of MNIST data

As previously stated, we can use these feature vectors for any type of machine learning task, such as classification, clustering, or finding similarity between given images, but for now, let's concentrate on clustering. However, in order to find K for clustering an unknown dataset, we will use one of the elbow methods.

However, before feeding feature vectors to the K-Means algorithm, we must first determine the appropriate K; otherwise, the results will be disastrous. As a result, there are numerous methods for determining K, such as the elbow method, dendrogram, and silhouette score analysis. We'll see how the elbow method helps us out.

According to the elbow method, you can find K where there is an elbow, i.e. before the elbow, the error gradually decreases and after the elbow, the decrease isn't significant, as shown in the graph, after K = 41, the error doesn't seem to decrease much, so we can conclude that there are mostly 5 groups in the given data.

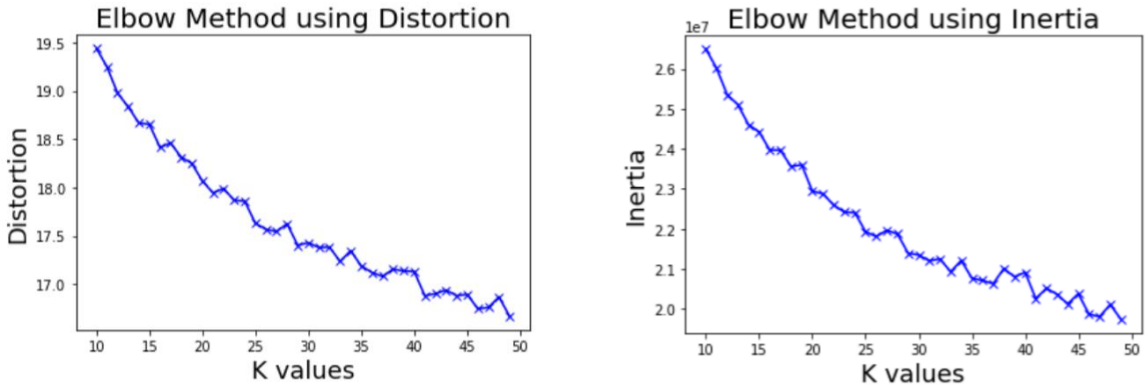

Figure 4: Elbow method using Distortion and Inertia

3D Visualization of the clusters

Using plotly, we will visualize the clusters in 3D. We will only use three of the 420 features in our dataset. This visualization aids in comprehending how well the clusters formed and how far out a single cluster is spread into other clusters.

Clustering does seem to group similar items together. A cluster either contains upper-body clothes(T-shirt/top, pullover, Dress, Coat, Shirt) or shoes (Sandals/Sneakers/Ankle Boots) or Bags. The clustering however performs poorly and seems to group it together with dresses.

Looking at the above results we can see the k -means algorithm distinguishes its clusters based on a The k-means algorithm distinguishes its clusters based on a reasonable pattern, as evidenced by the above results. As a result, we can conclude that the K-means clustering method, when combined with Principal Component Analysis, can produce acceptable results when classifying images without labels.

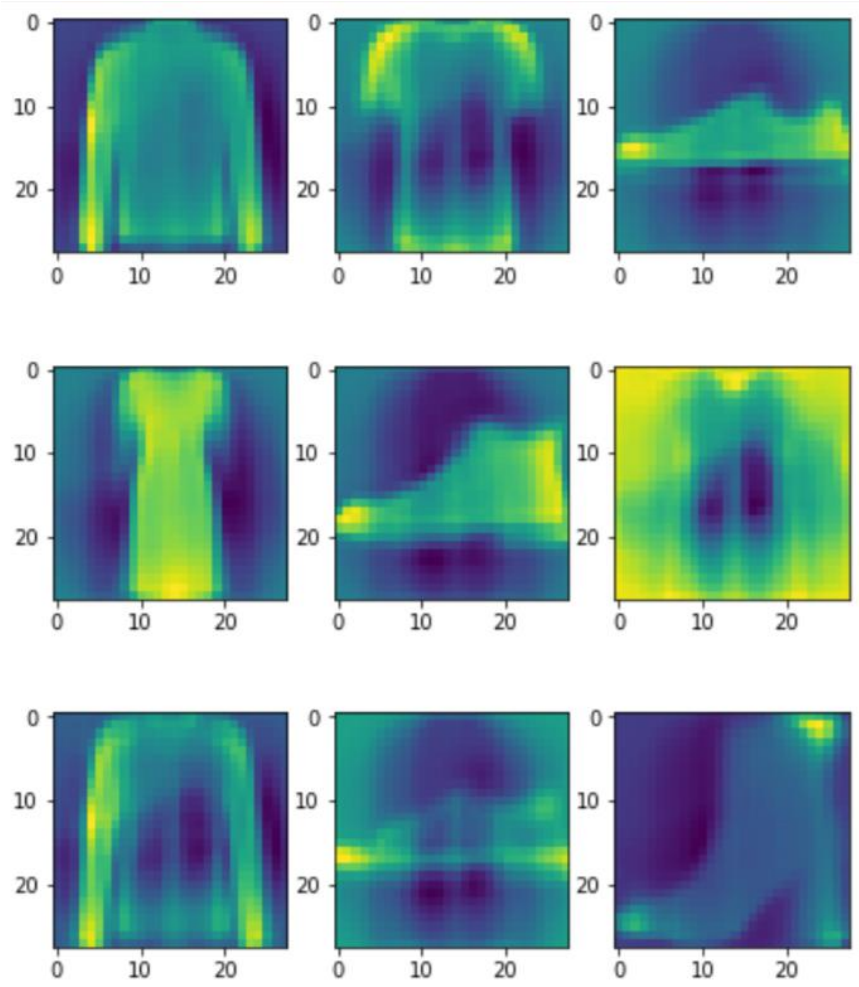

Figure 5: Cluster centroids

K-means Classification

Accuracy: 0.0315, Precision: 0.042881820875600295
Recall: 0.0225, F1: 0.028659745110531803

I've shared the K-means classification Confusion Matrix plotted across predicted and true labels. The outcome shows a poor precision score of 0.04 with low accuracy, recall, and F1.

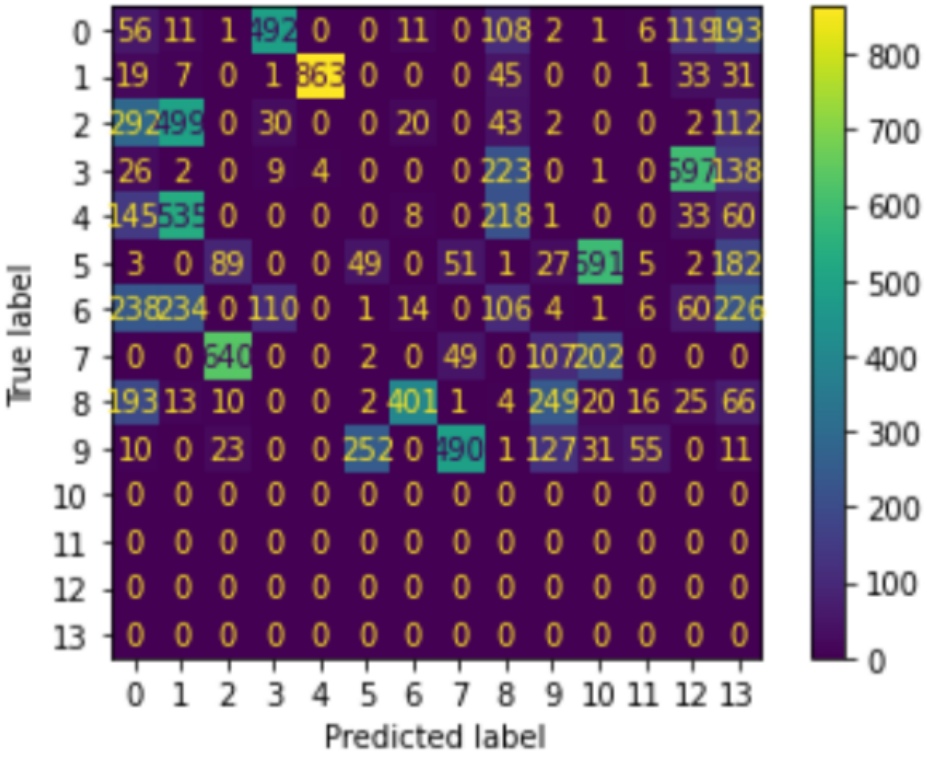

Figure 6: Confusion Matrix

Conclusion

Clustering appears to group similar items together. A cluster contains either all of the clothes (T-shirt/top, pullover, Dress, Coat, Shirt) or all of the shoes (Sandals/Sneakers/Ankle Boots) or all of the bags. Clustering, on the other hand, performs poorly and appears to group it with dresses.