STAT 542: Final Project

Karthik Vasu (kvasu2), Yining Lu (yining13)

Due: 05/05/2022

Contents

Literature review

The Fashion-MNIST data set has been created by researchers at Zalando for the purposes of benchmarking ML algorithms. It consists of 70000 grayscale images of dimension 28*28. These are images of clothing articles like T-shirt, Trouser, Pullover etc with 60000 training samples and 10000 testing samples. The purpose of this data set is to provide a more challenging classifying compared to the original MNIST data. There are algorithms which 99% accuracy on this making it too easy for modern algorithms.

The best accuracy we found was by a GitHub user named *Andy Brock* who was able to achieve an accuracy of 96.7% using wide residual networks. A lot of people have implemented algorithms with high accuracy. They can be for on *Zalando Research's GitHub page*.

Xiao, Rasul, and Vollgraf (2017) test out a variety of classifiers including Decision Tree ,Gradient Boosting, K Neighbors, Linear SVC, Logistic Regression and many more. They achieve the best result using the SVC classifier with C=10 and the rbf kernel. The testing accuracy for this algorithm is 89.7% on the fashion data set and 97.3% on the original MNIST data. Gradient boosting performs well with testing accuracy at 88% and 96.9% respectively. This is achieved for n_estimators=100 and max_depth=10.

Meshkini, Platos, and Ghassemain (2019) perform classification on the Fashion-MNIST data set using convolutional neural networks. They compare the performance of several well-known deep learning frameworks, such as AlexNet, GoogleNet, VGG and ResNet, DenseNet and SqueezeNet. The authors also propose an additional step of batch normalization to enhance the training speed and accuracy of the model. The best results are achieved by ResNet44 and SqueezeNet with batch normalization with testing accuracy at 93.39% and 93.43% respectively.

Summary Statistics

Data table

```
## Ytrain
##    0    1    2    3    4    5    6    7    8    9
## 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000
```

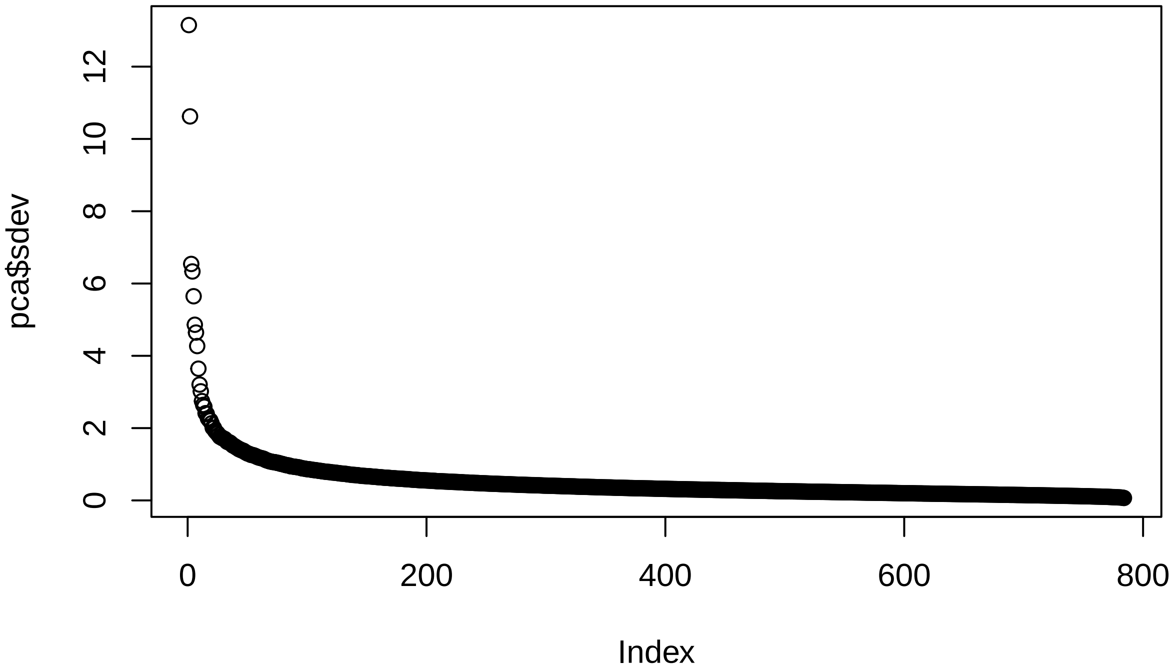
```
## Ytest
##    0    1    2    3    4    5    6    7    8    9
## 1000 1000 1000 1000 1000 1000 1000 1000 1000 1000
```

Pre processing

Before using PCA we have scaled and centered the data.

1.PCA

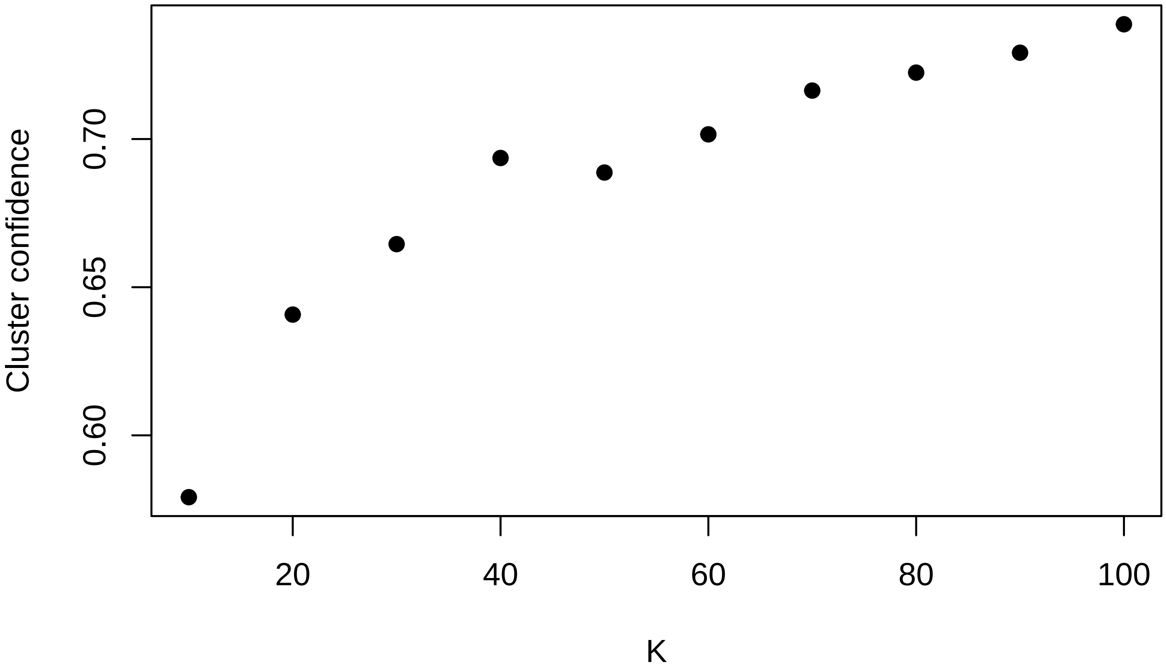
We perform PCA on the training data set and plot the standard deviation of each of the components in descending order. Using the elbow method the cutoff we choose is 1.75 and take all the components which have standard deviation above that value. It gives 28 components.



# Clustering

Kmeans

To first determine which K to use we run the kmeans algorithm for k =10,20,. . . ,100 and measure for each cluster the ration # of votes the majority label got/ total number of elements in the clustes. We also sum over all the clusters by weighing the ratios according to the cluster size. The plot of this cluster confidence vs K is as follows.



From this plot we can see that as we increase K the confidence increases. We choose K=100 so that to get the best clustering while keeping computation time low.

The majority label in each cluster and the cluster confidence % are as follows

```
##    [1] 1 4 1 9 0 8 8 8 6 0 8 6 8 3 7 8 8 6 2 9 8 5 3 3 6 9 0 4 8 2 9 3 8 9 8 5 6
##   [38] 6 9 3 0 2 0 7 8 9 4 3 3 8 2 4 5 8 8 8 2 8 0 0 4 1 8 1 4 5 7 8 0 9 4 2 4 0
##   [75] 0 2 9 2 4 6 8 4 7 5 3 6 9 5 1 1 2 8 4 7 9 8 6 5 4 9
```

```
##          [,1]
## [1,] 73.58667
```

References

Meshkini, Khatereh, Jan Platos, and Hassan Ghassemain. 2019. “An Analysis of Convolutional Neural Network for Fashion Images Classification (Fashion-MNIST).” In *International Conference on Intelligent Information Technologies for Industry*, 85–95. Springer.

Xiao, Han, Kashif Rasul, and Roland Vollgraf. 2017. “Fashion-MNIST: A Novel Image Dataset for Benchmarking Machine Learning Algorithms.” August 28, 2017. https://arxiv.org/abs/cs.LG/1708.07747.