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.

%%one more paper to be added%%

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

```
X.train.sca = scale(Xtrain)
```

Clustering

1.KMeans

```
library(plyr)
set.seed(1)

k=20
kmeans = kmeans(Xtrain, centers =k)

cluster_prediction = rep(NA,k)
#cluster_accuracy = matrix(rep(NA,2*k), nrow = k)
cluster_accuracy = matrix(nrow = k,ncol = 2)
clusters = split(Ytrain,kmeans$cluster)
for (i in 1:k) {
  x = as.data.frame(unlist(clusters[i]))
  y = count(x)
  max = which.max(y[,2])
  cluster_prediction[i] = y[max,1]
  s = sum(y[,2])
  cluster_accuracy[i,1] = y[max,2]/s
  cluster_accuracy[i,2] = s
}

weighted_cluster_accuracy =100* (t(cluster_accuracy[,1]) %*% cluster_accuracy[,2])/sum(cluster_accuracy
```

```
cluster_prediction
```

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

```
weighted_cluster_accuracy
```

```
##          [,1]
## [1,] 66.45667
```

2.PCA

```
pca = princomp(X.train.sca)
```

```
X.pca = pca$scores[,1:25]
k=30
kmeans = kmeans(X.pca, centers =k)

cluster_prediction = rep(NA,k)
#cluster_accuracy = matrix(rep(NA,2*k), nrow = k)
cluster_accuracy = matrix(nrow = k,ncol = 2)
clusters = split(Ytrain,kmeans$cluster)
for (i in 1:k) {
  x = as.data.frame(unlist(clusters[i]))
  y = count(x)
  max = which.max(y[,2])
  cluster_prediction[i] = y[max,1]
  s = sum(y[,2])
  cluster_accuracy[i,1] = y[max,2]/s
  cluster_accuracy[i,2] = s
}

weighted_cluster_accuracy =100* (t(cluster_accuracy[,1]) %*% cluster_accuracy[,2])/sum(cluster_accuracy
cluster_prediction
```

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

```
weighted_cluster_accuracy
```

```
##          [,1]
## [1,] 66.00333
```

References

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.