Unsupervised and Supervised Methods for Image Classification

Leon Wu

Imperial College London, Department of Mathematics

Introduction

We begin with the dataset, consisting of 70000 images of 10 different kinds of clothing items.

Figure 1:Fashion MNist sample data

K-Means Clustering

- The K-Means Clustering algorithm is an unsupervised method for clustering data.
- The following silhouette gives us a way to visualize the size of density of the clusters.
- For k=10 the silhouette plot show reasonably well-balanced sizes of clusters, with not too much overlap.

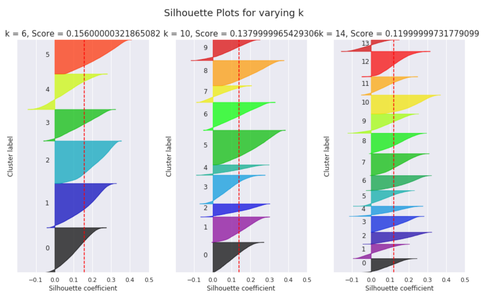

Figure 2:Silhouette Plot of K-Means Clustering

PCA with K-Means

- Principle Component Analysis lets us map the images onto a 2D plane.
- Using this, we can easily cluster and visualise the clustering performed on the reduced data.

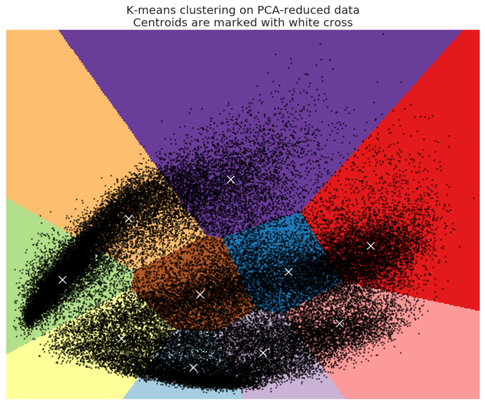

Figure 3:K-Means clustering on reduced data

- The K-Means clustering can also be fit on the full data, and then only visualized using PCA.
- In this way, we can visualize the actual predictions from our clustering by colouring the points, without losing information from the PCA transformation.

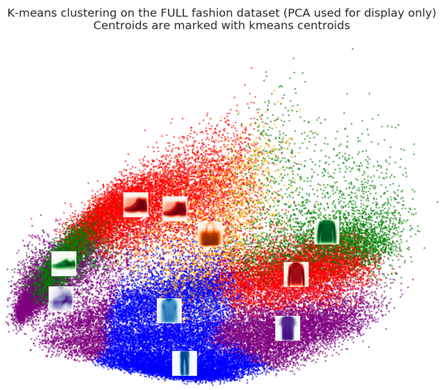

Figure 4:K-Means clustering on full data

Centroids

- The 'centroids' can also be interpreted as images - there are the centers of the K-Means clusters.
- The resulting centroids can be interpreted as the generalised images for each class that the K-Means algorithm worked out.
- The labels mostly bear a good resemblance to the actual 10 classes, which is impressive since this method is unsupervised.

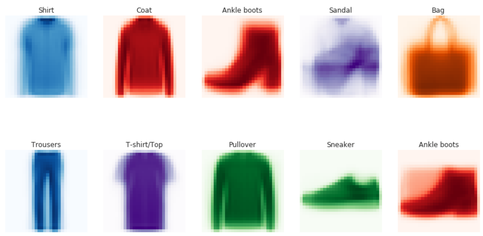

Figure 5:Centroids of K-Means Clustering

kNN Classification

- Using the K-Means clustering, we can fit a classifier to predict our samples
- The confusion matrix helps to interpret the results of this classifier; high scores on the diagonals mean that the classifier is correctly labelling the images

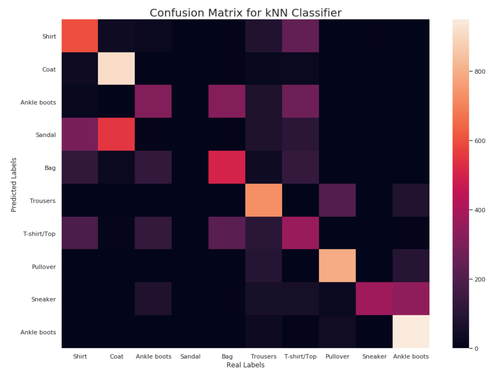

Figure 6:kNN confusion matrix

Supervised Classification

- So far, all the methods discussed have been 'unsupervised', meaning that the methods don't 'train' on the real labels.
- To train a classifier that performs better than our kNN classifier from the KMeans clusters, we can use supervised learning methods.

Convolutional Neural Networks

- Perhaps the most widely used model for image classification, the CNN is the preferred model for many reasons
- Most important is the fact that a CNN model can preserve spatial correlations between pixels, which is desirable in order to maximise the information gained from your data.
- The confusion matrix for a CNN classifier shows that this is much more effective than the kNN classifier.

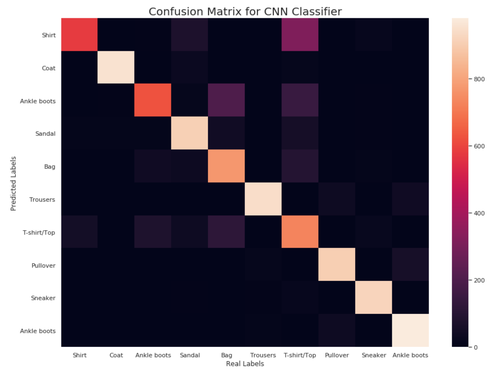

Figure 7:CNN confusion matrix

Conclusion

- Overall, unsupervised learning for image classification is a powerful way to quickly and easily cluster data in a way that is interpretable and useful for data exploration, whereas supervised learning is better for training a more accurate model for predictions.