IT UNIVERSITY OF COPENHAGEN

Zalando Clothing Classificaton

December 13, 2022

Authors:

Kristian Graven Hansen (krgh@itu.dk)

Lukas Sarka (lsar@itu.dk)

Mikkel Geisler (mgei@itu.dk)

1 Introduction

We are doing this cuz of that... We are doing this cuz of that... We are doing this cuz of that... We are doing this cuz of that... We are doing this cuz of that... We are doing this cuz of that...

2 Data and Preprocessing

2.1 Data-set

The Fashion-MNIST is a data-set of Zalando's article images—consisting of a training set of 60,000 samples and 10,000 test samples. The data-set used here is a small portion of the original data-set, 10,000 training and 5,000 test samples. Each sample is a 28x28 pixels gray-scale image with a label indicating the type of clothing item associated with each.

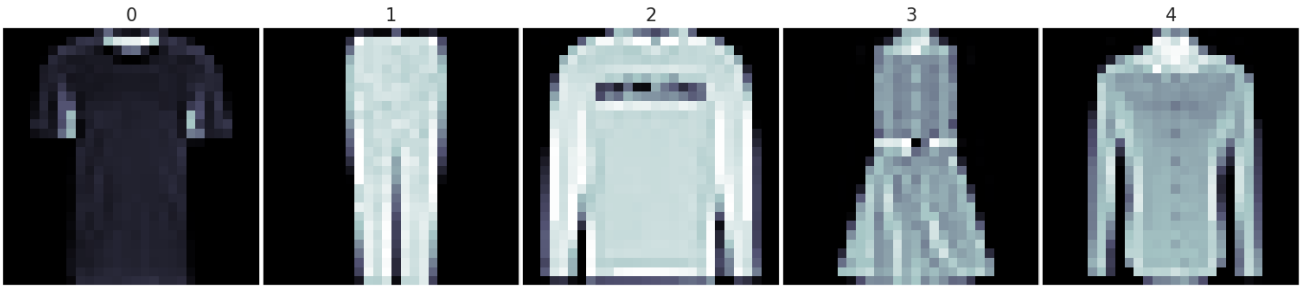

Figure 1: One sample from each class (reconstructing images from pixels)

2.2 Naming Conventions

The exploration of samples within each class gave rise to more appropriate names. Class 0 became T-shirt etc. Below is our naming conventions, which we will use throughout the report for better readability.

0	**1**	**2**	**3**	**4**
T-shirt	Pants	Sweatshirt	Dress	Shirt

Table 1: Mapping from class-labels to clothing type

2.3 Data Cleaning

The data-set provided was already in a cleaned state, which was verified by checking for missing values, and checking that the pixel values were no greater than 255 and no smaller than 0.

2.4 Preprocessing

The pixels were in the range $[0, 255]$. This range was normalized to $[0, 1]$ by dividing each pixel by 255. This was done primarily to improve training time.

2.5 Class Distribution

Whether or not a machine learning model can learn to predict classes well depends to a high degree on how those classes are distributed within our training and training data-set. Both our data-sets are extremely balanced, with the test being fully balanced (1000 of each class). This is illustrated on the plots below

Figure 2: Distribution of clothing items in our training and test data-set

3 Exploratory Data Analysis

One simple yet useful way to explore a data-set is to visualize the feature distribution for each class. In our situation this is simply not possible and not very insightful as we have too many features. To be precise $728 = 28 \times 28$ features/pixels. This doesn't mean the data-set can't be visualized, which we will discuss in the next section.

3.1 Principal Component Analysis

As mentioned before due to our large data-set it is hard to visualize feature distributions, this is where principal component analysis or PCA comes to the rescue. Principal Component Analysis (PCA) is a dimensionality reduction technique used to reduce the number of features in a dataset, while still preserving the most important information. It does this by transforming variables into a new set of variables, called principal components, which are uncorrelated from each other and explain the maximum amount of variance in the data. Below is plotted PCA1 vs PCA2 which to put it simply, represent the two directions of most variance in the data

Todo discuss what the pca plot means

Figure 3: Scatter-Plot of the two first principal components of training data