# Foundations and Frontiers of Machine Learning Group Assignment 2 Group 1

Contents

[Foundations and Frontiers of Machine Learning Group Assignment 2 Group 1 1](#_Toc179642033)

[1. Table of Contributions 1](#_Toc179642034)

[2. Data visualization (Task 1) 1](#_Toc179642035)

[3. Perceptrons (Task 2) 2](#_Toc179642036)

[4. Multilayer Perceptrons (Task 3) 2](#_Toc179642037)

[5. CNN (Task 4) 2](#_Toc179642038)

[6. Visualising CNN Outcomes (Task 5) 2](#_Toc179642039)

[7. Multitask learning – Fashion MNIST (Task 6) 2](#_Toc179642040)

## Table of Contributions

1. Data visualization - Emmanouil
2. Perceptrons – Emmanouil
3. Multilayer Perceptrons – Emmanouil
4. CNN
5. Vicualising CNN Outcomes

## Data visualization (Task 1)

Machine learning models use data to finetune the values of their internal parameters in a process called training. The power of the classification model is closely related to the amount of relevant information contained in the dataset, thus bigger datasets are usually preferred. However, not all information within the dataset is useful. Oftentimes, a feature contains information that has already been made available by another feature or a combination thereof, and as number of features within a dataset grows larger, the chances that a feature consists of the linear combination of other features of the dataset increase. At the same time, the mediums of visualising these datasets are usually books or electronic devices that are best suited for up to two features, as they physically have two dimensions. Using techniques such as video or colour/shape encoding can increase that by some amount, but datasets can have dimensions in the order of millions, orders of magnitude higher than what humans can visualise or comprehend (Sarkar, 2018).

A diagram of a colorful circle

Description automatically generated with medium confidenceThe issue of repeated information within the dataset can be addressed by constructing new features that encompass the dataset’s information without repetition, using with the Principal Components Analysis (PCA) technique. The first new feature is engineered to explain as much variance as possible from all features of the dataset (Maćkiewicz, 1993). Each next one is engineered to contain as much of the information that is left, without incorporating any information of the previously engineered features. This often means that the last features contain little to no additional information (depending on how linearly independent the dataset was) and can be discarded, allowing for easier visualization and reducing the overall noise. What is more, the features are now in descending order in terms of variance explained so we can select the first n ones, knowing that we have retain maximum information per feature. We can now select a number that is suitable for visualization. The new features have no actual meaning but can be used to show the distribution of the samples in the space, informing us about the separability of the classes.

Figure 1. Plot of 2 first components of PCA for all digits of the MNIST dataset

If we plot the first two components with different colour for each class, we can see which classes overlap, and which ones are linearly separable. Thus we have that, for example, the pairs 0 and 1, 0 and 2, 2 and 9 are some of the most easily to differentiate since the occupy opposing parts on the plane. Still, due to the significant overlap caused by 10 different classes, the distribution of some of the classes is obscured.

A screenshot of a computer game

Description automatically generatedA screenshot of a computer game

Description automatically generated

Figure 2. Plot of the first 2 PCA Components for all pairs of digits

Figure 3. Pair-wise PCA plot

Figure 2 provides a pairwise visualisation of the distributions. It Is clear that some pairs of classes, e.g. (0, 1), (0, 7), (3, 7) are relatively distinct, while others (5, 8), (7, 9) are virtually impossible to separate. However it must be kept in mind that this is the result of considering just 2 components and not all of the information contained in the dataset. That means distributions that do not seem differentiable now, may actually be when the model is trained on the original dataset. What is more, the above results are obtained by performing PCA on the whole dataset. However, when it comes to distinguishing classes in pairs, the datapoints of other classes introduce variance and noise that may be irrelevant. Thus, it is worth producing the same plot with the PCA calculated for each pair. This is shown in figure 3 and it becomes apparent that the results are better in terms of separability. To objectively measure this, we use the silhouette score which takes into account the average distance of a clusters samples between each-other and between the samples of the other class. Higher scores mean better clustering and for the generic PCA we get 0.337 where as the same score for the pairwise PCA is 0.407. This proves that we can achieve better clustering (and hence prediction performance) by performing pairwise PCA.

### Math of PCA

Before applying and transformations on the dataset, it is important to normalize it so the features with larger ranges do not affect the calculations disproportionately.

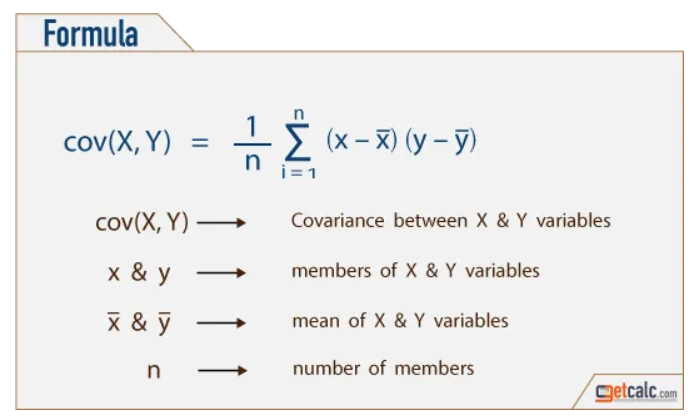
The next step is to identify how much each feature correlates with the other ones, as the more they correlate, the less useful information they add. Figure 4 shows the steps of calculation, which essentially shows that the more two variables follow the same trend, then larger the product of their difference to the mean will be.

Figure 4. Calculation of Covariance Matrix (Dubey, 2018)

From Linear Algebra we know that a matrix can represent a transformation and its eigenvectors represent the directions in which the data has the most variance. Thus these eigenvectors will form the basis of the new vector space which is now built so that each new dimensions contains the most variance

The actual amount of variance explained by each dimension is given away by the eigenvalues of the eigenvectors. We can create a new matrix by concatenating the desired number of eigenvectors which, when multiplied with our data, produces the PCA-Data, which is our data transformed to the new base.

## Perceptrons (Task 2)

Asdfasdf

## Multilayer Perceptrons (Task 3)

asdfasdfasdf

## CNN (Task 4)

asdfasdfasdfasdfasd

## Visualising CNN Outcomes (Task 5)

asdfasdfasdfasdfasdfasdfasdf

## Multitask learning – Fashion MNIST (Task 6)

Asdfasdfasdfasdfasdfqwerqwerqwer

References

<https://towardsdatascience.com/the-art-of-effective-visualization-of-multi-dimensional-data-6c7202990c57#:~:text=Visualizing%20data%20in%20Three%20Dimensions,values%20in%20a%20categorical%20dimension>.

https://www.sciencedirect.com/science/article/pii/009830049390090R

https://towardsdatascience.com/the-mathematics-behind-principal-component-analysis-fff2d7f4b643