# CE888 Assignment 1

Ogulcan Ozer

***Abstract*—The traditional unsupervised machine learning algorithms like k-Means, PCA and other linear factor models are proven to be very useful for many important applications. But in this day and age we are able to collect very complex, high dimensional data and these simple but efficient algorithms are failing to learn or represent it [1]. In this work we are going to use auto encoders to reduce the dimensions of three different datasets and try to extract good features using different activation functions. Then finally, we will cluster the data by using the extracted features and compare the results with the traditional techniques.**

## I. Introduction

UNSUPERVISED learning is a branch of machine learning that deals with unlabeled data to find patterns, groups and relations. This sub section of machine learning is useful to describe and understand unknown data, which also helps us understand the properties of future data samples. Right now we are collecting data much faster than we can process it. And most of the time the data we acquire is unlabeled, and these huge amounts of data are complex, noisy and it might include unnecessary information.

There are different ways to reduce the complexity of a given data and learn good features from it. In this work, we will look at auto encoders, one of those procedures used to learn features from a given data. Auto encoders are a part of representational learning, which can be trained with or without labels. They play a key role in deep learning applications. They are used in dimensionality reduction and feature extraction, and together with deep learning they gained a lot of traction after Hinton proposed a method to train deep belief networks one layer at a time [2].Another addition to the auto encoder will be the softmax activation function. Since softmax gives us the probability distribution over-the classes of our data- the output vector [3], we can use the output of our encoder to train a clustering algorithm. The motivation for using auto encoders is, in what they do they are very similar to traditional feature selection algorithms, but they are also capable of learning good features from complex and high dimensional data. This aspect of auto encoders and unsupervised learning in general is important. As LeCun states in [3], people are not paying attention to unsupervised learning because of the success of the supervised learning in recent years. But it will be far more important in the future because of how living beings perceive and learn, they do not need labels to recognize the world around them.

In this paper we are going to look at previous work (Section II), look at different methods and present their results to show how they compare. In Section III, we are going to present our goals and what we want to accomplish. Then we will talk about the datasets that we are going to use for the task and give detailed information about them. The planned experiments and analyses for the task will be mentioned in Section IV. Evaluation methods that are going to be used after the experiments will be presented in Section V.

## II. Background

## III. Methodology

Our main goal in this study is to correctly cluster data using unsupervised machine learning techniques. To accomplish this task we will go through three different subtasks. First, we will use traditional clustering algorithms to see how they perform on our datasets. In the second part, we will use auto encoders to learn good features from our datasets, then we will pass the-extracted features- outputs of the auto encoder to a clustering algorithm to see the difference in their performance. Finally, we will change the output activation function of the auto encoder to the softmax activation function, which will give us the probability distribution of the output vector. Then we will use the maximum weighted class as our cluster assignment.

Before starting the experiments the datasets will be examined and processed appropriately for each sub-task. For the first sub task, depending on the complexity and dimensions of the given data, a suitable clustering algorithm will be chosen. For the second sub-task, features of the datasets will be standardized. As LeCun concludes in [3], networks can learn better and faster if the inputs are centered around zero with equalized covariance. Another consideration for sub-tasks two and three is using ensemble of auto encoders. Since the optimization of the randomly initialized weights is difficult [4], we can have multiple auto encoders in our model to mitigate the effects of random weight initialization. For the last sub-task, since we will be using softmax as our activation function, the targets of our datasets will be converted to one hot coded versions.

The three datasets chosen to be clustered in this study are Modified NIST, Human Activity Recognition Using Smartphones (HAR) and High Time Resolution Universe Survey 2 (HTRU2).

MNIST is a image dataset of hand written digits. It is a modified version of the NIST dataset collected from Census Bureau employees and high-school students and processed by LeCun et.al..The original NIST set was normalized, so that each sample would be centered according to the weights of it's pixels in a 28x28 box. Images were also converted to grey-scale and shuffled to make it easy to use in machine learning applications [5]. The dataset consists of 60.000 training images and 10.000 test images. Since each sample is represented as a 28x28 pixel binary image, it can be said that samples have 784 features. MNIST dataset is frequently used in machine learning research[–], which makes it easier for researchers to compare their models and solutions to the others in the field.

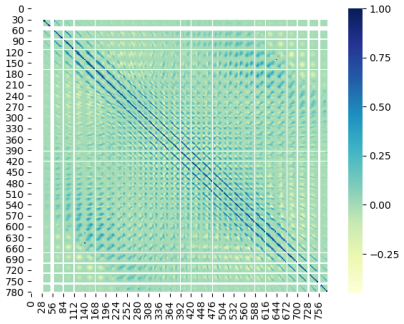

Fig. 1. Corraletion heatmap of MNIST features.

HAR is a dataset of multiple classes of human activities. It was created by recording different classes of activities of 30 volunteers by using the accelerometer and the gyroscope in a smartphone. These sensors were used to record the linear acceleration and angular velocity on three axes at 50 Hz. Volunteers were also recorded by a camera for the labeling of the activities(walking, walking upstairs, walking downstairs, sitting, standing, laying)[7]. The dataset consists of 7352 training samples and 2947 test samples. Each sample has 561 features which are- extracted- estimated from the recorded accelerometer and gyroscope signals.These features are normalized, scaled and shifted to be within [-1,1]. Subject id labels for the features and mentioned raw accelerometer and gyroscope signals are also provided with the dataset. Dimensions of the subject ids are the same with their training and testing parts. On the other hand, the raw signals are seperated into nine different files grouped under three axes, which are body acceleration (X,Y,Z), body gyroscope (X,Y,Z) and total acceleratin (X,Y,Z).

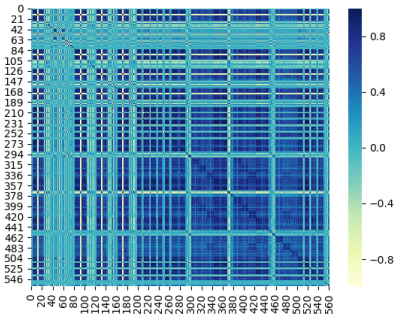

Fig. 2. Corraletion heatmap of HAR features.

Our last dataset HTRU2 was obtained by Bates[] using ANNs to first filter out the bad candidates,then manually inspecting the outputs of the ANN.Each sample of HTRU2 is a set of features extracted from the data acquired by large radio telescopes during High Time Resolution Universe Survey. HTRU2 has 8 features and consists of 17,898 total samples. Only 1,639 of them are positive pulsar samples and the 16,259 of them are negative samples. The first four features are simple statistics of the pulsar profiles and rest of the features were obtained from DM-SNR curve [].These compact features were obtained from 90.000 labeled pulsar candidates by using Pulsar Feature Lab [].

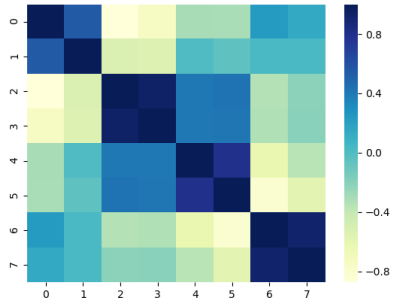

Fig. 3. Corraletion heatmap of htru features.

## IV. Experiments

In the experiments the implementation of the tasks and data analysis will be done using python. In the imlementation of the traditional clustering algorithms, sckit-learn library will be used because of its easy to use nature and rich functions. For the rest of the tasks, tensorflow and keras libraries will be used to implement auto encoders, Since tensorflow converts the python code into C code in runtime, it is fast and it can run on GPUs, making it one of the best neural network library. Finally, the data analysis and plotting will be done using matplotlib and seaborn libraries.

### A. Clustering Algorithms

In the core part of the clustering experiments, k-Means and DBSCAN algorithms will be used. These algorithms will be first compared to each other, then their results will be analysed, evaluated and recorded for future comparisons. If there is enough time to deviate from the core path,agglomerative clustering algorithm will also be used and these clustering algorithms will be tested again after applying dimensionality reduction methods like PCA/kernel PCA, LDA to the datasets. Otherwise, the experiments will continue with the auto encoders.

## V. Conclusion

The conclusion goes here.

## Appendix A
## Proof of the First Zonklar Equation

Appendix one text goes here.

## Appendix B

Appendix two text goes here.

## Acknowledgment

The authors would like to thank...