Introduction to Data Science Tools & Techniques

# Assignment – 2

## Ali Nauman – 18L1863

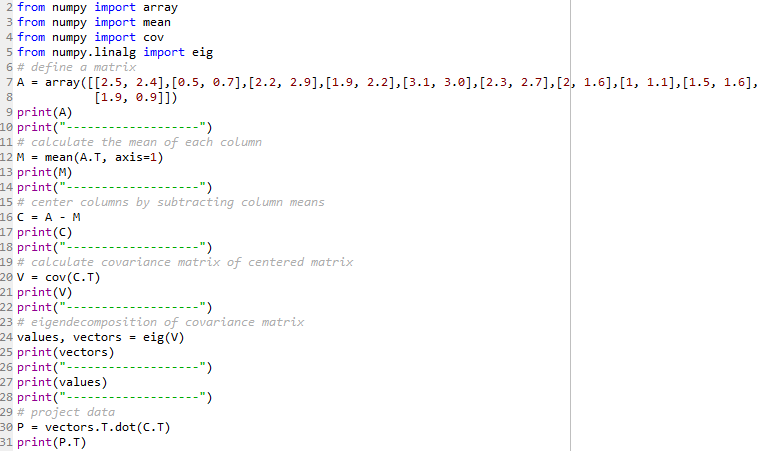
## Question No. 1

## Principal Component Analysis

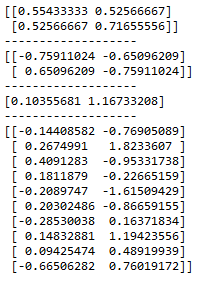
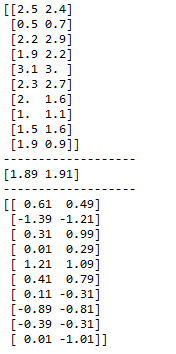
Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. This transformation is defined in such a way that the first principal component has the largest possible variance, and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components.

PCA is mostly used as a tool in exploratory data analysis and for making predictive models. It is the simplest of the true eigenvector-based multivariate analyses.

## Implementation of PCA using Python



### Results

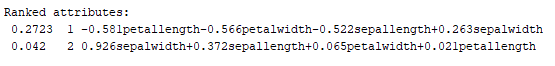


## Implementation of PCA using WEKA

### Iris Plants Data Set

This data set is probably one of the best known database to be found in the pattern recognition literature. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other. It has a total of 4 numeric, predictive attributes and class. The class distribution is at 33.3% for each of the 3 classes.

### Result & Analysis

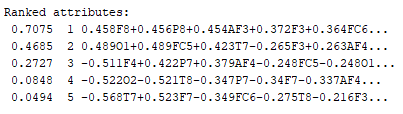


This shows the ranked attributes as the original data is projected onto a much smaller space. In this particular scenario it has performed *Feature Elimination*. It has reduced the feature space by eliminating features. By eliminating features, we’ve also entirely eliminated any benefits those dropped variables would bring.

EEG Eye State Data Set

All the data in this data set has been extracted from one continuous EEG measurement with the *Emotiv EEG Neuroheadset*. The duration of the measurement was 117 seconds. The eye state was detected via a camera during the EEG measurement and added later manually to the file after analyzing the video frames. '1' indicates the eye-closed and '0' the eye-open state. All values are in chronological order with the first measured value at the top of the data. This data set has a total of 15 Attributes and a total of 14980 instances for the algorithms to be implemented.

### Result & Analysis

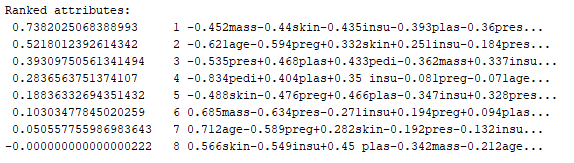


This shows the ranked attributes as the original data is projected onto a much smaller space. In this particular scenario it has performed Feature Elimination. It has reduced the feature space by eliminating features.

### Pima Indians Diabetes Data Set

This data set was collected by using the ADAP learning algorithm to forecast the onset of diabetes mellitus. The diagnostic, binary-valued variable investigated is whether the patient shows signs of diabetes according to World Health Organization criteria. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. ADAP is an adaptive learning routine that generates and executes digital analogs of perceptron-like devices. This data set includes 8 plus class attributes, and a total of 768 instances.

### Result & Analysis

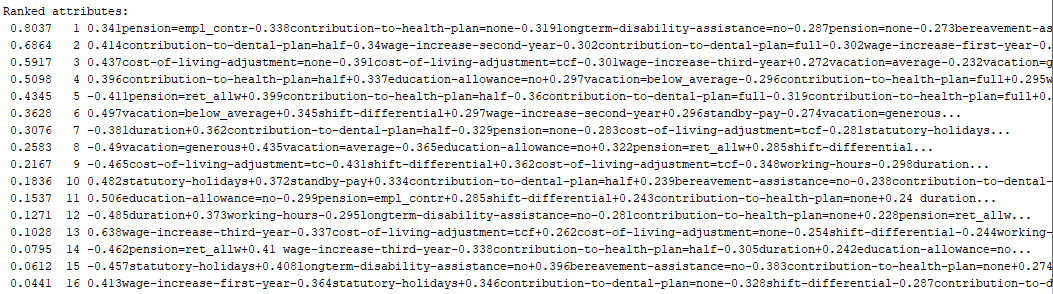


This shows the ranked attributes as the original data is projected onto a much smaller space. In this particular scenario it has performed Feature Elimination. It has reduced the feature space by eliminating features.

### Final settlements in labor negotiations in Canadian industry Data Set

The data includes all collective agreements reached in the business and personal services sector for locals with at least 500 members (teachers, nurses, university staff, police, etc.) in Canada in 87 and first quarter of 88. The data was used to learn the description of an acceptable and unacceptable contract. Data was used to test 2tier approach with learning from positive and negative examples. It has a total of 16 attributes and 57 instances.

### Result & Analysis



This shows the ranked attributes as the original data is projected onto a much smaller space. In this particular scenario it has performed Feature Elimination. It has reduced the feature space by eliminating features.

### Tamilnadu Electricity Board Hourly Readings Data Set

This data set is collected from the real time readings from residential, commercial, industrial, agriculture sectors to find the accuracy consumption in Tamil Nadu around Thanajvur. The data set has a total of 5 attributes and 45781 instances of data.

### Result & Analysis



In this scenario, *Feature Extraction* has been performed. Principal component analysis is a technique for feature extraction. It combines our input variables in a specific way, then we can drop the “least important” variables while still retaining the most valuable parts of all of the variables. We keep as many of the new independent variables as we want, but we drop the “least important ones.” Because we ordered the new variables by how well they predict our dependent variable, we know which variable is the most important and least important. These new independent variables are combinations of our old ones, we’re still keeping the most valuable parts of our old variables, even when we drop one or more of these “new” variables.

### Absenteeism at work Data Set

The database was created with records of absenteeism at work from July 2007 to July 2010 at a courier company in Brazil. This data set is a collection of employee absenteeism data. The data set allows for several new combinations of attributes and attribute exclusions, or the modification of the attribute type (categorical, integer, or real) depending on the purpose of the research. The data set (Absenteeism at work - Part I) was used in academic research at the Universidade Nove de Julho - Postgraduate Program in Informatics and Knowledge Management. It has a total of 21 attributes and 740 instances.

### Result & Analysis



It had a total of 21 attributes. It ranked a total of 63 attributes, creating new ones from the old ones, alongside removing the least useful ones. In this scenario, *Feature Extraction* has been performed. Principal component analysis is a technique for feature extraction. It combines our input variables in a specific way, then we can drop the “least important” variables while still retaining the most valuable parts of all of the variables. We keep as many of the new independent variables as we want, but we drop the “least important ones.” Because we ordered the new variables by how well they predict our dependent variable, we know which variable is the most important and least important. These new independent variables are combinations of our old ones, we’re still keeping the most valuable parts of our old variables, even when we drop one or more of these “new” variables.

# A Method to Facilitate Cancer Detection and Type Classification from Gene Expression Data using a Deep Autoencoder and Neural Network

## Critique

Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product. Due to increased affordability and availability of the gene expression sequence data, it can be utilized to characterize diseases such as cancer. It is the process by which the heritable information in a gene, the sequence of DNA base pairs, is made into a functional gene product, such as protein or RNA. The basic idea is that DNA is transcribed into RNA, which is then translated into proteins.

The issue that has been worked on in this paper is to present a models of deep learning (DL) and apply them to gene expression data for the diagnosis and categorization of cancer. This is a task which is difficult to present as establishing specificity in cancer diagnosis using gene expression data continues to pose challenges due to the high dimensionality and complexity of the data. The authors in this paper have developed two DL models using messenger ribonucleic acid (mRNA) datasets available from the *Genomic Data Commons* repository.

The authors have thoroughly gone through the basic idea of the issue in hand. They set the narrative of their work by how gene expression profiling technology has led to many significant biological discoveries. All the current solutions for disease detection, especially cancer detection methods, still rely on traditional feature selection and dimension reduction approaches. And further solidifying the claim that the models in hand cannot take full advantage of data from different cancer types due to the small sample sizes available for individual cancers.

To deal with these problems and exceed the limits of the dimension-reduction paradigm in order to develop a more generalized model of cancer detection and typing classification, has been proposed in this paper. The authors have worked on deep learning approach without the pains of feature space reduction/generation and that avoids the loss of potentially valuable information. And this allows their method to work on typing classification of multiple cancers.

They first introduced the computational configuration, then presented the deep learning algorithms used to explain a few keep DL techniques. These techniques are used to train model’s convolution and deep Autoencoder model. Feature selection is done by learning the feature representation through coupling encoder layers symmetrically with decoder layers. They have utilized 5 hidden layers with two encoder layers and thus, two decoding layers. They extracted the encoder layers and transformed them to a fully connected neural network with multi-classification as its output. In order to perform the cancer detection and type classification, they froze the encoder layers learned from the deep Autoencoder step and transformed them to a fully connected neural network.

For purposes of comparison, they also employed other machine and statistical learning techniques to analyze the gene expression data. The resulting extracted features are simply a linear function of the original input data, which lost all non-linearity of the relations between expressions of different genes. They have utilized top 40 principle components from the dimension reduction. They used the data on 5 machine and statistical learning models: linear discriminant analysis (LDA), neural network, K-nearest neighbor, random forest, and extremely randomized tree.

There models achieved 98% accuracy in cancer detection, with false negative and false positive rates below 1.7%. The results, demonstrated that 18 out of 32 cancer-typing classifications achieved more than 90% accuracy. Due to the limitation of a small sample size (less than 50 observations), certain cancers could not achieve a higher accuracy in typing classification, but still achieved high accuracy for the cancer detection task.

The facts covered in this paper are of the highest quality. The authors have diligently explained in detail the issue at hand and how they have worked on to cater the problem. The results show that for cancer detection, a single-classification task, the proposed model can achieve higher accuracy and lower false positive and negative rate than traditional algorithms. Finally, for cancer type classification, a multi-classification task, it performs very well when sufficient sample data are available to train the model.