# Applying Principal Component Analysis on Breast Cancer Wisconsin dataset

Shashank Pathak(15BCE1287)*,
*School of Computing Sciences and Engineering, VIT Chennai, Tamilnadu, India 600127
Email: shashank.pathak2015@vit.ac.in*

***Abstract*—Principal Component Analysis is used to find strong patterns in the dataset and for visualization of data.It also helps in finding the principal and important attributes of the dataset.visualize. .We perform the PCA on the breast cancer dataset and get the principal components with the help of eigen values**

## 1. Introduction

Principal Component Analysis is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. Since patterns in data can be hard to nd in data of high dimension, where the luxury of graphical representation is not available, PCA is a powerful tool for analysing data. The other main advantage of PCA is that once you have found these patterns in the data, and you reduce the number of dimensions, without much loss of information.PCA is used to rank features in terms of their variances.

## 2. Methodology

Technically, a principal component can be defined as a linear combination of optimally-weighted observed variables. The output of PCA are these principal components, the number of which is less than or equal to the number of original variables. Less, in case when we wish to discard or reduce the dimensions in our dataset. The PCs possess some useful properties like the PCs are essentially the linear combinations of the original variables, the weights vector in this combination is actually the eigenvector found which in turn satisfies the principle of least squares.The PCs are orthogonal.The variation present in the PCs decrease as we move from the 1st PC to the last one, hence the importance.The least important PCs are also sometimes useful in regression, outlier detection.

## 3. Database - Breast Cancer Wisconsin Database(Original)

The Wisconsin Breast Cancer datasets from the UCI Machine Learning Repository is used,to distinguish malignant (cancerous) from benign (non-cancerous) samples.This dataset consist of 699 instances and 11 attributes that would help the classiy the data into the two classes.The attributes are described in the fig.1

| # | Attribute | Domain |
|---|---|---|
| 1. | Sample code number | id number |
| 2. | Clump Thickness | 1 - 10 |
| 3. | Uniformity of Cell Size | 1 - 10 |
| 4. | Uniformity of Cell Shape | 1 - 10 |
| 5. | Marginal Adhesion | 1 - 10 |
| 6. | Single Epithelial Cell Size | 1 - 10 |
| 7. | Bare Nuclei | 1 - 10 |
| 8. | Bland Chromatin | 1 - 10 |
| 9. | Normal Nucleoli | 1 - 10 |
| 10. | Mitoses | 1 - 10 |
| 11. | Class: | (2 for benign, 4 for malignant) |

Figure 1.

## 4. Algorithm

The algorith for PCA algorithm undergoes these steps to identify the variance and principal components.We take the whole dataset consisting of d-dimensional samples ignoring the class labels. First we will compute the d-dimensional mean vector (i.e., the means for every dimension of the whole dataset) then we will compute the scatter matrix (alternatively, the covariance matrix) of 3)the whole data set.After that we will compute eigenvectors (e1,e2,...,ede1,e2,...,ed) and corresponding eigenvalues (1,2,...,d1,2,...,d) then sort the eigenvectors by decreasing eigenvalues and choose k eigenvectors with the largest eigenvalues to form a dk dimensional matrix W(where every column represents an eigenvector). Use this dk eigenvector matrix to transform the samples onto the new subspace. This can be summarized by the mathematical equation: y=WTxy=WTx (where x is a d1-dimensional vector representing one sample, and y is the transformed k1-dimensional sample in the new subspace.)

## 5. Experiments

In this experiment we perform PCA on chosen dataset.PCA is mainly done in order to extract the important parameters and remove the redundant parameters which exist in the database.The covariance matrix is constructed and eigen values calculated which further give the principal

components. The components then obtained are used to classify the dataset by making use of logistic regression for providing the probability for the classes.So the chosen dataset was taken and all the steps were performed on them.

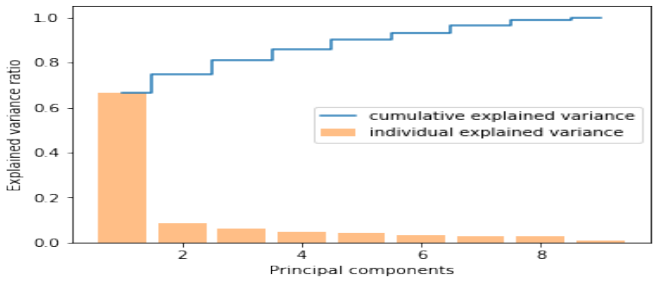


Figure 2.

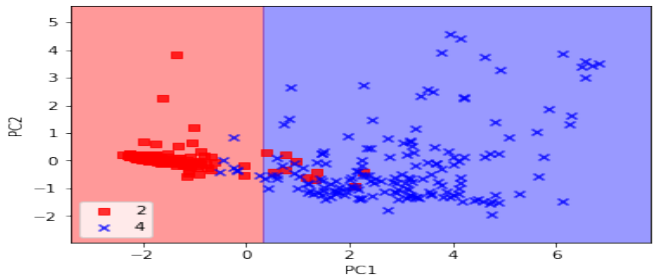


Figure 3. ”On Training Dataset”

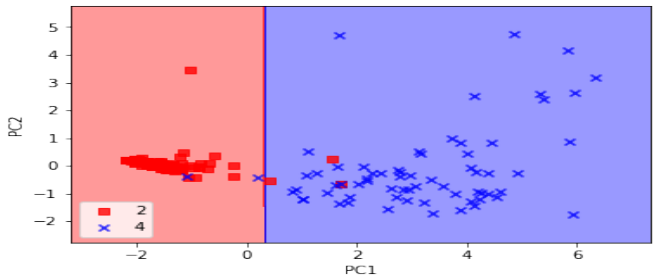


Figure 4. ”On Testing Dataset”

## 6. Conclusion

After applying PCA on the dataset we can say that PCA worked good on the dataset when compared to other algorithms and we were able to identify the important components of the dataset and it came out that the variance of the most important component was 0.66 and then the components were arranged in descending order.PCA works well because it extracts the important parameters instead of taking all the parameters which results in good accuracy.We can see that though PCA performs well when compared to other but still there are many misclassified points.So overall PCA performs well on the Breast Cancer dataset when compared to other algorithms,