Kernel PCA on classification of Breast Cancer Wisconsin (Diagnostic) Data Set

Anushree Gupta, Gautami Mudaliar, Rakesh Rathod, Padmasini Krishnan Venkat

I. Abstract

Dimensionality reduction is a data preparation technique performed on data prior to modeling. It might be performed after data cleaning and data scaling and before training a predictive model. Principal component analysis (PCA) is a popular tool for linear dimensionality reduction and feature extraction. Kernel PCA is a nonlinear variation of PCA that more effectively takes advantage of the intricate spatial organization of high-dimensional information. In this essay, we first go over the fundamental concepts of dimensionality reduction, why it is used and methods to achieve it. Furthermore, we talk about the concepts of PCA and Kernel PCA. The prediction of breast cancer from images of the Wisconsin Dataset using kernel PCA is thereafter our main project.

II. Introduction

In this section, we provide a quick overview of the Dimensionality Reduction procedure and a few approaches to achieving it.

A. Dimensionality Reduction

Advantages of using Dimensionality Reduction:

- Reduces calculation time: Converting an n-dimensional space to a smaller dimension space requires less computation, which lowers the temporal complexity.
- Reduces space complexity: Converting an n-dimensional space to a space of lower dimensions results in a reduction in the complexity of the space due to the reduction in the number of factors, features and parameters, affecting the distribution of data.
- Data visualization is simpler, since many of the features irrelevant/similar features are discarded thus, making it simpler to comprehend and apply the model.

Brief execution of Dimensionality Reduction

- Dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables. It can be divided into feature selection and feature extraction.
- By projecting the original data to a lower dimension sub-space via linear or non-linear transformations, high dimensional data is reduced to lower dimensions.

B. Types of Transformations

1) Linear Method: The original data is linearly projected onto a low-dimensional space when using linear methods. We'll talk about linear approaches PCA, FA, LDA, and Truncated SVD. These techniques only work effectively with linear data.

a) Principal Component Analysis: The principal component analysis (PCA) is a linear dimensionality reduction technique (algorithm) that converts a collection of correlated variables (p) into a smaller collection of uncorrelated variables (k) (kp), known as principal components, while preserving as much variation in the original dataset as is possible.

b) Factor Analysis: Factor analysis's main goal is not merely to make the data's dimensions smaller. Finding latent variables, which are inferred from other variables in the dataset rather than directly measured in a single variable, is a useful use of factor analysis. The term "factors" refers to these latent variables.

c) Linear Discriminant Analysis: The most common application of LDA is multi-class categorization. It can also be applied as a method of dimension reduction. The LDA algorithm separates or discriminates training instances by their classes the best (hence the name LDA). The LDA has various restrictions. The data must be regularly distributed in order to use LDA. Also included in the collection should be recognized class labels. The number of classes minus one is the maximum number of components that LDA may discover. In dimensionality reduction, LDA can only detect 2 (3-1) components if your dataset has just 3 class labels. Applying LDA does not need feature scaling.

d) Truncated Singular Value Decomposition: Through the use of truncated singular value decomposition, this technique reduces the number of linear dimensions (SVD). With sparse data, where many of the row values are zero, it performs well. Truncated SVD is much simpler to build with Scikit-learn. The TruncatedSVD() function can be used to do this.

2) Non Linear Method: The linear approaches presented thus far do not effectively reduce dimensionality when dealing with non-linear data, which are often employed in practical applications.

a) KPCA: A non-linear dimensionality reduction method that makes use of kernels is called kernel PCA. It may also be thought of as the non-linear variation of conventional PCA. When dealing with non-linear datasets, kernel PCA performs effectively where conventional PCA falls short.

b) t-distributed Stochastic Neighbor Embedding: The main application of this non-linear dimensionality reduction technique is data visualization. It is also extensively utilized in NLP and image processing. If there are more than 50 features in the dataset, the Scikit-learn documentation advises using PCA or Truncated SVD instead of t-SNE.

III. Principal Components Analysis

There are five steps that make up principal component analysis. I'll go over each stage, explaining PCA's actions logically while also demystifying complex mathematical ideas like standardization, covariance, eigenvectors, and eigenvalues without concentrating on how to calculate them.

$$C = \frac{1}{l}\sum_{j=1}^{l} x_j x_j^T$$

1) *Standardization:* In order for each continuous beginning variable to contribute equally to the analysis, this phase standardizes the range of the variables.

$$z = \frac{value - mean}{standard\ deviation}$$

2) *Covariance Matrix Computation:* The goal of this step is to determine the relationship—if any—between the variables in the input data set and how they differ from the mean in relation to one another. Because variables can occasionally be highly connected to the point where they include redundant data. We compute the covariance matrix in order to find these associations.

$$\begin{bmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{bmatrix}$$

The sign of covariances matters the most:

- If positive, then the two variables increase or decrease together (correlated)
- If negative, then one increases when the other decreases (Inversely correlated)

3) *Compute the Eigenvalues and Eigenvectors:* Eigenvectors and eigenvalues are the linear algebra concepts that we need to compute from the covariance matrix in order to determine the principal components of the data.

The new variables created as a result of the basic variables' linear combinations or mixtures are known as principal components. These combinations are made in a way that most of the information included in the original variables is condensed or squeezed into the first components, which are the new variables (i.e., principal components), which are uncorrelated. The premise is that 10-dimensional data provides you 10 principal components, but PCA seeks to fit as much information as possible into the first component, then as much information as is left in the second, and so on.

Principal components, or lines that encapsulate the majority of the information in the data, are the directions of the data that, geometrically speaking, explain the most variance. The relationship between variance and information in this case is that a line's dispersion of data points along it increases as variance increases, and a line's information content increases as dispersion increases. Simply expressed, consider of primary components as new axes that offer the greatest perspective for observing and assessing the data in order to make the contrasts between the observations more obvious.

4) *Feature Vector:* In this stage, we decide whether to keep all of these components or toss out any that have low eigenvalues and create a matrix of vectors that we refer to as the "Feature vector" using the ones that are left.

Therefore, the feature vector is just a matrix with the eigenvectors of the components that we choose to maintain as columns. This makes it the first step towards dimensionality reduction since the final data set will only have p dimensions if we decide to keep only p eigenvectors (components) out of n.

5) *Recast the data along the principal component axes:* Prior to standardization, you did not alter the data in any way; instead, you just chose the primary components and created the feature vector. However, the input data set remained constant in terms of the original axes (i.e, in terms of the initial variables).

The objective of this final step is to use the feature vector created using the eigenvectors of the covariance matrix to reorient the data from the original axis to those indicated by the principal components (hence the name Principal Components Analysis). The transpose of the original data set and the feature vector can be multiplied to achieve this.

$$FinalData = FV^T \times StandardizedOrigData^T$$

IV. KPCA

It's interesting how Kernel PCA operates intuitively. The classes become linearly separable after the data are temporarily projected into a new, higher-dimensional feature space using a kernel function (classes can be divided by drawing a straight line). The data is then projected back into a lower-dimensional space by the algorithm using the standard PCA. By converting non-linear data into a lower-dimensional space of data, Kernel PCA makes it possible to employ linear classifiers on non-linear data.
The number of components we wish to preserve, the kind of kernel, and the kernel coefficient are three crucial hyperparameters that must be specified in the Kernel PCA (also known as the gamma). We have the options of using a linear, poly, rbf(radial basis function), sigmoid, or cosine kernel.

$$C = \frac{1}{l}\sum_{j=1}^{l} \phi(x_j)\phi(x_j^T)$$

Algorithm for Kernel PCA:

- Clean the data so that it does not have any missing entries. If there are n samples and m features, the data matrix becomes an mxn matrix
- Find the gram matrix.
- The gram matrix is obtained by using the kernel trick. Kernel function is used to perform an operation (inner product) on data in higher

dimensional space and the resultant of that operation is the gram matrix
- Find the eigen values and eigen vectors of the gram matrix
- Find the normalized eigen vectors
- The eigen vectors associated with decreasing order of eigen values will give principal components.

Mercers Theorem: Suppose K is a continuous symmetric non-negative definite kernel. Then there is an orthonormal basis {ei}i of L2[a, b] consisting of eigenfunctions of TK such that the corresponding sequence of eigenvalues {λi}i is nonnegative. The eigenfunctions corresponding to non-zero eigenvalues are continuous on [a, b] and K has the representation

$$K(s,t) = \sum_{j=1}^{\infty} \lambda_j e_j(s) e_j(t)$$

V. FLOWCHART

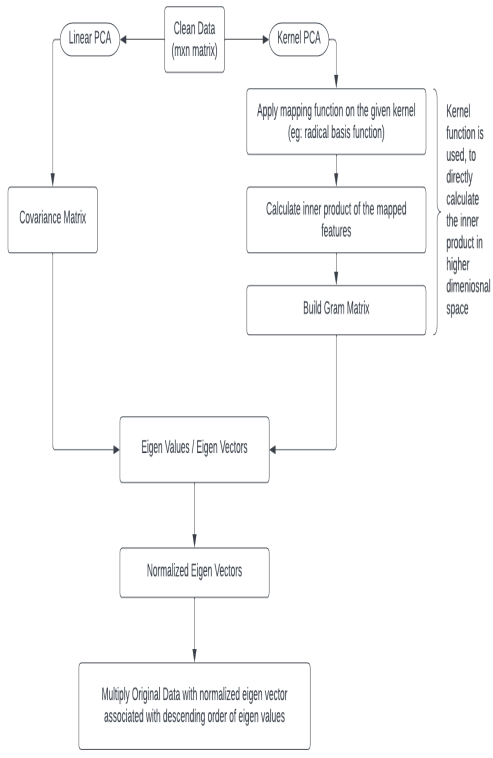

Figure 1: Flowchart

VI. DATASET

Wisconsin Breast Cancer Dataset is downloaded from the UCI repository. It is a cleaned dataset, as shown in the graph below. The absence of missing values in the dataset indicates the accuracy of the data. No incompatible values. There are 568 samples and 30 features in it. The features are calculated using a picture of breast mass that looks like a microneedle that has been digitally altered. The information describes the properties of the visible cell nuclei in the image. The forecast is made based on the diagnosis column [D], namely whether the cancer is benign or malignant.

A. PEARSON'S CORRELATION COEFFICIENT

The most popular approach to assess a linear connection is to use the Pearson correlation coefficient (r). The intensity and

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

direction of the link between two variables is expressed as a number between -1 and 1.

Pearson correlation coefficient (*r*) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

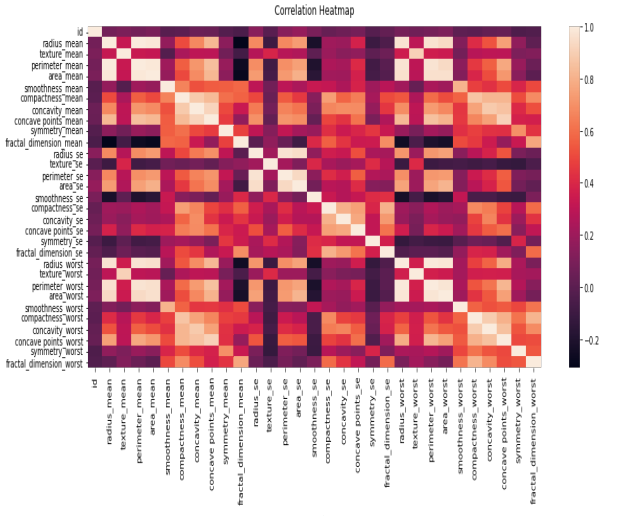

Figure 2: Correlation Matrix

The degree and direction of the linear relationship between two quantitative variables are specifically described. The stronger the correlation between the features, the greater the coefficient's value. Only data that is quantitative, linear, and free of outliers can be used. It is used to determine how closely two features are related to one another, and based on the degree of correlatability, they may be eliminated from or retained in the computation.
If the features are closely linked, they can be removed from the calculation because using both features would not add any new information. Lower dimensions are the result of the feature reduction. This correlation matrix can be denoted using a heat map. E.g. As observed in the heat map given below; radius mean and parameter worst have a high degree of co-relation thus having a value close to 1.

VII. Results

Eigen values and their corresponding eigen vectors are calculated using the Gram matrix. These are then arranged in the descending order to obtain the Principal Components.
The maximum amount of information about all the features combined can be found in the Principal Component in the eigen vector derived from the largest eigen value.
From the plot, it is observed that 99% of the variance in the original data is encompassed within the first two Principal Components. Hence, the upcoming PCs are negligible.

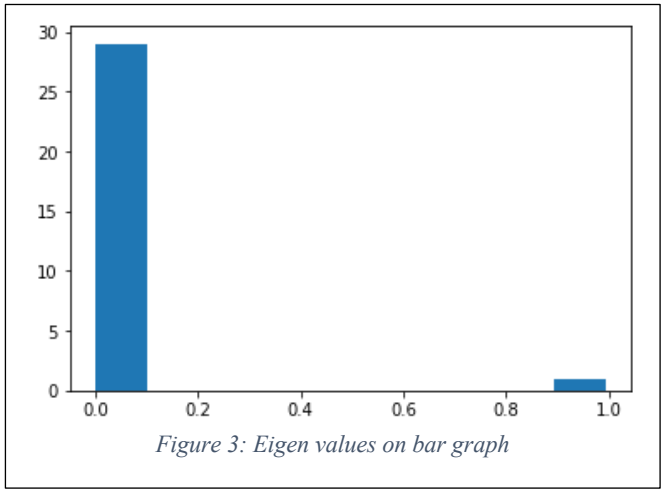

Figure 3: Eigen values on bar graph

Dimension Reductional techniques, namely PCA and Kernel PCA are now used on the obtained PCs. Using PCA:

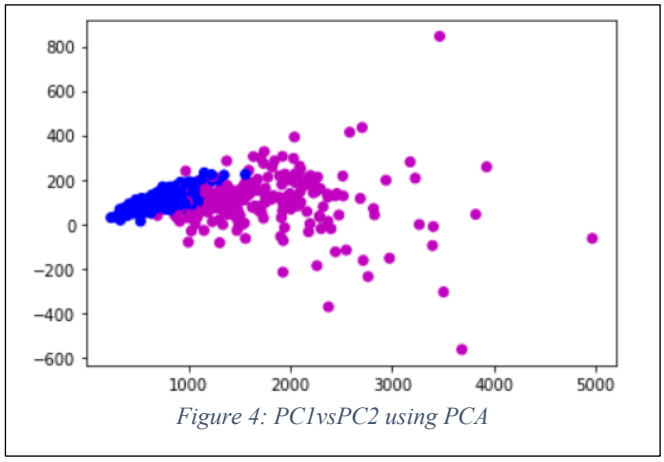

Figure 4: PC1vsPC2 using PCA

It is observed that this method is not so efficient since data is not linearly separable.
Using Kernel PCA: There are various mapping functions that can be utilized when working with Kernel PCA.

$$\varphi(r) = e^{(-\epsilon r)^2}$$

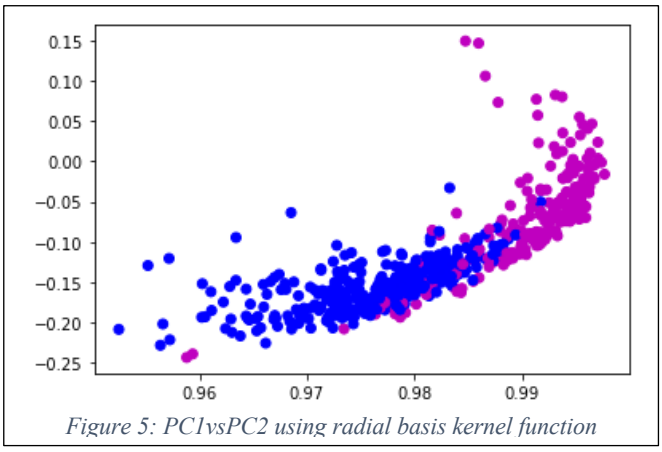

Figure 5: PC1vsPC2 using radial basis kernel function

$$K(x, y) = (x^T y + c)^d$$

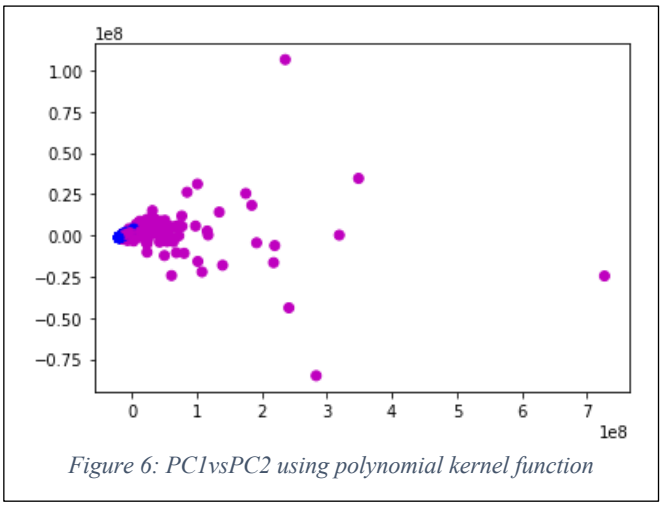

Figure 6: PC1vsPC2 using polynomial kernel function

$$\cos\theta = \frac{A.B}{\| A \|.\| B \|} = \frac{\sum A.B}{\sqrt{\sum A^2}.\sqrt{\sum B^2}}$$

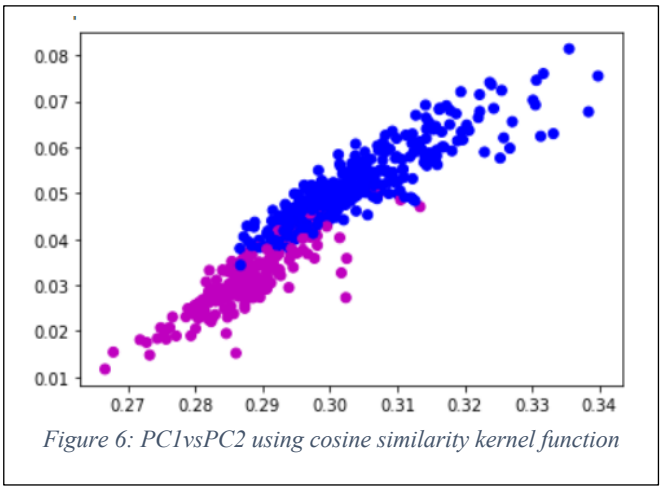

Figure 6: PC1vsPC2 using cosine similarity kernel function

Comparing the results obtained all three of the mapping functions, it is evident that Cosine similarity function yields the best results, since the data is (almost) perfectly separable, linearly.

VIII. Conclusion

As a result of the dimensionality reduction of our initial data being decreased from 30 features to two principal components, it is simple to visualize our data. The graph makes it evident that the Kernel PCA approach of non-linear data reduction performs well. It is clear that the optimal result, a linear separation of the data, is provided by the [4] [5] [6] [7]cosine mapping function.

IX. Bibliography

[1] D. Pelliccia, "PCA and Kernel PCA explained," [Online]. Available: https://nirpyresearch.com/pca-kernel-pca-explained/.

[2] "Kernel PCA," [Online]. Available: https://ml-explained.com/blog/kernel-pca-explained.

[3] S. K. M.-S. S. A. J. T. PALLE E.T. JORGENSEN, "AN INFINITE DIMENSIONAL ANALYSIS OF KERNEL PRINCIPAL COMPONENTS".

[4] Q. Wang, Kernel Principal Component Analysis and its Applications in Face Recognition and Active Shape Models.

[5] Z. Jaadi, "A step-by-step explaination of principal component analysis," [Online]. Available: https://builtin.com/data-science/step-step-explanation-principal-component-analysis.

[6] G. Tanner, "Principal Component Analysis," [Online]. Available: https://ml-explained.com/blog/principal-component-analysis-explained.