Department of Computer Science and Engineering

Institute of Engineering

JK Lakshmipat University, Jaipur, India

[Ritvik Tak]

[ritviktak@jklu.edu.in

Abstract

[In this case study we are using the unsupervised learning method i.e., “Principal Component Analysis” to determine the variance of given dataset.]

Variations of gene expression revealed by Principal component analysis of tumor and normal colon tissues

A case study on [Gene Expressions variances]

1. Introduction
2. Learning Objective of the case
3. Study Questions and suggested answers of questions

3.1. Criteria for choosing component

3.2. Scaling of features in PCA

3.3. Affects of scaling in covariance matrix

3.4. Look of Variance

Appendix A: Description of Utilized technique: [Principal Component Analysis]

Appendix B: CRISP Model Steps:

B.1. Business understanding

B.2. Data understanding

B.3. Data preparation

B.4. Modelling

B.5. Evaluation

References

1. **INTRODUCTION**

Recently introduced experimental techniques based on oligonucleotide or cDNA arrays now allow the expression level of thousands of genes to be monitored in parallel (1–9). To use the full potential of such experiments, it is important to develop the ability to process and extract useful information from large gene expression data sets. Elegant methods recently have been applied to analyse gene expression data sets that are comprised of a time course of expression levels. Examples of such time-course experiments include following a developmental process or changes as the cell undergoes a perturbation such as a shift in growth conditions. The analysis methods were based on finding the variation of genes according to similarity in their temporal expression (5, 6, 9–11). Such analysis has been demonstrated to identify functionally related families of genes, both in yeast and human cell lines (5, 6, 9, 11). Other methods have been proposed for analysing time-course gene expression data, attempting to model underlying genetic circuits (12, 13). The data set used is composed of 40 colon tumour samples and 22 normal colon tissue samples, analysed with an Affymetrix oligonucleotide array complementary to more than 6,500 human genes and expressed sequence tags. The correlation in expression levels across different tissue samples is demonstrated to help identify genes that regulate each other or have similar cellular function. To detect variations of related genes and tissues we applied Principal Component Analysis an effective technique for detecting variations in data sets . The main result is that an efficient unsupervised algorithm revealed variations of genes whose expression is correlated, suggesting a high degree of organization underlying gene expression in these tissues. It is demonstrated, for the case of ribosomal proteins, that clustering can classify genes into coregulated families.

So here we have used the dataset composed of 40 colon tumour samples and 22 normal colon tissue samples, which can be used to determine the variation between these tissues with respect to all the 6500 human gene.

1. **LEARNING OBJECTIVE OF THE CASE**

In this case, we have used an unsupervised learning algorithm in the machine learning.

The algorithm name “Principal Component Analysis” gives the following objectives:

1. PCA can be used in industries to determine the most important traits we can use to differentiate something.
2. The attribute we want to represent in the end result are the ones with the greatest variance.
3. We can start with finding the linear combination of the variables and to rank them according to their importance.
4. It is the persisting method we can use for dimension reduction and this is how it allows us to collect continue to grow.
5. **STUDY QUESTIONS**
6. What is the best criterion for choice the number of component for interpretation of Principal Component Analysis(PCA)?
7. How this PCA can be used to scale the features of data?
8. Does this scaling affects the covariance matrix?
9. How does the variance look like at the end of representation?

Before starting the analysis we have to look at the [data](https://github.com/ramhiser/datamicroarray/wiki/Alon-(1999)). After understanding the target and feature variable we will look forward to analysis.

* 1. **What is the best criterion for choice the number of component for interpretation of Principal Component Analysis(PCA)?**

Examine the components. The first PC often reflects a global quality that all variables share. For example, I’m looking at gene expression of colon cancer, and the first component simply reflects how many human genes were there. The subsequent components are more interesting, in that they can identify groups of variables within the whole that tend to vary together. But statistical considerations have to go together with a theory in order to look meaningful component.

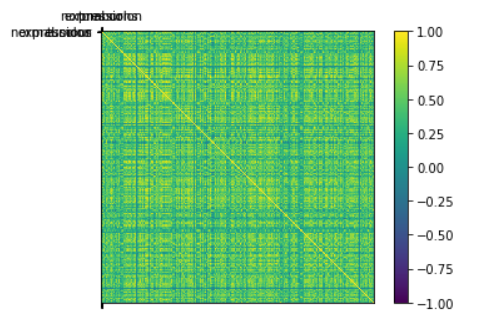
* 1. **How this PCA can be used to scale the features of data?**

Feature scaling through standardization can be an important pre-processing step for many machine learning algorithms. Standardization involves rescaling the features such that they have the properties of a standard normal distribution with a mean of zero and a standard deviation of one. To illustrate this, PCA is performed comparing the use of data with “Standard scaler” applied, to unscaled data.

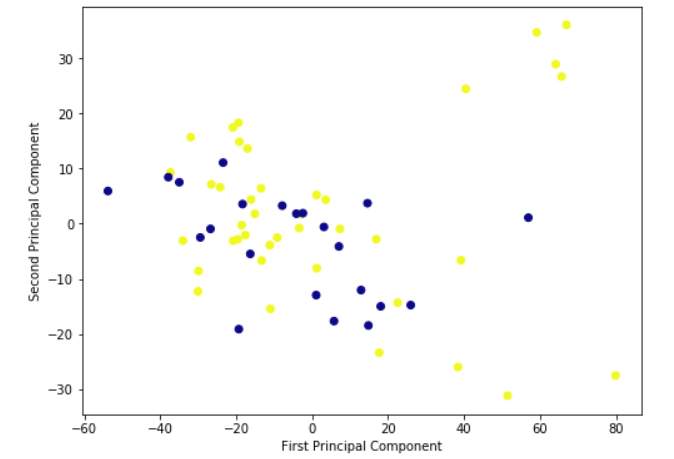
The transformed data is then used to decomposition i.e. the process of finding the variance by applying the “Principal Component Analysis”.

* 1. **Does this scaling affects the covariance matrix?**

If features got scaled, e.g., standard deviation of one , it does affect the covariance and therefore influences the results of a PCA.

****

* 1. **How does the variance look like at the end of the representation?**

****

Above diagram refers to the variance between the samples of tumour and normal colon tissues causing colon cancer. As we have seen that these two are the target variable and can be used to determine the gene expressions.

**Appendix A:**

Description of Utilized technique: [Principal Component Analysis]

**A.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.

Principal components analysis (PCA) produces a low-dimensional representation of a data set. It finds a sequence of linear combination of the variables that have maximal variance and are mutually uncorrelated. Apart from producing derived variables for use in supervised learning problems, PCA also serves as a tool for data visualization.

The first principal component of a set of features *X*1, *X*2,…, *Xp* is the normalized linear combination of the features

Z1=ϕ11X1+ϕ21X2+…+ϕp1Xp

with the largest variance. By normalized, we mean that ∑j=1pϕj12=1; the elements *ϕ*11,…, *ϕp*1 are the loadings of the first principal component; together, the loadings make up the principal component loading vector, *ϕ*1 = (*ϕ*11, *ϕ*21, … *ϕp*1)T.

PCA is either done by singular value decomposition of a design matrix or by doing the following 2 steps:

1. calculating the data covariance (or correlation) matrix of the original data
2. performing eigenvalue decomposition on the covariance matrix

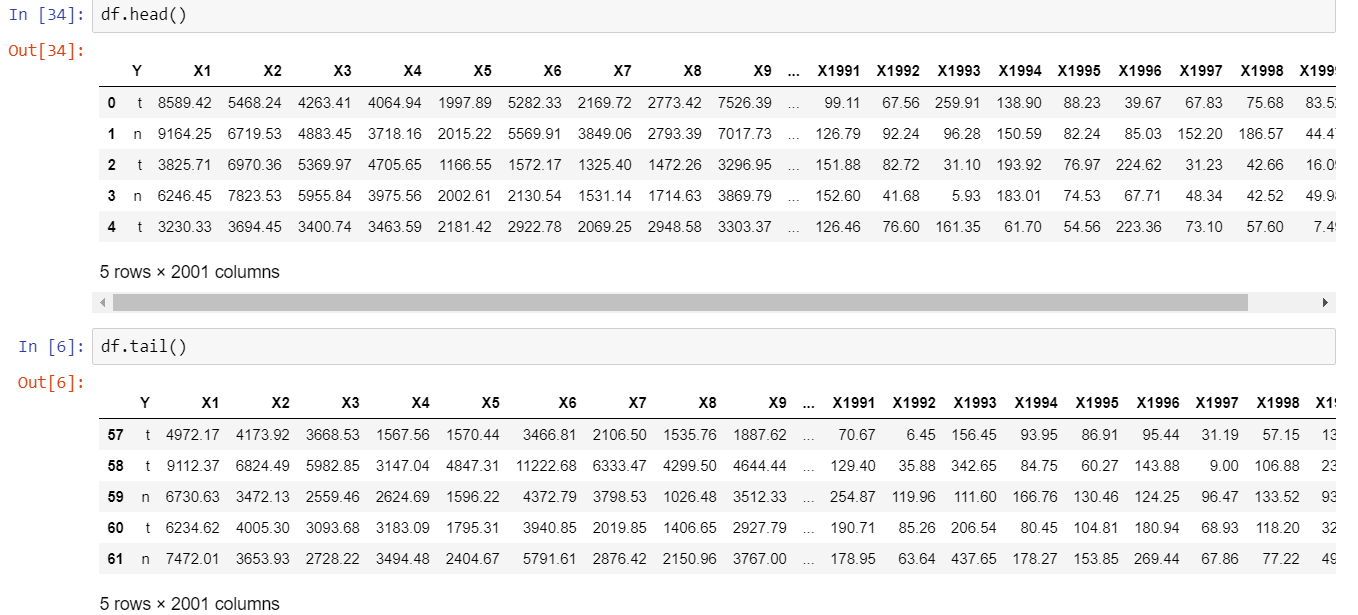
**Appendix B:**

**B.1. Business understanding**

In this case study, its been observed that how we can use the large-scale data to determine the variance. Today in our industry there is a lot of data available for such machine learning algorithms. But this unsupervised learning method is very unique to do with data.

**B.2. Data understanding**

This is how data looks like , containing 2001 columns



“Colon cancer Tissue dataset” contains 1 csv file.

As you can see the above diagram,

● The column ‘Y’ contains the samples of tumour and normal colon tissues.

● Columns ‘X1’ to ‘X1999’contains the genes expression levels of cancer-causing tissues.

● There are 6500 human genes obtained with an Affymetrix oligonucleotide array.

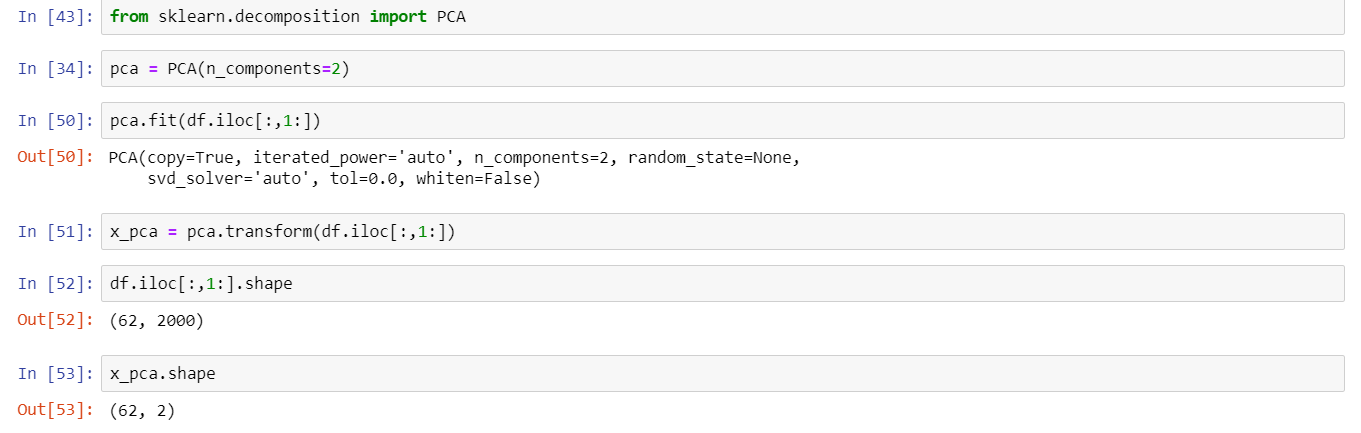
**B.3. Data preparation**

I have done standard scaling in standard deviation of one for better results.

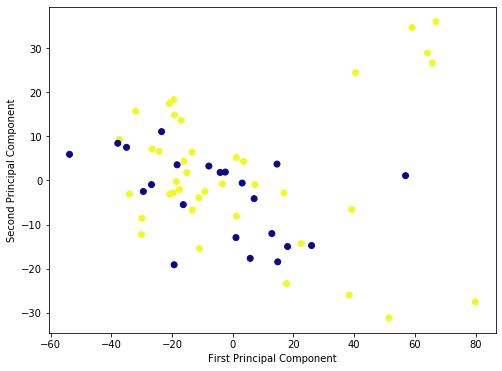
****

**B.4. Modelling**

After scaling the data, I applied PCA decomposition and transformed the data.

****

**B.5. Evaluation**

****

After transforming the data when we plot the variance it shows the above result.

To check the code, please find ML(RitvikTak)PCA.py file in the below drive link

<https://drive.google.com/open?id=1YMdj0jKPd_hmhJmDvJSAvXA_4FfdU3yU>

**References**

* <https://www.pnas.org/content/96/12/6745>
* <https://github.com/ramhiser/datamicroarray/wiki/Alon-(1999)>