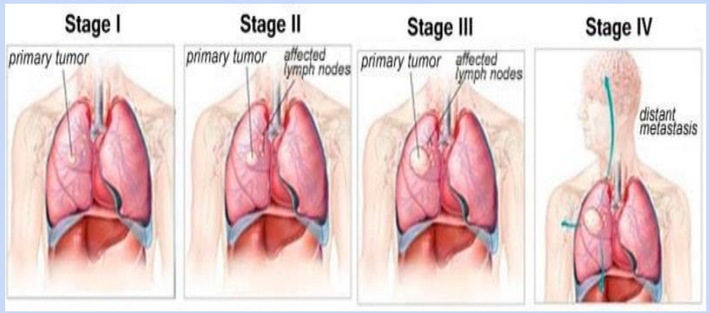
Machine Learning Classification of Stages in Lung Cancer Patients

Theodore Berger[1], Ting Jin[2], Daifeng Wang, PhD[1]
[1] North Shore Hebrew Academy High School , [2]Department of Biomedical Informatics, Stony Brook University

Stony Brook University

Introduction

- There are four main stages of lung cancer each containing two sub stages. This project explores multiple machine learning classification methods using Python, which predict the cancer stage the patient is enduring based on gene expression data. Additionally this program will show the accuracy of the different classifications methods as well as the accuracy when using different amounts of groups to describe the stage of cancer.



M. (2018, June 05). What is Lung Cancer? Retrieved from http://www.ourhealthpage.com/what-is-lung-cancer/

Methods

Feature Selection

- The first step in building the classification model was doing feature selection. The gene expression data contains to many genes to process without feature selection. The method chosen for feature selection was Principal Component Analysis(PCA). Principal component analysis is used to emphasize variation and strong patterns in a dataset. It's often used to make data easy to explore and visualize.

Classification

- For the classifications and predictions three different methods where used. The models were GaussianNB, KNN and SVC. The next step for all three models was Splitting the dataset into train and test using the classification methods, the train set is used to build the classification model and the test dataset is used to test the performance of our model.

Results

Figure 1

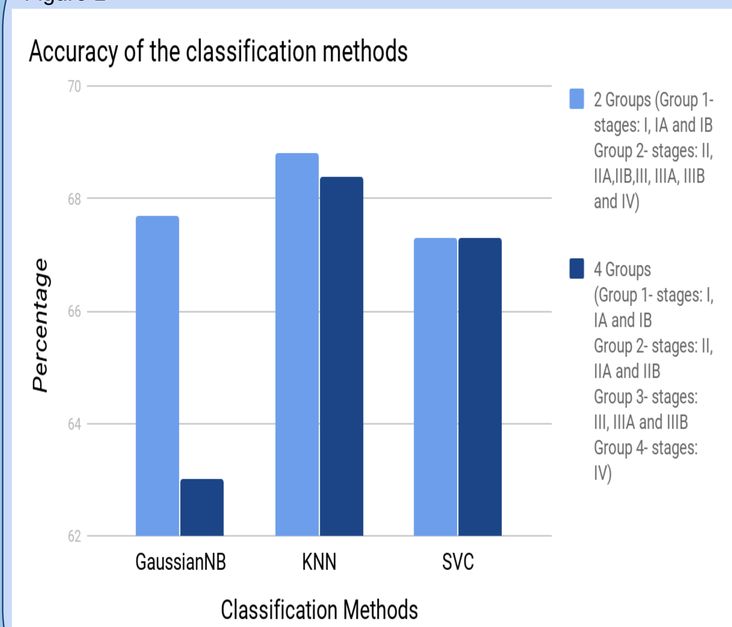


Figure 1 - The above figure is a bar graph showing the effectiveness of each classifier used to predict the stage of cancer.

Goals and Challenges

Goals

- The main goal is to successfully predict the stage of Lung cancer based on the gene expression data of each patient.
- To compare the efficiency and accuracy of different machine learning methods that predict the cancer class.

Challenges

- The Gene expression data, contained over 1000 patients and over 10000 genes which is to big for classification so feature selection is imperative.

Results

- These graphs and table show the accuracy of the three classification methods. Each classifier was used to predict the cancer stage into two groups and four groups. The two groups method put stages I, IA and IB in group one and stages II-IV in group two. In the four group method stages I- IB is put into group one and stages II-IIB is in groups two and so on.

Accuracy	2 Groups	4 Groups
GaussianNB	67.7%	63%
KNN	68.8%	68.4%
SVC	67.3%	67.3%

Conclusions & Future Work

- All the classification methods worked very well and for the most part successfully predicted the cancer class of each patient.
- Although, one method clearly stands out as the most accurate, the KNN classifier when classifying the data into two groups.
- Ideally in the future the goal would be to make each classifier or at the very least one more accurate when predicting the stage of lung cancer.

References

- Markham, K. (2018, July 04). Introduction to machine learning in Python with scikit-learn (video series). Retrieved from https://www.dataschool.io/machine-learning-with-scikit-learn/
- V. (n.d.). Principal Component Analysis explained visually. Retrieved from http://setosa.io/ev/principal-component-analysis/ https://www.dataschool.io/machine-learning-with-scikit-learn/
- M. (2018, June 05). What is Lung Cancer? Retrieved from http://www.ourhealthpage.com/what-is-lung-cancer/