**Design & Implement SVM on Parkinson's Disease Classification**

**Overview**

Classify whether or not a patient has Parkinson’s disease based on the dataset. For context, Parkinson’s is a progressive disease that causes the degeneration of the brain, leading to both motor and cognitive problems. It is thus reasonable to assume a correlation between a patient’s ability to speak and their progression into Parkinson’s as these capabilities regress. The dataset I worked with was created by Max Little of the University of Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado, who recorded the speech signals. The original study published the feature extraction methods for general voice disorders.

**Problem Statement**

This project attempts to prove that data from a patient can help diagnose whether or not they suffer from Parkinson’s. I will attempt to run Support Vector Machine on the data in hopes to reach a high predictability rate that is matched with a reasonable runtime. The study itself obtained a predictability rate of 91.4% and so I hope to reach a rate close to this or possibly to surpass it.

**Data Set Description**

The dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). Each column in the table is a particular voice measure, and each row corresponds one of 195 voice recording from these individuals ("name" column). The main aim of the data is to discriminate healthy people from those with PD, according to "status" column which is set to 0 for healthy and 1 for PD.

Attributes :-

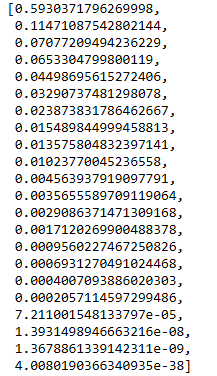
Name - ASCII subject name and recording number  
MDVP:Fo(Hz) - Average vocal fundamental frequency  
MDVP:Fhi(Hz) - Maximum vocal fundamental frequency  
MDVP:Flo(Hz) - Minimum vocal fundamental frequency  
MDVP:Jitter(%),MDVP:Jitter(Abs),MDVP:RAP,MDVP:PPQ,Jitter:DDP - Several measures of variation in fundamental frequency  
MDVP:Shimmer,MDVP:Shimmer(dB),Shimmer:APQ3,Shimmer:APQ5,MDVP:APQ,Shimmer:DDA - Several measures of variation in amplitude  
NHR,HNR - Two measures of ratio of noise to tonal components in the voice  
status - Health status of the subject (one) - Parkinson's, (zero) - healthy  
RPDE,D2 - Two nonlinear dynamical complexity measures  
DFA - Signal fractal scaling exponent  
spread1,spread2,PPE - Three nonlinear measures of fundamental frequency variation

**Dimensionality Reduction:**

**Principal Component Analysis**

The main idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of many variables correlated with each other, either heavily or lightly, while retaining the variation present in the data set, up to the maximum extent. PCA extracts new independent variables from our data set that explain the most of the variance of the data set.

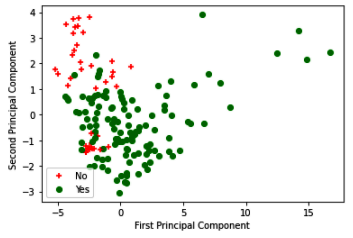
Variance explained by each dimension:



So here we can see that the first 2 components contribute to 70% of the total variance. So it’s good enough to choose only 2 components.

**Exploratory Visualization**

To get a better grasp of the data, I developed a scatter graph that allows me to see the spread of the data with different labels for healthy and affected patients.



From the graph, it seems that the data is not linearly separable. Hence, I used RBF kernel to classify the data in higher dimensions.

**Data Preprocessing**

There was no need to implement feature transformation, as all of the data is continuous. The data itself does not project any outliers that I feel would have an enormous impact on the classifiers I have created. Thus no effort has been made to seek out and eliminate data points.

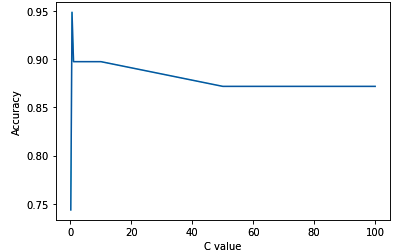
**Support Vector Machine**

I investigated the use of SVM using RBF kernel with varying C value (Soft Margin to Hard Margin) and Gamma values.

C = [0.1, 0.5, 1, 10, 50, 100]

Gamma = [0.01, 0.05, 0.07, 0.1, 1, 0.5, 5, 10, 50]

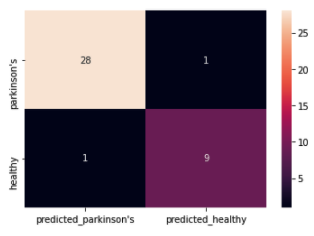
Accuracy chart for RBF Kernel (Gamma =1)



As it is visible, the highest accuracy was found at almost C= 0.5 i.e. soft margin. Hence, the optimum solution:

|  |  |  |  |
| --- | --- | --- | --- |
| Kernel | C | Gamma | Accuracy |
| RBF | 0.5 | 1 | 94.9% |

**Results:**

****

Confusion Matrix