Report 1: Regression on Parkinson’s disease data

Mohammad Eftekhari pour, s307774,
ICT for Health attended in A.Y.

2022/23 January 19th 2023

1 Introduction

Parkinson’s disease (PD) is a nervous system disorder that grows slowly step by step in a long term and affects movements. The cause of the PD is unknown, but it makes certain nerve cells (neurons) in the brain break down or die little by little. The severity of the PD is measured by neurologist during different sections and interviews. This process is time consuming and results can be different from one doctor to another. Because of that an automatic way is needed to estimate the severity of the PD faster and more accurately.

Many of them cannot speak correctly, since they cannot control the vocal chords and the vocal tract. It has been shown that they overcome the illness if they dance or have an external clock that gives the time.[1].

2 Data Analysis

The dataset chosen for this lab was Parkinson diseases dataset. The dataset is available on University of California Machine Learning Website[1]. It contains information related to patients who each patient was monitored for six months and the features that doctors were interested in were recorded automatically. Table 1 shows those attributes.

1	subject	2	age	3	sex
4	test time	5	motor UPDRS	6	total UPDRS
7	Jitter(%)	8	Jitter(Abs)	9	Jitter:RAP
10	Jitter:PQ5	11	Jitter:DDP	12	Shimmer
13	Shimmer(db)	14	Shimmar APQ3	15	Shimmar APQ5
16	Shimmar APQ11	17	Shimmar DD	18	NHR
19	HNR	20	RPDE	21	DFA
22	PPE				

Table 1: List of features

[1] https://archive.ics.uci.edu/ml/ machine-learning-databases/parkinsons/telemonitoring/

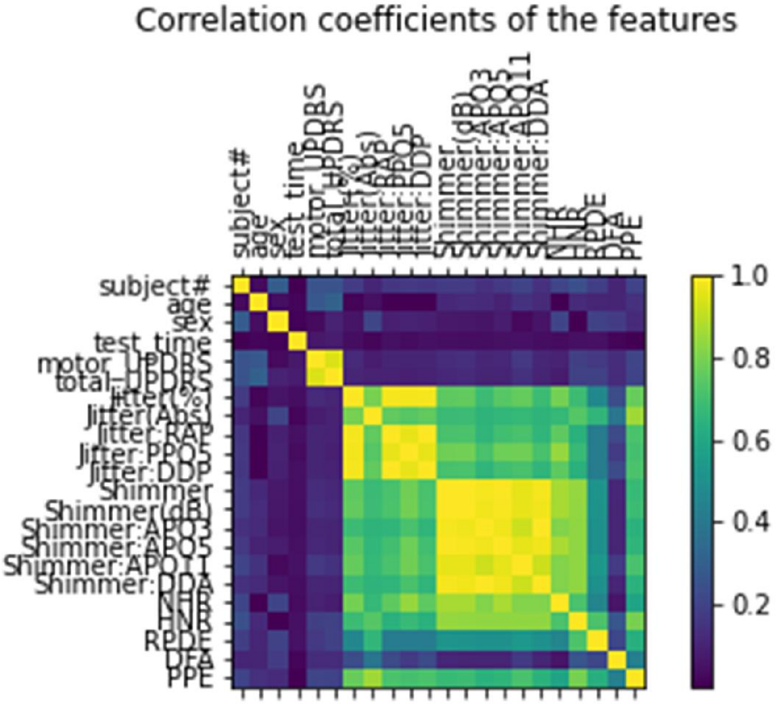

Figure 1: Covariance matrix of the features

Figure 1 shows the measured covariance matrix for the entire normalized dataset: corre-lation between total and motor UPDRS is evident, and strong correlation also exists among shimmer parameters and among jitter parameters (possible collinearity); on the other hand only a weak correlation exists between total UPDRS and voice parameters.

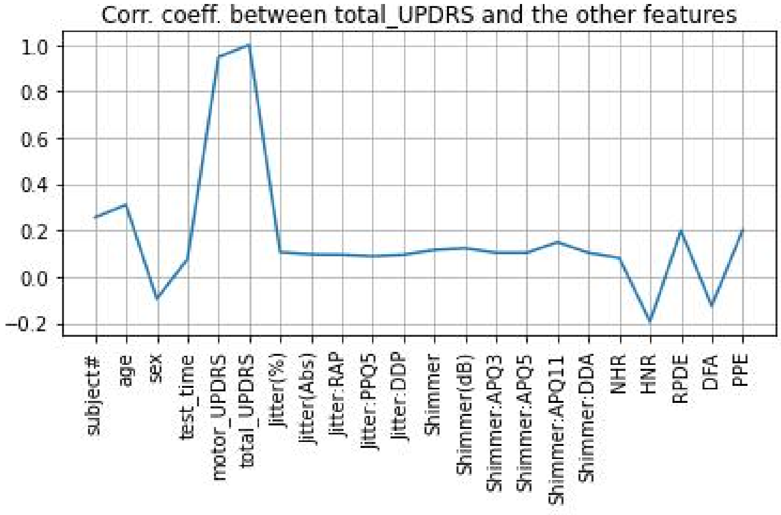

Figure 2- correlation between total_UPDRS and the other features

In figure 2 we can see that correlation coefficient between total_UPDRS and motor_UPDRS is big and it is obvious that total_UPDRS has a big correlation coefficient with itself. On other hand almost the rest of features do not have big correlation coefficient, and the last four features their correlation coefficient have some fluctuation with total_UPDRS.

The 22 features which are used in this experiment are listed in table 1. Due to the collinearity of two of the features with Jitter:DDP, and Shimmer:DDA- are removed, because they have lowest correlation coefficient with the regressand(total UPDRS in our case). Moreover, subject ID is the patient number and it has no effect on our dataset so will be removed. So, we have 18 regressors.
$(X_{1,}X_2, \dots, X_{18})$

Linear regression

In general, the formula of Linear Regression is:

$$Y = w_1 X_1 + w_2 X_2 + \cdots + w_F X_F \tag{1}$$

But because of errors occurring in measurement process the real model will be:

$$Y = w_1 X_1 + w_2 X_2 + \cdots + w_F X_F + \nu \tag{2}$$

Y is the regressand, $X_1, X_2, \dots, X_F$ are the regressors(independent variable), ν is the unknown error and $w_1, w_2, \dots, w_F$ are the weights to be found.
ν is the Greek letter that corresponds to "n".

Then we can write:

$$\begin{bmatrix} y(1) \\ y(2) \\ . \\ . \\ . \\ y(N) \end{bmatrix} = \begin{bmatrix} x_1(1) & x_2(1) & x_3(1) \dots x_F(1) \\ x_1(2) & x_2(2) & x_3(2) \dots x_F(2) \\ & & . \\ & & . \\ & & . \\ x_1(N) & x_2(N) & x_3(N) \dots x_F(N) \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ . \\ . \\ . \\ w_F \end{bmatrix} + \begin{bmatrix} \nu_1 \\ \nu_2 \\ . \\ . \\ . \\ \nu_F \end{bmatrix} \tag{3}$$

i.e. $\boldsymbol{y} = \boldsymbol{Xw} + \boldsymbol{\nu}$

We assume that errors $\nu(n)$ are small. Therefore it makes sense to find **w** as the solution of the problem $min_w \; ||\boldsymbol{y} - \boldsymbol{Xw}||^2$.

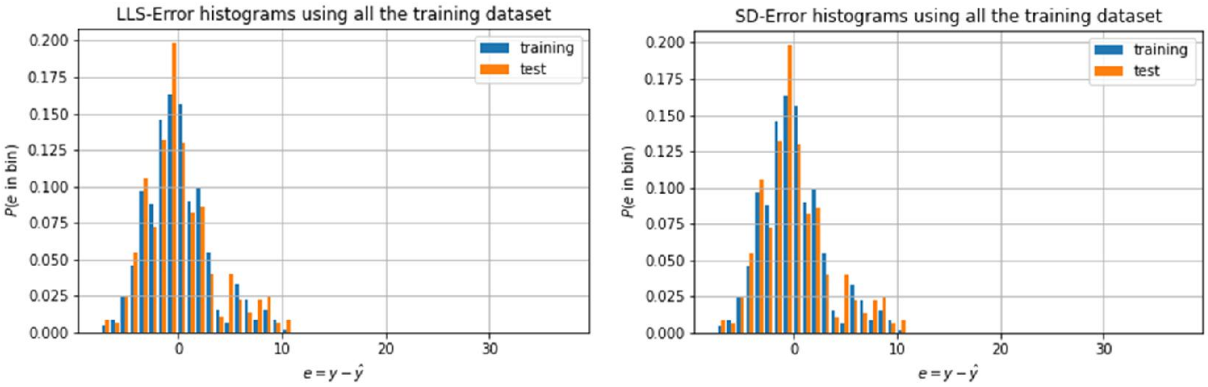

Figure 3 - comparison between errors of two used algorithms

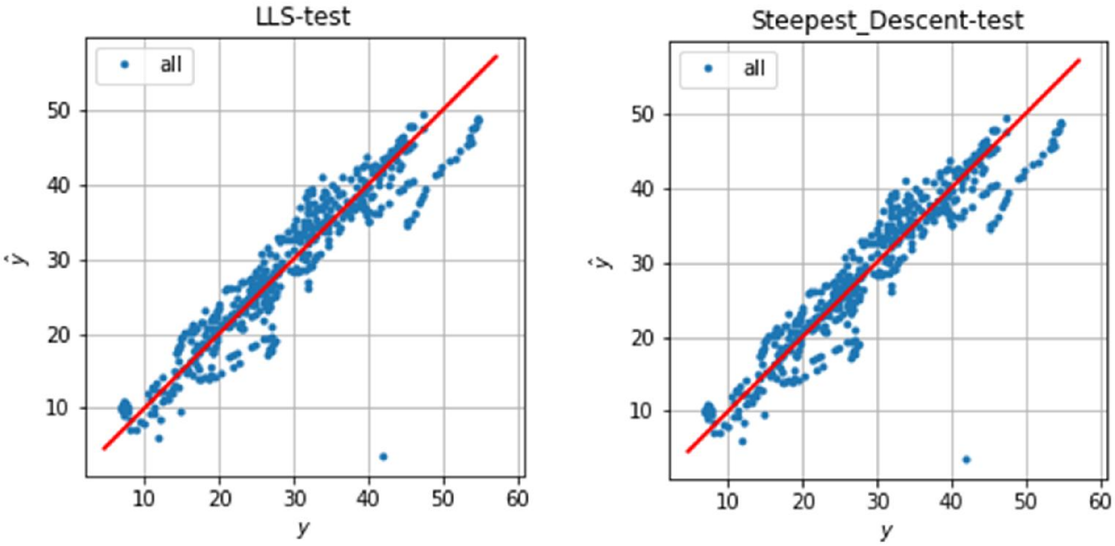

Figure 4 - Differences between Data test in two utilized algorithms

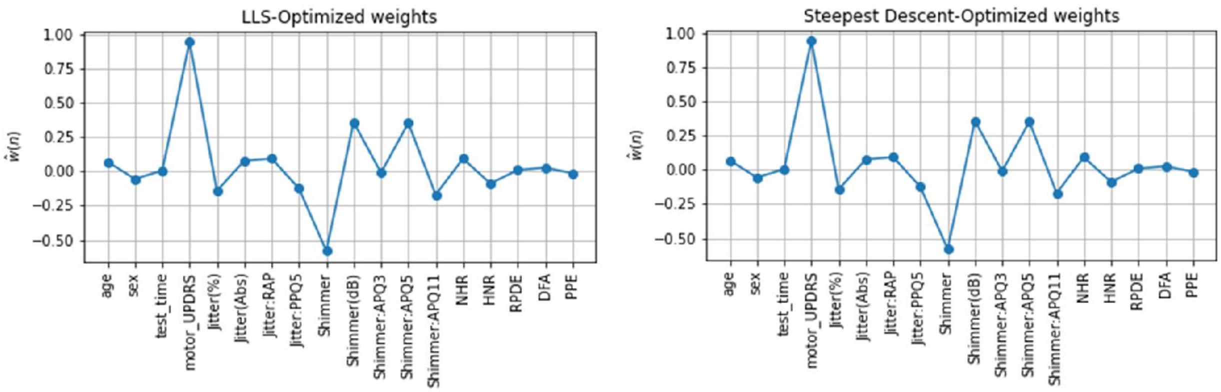

Figure 5 - Calculated Weights of used algorithms

Based on Figure 3,4, and 5, the two algorithm work the same and there is not different between two algorithm.

3 Conclusions

According to gained results, the data of minimum, maximum, and other features foe LLS algorithm listed in table 2

	Min	Max	Mean	Std	MSE	R^2	corr_coeff
Training set	-7.497	10.424	7.737×10^{-15}	3.056	9.339	0.917	0.957
Test set	-7.438	38.383	1.483×10^{-1}	3.777	14.285	0.877	0.937

Table 2 – LLS

According to gained results, the data of minimum, maximum, and other features foe Steepest descent algorithm listed in table 3

	Min	Max	Mean	Std	MSE	R^2	corr_coeff
Training set	-7.497	10.424	7.737×10^{-15}	3.056	9.339	0.917	0.957
Test set	-7.438	38.383	1.482×10^{-1}	3.777	14.284	0.877	0.937

Table 2 – Steepest descent

4 Conclusion

Both algorithm work working well and there is no different between them.

References

https://www.parkinsons.org.uk/information-and-support/parkinsons-symptoms