Report 2: Gaussian Process Regression on Parkinson's disease data

Seyed Mohammad Mehdi Hosseini, s301769,
ICT for Health attended in A.Y. 2021/22

November 25th, 2021

1 Introduction

Patients affected by Parkinson's disease cannot perfectly control their muscles. In particular they show tremor, they walk with difficulties and, in general, they have problems in starting a movement. Many of them cannot speak correctly, since they cannot control the vocal chords and the vocal tract.

Levodopa is prescribed to patients, but the amount of treatment should be increased as the illness progresses and it should be provided at the right time during the day, to prevent the freezing phenomenon. It would be beneficial to measure total UPDRS ((Unified Parkinson's Disease Rating Scale) many times during the day in order to adapt the treatment to the specific patient. This means that an automatic way to measure total UPDRS must be developed using simple techniques easily managed by the patient or his/her caregiver.

One possibility is to use patient voice recordings (that can be easily obtained several times during the day through a smartphone) to generate vocal features that can be then used to regress total UPDRS.

Gaussian Process Regression (GPR) was used on the public dataset at [1] to estimate total UPDRS, and the results were compared to those obtained with linear regression, showing the superiority of GPR.

2 Data analysis

The 22 features available in the dataset at [1] are listed in table 1: of these, subject ID and test time were removed, total UPDRS is the regressand. All the remaining 19 features were used as regressors in linear regression, but only 3, namely motor UPDRS, age and PPE, were used in GPR.

The number of points in the dataset is 5875; data are shuffled and the first 50% of the points are used to train the linear model, 25% of the points are used for the validation and

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:PPQ5	11	Jitter:DDP	12	Shimmer
13	Shimmer(dB)	14	Shimmer:APQ3	15	Shimmer:APQ5
16	Shimmer:APQ11	17	Shimmer:DDA	18	NHR
19	HNR	20	RPDE	21	DFA
22	PPE				

Table 1: List of features

the remaining 25% are used to test the model performance. Data are normalized using mean and standard deviation measured on the training dataset.

3 Gaussian Process Regression

In GPR, it is assumed that $N-1$ measured datapoints $(\mathbf{x}_k, y_k)$ are available in the training dataset, and that a new input $\mathbf{x}_N$ is present, whose corresponding output y_N has to be estimated.

In the following, $\mathbf{Y}_L = [Y_1, \ldots, Y_L]$ is the L-dimensional random vector that includes the random variables Y_ℓ and $\mathbf{y}_L = [y_1, \ldots, y_L]$ is the L-dimensional vector that stores the measured values of Y_ℓ. Vector $\mathbf{x}_\ell$ stores instead the measured regressors for Y_ℓ. The random variable to be estimated is Y_N, knowing the corresponding regressors $\mathbf{x}_N$, and the training dataset made of $N-1$ measured couples $(\mathbf{x}_\ell, y_\ell)$, $\ell = 1, \ldots, N-1$.

- The $N \times N$ covariance matrix $\mathbf{R}_{Y,N}$ of $\mathbf{Y}_N$ has n, k value:

$$\mathbf{R}_{Y,N}(n,k) = \theta \exp\left(-\frac{\|\mathbf{x}_n - \mathbf{x}_k\|^2}{2r^2}\right) + \sigma_\nu^2 \delta_{n,k}, \quad n,k \in [1, N]$$

- $\mathbf{R}_{Y,N}$ can be rewritten as

$$\mathbf{R}_{Y,N} = \begin{bmatrix} \mathbf{R}_{Y,N-1} & \mathbf{k} \\ \mathbf{k}^T & d \end{bmatrix}$$

  where $\mathbf{R}_{Y,N-1}$ is the covariance matrix of $\mathbf{y}_{N-1}$.

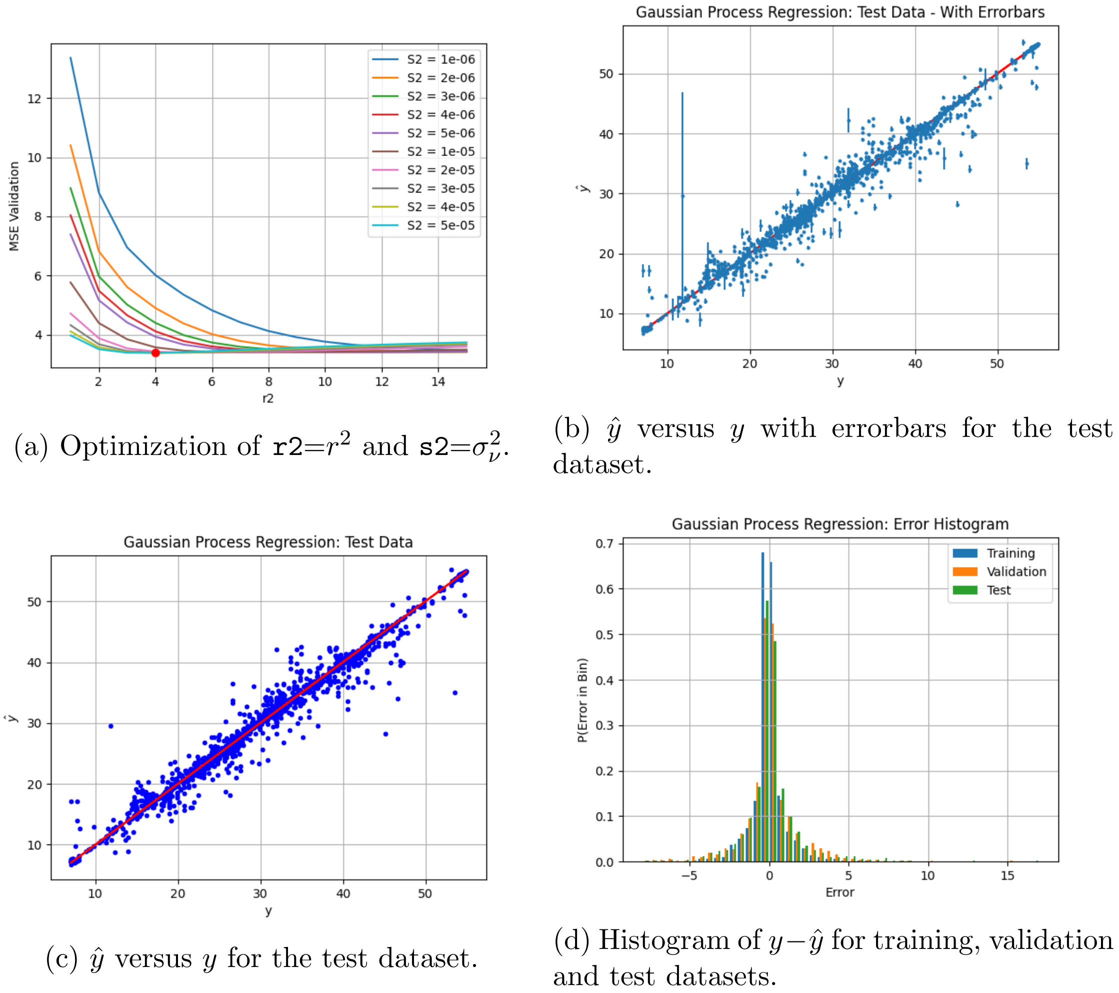
- Then the pdf of Y_N given the measured values $\mathbf{y}$ of $\mathbf{y}_{N-1}$ is

$$f_{Y_N|\mathbf{y}_{N-1}=\mathbf{y}}(z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}$$

$$\mu = \mathbf{k}^T \mathbf{R}_{Y,N-1}^{-1} \mathbf{y} \tag{1}$$

$$\sigma^2 = d - \mathbf{k}^T \mathbf{R}_{Y,N-1}^{-1} \mathbf{k} \tag{2}$$

  The point estimation of Y_N is $\hat{y}_N = \mu$.



(a) Optimization of $\texttt{r2}$=r^2 and $\texttt{s2}$=σ_ν^2.

(b) $\hat{y}$ versus y with errorbars for the test dataset.

(c) $\hat{y}$ versus y for the test dataset.

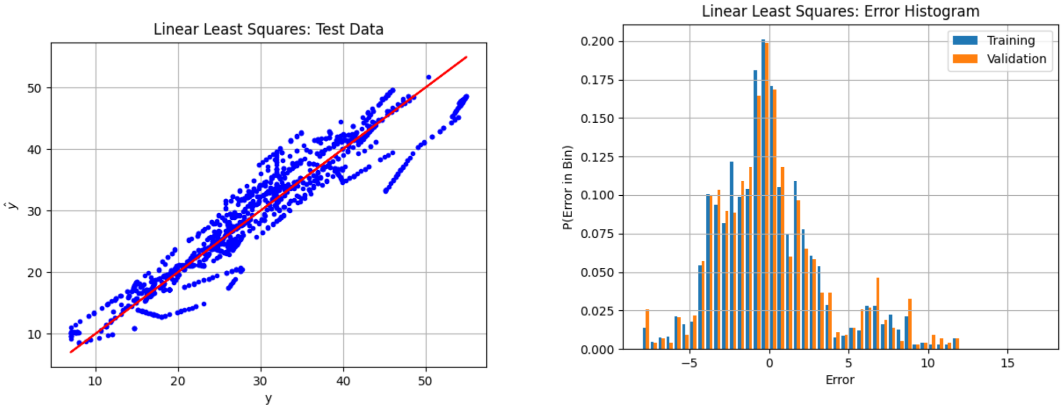
(d) Histogram of $y-\hat{y}$ for training, validation and test datasets.

Figure 1: Gaussian Process Regression results.

- In the above equations, couples $(\mathbf{x}_\ell, y_\ell)$ for $\ell = 1, \ldots, N-1$ belong to the training dataset, couple $(\mathbf{x}_N, y_N)$ belongs to the test or to the validation dataset.

The model hyperparameters are three: θ, r^2 and σ_ν^2. Since the training dataset stores normalized data, and σ_ν^2 is small, parameter $\theta = \mathbf{R}_{Y,N}(n,n)$ (variance of y_n) was set equal to 1. Hyperparameters r^2 and σ_ν^2 were set to minimize the mean square error $\mathbb{E}\{[y_N - \hat{y}_N]^2\}$ for the validation dataset. In particular, for each point $(\mathbf{x}_N, y_N)$ in the validation dataset, the $N = 10$ closer points in the training dataset were found, a set of possible values for r^2 ([1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]) and σ_ν^2 ([0.000001, 0.000002, 0.000003, 0.000004, 0.000005, 0.00001, 0.00002, 0.00003, 0.00004, 0.00005]) was tried using grid search and the optimum values were found among the considered cases (see Fig. 1a): these optimum values are $r_{opt}^2 = 4$ and $\sigma_{opt}^2 = 0.00004$ which have been marked in Fig. 1a with red dot.

Fig. 1c shows $\hat{y}$ versus y whereas Fig. 1b also shows the error bars ($\pm 3\sigma_y$ where σ_y is the denormalized version of σ in (2)). The estimation error histogram is shown in Fig. 1d.



(a) $\hat{y}$ versus y for test dataset.

(b) Histogram of $y-\hat{y}$ for training, validation and test datasets.

Figure 2: Linear Least Squares results.

Figs. 1b-1d were obtained using r^2_{opt} and σ^2_{opt}.

4 Linear regression based on Linear Least Squares

The model assumed in linear regression is

$$Y = w_1 X_1 + \ldots + w_F X_F = \mathbf{X}^T \mathbf{w} \tag{3}$$

where Y is the regressand (total UPDRS), $\mathbf{X}^T = [X_1, \ldots, X_F]$ stores the F regressors[1] and $\mathbf{w}^T = [w_1, \ldots, w_F]$ is the weight vector to be optimized. In (3), $Y, X_1, \ldots, X_F$ are all random variables.

Linear Least Squares (LLS) minimizes the mean square error (MSE) and the optimum weight vector $\mathbf{w}$ can be obtained in closed form as:

$$\hat{\mathbf{w}} = \arg\min \mathbb{E}\{(Y - \mathbf{X}^T\mathbf{w})^2\} = \left(\underline{\mathbf{X}}^T \underline{\mathbf{X}}\right)^{-1} \underline{\mathbf{X}}^T \mathbf{y} \tag{4}$$

where $\underline{\mathbf{X}}$ is the matrix that stores the (normalized) training regressor points and $\mathbf{y}$ is the (normalized) training regressand vector. Given $\hat{\mathbf{w}}$, the normalized regressand is estimated as

$$\hat{y}_N = \mathbf{x}_N^T \hat{\mathbf{w}} \tag{5}$$

Figure 2 shows the results obtained with LLS. Note that, to get a meaningful comparison with GPR, the training dataset and test datasets with the two regression models are the same; the validation dataset was only used for GPR, not for LLS regression.

[1] $\mathbf{X}$ is a column vector and $\mathbf{X}^T$ is its transpose

5 Comparison

It is evident, by comparing Figs. 1c and 2a that, with the Parkinson's dataset, Gaussian Process Regression (GPR) is more precise than linear regression, and this is also confirmed by the estimation error histograms in Figs. 1d and 2b.

Table 2 lists the main statistical properties of the estimation error $e = y - \hat{y}$ for the training, validation and test datasets. The mean square error of GPR is about 1/3 than that of LLS.

	Dataset	Err. Mean	Err. St. dev.	MSE	R^2
LLS	Training	0.0	3.153	9.943	0.9847
	Test	0.1272	3.394	11.539	0.9834
GPR	Training	0.0104	1.029	1.059	0.9984
	Validation	0.0483	1.840	3.388	0.9950
	Test	0.0620	1.955	3.827	0.9945

Table 2: Numerical comparison between GPR and LLS.

6 Conclusions

From all races and cultures, there are a huge number of people worldwide who are suffering from Parkinson's disease today, and according to the Global Burden of Disease Study 2015, nearly 13 million people will be involved in the prevalence of Parkinson's disease by 2040 [2]. Pain, anxiety, and depression are just some non-motor symptoms of people who are affected by Parkinson's condition but more importantly, the motor symptoms like tremor, muscle rigidity, slowness of movement, loss of control over vocal cords and tract, and etc. are the ones which directly affect the daily life tasks of Parkinson's patients. Accordingly, Daily and regular total UPDRS measurement could help to allocate treatments to patients in a scheduled manner to prevent the disease progress. Fortunately, within the advancement of electronic and telecommunication technology and the presence of smartphones, this process has been facilitated. In this study, the patients' remote voice recordings have been used to generate the vocal features to predict their total UPDRS. To regress this score, Gaussian Process Regression (GPR) [3] and Linear Regression based on Linear Least Squares (LLS) have been investigated. Bringing Gaussian Process into the regression task, it calculates the possible functions with respect to a prior space that can fit the data. On the contrary, other regression alternatives mostly need a specific model for the data just like linear regression. Moreover, hypermeter-tuning has been included to find the best r^2 and σ^2 hyperparameters to minimize the validation set mean squared error using grid search. Investigation through the results shows that the GPR could significantly outperform LLS considering both mean squared error and R2 score on the test set. GPR has three times less error on the test set in comparison with LLS which is really promising.

References

[1] https://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring

[2] Global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016 https://www.thelancet.com/journals/laneur/article/PIIS1474-4422(18)30295-3

[3] Gaussian Processes for Machine Learning, Carl Edward Rasmussen and Christopher K. I. Williams, The MIT Press, 2006. ISBN 0-262-18253-X. http://www.gaussianprocess.org/gpml/