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

Francesco Donato, s259358,
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 model, 25% of the points are used for the validation and the

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

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 $\mathbf{y}_{N-1}$.

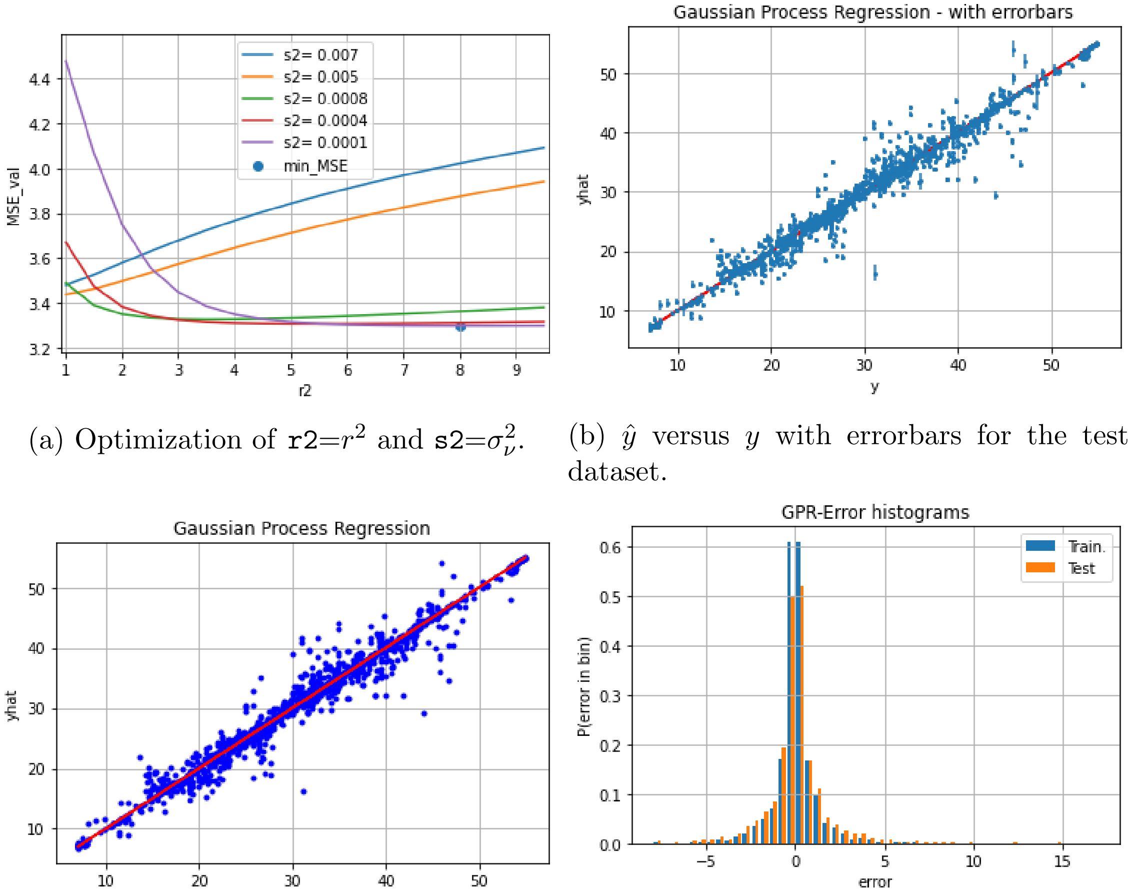
- Then the pdf of Y_N given the measured values $\mathbf{y}$ of $\mathbf{y}_{N-1}$ is

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

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

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

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



(a) Optimization of r2=r^2 and s2=σ_ν^2.

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

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

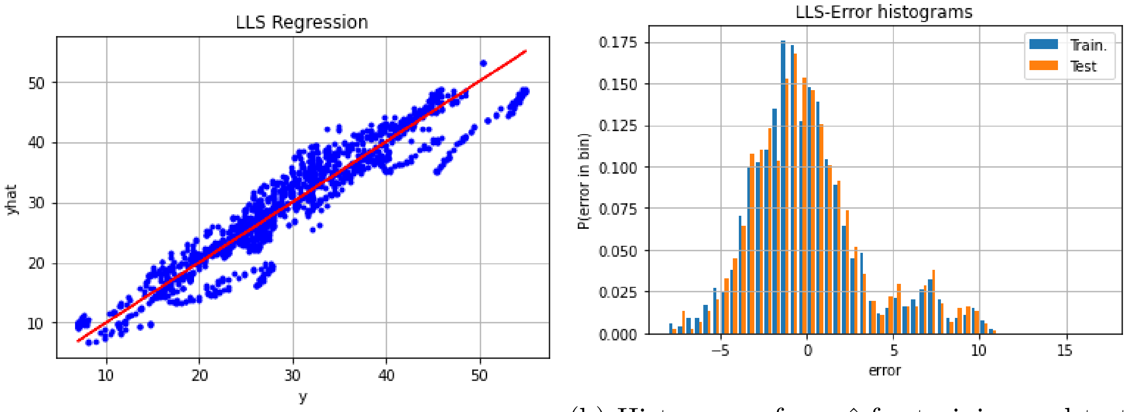
(d) Histogram of $y - \hat{y}$ for training and test datasets.

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 and σ_ν^2 was tried and the optimum values were found among the considered cases (see Fig. 1a): these optimum values are $r^2_{opt} = 8$ and $\sigma^2_{opt} = 0.0001$.

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



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

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

Figure 2: Linear Least Squares results.

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.

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.

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

Table 2 lists the main statistical properties of the estimation error $e = y - \hat{y}$ for the training 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	3.256	10.6	0.984
	Test	0.0665	3.319	11.022	0.983
GPR	Training	-0.0041	1.201	1.443	0.998
	Test	0.0242	1.871	3.502	0.995

Table 2: Numerical comparison between GPR and LLS.

6 Conclusions

Comparing the two regression models, it can be observed that the GPR (Gaussian Process Regression) has a better performance as the MSE (mean square error), for both training and test datasets, is lower than that of LLS (Linear Least Squares) method (Table 2). For both methods, the error mean for the training data is practically zero (as it must be), while for the test data is higher because the test dataset is another subset, with a slightly different value for each feature. The standard deviation is around the same for the training and test sets, which means there is no overfitting (Table 2), but this consideration regards only the LLS method because gaussian processes are not subjected to overfitting (few parameters define the model). Moreover, the coefficient of determination for the GPR is higher in both test and training datasets, compared to LLS, as the error standard deviation is lower (Table 2). Regarding the LLS regression, it can be seen that, in the error histogram (Figure 2b), there is a mixture of two gaussians because, analyzing the regression line graph (Figure 2a), significant errors are present in the range where true UPDRS value is large (from around 40 to almost 60). This is an issue that does not occur with GPR (Figure 1c), in fact the error histogram is more compact and does not show a mix of gaussians (Figure 1d). An advantage of implementing GPR is the possibility to know the range for predicted values, as shown in the regression with error bars (Figure 1b).

As the estimation should be compared to that of medical doctors, it is important to have an error standard deviation that is lower than the one coming from the doctor measurements of total UPDRS. For the dataset employed in this laboratory, the total UPDRS standard deviation is around 9, while the error standard deviation for the test dataset, resulting from GPR, is significantly lower (around 2, see Table 2), which is a positive result. Overall, the GPR is more reliable and precise in predicting total UPDRS.

It is important to highlight that, in GPR, only three features have been used: age, motor UPDRS and PPE (Perceived Phonatory Effort)[2], which is a feature related to the vocal fatigue of the patient.

The purpose of this regression is to take advantage of ICT devices in order to collect data and predict total UPDRS in an easy manner, so that the patient feels less stressed and also saving precious time for the doctor. As patients affected by Parkinson's disease cannot perfectly control their muscles, as well as their vocal chords and vocal tract, a way for collecting data easily is through voice recordings (using a smartphone for example). This is the reason why it could be useful to remove motor UPDRS from the features, because for measuring it, the patient has to visit a doctor, hence making useless the objective of the prediction, but with these regression methods and dataset, removing motor UPDRS makes the results to be significantly worse.

Parkinson's symptoms, including tremors (shaking), stiffness, and slowness of movement, are caused by a lack of dopamine, a natural substance usually found in the brain. Levodopa is in a class of medications known as central nervous system agents. It works by being converted to dopamine in the brain (only about 1 % reaches the brain tissue). Carbidopa is in a class of medications known as decarboxylase inhibitors. It works by preventing levodopa from being broken down before it reaches the brain. This allows for a lower dose of levodopa, which causes less nausea and vomiting.

In conclusion, this study suggests how, between the two evaluated methods, the GPR is the better choice in finding an automatic way to give the patient an objective and more reliable UPDRS score. It better helps the neurologist decide whether the patient should take levodopa or not and when, hence optimizing the treatment and dealing with the progression of the illness.

References

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

[2] https://reader.elsevier.com/reader/sd/pii/S0892199704000190?token=
6539DAC6A09A6ED0C58EFF5DA1302BE612A67D810D4DBC5A39D42EED8B0057183630EA68D20368CA978DB
originRegion=eu-west-1&originCreation=20211124161029