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

Francesco Conforte, s277683,
ICT for Health attended in A.Y. 2020/21

November 22nd, 2020

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 are removed, total UPDRS is considered as 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, $\mathbf{y}_N = [y_1, \ldots, y_N]$ is the N-dimensional random vector that includes y_N (to be estimated) and $\mathbf{y}_{N-1} = [y_1, \ldots, y_{N-1}]$ is the $N-1$-dimensional random vector whose values have been already measured (training data).

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

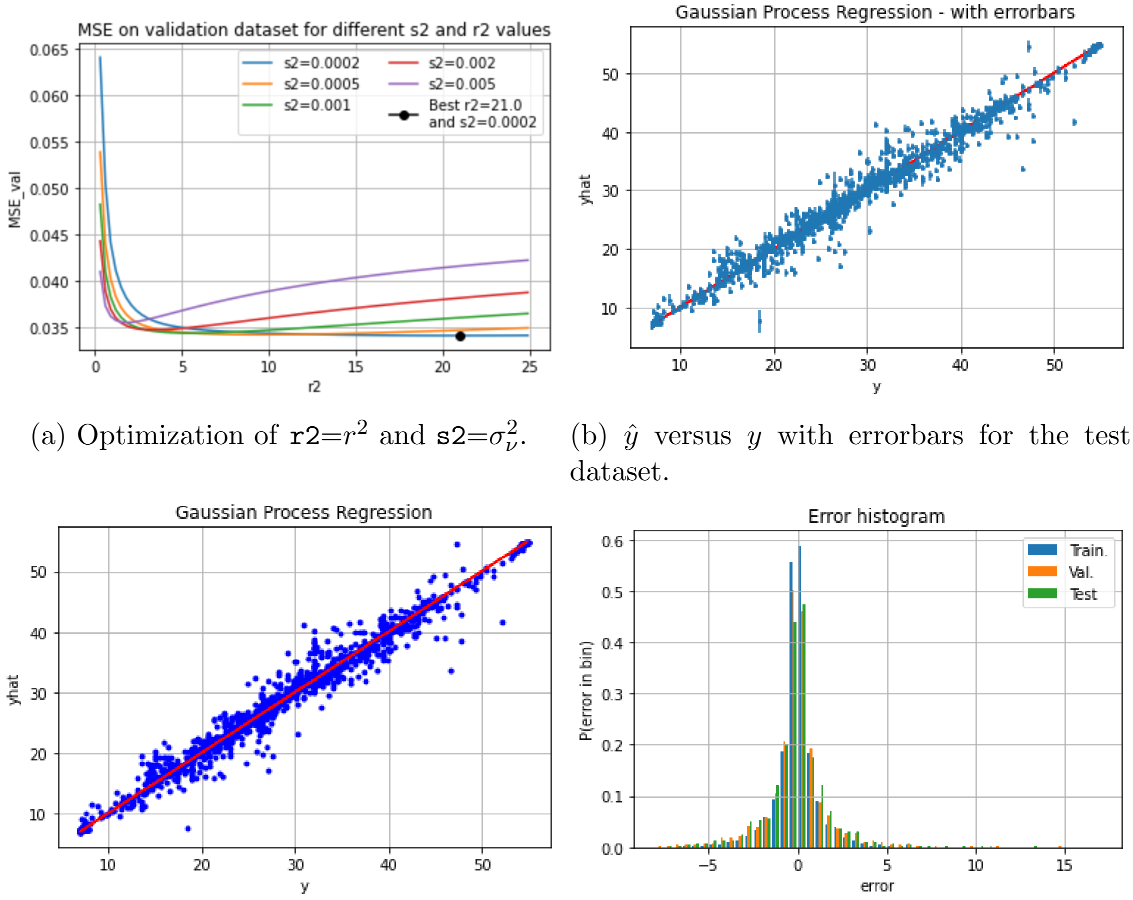
- Then the pdf of y_N given the measured values $\mathbf{y}$ of $\mathbf{y}_{N-1}$ is

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

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

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

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



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

(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 $(\mathbf{x}_n, y_n)$ for $n = 1, \ldots, N-1$ belong to the training dataset, couple $(\mathbf{x}_N, y_N)$ belongs to the test or to the validation dataset.

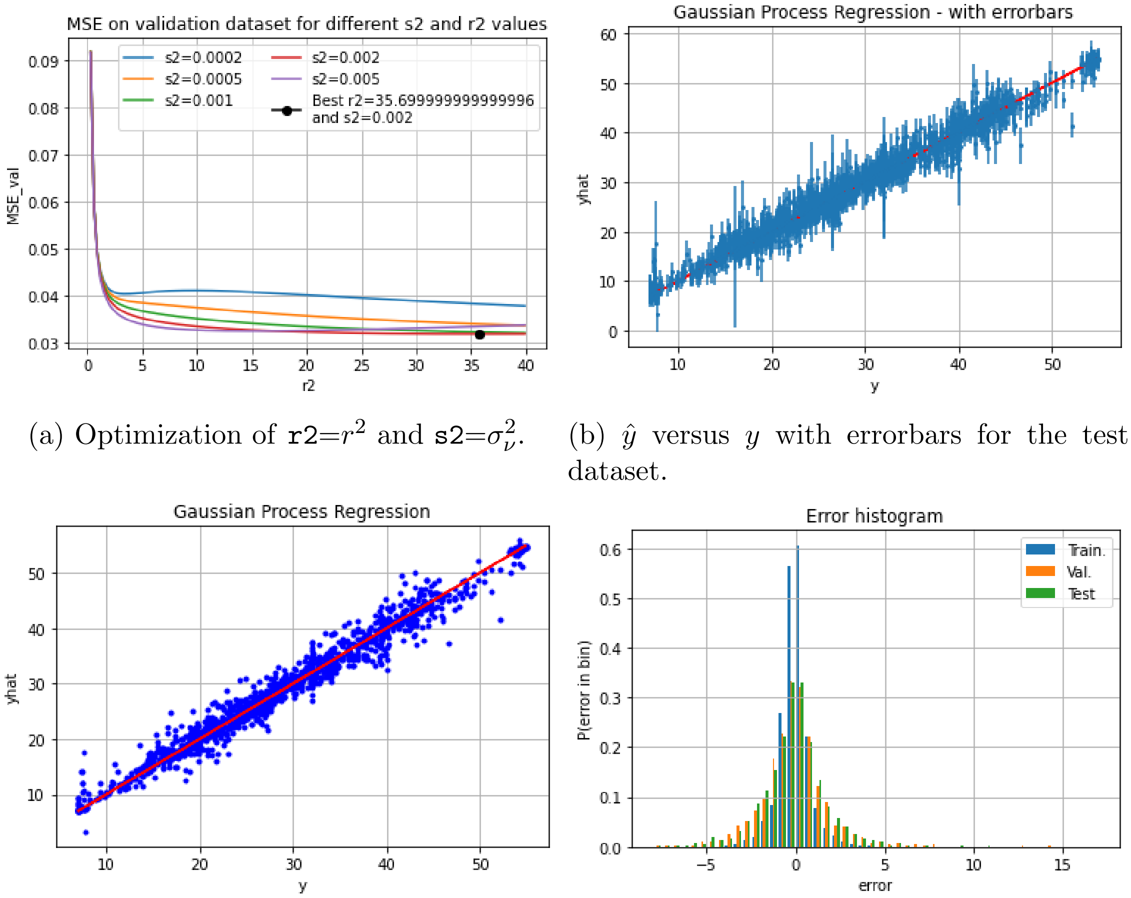
The model hyperparameters are three: θ, r^2 and σ_ν^2. Since the training dataset stores normalized data, parameter $\theta = \mathbf{R}_{Y,N}(n,n)$ (variance of y_n) was set equal to 1. Hyperparameters r^2 and σ_ν^2 were set to minimize the mean square error $\mathbb{E}\{[y_N - \hat{y}_N]^2\}$ for the 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 $r_{opt}^2 = 21$ and $\sigma_{opt}^2 = 0.0002$.

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_{opt}^2 and σ_{opt}^2.

4 GPR with further features as regressors

GPR was performed again by using 5 further features as regressors in addition to the ones used in Section 3.

They were: Jitter(Abs), Shimmer, NHR, HNR and DFA. Such features were chosen because they "complement each other with minimal overlapping information, and at the same time capture practically the entire range of possible differentiating features of the speech signals useful in determining UPDRS values". (See [2]).



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

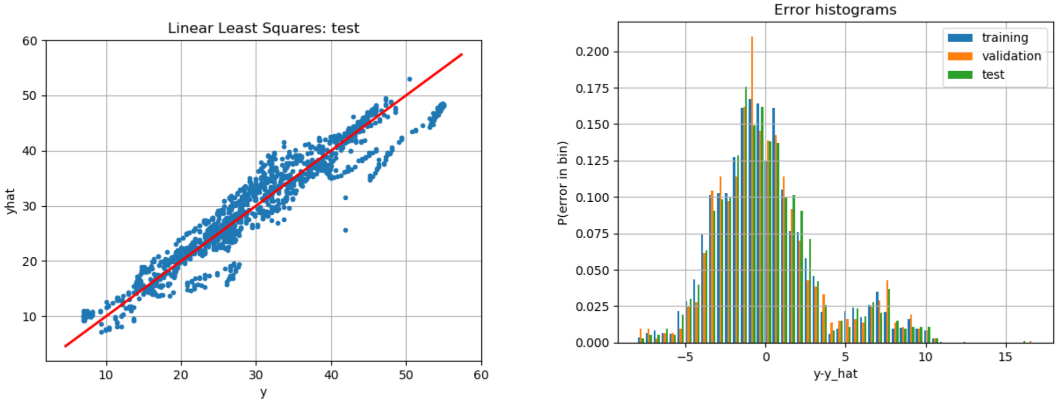
(b) $\hat{y}$ versus y with errorbars for the test dataset.

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

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

Figure 2: Gaussian Process Regression results

In this case, the new optimum values of r^2 and σ_ν^2 are respectively 35.7 and 0.002, as shown in Fig. 2a. They were used to produce Figs. 2b-2d.



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

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

Figure 3: Linear Least Squares results

5 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), $\mathbf{x}^T = [x_1, \ldots, x_F]$[1] stores the F regressors 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 eright vector $\mathbf{w}$ can be obtained in closed form as:

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

where $\mathbf{X}$ 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

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

Figure 3 shows the results obtained with LLS.

6 Comparison

It is evident, by comparing Figs. 1c and 2c with Fig. 3a, that with the Parkinson's dataset, Gaussian Process Regression is more precise than linear regression, and this is also confirmed by the estimation error histograms in Figs. 1d, 2d and 3b.

[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, 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	3.01×10^{-13}	3.244	10.519	0.989
	Validation	0.134	3.299	10.892	0.989
	Test	0.133	3.266	10.674	0.989
GPR (3 features)	Training	-0.001	1.260	1.588	0.998
	Validation	-0.056	1.988	3.957	0.996
	Test	-0.070	1.797	3.234	0.997
GPR (8 features)	Training	-0.025	0.856	0.733	0.999
	Validation	-0.133	1.922	3.710	0.996
	Test	-0.139	1.840	3.405	0.997

Table 2: Numerical comparison between GPR and LLS.

7 Conclusions

It is necessary to give particular attention to values of error standard deviation. Indeed, LLS provides an error on the test set of around 3.2 points, which means that most of the times the regression error is around 3-6 points. Whereas, GPR provides an error on the test set of around 1.8 points in both cases, which means that most of the times the regression error of total UPDRS is around 2-4 points. From a medical point of view, these results are acceptable both in case of linear regression and in case of Gaussian Process regression, since the values of error standard deviations are lower than the standard deviation of true total UPDRS values stored in the test set, which is around 10 points. However, it is important to recall that it is a prediction for medical purpose. Thus, the accuracy of results is a factor which cannot be negligible at all. For this reason, Gaussian Process Regression may be preferred to the linear one.

In addition, a further comment needs to be done about the 5 more features used to perform again GPR. As it can be noted from the main statistical properties of Table 2, there is not an evident difference between GPR performed by using only motor UPDRS, age and PPE and GPR performed by using also other dysphonia measures. Therefore, in conclusion, it can be claimed that the most relevant feature to use to regress total UPDRS is necessarily motor UPDRS, whereas voice parameters do not add relevant information to regression.

References

[1] Athanasios Tsanas, Max A. Little, Patrick E. McSharry, and Lorraine O. Ramig. Accurate telemonitoring of parkinson's disease progression by non-invasive speech tests. `https://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring`.

[2] A. Tsanas, M.A. Little, P.E. McSharry, and L.O. Ramig. Accurate telemonitoring of parkinson's disease progression by non-invasive speech tests. *IEEE Transactions on Biomedical Engineering*, 57(4):884–893, April 2010.