Report 1: Regression on Parkinson's disease data

Marco Colocrese, s301227,
ICT for Health attended in A.Y. 2021/22

November 7th 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 should 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.

In the following, linear regression is used, with three different algorithms, and it is applied on the public dataset that can be downloaded at [1].

2 Data analysis

The 22 features available in the dataset are listed in table 1: of these, subject ID and test time are removed, total UPDRS is considered as regressand and the remaining 19 features are used as regressors. In particular, regressors include many voice parameters (jitter and shimmer measured in different ways, Noise to Harmonic Ratio NHR, etc) and motor UPDRS. The number of points in the dataset is 5875; data are shuffled and the first 75% (74% for GAM due to compatibility with matrix dimensions and the algorithm implementation) of the points are used to train the linear model, and the remaining 25% (26% for GAM) are used to test its performance. Each regressor X_m[1] is normalized using its mean μ_m and standard

[1]Capital letters are used to identify random variables, small letters instead identify measured values: X_m is the random variable, x_{nm} is the n-th measured value of X_m.

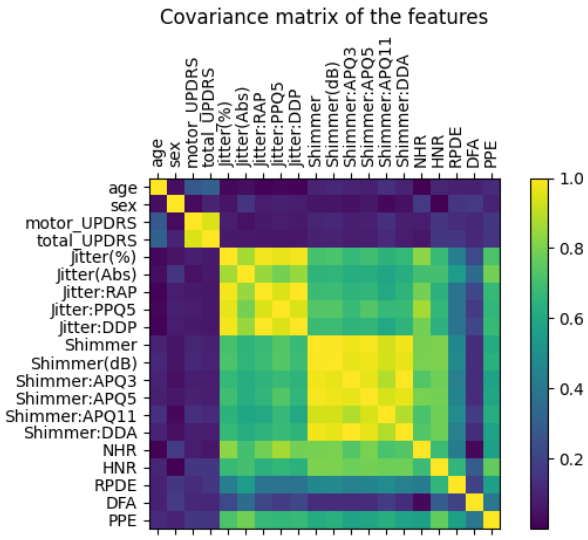


Figure 1: Covariance matrix of the features

deviation σ_m, measured on the training dataset:

$$X_{m,N} = \frac{X_m - \mu_m}{\sigma_m}. \tag{1}$$

Similarly, the regressand is normalized

$$Y_N = \frac{Y - \mu_Y}{\sigma_Y}. \tag{2}$$

Figure 1 shows the measured covariance matrix for the entire normalized dataset: correlation 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.

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

3 Linear regression

The model assumed in linear regression is

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

where Y is the regressand (total UPDRS), $\mathrm{X}^T = [X_1, \ldots, X_F]$[2] is the row vector that stores the F regressors (random variables) and $\mathbf{w}^T = [w_1, \ldots, w_F]$ is the weight vector to be optimized. The optimum vector $\mathbf{w}$ is the one that minimizes the mean square error

$$e(\mathbf{w}) = \mathbb{E}\left\{\left[Y - \mathrm{X}^T \mathbf{w}\right]^2\right\}. \tag{4}$$

The algorithms described in the following sections use the normalized training dataset: the measured values of regressors are stored in matrix $\mathbf{X}_N$ (size 4406x19 for SG and 4352x19 for GAM) and the corresponding measured values of the normalized regressand are in vector $\mathbf{y}_N$.

3.1 Linear Least Squares (LLS)

Linear Least Squares (LLS) directly finds $\mathbf{w}$ by setting to zero the gradient of $e(\mathbf{w})$:

$$\hat{\mathbf{w}} = \left(\mathbf{X}_N^T \mathbf{X}_N\right)^{-1} \mathbf{X}_N^T \mathbf{y}_N \tag{5}$$

Given $\hat{\mathbf{w}}$, the normalized regressand for the test dataset is estimated as

$$\hat{y}_{N,te} = \mathbf{X}_{N,te}^T \mathbf{w} \tag{6}$$

where $\mathbf{X}_{N,te}$ has size 1469x19 for SG and 1523x19 for GAM. The denormalized regressand is instead

$$\hat{y}_{te} = \sigma_Y \hat{y}_{N,te} + \mu_Y. \tag{7}$$

Figure 2 shows the results obtained with LLS, using denormalized data.

3.2 Stochastic gradient algorithm with Adam optimization (SG)

Minimization of eq. (4) can be obtained iteratively using the stochastic gradient algorithm. At the i-th step

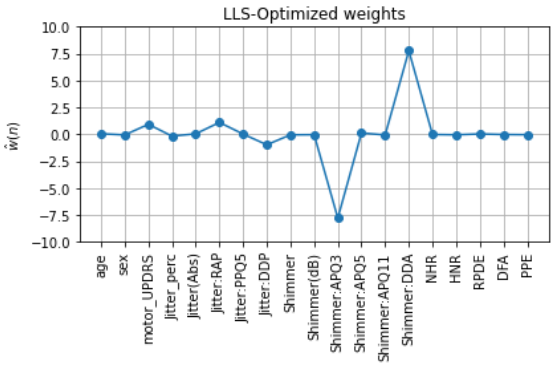
$$\hat{\mathbf{w}}_{i+1} = \hat{\mathbf{w}}_i - \gamma \nabla f_i(\mathbf{x}(i))$$

where

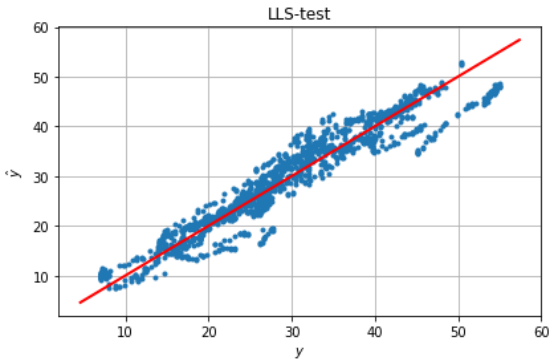
$$f_i(\mathbf{x}(i)) = [\mathbf{x}^T(i)\hat{\mathbf{w}}_i - y(i)]^2, \quad \nabla f_i(\mathbf{x}(i)) = 2[\mathbf{x}^T(i)\hat{\mathbf{w}}_i - y(i)]\mathbf{x}(i)$$

being $\mathbf{x}^T(i)$ the i-th row of matrix $\mathbf{X}_N$ and $y(i)$ the i-th element of the regressand vector $\mathbf{y}_N$ for the normalized training dataset. Adam optimization was applied and therefore $\nabla f_i(\mathbf{x}(i))$ was substitued with its "mean" value, according to [2], using exponential decay rates β_1=0.9

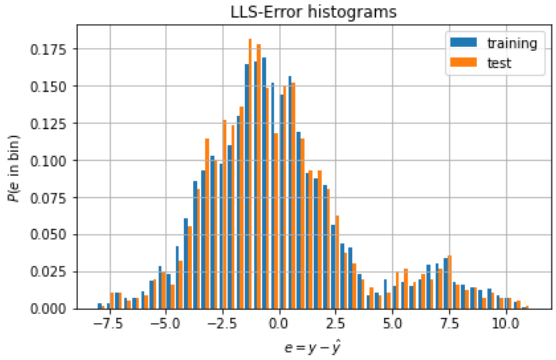
[2]Note that $\mathbf{x}$ is a column vector, and x^T is its transpose



(a) $\hat{\mathbf{w}}$.

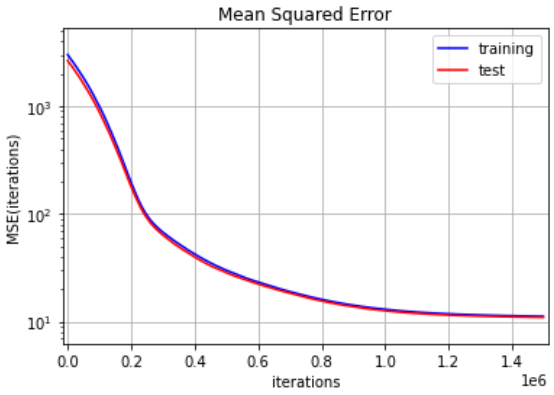


(b) $\hat{y}$ versus y for test dataset.

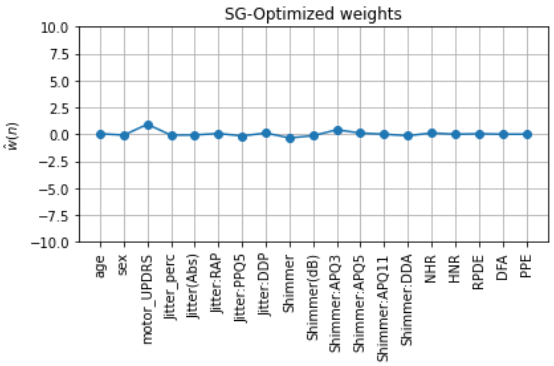


(c) Histogram of $y - \hat{y}$ for training and test datasets.

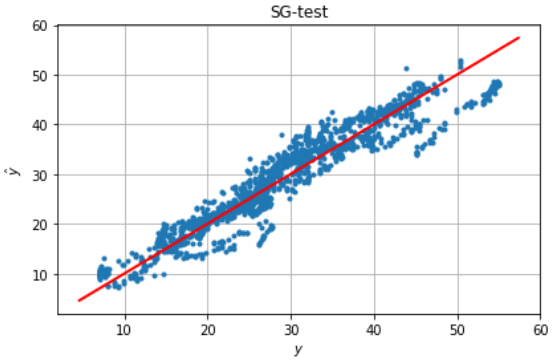
Figure 2: Results for Linear Least Squares (LLS).



(a) $\mathbb{E}\{(y - \mathbf{x}\hat{\mathbf{w}}_L)^2\}$ for training and test datasets as a function of the iteration step i.



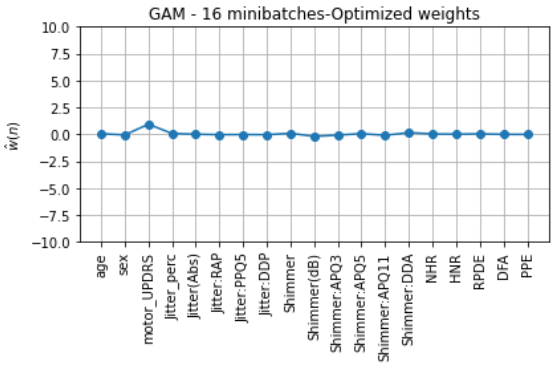
(b) $\hat{\mathbf{w}}$.



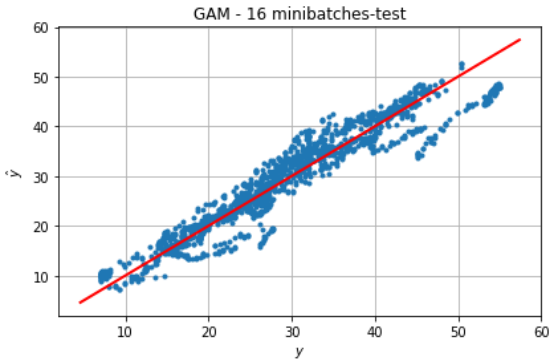
(c) $\hat{y}$ versus y for test dataset.

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

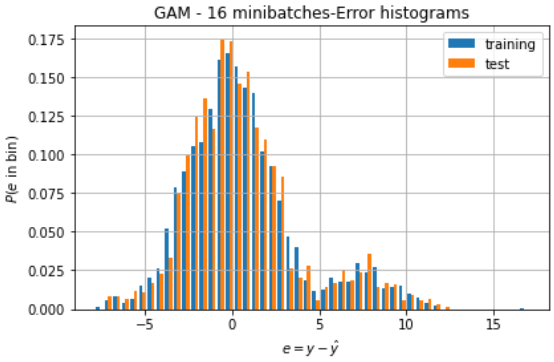
Figure 3: Results for stochastic gradient (SG).



(a) $\hat{\mathbf{w}}$.



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



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

Figure 4: Results for gradient algorithm with minibatches (GAM).

(for the mean) and β_2=0.999s (for the mean square value). The used value of the learning coefficient γ was 1e-5 and the stopping condition for the algorithm was based on the error —y-$\mathbf{X}_N$@$\hat{\mathbf{w}}$: the algorithm stopped when this value was the minimum, having noticed that there was only one global minimum among the analyzed iterations. The resulting number of iterations was 1.493.634 (339 epochs).

Results are shown in Figure 3.

3.3 Gradient algorithm with minibatches (GAM)

The gradient algorithm with minibatches iteratively finds $\hat{\mathbf{w}}$ dividing the training dataset into minibatches. Being $(\mathbf{y}_i, \mathbf{X}_i)$ the regressand and regressor values of the i-th minibatch (from the normalized matrix $\mathbf{X}_N$), the algorithm finds the next solution $\hat{\mathbf{w}}_{i+1}$ as

$$\hat{\mathbf{w}}_{i+1} = \hat{\mathbf{w}}_i - \gamma \left(2\mathbf{X}_i^T\mathbf{y}_i + 2\mathbf{X}_i^T\mathbf{X}_i\hat{\mathbf{w}}_i\right)/L_B \tag{8}$$

where γ is the learning rate and L_B is the number of samples in a minibatch. The minibatch size L_B was set equal to 272 because it resulted to give the same (or better) results as different attempts made with smaller minibatches, γ was set equal to 1e-5, and the stopping condition was to find the iteration with the minimum error —y-$\mathbf{X}_N$@$\hat{\mathbf{w}}$ resulting in 2912 iterations (182 epochs). In this case adam optimization was not used.

Results are shown in Figure 4.

3.4 Numerical comparison

The regression error $e = y - \hat{y}$ for the training and test datasets can be seen as a random variable, which can be statistically described. Table 2 lists its main statistical parameters: mean μ_e, standard deviation σ_e, mean square value MSE, and coefficient of determination R^2, for the three analyzed methods.

	Dataset	μ_e	σ_e	MSE	R^2
LLS	Training	7.33×10^{-15}	3.264	10.653	0.989
	Test	-6.24×10^{-2}	3.226	10.401	0.989
SG	Training	-7.88×10^{-15}	3.353	11.239	0.988
	Test	-4.67×10^{-2}	3.318	11.004	0.988
GAM	Training	5.88×10^{-15}	3.309	10.949	0.989
	Test	-8.46×10^{-2}	3.257	10.614	0.989

Table 2: Comparison among the three methods: LLS, SG and GAM.

4 Conclusions

Being the Unified Parkinson's Disease Rating Scale widely used and extensively tested for its clinimetric properties, Regression methods can represent a great relief for medical doc-

tors. Parkinson's is a neurological movement disorder which progress slowly, therefore it is important to have a way to constantly and objectively (without neurologist's subjectivity) evaluate UPDRS values in a reliable way. Researches say about 89% of people affected by Parkinson's will have speech and voice symptoms. This observation obtains a fundamental value in comparison to other symptoms such as resting tremor, rigidity, akinesia, bradykinesia (which are due to loss of function of the basal ganglia involved in the coordination of body movement) which are clearly more difficult to be evaluated without a medical doctor intervention. Speech symptoms are the simpliest symptoms to be collected and analyzed as described in this report. Furthermore, to have more measurements is useful to neurologist who can optimize when and how much levodopa the patient should take in order to increase dopamine in the brain and decrease motor dysfunction. A correct and recurrent evaluation is fundamental because as the disease progress, more dopaminergic neurons in the substantia nigra are lost and conversion of levodopa to dopamine decreases; moreover as the movements become slower and slower, levodopa stays more and more in the stomach without reaching the intestine where it should be absorbed.
Examining the optimized weights (in particular referred to the LLS), the presence of collinearity emerges caused by the correlation between different features (such as Shimmer:APQ3 and DDA). From the analysis of the comparison between $\hat{y}$ and y (figures 2.b, 3.c and 4.b) it is clear that the applied regression methods perform well for y values smaller than 40. For values greater than 40 it can be observed that predicted values are more distant from the real values. This observation also explains the presence of two gaussian pdf in the histograms of $e = y - \hat{y}$: the one with the higher probabilities is referred to values<40, in fact errors are close to zero; the other one for errors related to y>40 (except for some outliers observable in the range 20<y<30). Observing Table 2 it can be noticed that LLS and GAM give almost the same results, SG slightly worst. In particular R^2 is pretty much the same for all methods (0.989 for LLS and GAM and 0.988 for SG): it is a clear indicator of the quality of the regressions. The error means are zero in the training subset. However error means are not zero (but slightly different) for the test set due to the fact that total_UPDRS means in the test subset might not be exactly zero, after normalization. The error standard deviations, being almost equal (between test set and training set) for the three methods show the absence of overfitting. This particular aspect is also proved by the shapes of the histograms relating to training test sets which are early the same. The error standard deviation makes possible to estimate that the regression error is almost always between 3-6. This value, compared to the total_UPDRS scale (regressand), means that the results might be accepted by medical doctors. By the reliability point of view, all three methods are valid.

References

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

[2] D.P. Kingma, J. Ba, "Adam: A Method for Stochastic Optimization", *arXiv:1412.6980 [cs.LG]*