DATASET

Machine learning for medical data course from Mme. N. Sokolovska

Yunfei Zhao & Lise Le Boudec

M2A - Sorbonne Université

Sommaire

Introduction and context

Chosen dataset

SCIENCES SORBONNE UNIVERSITÉ

The Data

- Kaggle dataset [3] *"Classify gestures by reading muscle activity"*
- 64 features (8 sensors with 8 consecutive readings) and 1 class. Recorded under 200HZ that takes 40ms for each line of recording.
- 120 second recording for each gesture, nearly 3000 recordings for each gesture in different CSV files respectively.
- 4 classes for each considered gestures : rock - 0, scissors - 1, paper - 2, ok - 3. 1 file per class.

Figure 1: Data Structure illustration.

Objective

Predict the gesture based on the recorded muscle activity

Practical use

Learn to a mind-controlled robot arm to recognize some gestures from sensors on the muscles to replace a hand

Data preprocessing

preprocessing

SCIENCES SORBONNE UNIVERSITÉ

Steps

- Merge the datasets into one dataset with all inputs
- 80% training data and 20% testing data
- In training data, we use 20% for validation with 2 folds cross-validation.
- Standardization of the data
- PCA

Metrics for evaluation

Metrics

SCIENCES SORBONNE UNIVERSITÉ

F1-score

- For class sizes that are not balanced.
- Take into account of Precision and recall. (More credible than accuracy) $f1 = \frac{2PR}{P+R}$
- Micro integration for multi-classes cases.

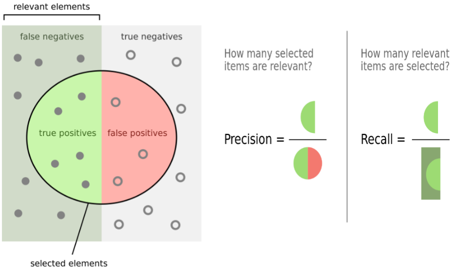


Figure 2: Precision recall curve, f1 score and AP.[2]

Unsupervised learning Method applied on the dataset

Principal component analysis (PCA)

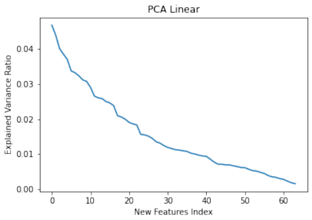


Figure 3: PCA with Linear Kernel

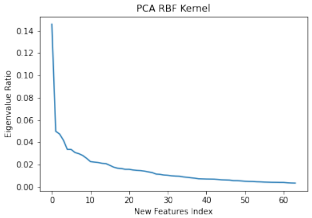


Figure 4: PCA with rbf Kernel

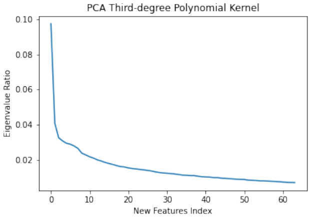


Figure 5: PCA with polynomial Kernel

Conclusion: all variables were important in out dataset. According to the observed results, we decided not to remove any of the features.

Supervised learning Method applied on the dataset

Methodology

- Compute a grid search on a set of chosen parameters on the training set
- Keep the best model according to the metric
- Compute the PR curves on the test set for comparison
- 6 models tested : Random Forest, SVM, Logistic Regression, LDA, QDA and MLP

Hypothesis

In these models, all consecutive readings are considered as independent variables.

Random Forest

SCIENCES SORBONNE UNIVERSITÉ

⇒ Parameters : criterion, max_depth, n_estimators

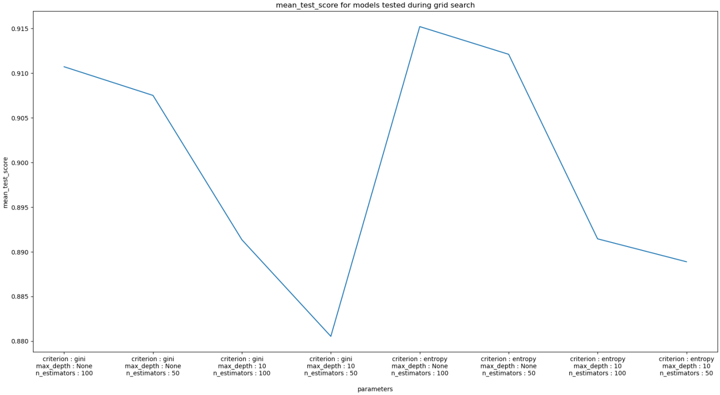
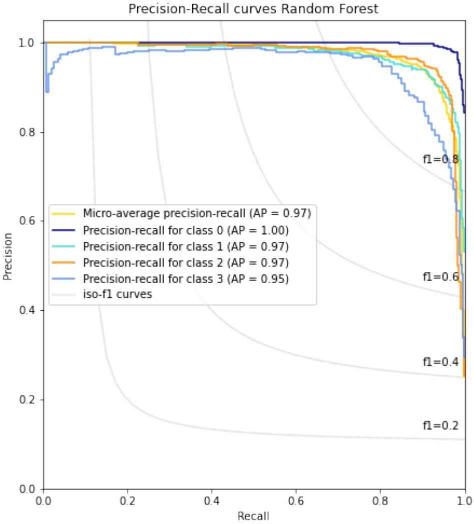


Figure 6: Grid search on Random Forest model

Random Forest

Final model :

- criterion: entropy
- max depth: None
- number of estimators: 100



SVM

$\Rightarrow$ Parameters : C, kernel

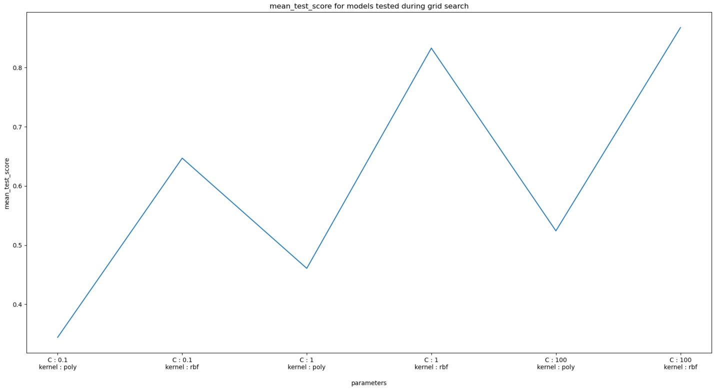
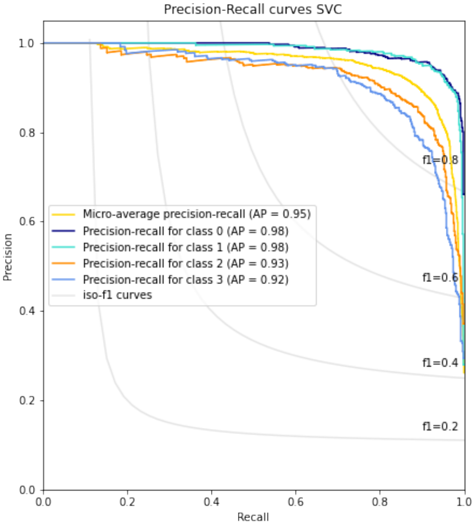


Figure 8: Grid search on SVM model

SVM

Final model :

- C: 100
- kernel: rbf



Logistic Regression

SCIENCES SORBONNE UNIVERSITÉ

⇒ Parameters : C, penalty

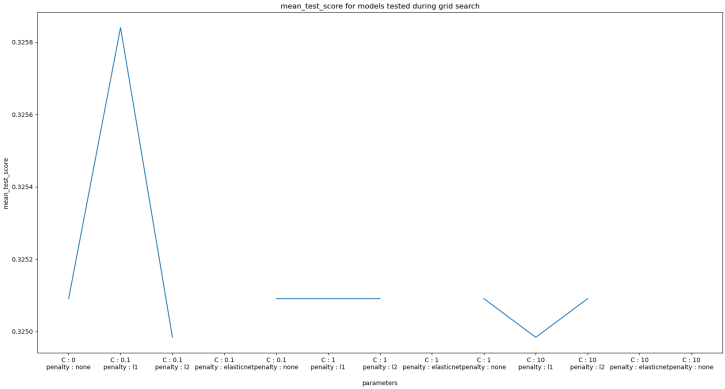
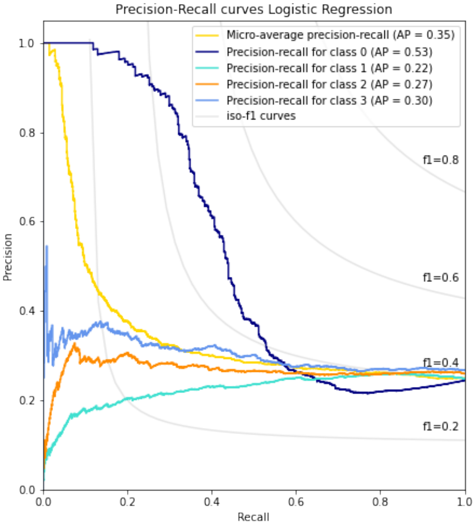


Figure 10: Grid search on logistic Regression model

Logistic Regression

Final model :

- penalty: l1



LDA/QDA

SCIENCES SORBONNE UNIVERSITÉ

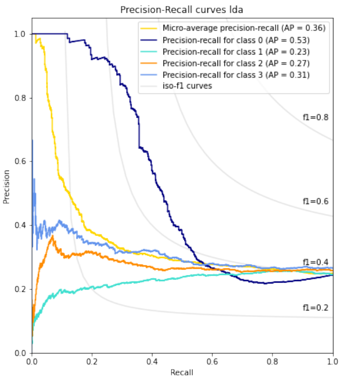


Figure 12: Precision recall curve for lda classifier on test dataset.

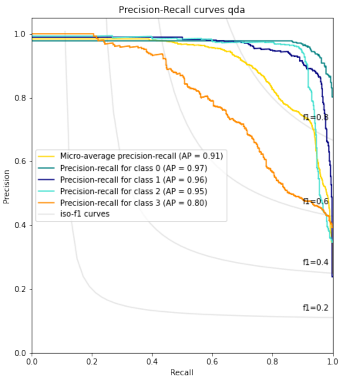


Figure 13: Precision recall curve for qda classifier on test dataset.

Neural Network

⇒ Parameters : alpha, hidden_layer_size

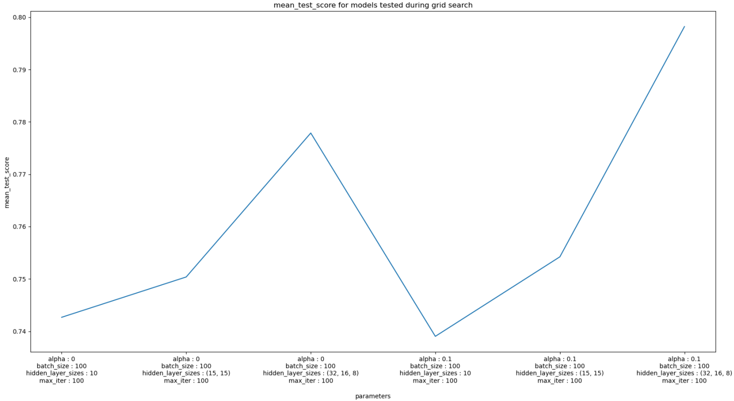


Figure 14: Grid search on sk learn MLP model

Neural Network

- batch size: 100, hidden layer sizes: (32, 16, 8).
- relu, batch noramlisation, dropout.

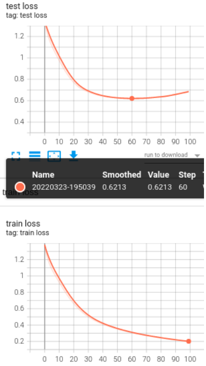


Figure 15: Loss curve on train dataset and test dataset.

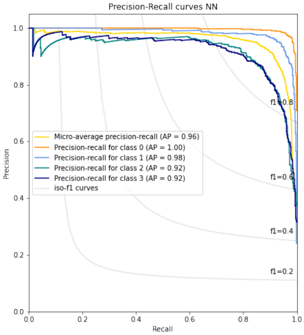


Figure 16: Precision recall curve for NN classifier on test dataset.

Improvement

Long short-term memory

⇒ Long short-term memory (LSTM) is an architecture of neural network used to deal with sequences [1].

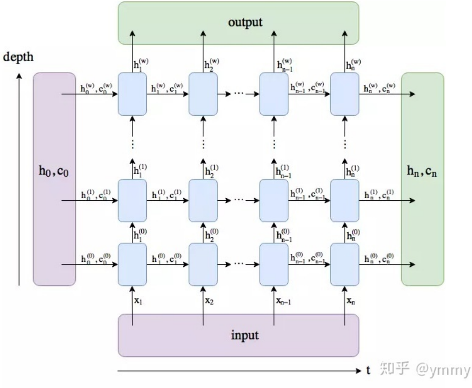


Figure 17: Schema for multi-layers LSTM.

Hypothesis

In this model, we consider the data as sequences of the features, i.e. we have sequences of size 8 for each of the 8 sensors (features).

Long short-term memory

- input size: 8, hidden vector sizes: 16, projection size: 4.
- softmax, batch size:100, dropout:0.2, number of layers: 4.

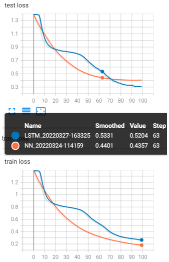


Figure 18: Loss curve on train dataset and test dataset.

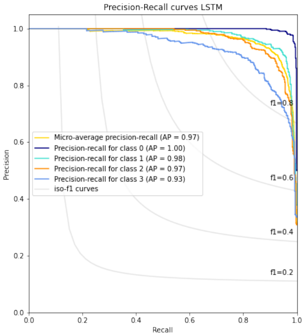


Figure 19: Precision recall curve for NN classifier on test dataset.

The Idea

SCIENCES SORBONNE UNIVERSITÉ

Our data are taken from 8 sensors with 8 timesteps. What if we compute one model per timestep and keep the major prediction ? ("Bagging") Example with the Random Forest model

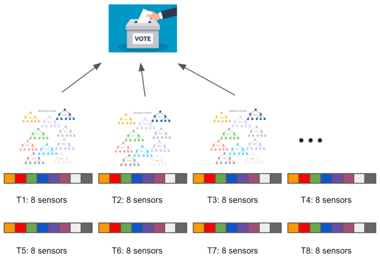


Figure 20: Schema of random forest with bagging method

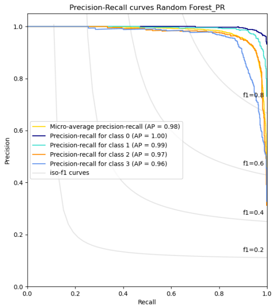


Figure 21: PR curve with the bagging method

Comparison of the methods

- LDA & Logistic Regression : bad results, not adapted to the data
- QDA, Random Forest, SVM, MLP and LSTM have all a average precision above 0.9
- Best models on average are LSTM and Random forest but they both are difficult to interpret (0.97).
- New Bagging Random forest get the highest average precision (0.98).

References

References I

[1] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-term Memory". In: *Neural computation* 9 (Dec. 1997), pp. 1735–80. doi: 10.1162/neco.1997.9.8.1735.

[2] wikipedia. *F-score*. https://en.wikipedia.org/wiki/F-score. 2019.

[3] Kirill Yashuk. *Classify gestures by reading muscle activity*. https://www.kaggle.com/datasets/kyr7plus/emg-4. 2019.

Appendix

Preprocessing results

preprocessing	None	Norm with 1 measurements/8	Norm	PCA + Norm	indep measurements : nb obs * 8
svm	0.87	0.68	0.91	X	X
random forest	0.93	0.71	0.93	0.63	0.73
logistic regression	0.34	0.32	0.35	0.41	0.28
lda	0.36	0.32	0.36	0.41	0.28
qda	0.93	0.68	0.93	0.64	0.69
sk mlp	0.90	0.74	0.90	0.65	0.74
torch mlp	0.90	X	0.85	0.81	X
torch lstm	0.88	X	0.86	0.90	X

Table 1: Different preprocessing results

Computing time

method	PCA	PCA rbf	SVM	RF	LR	LDA	QDA	sk MLP	torch MLP	LSTM	Bagging RF
training time (s)	0.9	53.9	20.1	4.5	0.6	0.08	0.03	7.7	16	102	12.0
execution time on test dataset (s)	0.5	0.6	1.4	0.05	0.002	0.001	0.006	0.002	0.014	0.2	0.4

Table 2: Training and executing time of the methods

Other ideas

- Bayesian Neural Networks for more explainability on the model
- Do more preprocessing on the data : take the average on each sensors on the 8 measurements for example
- k-means, knn...