MACHINE LEARNING PROJECT

Class: Advanced biostatistics

Submitted to : Aitor Gonzale & Pauline Brochet, PhD

University of Aix-Marseille, M2 DLAD 2021-2022

Summery:

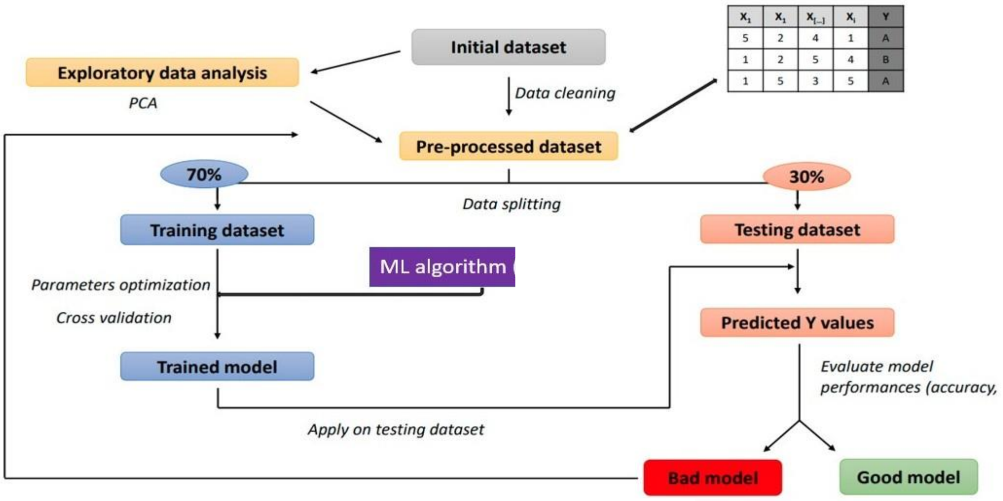
In this project, we were asked to experiment with a real world dataset, and to explore how machine learning algorithms can be used for prediction tasks. We were expected to gain experience using a common data-mining and machine learning library, Weka, and were expected to submit a report about the dataset and the algorithms used. After performing the required tasks on then chosen dataset, herein lies my final report.

Keywords: Machine Learning, Pattern Recognition, Classification, Supervised learning, Artificial Intelligence.

I- Introduction :

Machine learning is a sub-domain of computer science which evolved from the study of pattern recognition in data, and also from the computational learning theory in artificial intelligence. It is the first-class ticket to most interesting careers in data analytics today[1]. As data sources proliferate along with the computing power to process them, going straight to the data is one of the most straightforward ways to quickly gain insights and make predictions.

Machine Learning can be thought of as the study of a list of sub-problems, viz: decision making, clustering, classification, forecasting, deep-learning, inductive logic programming, support vector machines, reinforcement learning, similarity and metric learning, genetic algorithms, sparse dictionary learning, etc. Supervised learning, or classification is the machine learning task of inferring a function from a labeled data [2]. In Supervised learning, we have a training set, and a test set. The training and test set consists of a set of examples consisting of input and output vectors, and the goal of the supervised learning algorithm is to infer a function that maps the input vector to the output vector with minimal error. In an optimal scenario, a model trained on a set of examples will classify an unseen example in a correct fashion, which requires the model to generalize from the training set in a reasonable way. In layman’s terms, supervised learning can be termed as the process of concept learning, where a brain is exposed to a set of inputs and result vectors and the brain learns the concept that relates said inputs to outputs. A wide array of supervised machine learning algorithms are available to the machine learning enthusiast, for example Neural Networks, Decision Trees, Support Vector Machines, Random Forest, Naïve Bayes Classifier, Bayes Net, Majority Classifier[4,7,8,9] etc., and they each have their own merits and demerits. There is no single algorithm that works for all cases, as merited by the No free lunch theorem [3]. In this project, we try and find patterns in a dataset [2], which is a sample of males in a heart-disease high risk region of South Africa, and attempt to throw various intelligently-picked algorithms at the data, and see what sticks.



X_1	X_1	$X_{[...]}$	X_i	Y
5	2	4	1	A
1	2	5	4	B
1	5	3	5	A

Steps to achieve a ML model

II- Materials and methods

We used Python3 3.8, SciKit Learn5 1.0.2, pandas7 1.3.5, NumPy8 1.22.1 on Pycharm IDE.

The data collected was assembled to create a unified feature and label dataframe. Before creating the training and testing sets we started by the dimensionality reduction for data exploration and visualization. Finally, we mixed and split the data into training and testing sets with a 30% testing and 70% training ratio to perform the chosen algorithms.

A. Code implementation (python)

Github link: https://github.com/Chaima-Bouchenak/projectML

III- Discussion

A. Data set description

The dataset used in our Project is the gene expression cancer RNA-Seq Data Set .

This dataset is originally from the University of Genoa. This collection is part of the RNA-Seq (HiSeq) PANCAN data set, a random extraction of gene expressions of patients having different types of tumor (details below).

The objective of the dataset is to predict which kind of cancer a patient has, based on certain gene expression measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database.

You can learn more about this dataset description at the UCI Data Repository :
https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq

The CSV files contain: "**data**" of genetic expression of 20531 genes of 801 patients & "**labels**" of the type of tumor the patients had. Globally, there is 5 types of cancer:

-> BRCA: Breast Invasive Carcinoma

-> LUAD: Lung Adenocarcinoma

-> PRAD: Prostate Adenocarcinoma

-> KIRC: Kidney Renal Clear Cell Carcinoma

-> COAD: Colon Adenocarcinoma

B. Data set exploration

The dataset has 801 rows and 20532 columns and tumor classes are distribute as following:

BRCA 300, KIRC 146, LUAD 141, PRAD 136 and COAD 78

C. Data set splitting

In a supervised model, we first processed by dividing our dataset into two subsets to create a model that generalizes well to new data:

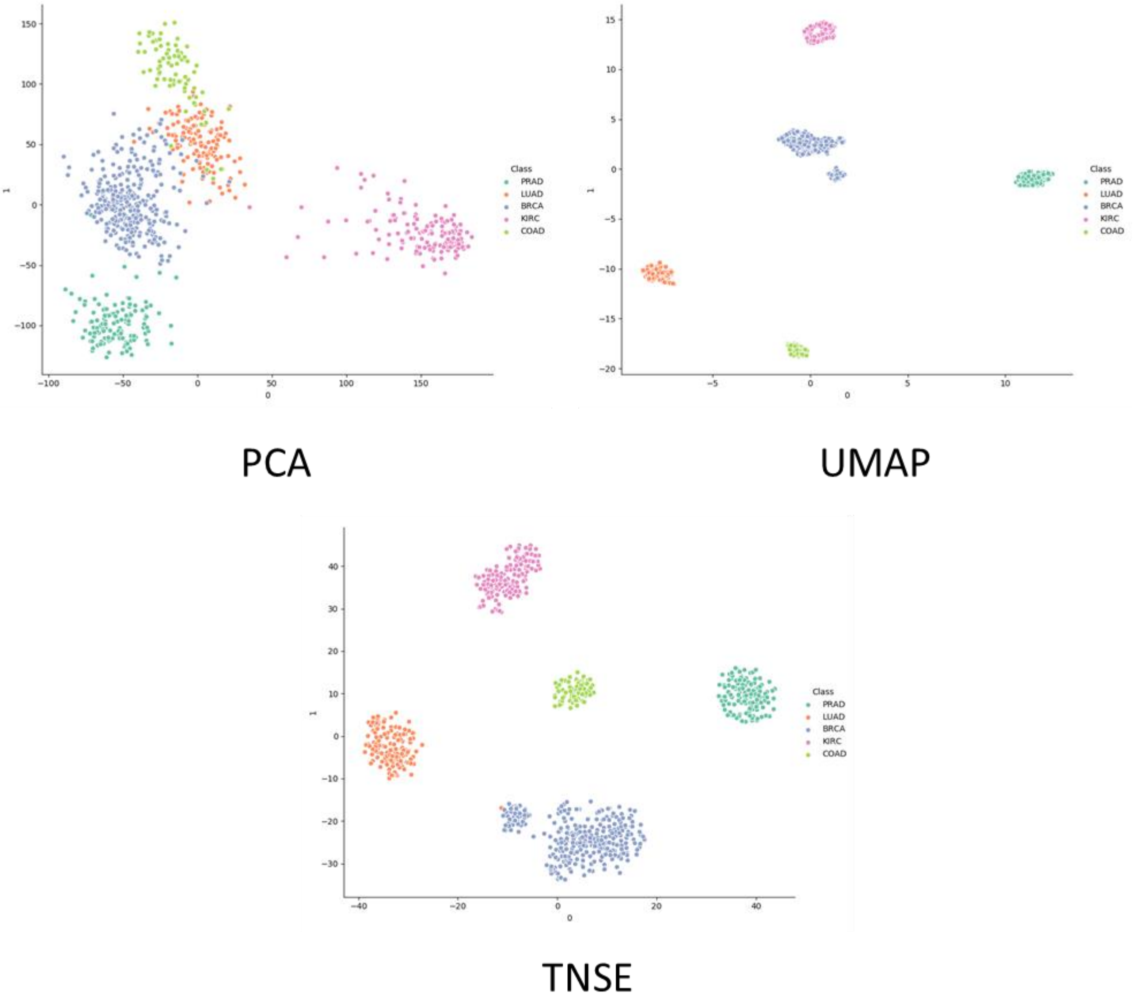
Training set, a subset to train a model (70%)

Test set, a subset to test the trained model (30%)

During this process, we make sure that the test set is representative of the data set as a whole (I.e same characteristics)

D. Dimensionality Reduction discussion

Many Machine Learning problems involve thousands of features, having such a large number of features bring along many problems, like making the training extremely slow or difficulty to find a good solution**. Dimensionality Reduction** is the process of reducing the number of features to the most relevant ones, it filter out some of the noise present and some of the unnecessary details and provide a meaning insights for data visualization.



PCA

UMAP

TNSE

Figure 1: Dimensionality reducing technics

PCA was not able to do such a good job in differentiating the signs. The main drawback of PCA is that it is highly influenced by outliers present in the data. PCA is a linear projection, which means it can't capture non-linear dependencies, its goal is to find the directions (the principal components) that maximize the variance in a dataset.

t-SNE on the other side does a better job (preserve topology neighbourhood structure) as compared to PCA when it comes to visualising the different patterns of the clusters. Similar labels are clustered together, even though there are big agglomerates of data points on top of each other, certainly not good enough to expect a clustering algorithm to perform well.

UMAP outperformed t-SNE and PCA, we can see mini-clusters that are being separated well. It is very effective for visualizing clusters or groups of data points and their relative proximities. However, for this use case certainly not good enough to expect a clustering algorithm to distinguish the patterns.UMAP is much faster than t-SNE.

E. Algorithm’s discussion

Algorithm	Precision	Variation
RandomForestClassifier	0.989286	0.032143
Logistic Regression	0.**998214**	0. 005357
Gaussian Naive Bayes	0. 978786	0. 002999
Decision Trees Classifier	0. 988315	0. 003348
Support Vector Machine	0. 995340	0. 005299

The model with the best precision is: LR

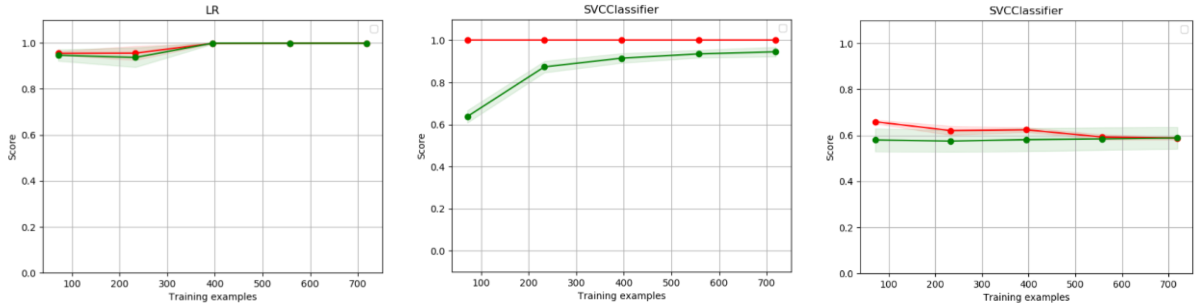


Figure 2: Examples of measurement of the accuracy for train (red) and validation data (green). A. LR algorithm (good fitting) B. SVM algorithm (Overfitting) C. SVM (Underfitting)

On figure 2, we can visualize the learning curves in accuracy for training and testing data. We see that within the first 400 examples the model reaches near 1 for accuracy, both training and testing. However for in the SVM model for example it shows signs of overfitting as expected from an accuracy of 99,53%. We noticed also that,
As the number of epochs increases, the score can go from underfitting to optimal to overfitting.

IV- Conclusion and perspective

In summary, in order to implement a ML algorithm, one can proceed in the following way by making sure to respect the settings of the hyperparameters of algorithms in particular the learning rate and the good preparation of the data set in advance.

To conclude with, we found out that all chosen ML algorithms allow for excellent prediction accuracy. In our case, the Multinomial logistic regression takes the lead by a very small margin.

V- References

[1] "Intro to Machine Learning | Udacity." Intro to Machine Learning | Udacity. Accessed April 27, 2016. https://www.udacity.com/course/intro-to-machine-learning--ud120

[2] "Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Edition. Datasets:Coronary Heart Disease Dataset." Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Edition. Accessed April 27, 2016. http://statweb.stanford.edu/~tibs/ElemStatLearn

[3] "No Free Lunch Theorems." No Free Lunch Theorems. Accessed April 27, 2016. http://www.no-free-lunch.org/.

[4] Hastie, Trevor, Robert Tibshirani, and J. H. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction: With 200 Full-color Illustrations. New York: Springer, 2001.

[5] "Weka 3: Data Mining Software in Java." Weka 3. Accessed April 27, 2016. http://www.cs.waikato.ac.nz/ml/weka/.

[6] Bozhinova, Monika, Nikola Guid, and Damjan Strnad. Naivni Bayesov Klasifikator: Diplomsko Delo. Maribor: M. Bozhinova, 2015.

[7] Schölkopf, Bernhard, Christopher J. C. Burges, and Alexander J. Smola. Advances in Kernel Methods: Support Vector Learning. Cambridge, MA: MIT Press, 1999.

[8] Norving, Peter, and Stuart Russel. Artificial Intelligence: A Modern Approach. S.l.: Pearson Education Limited, 2013.

[9] Witten, I. H., and Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques. Amsterdam: Morgan Kaufman, 2005.