Multi-class classification on gene expression data

Simone Daniotti[1] and Riccardo Castelli[2]

[1] *Physics Department, University of Milan*
[2] *Informatics Department, University of Milan*

September 26, 2019

Abstract

Your abstract goes here... ...

Contents

1 Dataset description

The dataset for this work is taken from UCI Machine Learning Repo (available at https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq) [1]; this is part of the RNA-Seq (HiSeq, a tool for measuring gene expression levels) PANCAN data set. It is a collection of gene expression levels of patients having different types of tumor: BRCA(breast), KIRC(kidney), COAD(colon), LUAD(lung) and PRAD(prostate). These data represent the quantity of gene information used in the synthesis of a functional gene product. For further information, we refer to [2].

2 Dataset Manipulation

2.1 Preliminary manipulation

Both dataset and labels can be downloaded in a csv ('comma separated value') format, then can be easily imported in a Pandas Dataframe (https://pandas.pydata.org/). The first column represents patient's ID: it has been removed because it is useless for our purposes.

2.2 Label handling

The labels are strings representing the five types of cancer. Learning models can be created using raw features(in the case of trees and forests), using Label Encoding or One-hot Encoding. Label Encoding creates a map between the string and an ordered sequence of natural numbers, from 0 to 4 in our case. One-hot encoding creates a single binary label for each class, changing the task of the learning model to a multi-label problem.

2.3 Train and Test Set

For training the net and then evaluating it, we split the dataset in training and test set using the Sklearn library *train_test_split* (https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html). We set the seed of the random split and the proportions between the two sets: test set is 0.15 of the entire database.

2.4 Class Weighting

3 Architectures

There are lots of books reviewing these architectures and concepts. Here we refer to [3] , [4] [5].

3.1 Support Vector Machines (SVM)

3.2 Decision Tree Classifier and Random Forests

Decision Trees are versatile Machine Learning algorithms that can perform both classification and regression tasks, and even multioutput tasks. They are particularly useful in treating with complex data, such as a dataset that can hardly be represented by a vector in a multi-dimensional space: this is not our case, but we think it is useful to approach our problem with more simple models and evaluating them before *deeping* in more difficult models. Tree Classifiers have the structure of an *ordered and rooted tree*. It is *ordered* because the children of any internal node are numbered consecutively, and *rooted* because splitting starts from only one node. From that node, the model is built following the attribute selection measure. Attribute selection measure is a heuristic for selecting the splitting criterion that partition data into the best possible manner.

Scikit-Learn uses the *Classification And Regression Tree (CART) algorithm* to train Decision Trees.It works as follows: it first splits the training set in two subsets using a single feature k and a threshold t_k . How does it choose k and t_k? It searches for the pair (k, t_k) that produces the purest subsets. The cost functions that the algorithm tries to minimize are different: most used are *Gini Impurity* and *Entropy*, tunable in Scikit-learn by the hyperparameter *criterion*. Entropy hyperparameter measures Shannon's Entropy, a concept taken from information theory. Each leaf of the tree corresponds to a possible classification label, so inserting datas of a patient from the root, the model *spits* its classification. This can be a modality for building a predictor.

If you aggregate the predictions of a group of predictors (regulated by certain rules, such as majority rule..), you will often get better predictions than with the best individual predictor. A group of predictors is called an *ensemble*; thus, this technique is called *Ensemble Learning*, and an Ensemble Learning algorithm is called an *Ensemble method*. Training a group of Decision Tree Classifiers and gathering into one single predictor is called a *Random Forest*. The Random Forest algorithm introduces extra randomness when growing trees; instead of searching for the very best feature when splitting a node (as in the tree case), it searches for the best feature among a random subset of features. This results in a greater tree diversity.

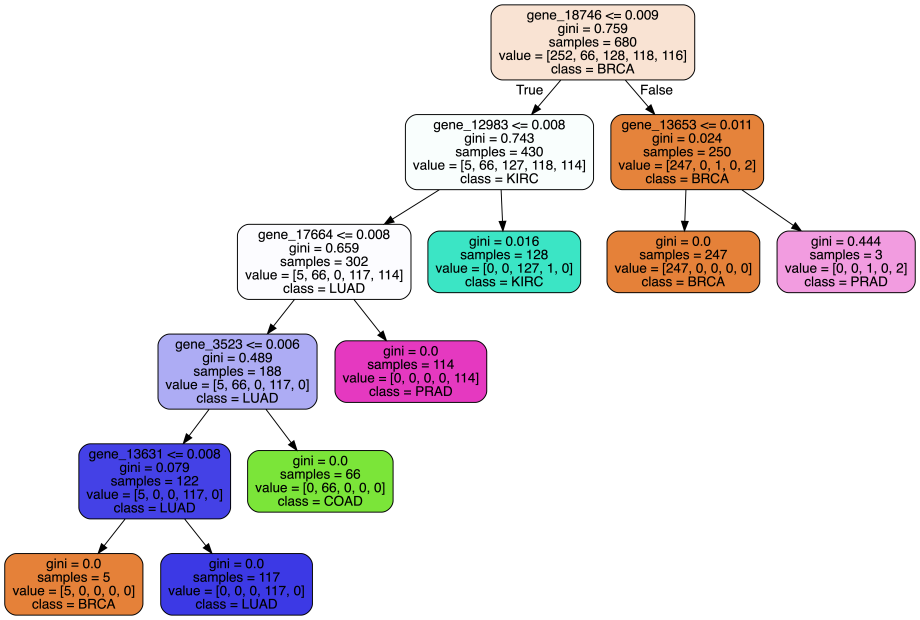

Figure 1: Shape of a tree classifier, with hyperparameters tuned by tecniques explained below.

3.3 Deep Learning

3.3.1 PyTorch

3.3.2

4 Parameter optimization and Validation

4.1 Cross-Validation

4.2 Grid Search CV (and Random Search)

4.3 Architecture setting

4.3.1 Parameters

4.3.2 Optimizers

4.3.3 Losses

5 Models Evaluation

5.1 PCA and Permutation Importance

5.2 Metrics

6 Conclusions and Outlook

References

[1] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[2] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113, 2013.

[3] Aurélien Géron. *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems.* " O'Reilly Media, Inc.", 2017.

[4] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[5] John Hertz, Anders Krogh, Richard G Palmer, and Heinz Horner. Introduction to the theory of neural computation. *Physics Today*, 44:70, 1991.