AIX-MARSEILLE UNIVERSITÉ
LE MASTER BIOINFORMATIQUE, DLAD
2018-2019

Compte rendu du Projet d'évaluation : Compléments de mathématiques pour la bioinformatique

Alexey Solovyev

Enseignants :
Quentin Ferré & Aitor González

Introduction

The purpose of this project is to demonstrate the methods of machine learning. Real figures are used as the source data (input data). The dataset is to be found at https://archive.ics.uci.edu/ml/machine-learning-databases/00401/ . The data contain the levels of gene expression and the different types of diagnosed cancer. The input data has 801 instances of observations and 20,531 attributes. Each instance corresponds to the category of the tumor, which is stored in a separate file. The data are cleared from insignificant columns — genes that exhibit zero expression in all experiments. The data is normalized by the column, so the average value of gene expression in the column is zero and the variance is one. In the present work will be used: Logistic regression, Cross-validation, Regularization, Learning curve, Grid search, Clustering, Principal component analysis (PCA), Gradient boosting and Artificial neural networks (ANN). Based on the data obtained, the author will conclude which method is best used to solve the given problem – to predict the cancer class.

Materials and Methods

Python and PyCharm was used to write the program code. For the implementation of the calculations next packages were used: sys, os, numpy, pandas, keras, sklearn, xgboost, matplotlib, seaborn.

1. Logistic regression

To solve this problem a script «Examples_scikit_learn.intro()» was written. The author of the work uses the separation of 70% and 30% for training and test data, respectively. For the purpose of reproducibility and verification of results, a random seed is established 42. When changing the value of a random seed or its removal, the results may differ from those shown in the work.
Using the sklearn.linear_model package, we train the model and test our model on test data. The script returns cross-tabulation and an accuracy level (score) calculated on test data that the model has not seen before. For input data we obtained the following values predicted by model:

Class	BRCA	COAD	KIRC	LUAD	PRAD
BRCA	90	0	0	0	0
COAD	0	28	0	0	0
KIRC	0	0	41	0	0
LUAD	0	0	0	40	0
PRAD	0	0	0	0	42

Table 1. cross-tabulation of predicted and test data. Score is 1.

2. Cross validation

To solve this problem a script «Examples_scikit_learn.cross_validation()» was written. After receiving the training and test data as described in the previous paragraph, the training data is divided into ten parts – folds. The number of parts can be set as a parameter in script. Next, a

model is trained in these parts: nine parts are used as training data and last one is used as test data. Thus, each fold participates 9 times as training data and 1 time – as test data. The script forms the report file «cross_validation.txt», in which, besides the header, there are ten lines. Each line contains: the number of the fold, the accuracy calculated by the test of fold, and the accuracy calculated by the test data of the initial sample - the model is trained on nine folds but is checked on the test data of the initial sample. For input data we obtained the following values of cross-validation:

#_of_fold	1	2	3	4	5	6	7	8	9	10
score_of_fold	1	1	1	1	1	1	0,98	1	1	1
score_of_test	1	1	1	1	1	1	1	1	1	1

Table 2. cross-validation of 10 folds

3. Regularization

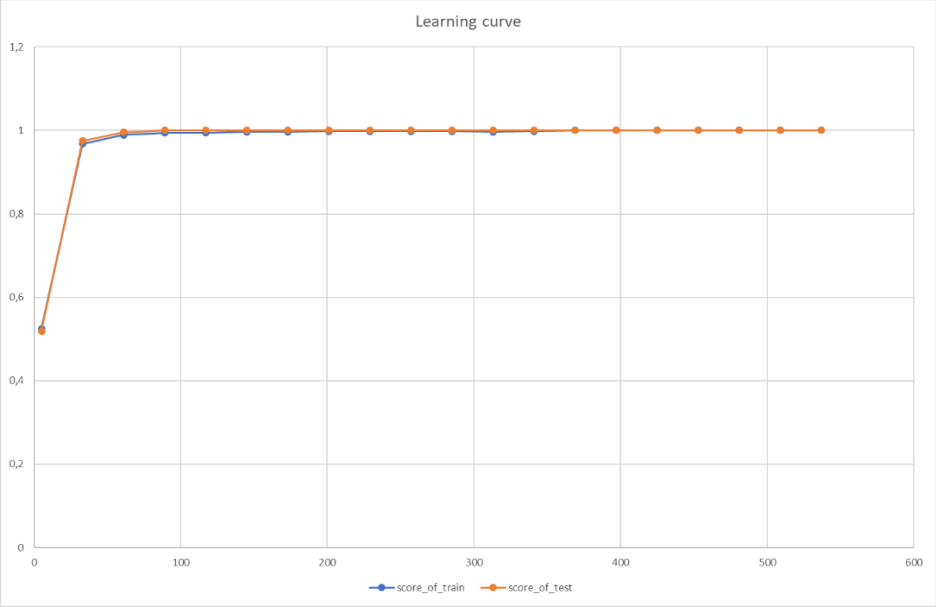
To solve this problem a script «Examples_scikit_learn.regularisation()» was written. After receiving the training and test data as described in the previous paragraph, the script selects the most successful regularization coefficient in the range from 10^{-20} to 10^{+20} in increments of x10. This interval can be set as a parameter. Thus, we determine the regularization coefficient which on the one hand shows the best accuracy (score) on the test data and on the other hand uses the minimum number of model features – attributes, fewer genes are involved in the calculation. The script generates a «regularisation.txt» report file, in which, in addition to the header, there is a line with the value of the regularization coefficient and the obtained accuracy for this coefficient. For input data we obtained the following values of scores:

exp_C	-20	-19	...	-8	-7	-6	-5	-4	-3	-2	-1	0	1	...	19	20
score_of_test	0,99	0,99	...	0,37	0,37	0,56	1,00	1,00	1,00	1,00	1,00	1,00	1,00	...	1,00	1,00

Table 3. scores for regularization coefficient in the range from 10^{-20} to 10^{+20}

4. Learning curve and Grid search

Two auxiliary functions that allow you to visually assess the quality of the model (Learning curve), as well as find the value of one of the model parameters (Grid search) were written.



Pic 1. Learning curve of model
In our case, the Grid search duplicated Regularization and received the same answer.

5. Clustering

In this part, we will show the possibilities of k-means clustering. We assume that we do not know the cancer class and we just want to group our patients by the pattern of gene expression. To solve this problem a script «Examples_scikit_learn.clusterisation()» was written. In our example, we know that the number of clusters is five. In this case, learning without teacher, there is no division into training and test data. For input data we obtained the following clustering:

Class	1	2	3	4	5	Sum
BRCA	0	0	55	0	245	300
COAD	74	0	4	0	0	78
KIRC	0	144	2	0	0	146
LUAD	0	0	139	0	2	141
PRAD	0	0	1	134	1	136
Sum	74	144	201	134	248	801

Table 4. cross-tabulation of real data and results of clustering

6. Principal component analysis (PCA)

In this section, we will try to reduce the number of attributes using the method Principal component analysis (PCA). To overcome the curse of dimension we will use the same number of attributes as we have observations. On the obtained data we will apply a Logistic regression as in paragraph 1. For input data we obtained the following reducing:

Class	BRCA	COAD	KIRC	LUAD	PRAD
BRCA	90	0	0	0	0
COAD	0	28	0	0	0
KIRC	0	0	41	0	0
LUAD	0	0	0	40	0
PRAD	0	0	0	0	42

Table 5. cross-tabulation of predicted and test data of reduced dataset. Score is 1.

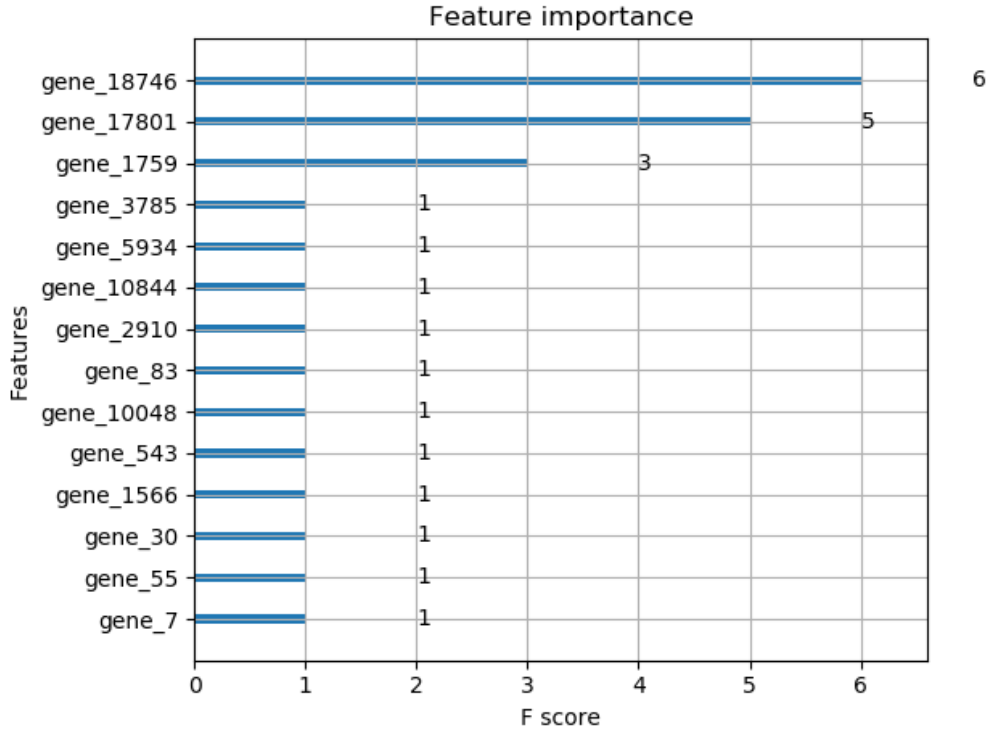
7. Gradient boosting

In this paragraph, we touch the concept of a decision tree. We modified our task and now our task has a binary solution, i.e. it suggests classification, but exclusively binary. For our experiment, we chose the «BRCA» tumor class and try to answer the question of whether our test data is in this class or not. To solve this problem a script «Examples_XGBoost.intro()» was written. The script provides the ability to change the cancer class. In addition, since the decision to classify as a cancer class is probabilistic, it is possible to construct a ROC-curve and calculate AUC. For simplicity, we assume that if the probability of hit into a class is more than 50%, then we assign this Instance to chosen class. The script provides an opportunity to vary this

parameter, for example, to indicate when 10% is reached, we are already ready to include our Instance in the chosen class. For input data we obtained the following decisions:

Prob	0,025	0,031	0,032	...	0,966	0,966	0,974
False	129	5	3	...	0	0	0
True	0	0	0	...	1	6	63

Table 6. Probabilities and decisions of classification



Pic 2. Feature importance for a decision tree. Score is 0.977. AUG is 0.986.

8. Artificial neural network (ANN)

We will solve the problem posed in the previous paragraph by the method of constructing Artificial neural network (ANN). For input data we obtained the following classification:

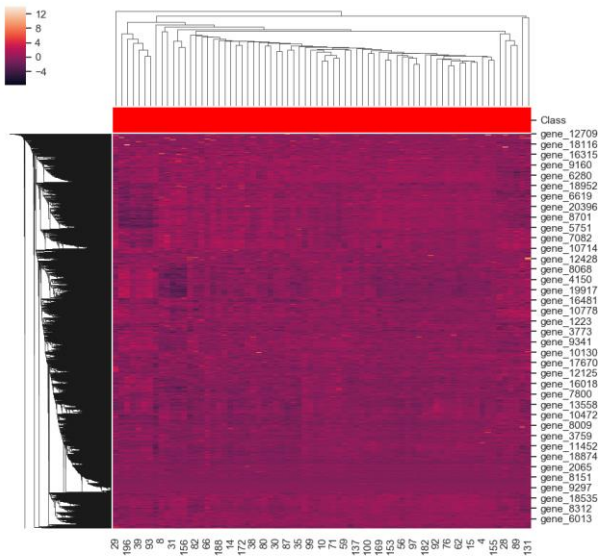
	0	1
0	151	0
1	0	90

Table 7. cross-tabulation of predicted and test data by ANN. Score is 1.

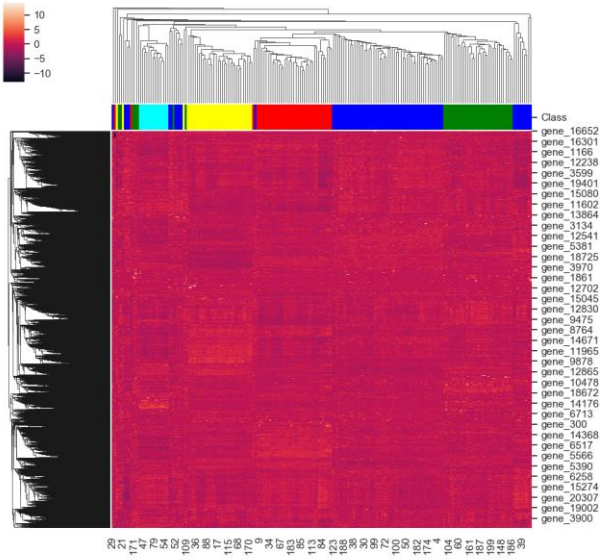
9. Heat-map graph

As a small visual report support, we will construct a heat-map graph for one cancer class and for the entire set of tumor classes. Generally speaking, on such massive data the direct meaning of the heat-map graph is lost. This graph can be used to determine the vector of further research. For example, try to establish a relationship between classes or, on the contrary, ask yourself how homogeneous a given class is. In our work, clustering of patients (cancer class) was shown, but nothing was said about gene clustering. Heat-map graph can be used for this - a preliminary estimate of the number of clusters or determining the depth of clustering. For input data reduced

to 200 patients we obtained the following heat-map graphs (columns – cancer class/patients, lines – genes):



Pic 3. Heat-map for «BRCA» tumor class



Pic 4. Heat-map for all tumor classes

Results and conclusions and questions for further analysis

In general, we can say that we got relatively light and smooth data, despite their massiveness. However, it should be noted several strange facts that were identified during the analysis and what will be discussed below.

Logistic regression has shown excellent results - a high predictive power of the model with a fixed random seed. The disadvantage of this method is that the decision-making method remains a mystery to the researcher. There is no solution that is intuitive. Of course, it is possible to extract regression coefficients from the model, but their semantic content will remain incomprehensible. In fact, we have developed a black box, which gives a good result, but does not give a clear explanation of why he does this. This disadvantage can be eliminated by combining regression with other methods.

Cross-validation made a surprise at the seventh fold and showed a score of less than one. Generally speaking, 0.98 is a very good result, but for our dataset it is a strange result, because the remaining cross-validation figures is one. Apparently, a patient or a group of patients with atypical gene expression got into the test data of 7-th fold and it makes sense to look at them more closely. This question can be investigated in the continuation of the analysis.

Regularization shows good values already at the level of 10^{-5} - 10^{-4}. This means that most regression coefficients will have very low values, around 10^{-10}. From a practical point of view, this means that the set of attributes can be dramatically reduced by removing the corresponding gene expression from the analysis. We can try the following analysis on data with the number of attributes reduced by this method.

The learning curve also presented a surprise: at a long sector, the test score is better than the training score — this is anomalous behavior. On the other hand, this anomalous behavior is

correlated with the outlier that we observed in cross-validation. It is interesting to know, and this is a question for the following analysis, how the learning curve without these outliers will behave.

Clustering in general shows a good result for the four classes, but it broke the BRCA into two weighty groups: 55 and 245 instances. 55 patients were assigned to the LUAD patient group. This means that the following hypothesis can be tested in a «wet» laboratory: the BRCA class actually consists of two types of cancer. It also makes sense to check the biochemical relations, metabolic pathways, etc. of two types of cancer, BRCA and LUAD.

The PCA did not give anything new and showed the same predictive power as regression. Of course, the amount of data was reduced by two orders of magnitude.

Busting showed a good, though not equal to one, predictive power, but its main achievement is the compilation of a list of genes whose expression has crucial role – these genes appear in the decision tree. This is the same element that was missing in the logistic regression. Now we know which genes to look for.

The ANN showed a one hundred percent classification result and can act as a predictive tool along with regression. Of course, it should be added that the deficiency, which we talked about in the case of regression. Here, with the ANN, this deficiency is increased by a multiplier. The magic of ANN is even more complicated and less clear than the magic of regression.

In the course of the analysis, various methods and approaches of machine learning were demonstrated and proposed. A model was built to solve a specific task – it is a regression, supported by the results of the ANN plus an interpretation obtained using a decision tree. In addition to solving the task, an array of interesting questions was formed. The questions that can be developed in further research work.