Machine Learning: Computer Exercise 1

机 53　陈声健　2015010509

1. Task Description

In this programming exercise, gene expression data of selected genes of cells in human embryonic cells of day 3 to day 7 (labeled as E3~7) is given. The first task is to use logistic regression and FLD method to classify samples of E3 vs. E5 and apply the trained classified to the samples of E4, E6, E7. There is also an optional task to find the relative importance of each gene in the trained classifiers. The second task is to Use linear regression to train a model to predict the day of the embryo cells from their gene expression data and find major metrics for assessing the performance of linear regression models.

2. Experiment Design

The two tasks are quiet straightforward. All the functions required in this exercise are available in the standard library. I can just search the documentation and use the function.

Date are read with the recommended way (read_csv). The column index is changed from cell code to integer number for the convenience of traversal. And then gene data are grouped by the corresponding day for later learning process.

3. Method

All the experiment is run on Anaconda Jupyter notebook with Python 3.6 and then migrated to a .py file. Python libraries Pandas, Numpy, matplotlib and sklearn are used in the program.

In task 1, *sklearn.linear_model.LogisticRegressionCV* [1] is used for logistic regression. ***cv*** is set to 10, corresponding to 10-folds cross validation. ***max_iter*** is 500 and ***random_state*** is 0. *sklearn.discriminant_analysis.LinearDiscriminantAnalysis* [2] is used for FLD classifier. There is no parameter need to be specified in FLD. In both logistic regression and FLD, binary classification output 0 or 1. In the experience, E3 corresponds to 0 and E5 is 1. The prediction of the classifier is represented as:

$$\frac{sum\ of\ prediction\ output}{number\ of\ prediction\ samples}$$

In a perfect classifier, we will get 0 if all the samples are E3.

An easy and widely used way to measure the relative importance of each feature (gene in our exercise) is to compare the coefficients of the trained model [3]. The feature with larger coefficient is thought to be more important. The relative importance of a feature in this report is measured as

$$100 * \frac{feature\ coefficient}{\max coefficient}$$

In task 2, *sklearn.linear_model.LinearRegression* is used for linear regression. No parameter need to be specified. A simple way to assess the performance of the linear regression model is to calculate the R^2 coefficient which is defined as

$$1 - \frac{sum\ of\ \left(y_{pred} - y_{true}\right)^2}{sum\ of\ \left(y_{ture} - mean(y_{true})\right)^2}$$

In a perfect linear regression model, R^2 equals to 1. To calculate R^2 of a model on certain set of sample, we can simply call *sklearn.linear_model.LinearRegression.score*.

Other way to assess the linear regression model include root mean square error method and plotting the learning curve. R^2 is used in this report.

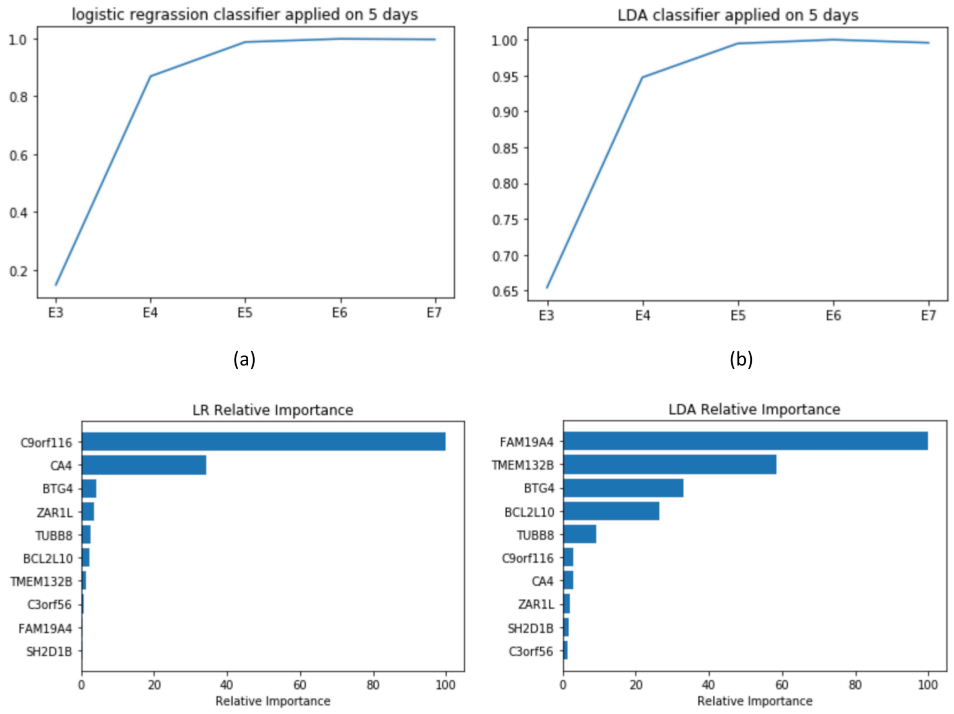
4. Result and Observation

Task 1

Result of applying the five days' data to the trained classifier is demonstrated below (Fig. 1a,1b). In logistic regression, I get 0.148(should be 0) on E3 and 0.987(should be 1) on E5. Since we have more data in E5 compared with E3, the model is doing pretty good on E5 but quiet poor on E3. FLD (also referred as LDA) performs even worse on E3(0.65 rather than 0) data but equal as logistic regression on E5. I think the model will do better if we have more data on E3.

Looking at the five days result as a whole, we can find that the result is increasing as the time goes on. We get number between E3 and E5 on E4, and get nearly 1 on E6 and E7. This may reveal some kind of development process of the selected genes in the growth of the human embryonic cells in the selected days.

Comparing the coefficient of the two models (fig. 1c, 1d), I find that the two models are quiet different. In logistic regression, ***C9orf116*** ranks the most important gene while ***CA4*** is the second but it is much less important than ***C9orf116***. And the rest are nearly neglectable. In the LDA model, however, ***FAM19A4*** is the most important gene, which is thought to be totally unimportant in logistic regression. And there also several genes that play critical roles, like ***TMEM1328***, ***BTG4***, ***BCL2L10***, ***TUBBB***.



(a)

(b)

(c)

(d)

Fig. 1

Task 2

I apply the whole dataset to linear regression and get a trained model. Then the R^2 value is calculated by calling the *score*() function. In our dataset, $R^2 = 0.118$. This is much smaller than 1. I can conclude that our dataset is linearly non-separable. We need more complicated model to predict the day according to the 10 genes.

5. Conclusion

On our dataset, logistic regression is doing better than FLD method. I also find that the number of training samples is very important to the performance of the model. If we get more data on E3, we may do better. Linear regression's poor performance on this dataset indicates that this dataset may be linearly non-separable. The underlying relationship is much more complicated.

From the result of logistic regression, we can find some kind of trend in the 10 gene in the development of the cells. This is a very good example to show the power of machine learning in processing genetic data.

The code is attached as supplementary material. To run the code, simply type:

python exercise1.py

6. Reference

1. sklearn.linear_model.LogisticRegressionCV. http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegressionCV.html
2. sklearn.lda.LDA. http://scikit-learn.org/0.15/modules/generated/sklearn.lda.LDA.html
3. Gradient Boosting regression. http://scikit-learn.org/stable/auto_examples/ensemble/plot_gradient_boosting_regression.html#sphx-glr-auto-examples-ensemble-plot-gradient-boosting-regression-py