Logistic Regression

Logistic regression is a common statistical analysis model, suitable for modeling discrete data. Its basic function is to perform binary classification and predict its possible labels for different input features. Logistic regression measures the relationship between a categorical dependent random variable (label) and one or more independent variables (features) by using a logistic regression function to estimate probability. Binary classification is a common task in machine learning. Linear classification performs well when the data is “linearly separable”. However, when we try to use the least square (LS) approach to fit parameter , because the features are imbalanced, this may cause the decision boundary to be pulled to a certain side of the data and lead to misclassification. Under normal circumstances, given a score z, calculate the probability that the corresponding output label is 0 or 1, the equations are as follows:

Where **,**

As z increases, it is more likely to be judged as class “1”, otherwise it tends to be “0”.

From this case, corresponding to a series of discrete data points, the most suitable logistic regression model is needed to make relatively accurate predictions. Unlike linear regression using LS to find the most suitable linear model, in logistic regression, it’s able to fit the parameters of the training data by using the maximum likelihood (ML) estimation, when , the ML weights for binary logistic regression are:

Which is a common alternative expression, it also known as binary cross-entropy loss.

Mice Protein Expression Data Set was found in the UCI Machine Learning Repository. This data set includes the expression levels of 77 proteins that produce detectable signals in the cortex nucleus. There were 38 control mice, 34 trisomy mice (Down’s syndrome), 72 in total. In the experiment, each sample/mouse has 15 measurement records for each protein. Therefore, for control mice, there are 570 measurements, and for trisomy mice, there are 510 measurements. The data set contains 1080 measurement values ​​for each protein, which are all independent of each other [1]. The original data is divided into 8 classes according to genotype, behavior, and treatment. This study only focuses on the relationship between genotype and protein expression level, so it is simplified to a binary classification of 2 classes.

According to the summary information of the website, there are missing values ​​in the data set. After re-retrieving the data set, the position of the missing value is located. In order to ensure the integrity of the data and facilitate subsequent research, a new data set was created in which the missing values ​​in each column were filled with the average of the non-missing values. At the same time, the strings of “Control” and “Ts65Dn” in the “Genotype” column are converted into numbers “0” and “1”, so that logistic regression can be applied more intuitively. The last step in the data preprocessing is to adjust the 77 columns of protein expression level data to a normal distribution, which is to avoid overfitting of the data caused by the excessive variance of some features.

In order to better reflect the relationship between data volume and accuracy, we randomly sampled the data and compared the data performance at 20%, 40%, 60%, 80%, and 100%. The results are shown in the table below.

Table 1. Data Performance

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Fraction | Precision | Recall | F1-score | Error rate  (1 - Accuracy) |
| 20% | 0.9361 | 0.9011 | 0.9135 | 0.0840 |
| 40% | 0.9405 | 0.9360 | 0.9361 | 0.0625 |
| 60% | 0.9409 | 0.9532 | 0.9467 | 0.0495 |
| 80% | 0.9508 | 0.9525 | 0.9514 | 0.0451 |
| 100% | 0.9542 | 0.9575 | 0.9554 | 0.0417 |

It can be seen that as the amount of data increases, data performance shows a steady upward trend, and the ability to capture the expression level of the protein with the highest correlation is also improved. For the training results under the complete data set (that is, the fraction is 100%), the finally drawn coefficient stem and leaf map is shown in Figure 1, and the two genes with the highest correlation are ITSN1\_N and BRAF\_N.

图表

描述已自动生成

Figure 1 Coefficient Stem Plot of Logistic Regression Model

Reference

[1] “Mice Protein Expression Data Set”, UCI Machine Learning Repository.