Model Evaluation

Taolan Jiaohaer

November 2022

Model evaluation is a very important part of our project.Model evaluation is the process of using different evaluation metrics to understand a machine learning model's performance, as well as its strengths and weaknesses. Model evaluation is important to assess the efficacy of a model during initial research phases, and it also plays a role in model monitoring. In this section, we will list some of the most popular methods for measuring classification performance.

1 Accuracy

The accuracy refers to the ratio of the number of correctly classified records to the total number of records when the test set is used to classify the model in the classification:

$$accuracy = \frac{n_{correct}}{n_{total}}$$

Accuracy seems very simple, However, the accuracy evaluation metric does not distinguish between different classes or treats each class equally. So this kind of evaluation is sometimes not enough. For example, sometimes it depends on the number of classification errors under category 0 and category 1, because the cost of misclassification under different categories is different, that is, the bias for different categories is different. For example, in the diagnosis of patients, the diagnosis of cancer is actually not cancer (False Positive) and the diagnosis of cancer is actually cancer. (False Negative) The importance of these two cases is not the same. Another reason is that the data distribution may be unbalanced, that is, there are too many samples in some categories, and the number of samples in some categories is too small, and the number of the two categories is quite different. In this way, the category with most samples dominates the calculation of the accuracy rate. In order to solve this problem, the accuracy rate is improved to obtain the average accuracy rate.

$$accuracy = \frac{TP + TN}{n}$$

2 Average Per-class Accuracy

In order to deal with the situation that the number of samples under each category is different, the accuracy rate is variated, the accuracy rate under each

category is calculated, and then their average value is calculated. For example, the accuracy rate of category 0 is 80%, and the accuracy rate of category 1 is 97.5%, then the average accuracy rate is

$$\frac{(80\% + 97.5\%)}{2} = 88.75\%$$

Because the number of samples under each category is different, that is, when calculating the accuracy rate of each category, the denominator is different, and the average accuracy rate is not equal to the accuracy rate. If the number of samples under each category is the same, the average Accuracy is equal to accuracy.

The average accuracy rate also has its own disadvantages. For example, if there is a category with a small number of samples, the variance of the accuracy rate of the category may be too large when using the test set for testing (such as k-fold cross validation) , which means that the reliability of the accuracy rate of this category is not strong.

3 Confusion Matrix

A Confusion Matrix is a specific table layout that allows visualization of the performance of an algorithm, and a table that describes the results of the classification in detail, whether it is classified correctly or incorrectly, and distinguishes different categories. For the binary classification, it is a 2×2 matrix, and for the n classification, it is $n \times n$ matrix. For binary classification, the first row is the number of records (number of samples) whose real category is "Positive", the second row is the number of records whose real category is "Negative", and the first column is the predicted value of "Positive" , and the second column is the number of records whose predicted value is "Negative". As shown in the table below:

Table 1: 2×2 Confusion Matrix

	Predicted as Positive	Predicted as Negative
Labeled as Positive	True Positive (TP)	False Negative (FN)
Labeled as Negative	False Positive (FP)	True Negative (TN)

As shown in the table above, the results can be divided into four categories:
* True Positive, (TP): positive samples classified correctly by the model;
* False Negative (FN): Positive samples misclassified by the model;
* False Positive, (FP): Negative samples classified by the model;
* True Negative (TN): Negative samples that are correctly classified by the model.

1. According to the table, Accuracy could be denoted as :

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

2. Recall/ Sensitivity / True Positive Rate(TPR):

$$Recall = \frac{TP}{TP + FN}$$

3. Specificity:

$$Specificity = \frac{TN}{FP + TN}$$

4. Precision:

$$Precision = \frac{TP}{TP + FP}$$

5. F-Measure

$$F1 - Score = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

4 ROC-AUC

A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.In the ROC curve, the y-axis is TPR/Recall/Sensitivity, and the x-axis is FPR/1-Specificity.

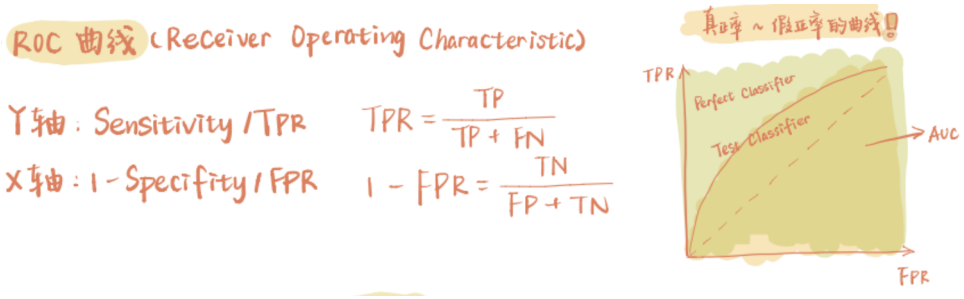

Figure 1: ROC-AUC

* For the ROC curve, The closer to the upper left corner (FPR=0, TPR=1), the better the model effect (the false positive rate is 0, and the true rate is 1) .

* AUC stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC curve. So for the AUC value, the bigger is the better.

5 Precision-Recall Curve

Precision-Recall curves summarize the trade-off between the true positive rate and the positive predictive value for a predictive model using different probability thresholds : the y-axis is Precision and the x-axis is TPR/Recall/Sensitivity.

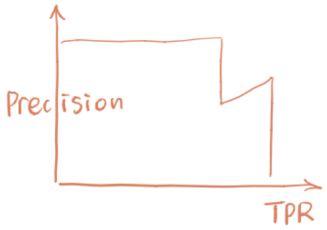

Figure 2: PR Curve

So, for the PR curve, if it is closer to the upper right corner (Precision=1, Recall=1), the model would be better(the precision and recall are both 1).

6 LIFT Curve

The LIFT curve shows the relation between the number of instances which were predicted positive and those that are indeed positive and thus measures the performance of a chosen classifier against a random classifier.For the LIFT curve, the y-axis is the Lift value, which is

$$LIFT = \frac{Recall}{\frac{TP+FP}{n}}$$

or

$$LIFT = \frac{Precision}{\frac{TP+FN}{n}}$$

While x-xis is

$$\frac{TP+FP}{n}$$

represents for the Predicted positive sample rate.

Figure 3: Lift Curve

So, for the LIFT curve, if the line is closer to the upper right corner, the model would be better.