Эксплуатационные ошибки

Терехов Олег

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

10 апреля 2019 г.

Precision-Recall

Основные метрики

1. Recall or Sensitivity or TPR (True Positive Rate): Number of items correctly identified as positive out of total true positives- TP/(TP+FN)
2. Specificity or TNR (True Negative Rate): Number of items correctly identified as negative out of total negatives- TN/(TN+FP)
3. Precision: Number of items correctly identified as positive out of total items identified as positive- TP/(TP+FP)
4. False Positive Rate or Type I Error: Number of items wrongly identified as positive out of total true negatives- FP/(FP+TN)
5. False Negative Rate or Type II Error: Number of items wrongly identified as negative out of total true positives- FN/(FN+TP)

F_β-measure

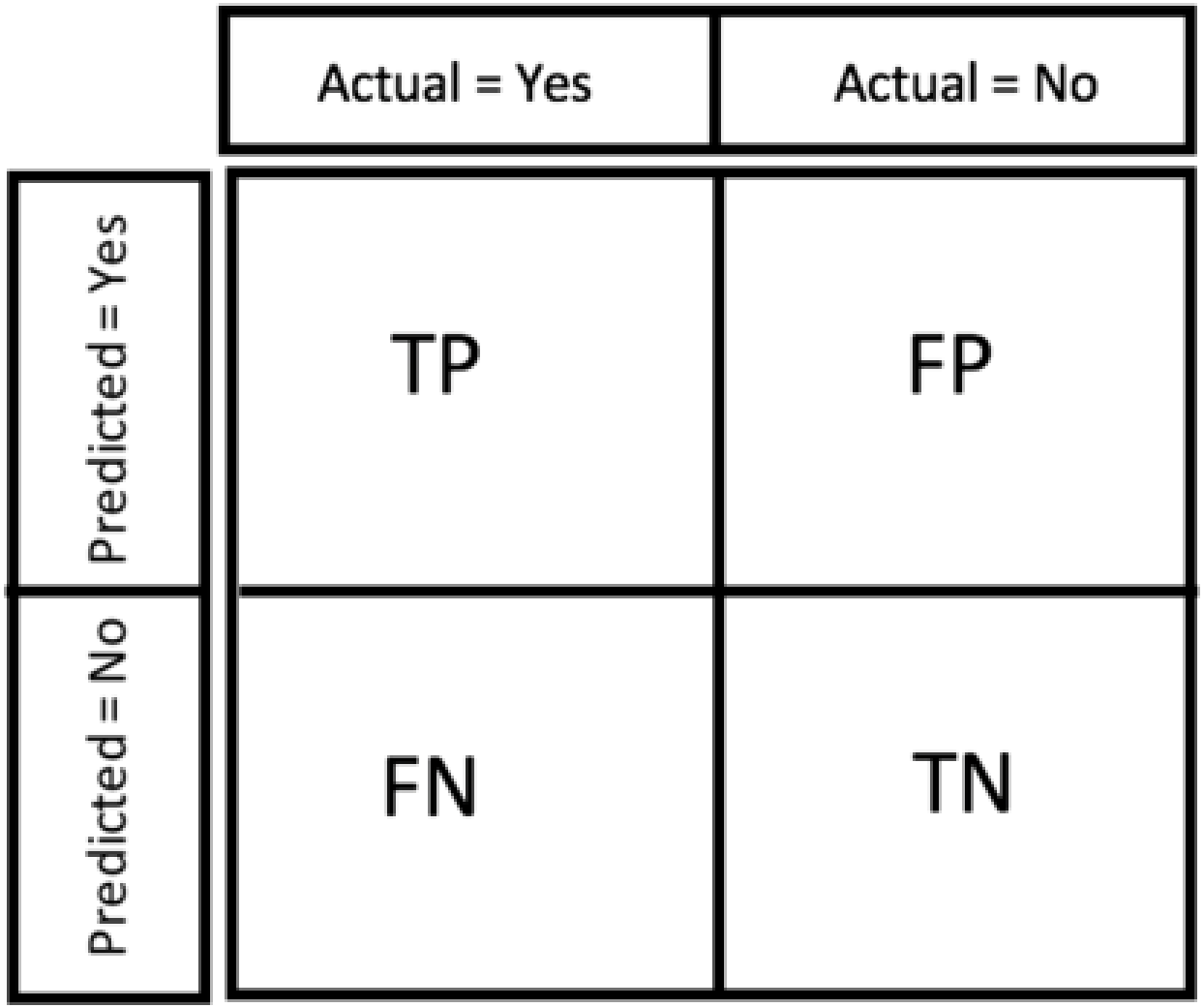
$$F_\beta = (1 + \beta^2)\frac{precision \times recall}{(\beta^2 precision) + recall}$$

The F-measure reaches a maximum with completeness and accuracy equal to one, and is close to zero if one of the arguments is close to zero.

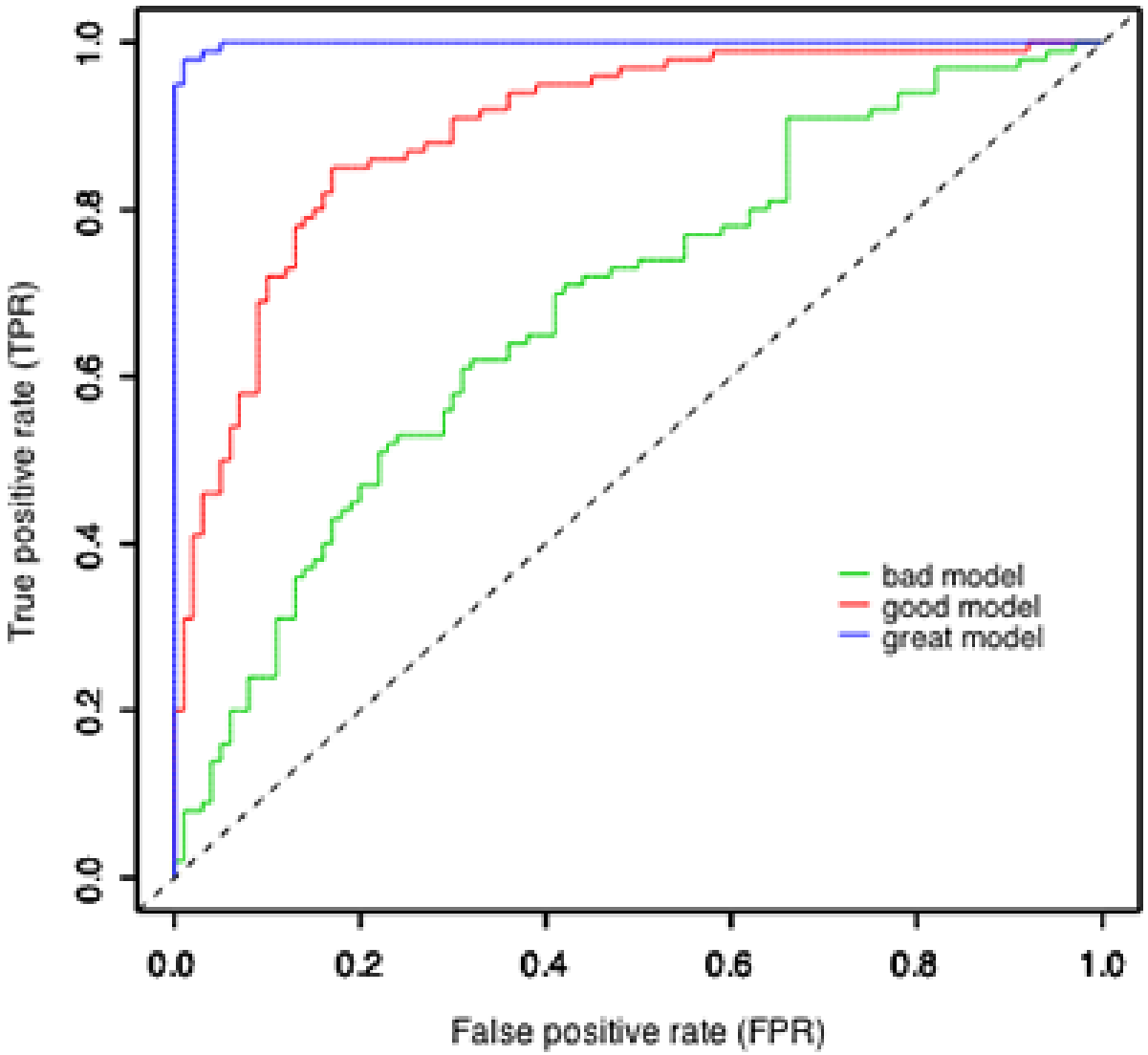
F1 Score: a harmonic mean of precision and recall

$$F1 = 2\frac{Precision \cdot Recall}{Precision + Recall}$$

Confusion Matrix



ROC-curve


True positive rate (TPR)
False positive rate (FPR)
0.0
0.2
0.4
0.6
0.8
1.0
bad model
good model
great model

Accuracy

Accuracy: Percentage of total items classified correctly

$$\text{Accuracy} = \frac{\#\text{correctly classified items}}{\#\text{all classified items}}$$

Defective Pairs

$$\mathrm{DP} = \frac{2}{n(n-1)} \times \sum_{i<j}^{n} [y_i > y_j]$$

Связь DP и AUC

$$\mathrm{DP} = \frac{2n_- n_+}{n(n-1}(1 - \mathrm{AUC})$$

Gini

$$\mathrm{Gini} = 2\mathrm{AUC} - 1$$

Log-Loss

Log-loss is a measurement of accuracy that incorporates the idea of probabilistic confidence given by following expression for binary class

$$logloss = -\frac{1}{n}\sum_{i=1}^{n}(y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i))$$

Definitions

1. SS be the number of pairs of items belonging to the same cluster and category
2. SD thenumber of pairs belonging to the same cluster and different category
3. DS the number of pairs belonging to different cluster and the same category
4. DD the number of pairs belonging to different category and cluster

SS and DD are “good choices”, and DS, SD are “bad choices”

Statistics

Rand statistic $R = \frac{SS+DD}{SS+SD+DS+DD}$

Jaccard Coefficient $J = \frac{SS}{SS+SD+DS}$

Folkes and Mallows $FM = \sqrt{\frac{SS}{SS+SD}\frac{SS}{SS+DS}}$

RMSE

RMSE

It represents the sample standard deviation of the differences between predicted values and observed values.

$$RMSE = \sqrt{\frac{1}{n} \times \sum_{j=1}^{n} (y_j - \hat{y}_j)^2}$$

MAE

MAE

MAE is the average of the absolute difference between the predicted values and observed value. The MAE is a linear score which means that all the individual differences are weighted equally in the average. For example, the difference between 10 and 0 will be twice the difference between 5 and 0. However, same is not true for RMSE which we will discuss more in details further.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

R Squared

R Squared and Adjusted R Squared are often used for explanatory purposes and explains how well your selected independent variable(s) explain the variability in your dependent variable(s).

$$\hat{R}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

Higher the MSE, smaller the R^2 and poorer is the model.

Adjusted R Square

Just like R^2, R_Adj^2 also shows how well terms fit a curve or line but adjusts for the number of terms in a model

$$R^2_{adj} = 1 - \frac{\left(1 - R^2\right)(n-1)}{n-k-1}$$

where n is the total number of observations and k is the number of predictors. R_Adj^2 will always be less than or equal to R^2

BLEU

Bilingual Evaluation Understudy Steps to compute BLEU score: 1. Convert the sentence into unigrams, bigrams, trigrams, and 4-grams 2. Compute precision for n-grams of size 1 to 4 3. Take the exponential of the weighted average of all those precision values 4. Multiply it with brevity penalty

$$BLEU = BP \times \exp\left(\sum \mathbf{w}_n \log(P_n)\right)$$

$$= \begin{cases} 1, & \text{если } c \geqslant r; \\ exp\left(1 - \frac{r}{c}\right), & \text{если } c < r. \end{cases}$$

Here BP is the brevity penalty, r and c is the number of words in reference and candidate respectively, w—weights, P—Precision values

ROUGE

1. ROUGE-N – measures unigram, bigram, trigram and higher order n-gram overlap
2. ROUGE-L – measures longest matching sequence of words using LCS. An advantage of using LCS is that it does not require consecutive matches but in-sequence matches that reflect sentence level word order. Since it automatically includes longest in-sequence common n-grams, you don't need a predefined n-gram length.
3. ROUGE-S – Is any pair of word in a sentence in order, allowing for arbitrary gaps. This can also be called skip-gram coocurrence. For example, skip-bigram measures the overlap of word pairs that can have a maximum of two gaps in between words. As an example, for the phrase "cat in the hat" the skip-bigrams would be "cat in, cat the, cat hat, in the, in hat, the hat".

Публикации по теме

1. Соколов Е. (Выбор моделей)
2. Hardt M., Price E., Srebro N. Equality of Opportunity in Supervised Learning
3. Amigo E., Gonzalo J., Artiles J., Verdejo F. A comparison of Extrinsic Clustering Evaluation Metrics based on Formal Constraints
4. (Metric's zoo part 1)