Data learning

Курс “Машинное обучение”
Лабораторная работа

Лаборатория анализа данных

Binary model-wide measures

Харитонов Е.А., М16-524
Вариант 1-09

2017

Исходные данные

Source data – binary classified (positive and negative classes) set of entities, described by two parameters x1 and x2.
Sample size – 350 (50 positive, 300 negative)
Representation – entities can be interpreted as points on Cartesian coordinate plane with axes x1 and x2 for ease.

x1	x2	label	score
1.5998	0.76857	-1	-2.9341
0.95352	-1.0024	-1	-4.7261
1.7588	0.86545	-1	-2.7575
2.679	1.9814	-1	-1.3024
1.6792	0.59262	-1	-3.0209
1.3089	0.15303	-1	-3.5991
2.4106	1.5793	-1	-1.7866
2.4737	1.1979	-1	-2.044
2.0354	1.0006	-1	-2.4764
2.8452	1.5663	-1	-1.5206
2.8618	1.4816	-1	-1.5761
2.6181	1.21	-1	-1.9428
1.4692	0.18198	-1	-3.4747
1.8713	-0.03735	-1	-3.3903
…	…	…	…

Figure 1. Piece of sample data

Исходные данные

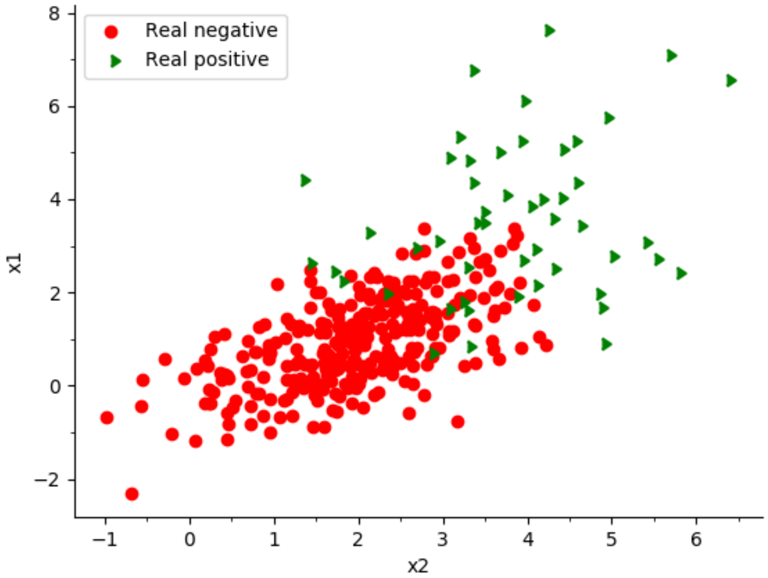

Figure 1. Visualization of source data set as points on the coordinate plane with real classification

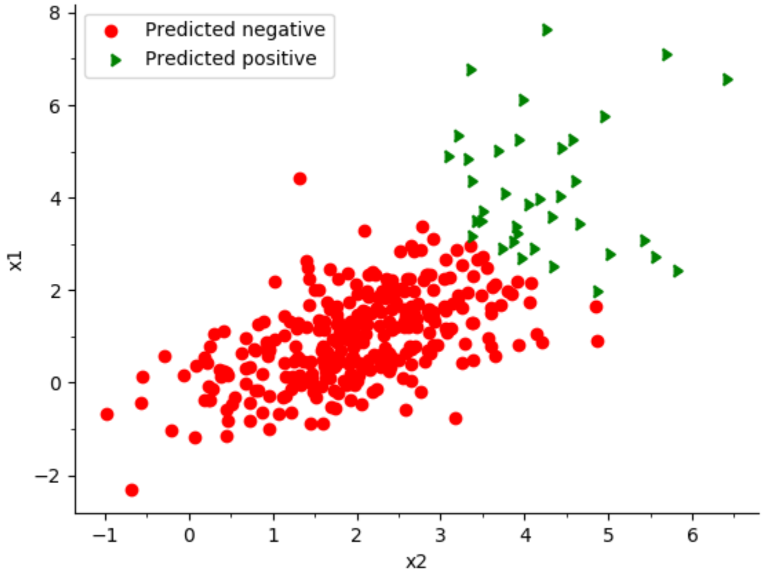

Figure 2. Visualization of source data set as points on the coordinate plane classified by proposed classifier

Используемые методы и формулы

ROC & PR curves

(1) TPR = REC = SENS = TP/(TP+FN)

(2) FPR = FP/(FP+TN)

(3) PREC = TP/(TP+FP)

(4) SPEC= TN/(TN+FP)

(5) $\mathrm{F}_1 = \frac{2*\mathrm{PREC}*\mathrm{REC}}{\mathrm{PREC}+\mathrm{REC}}$

(6) $\varkappa = \frac{\mathrm{ACC} - ACC_0}{1 - ACC_0}$

(7) Y = (SENS + SPEC - 1)

where

TP – True Positives count,

FN – False Negatives count,

FP – False Positives count,

TPR – True Positive Rate,

SENS – Sensitivity,

SPEC – Specificity,

FPR – False Positive Rate,

PREC – Precision,

REC – Recall

F_1 – F-score value for β = 1,

$\varkappa$ – Cohen’s kappa,

Y – Youden’s index.

Используемые методы и формулы

ROC (receiver operating characteristic) curve of classifier *h* is a graphical plot of its sensitivity (true positive rate) against the 1-specificity (false positive rate) at various thresholds $b \in \mathbb{R}$

PR (Precision-Recall) curve of classifier *h* is a graphical plot of its precision (PREC) against RECALL (REC) at various thresholds $b \in \mathbb{R}$

AUC (Area Under Curve) - the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one

Используемые методы и формулы

Mann–Whitney *U* test – is a test of the null hypothesis that it is equally likely that a randomly selected value from one sample will be less than or greater than a randomly selected value from a second.

(8) $\mathrm{U} = \mathrm{n_1 n_2} + \frac{\mathrm{n_x(n_x+1)}}{2} - \mathrm{T_x}$ - test statistic, where

n_1, n_2 - the sample sizes,

n_x , T_x - the size and the rank sum of the sample with bigger rank sum

Используемые методы и формулы

Correlation coefficient is a number that quantifies a type of correlation and dependence, meaning statistical relationships between two or more values in fundamental statistics.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (9)$$

where:

n - the sample size,

x_i, y_i are the single samples indexed with I,

$\bar{x}, \bar{y}$ are the sample mean

Результаты исследований

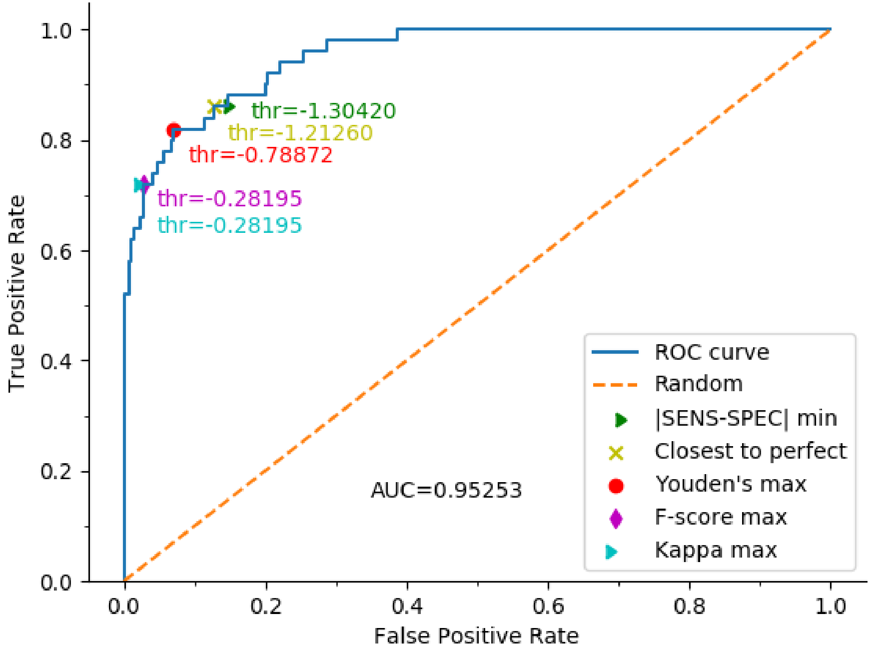

Figure 3. **ROC curve** of the classifier family with AUC value and optimal threshold values (thr) obtained by different methods

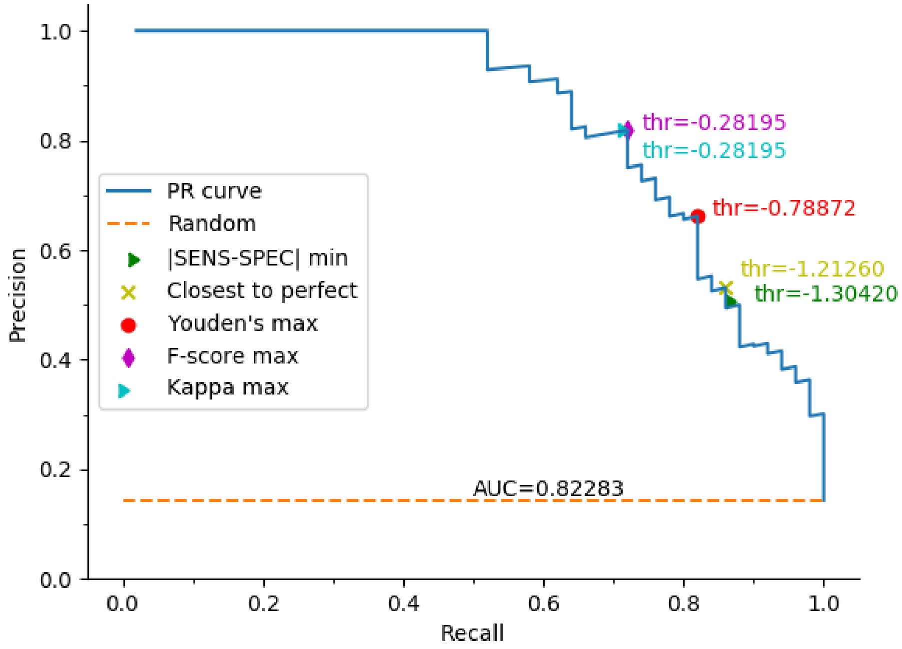

Figure 4. **PR curve** of the classifier family with AUC value and optimal threshold values (thr) obtained by different methods

Результаты исследований

Mann–Whitney *U* test

H0: scores of the positive sample is the same as scores of the negative sample

H1: scores of the positive sample is greater than scores of the negative sample

U_{emp} = 94

U_{cr} = 557 (*p* = 0.01)

$U_{emp} < U_{cr} \Longrightarrow$ H0 is rejected

p-value = 2.52 * 10^{-12}

x1	x2	label	score	rank
1.5998	0.76857	-1	-2.9341	2
0.95352	-1.0024	-1	-4.7261	1
3.8393	1.9242	1	-0.60974	5
4.2738	3.5801	1	0.95799	6
1.7588	0.86545	-1	-2.7575	3
1.3966	2.629	1	-1.6119	4

$T_n = 6$

$n_n = 3$

$T_p = 15$

$n_p = 3$

Figure 5. Simple example of Mann-Whitney’s sums of ranks calculation

Результаты исследований

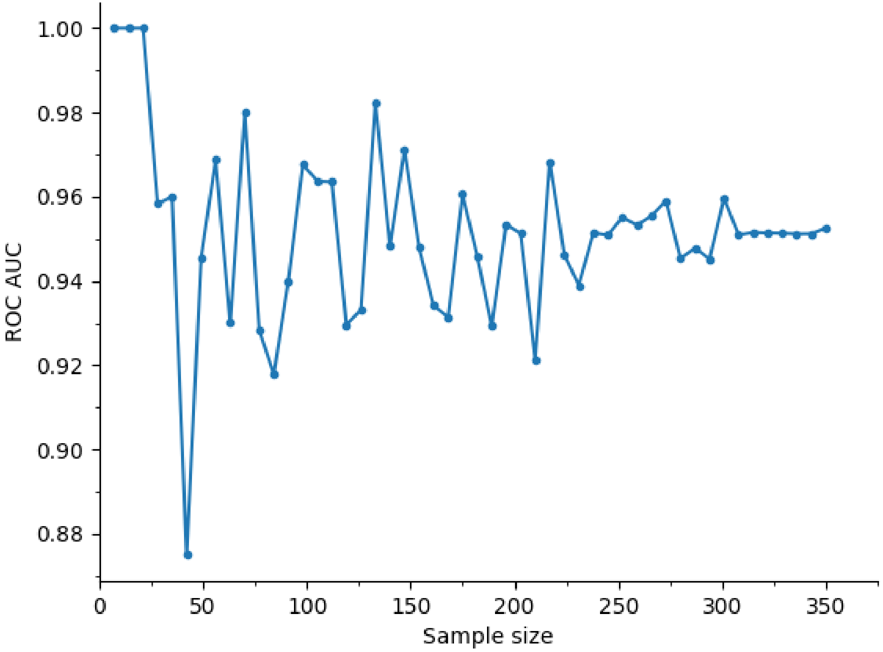

Figure 6. ROC curve AUC dependency from the Sample size plot

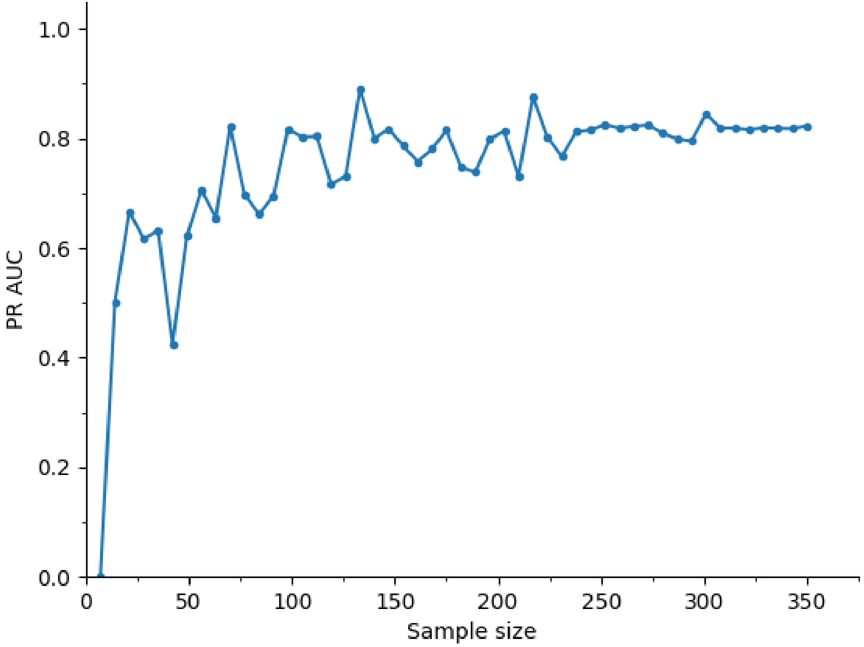

Figure 7. PR curve AUC dependency from the Sample size plot

Результаты исследований

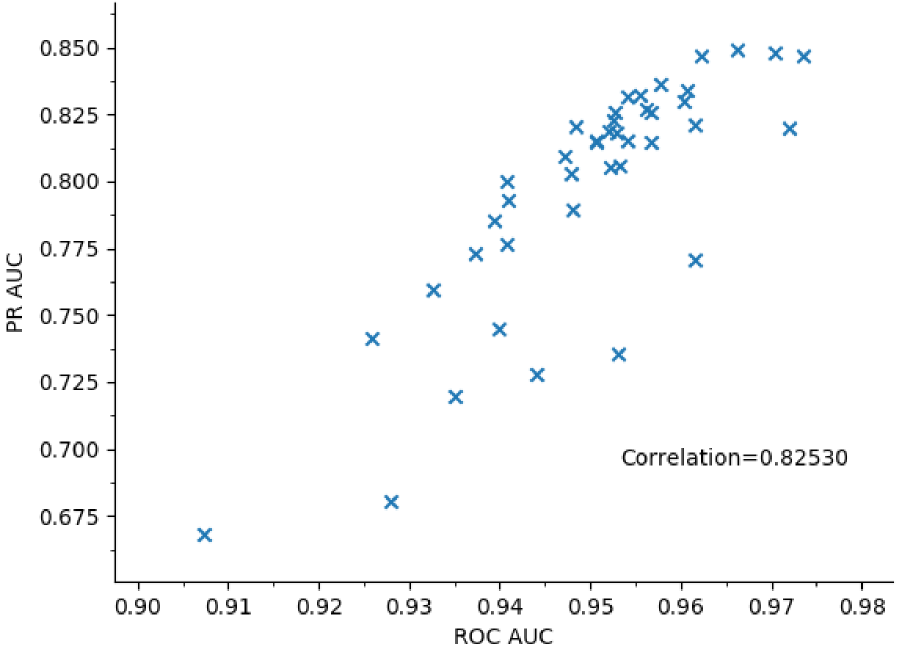

Figure 8. Scatter plot of ROC AUC and PR AUC with correlation coefficient value

Linear correlation between ROC AUC and PR AUC values with different sample sizes is strong enough (if samples with little sample sizes are ignored). On different random samples selection it varies in range [0.60; 0.95]

Выводы

ROC and PR curves can help to decide which classifier is more appropriate. Classifier h1 is better than classifier h2 if it's curve is closer to the ideal.

PR curve is more informative when dealing with "needle-in-haystack" type problems or problems where the positive class is more important than the negative class.

AUC of both curves can help to compare curves not relaying on their visualization. The closer AUC value to 1.0 the better is classifier.