Classifiers assessment methods

Jacek Kluska

Rzeszow University of Technology

Assumptions, notation

The **learning dataset**

$$D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_Q, y_Q)\} \subset \mathbb{R}^N \times L, \tag{1}$$

where

- $L = \{l_1, \ldots, l_r\}$ – a finite set of original labels (classes), $r \geqslant 2$,
- $\mathbb{R}^N$ – the set of N-dimensional real vectors of features (attributes),
- $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,j})$, $i = 1, \ldots, Q$, $j = 1, \ldots, N$,
- $y_i \in L$ – class label (a target or a category) represented by a natural number.
- If $r = 2$, the problem is called **binary classification**;
- If $r > 2$, the problem is called a **multiclass classification** problem.

The vectors of true and predicted labels

For the multiclass classification problem, we use **label-based metrics**.
Let D' be a **test dataset** containing n input-output pairs.

- $L^n = L \times \ldots \times L$ – the Cartesian product of n sets L of class labels as in (1).
- $\mathbf{t} = [t_1, \ldots, t_n]$ – a vector of *true* labels, $\mathbf{t} \in L^n$.
- $\mathbf{p} = [p_1, \ldots, p_n]$ – a vector of *predicted* labels, $\mathbf{p} \in L^n$.
- A subset T_l containing all indices of data records labeled by l that are coordinates of **the vector of true labels**,
$$T_l = \left\{ i \mid \exists\ \mathbf{x}_i \in \mathbb{R}^N, (\mathbf{x}_i, l) \in D',\ \text{and } l \in \mathbf{t} \right\}. \tag{2}$$
- A subset P_l containing all indices of data records labeled by l that are coordinates of **the vector of predicted labels**,
$$P_l = \left\{ i \mid \exists\ \mathbf{x}_i \in \mathbb{R}^N, (\mathbf{x}_i, l) \in D',\ \text{and } l \in \mathbf{p} \right\}, \tag{3}$$
where $D' \subset D$.
- $|X|$ - is the cardinality of the set X.

Confusion matrix – Accuracy measure (ACC)

Confusion matrix is defined as follows:

$$CM = \begin{array}{c|c|c|c|} & P_1 & \dots & P_r \\ \hline T_1 & |T_1 \cap P_1| & \dots & |T_1 \cap P_r| \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline T_r & |T_r \cap P_1| & \dots & |T_r \cap P_r| \\ \hline \end{array} . \tag{4}$$

Classification *accuracy* (ACC) is our usual meaning of the term "accuracy". It measures the number of times any class was predicted correctly, normalized by the number of data points, and is defined as follows:

$$ACC = \frac{1}{n} \sum_{l \in L} |T_l \cap P_l| . \tag{5}$$

Precision (“Pre” or positive predictive value “PPV”)

Weighted precision (or the *positive predictive value*) is defined by

$$Pre = \sum_{l \in L} v_l \cdot Pre_l, \tag{6}$$

where *precision by label* denoted by Pre_l considers only one class and measures the number of times a specific label $l \in L$ was predicted correctly, normalized by the number of times the label appeared in the output,

$$Pre_l = \frac{|T_l \cap P_l|}{|P_l|} \tag{7}$$

and the weight is given by

$$v_l = \frac{|T_l|}{\sum_{l \in L} |T_l|} = \frac{1}{n} |T_l| \,. \tag{8}$$

Sensitivity ("Sen" or "recall", "hit rate" or "true positive rate")

Weighted sensitivity (also called *recall*, *hit rate* or the *true positive rate*) is defined by

$$Sen = \sum_{l \in L} v_l \cdot Sen_l, \tag{9}$$

where *sensitivity by label* Sen_l considers only one class and measures the number of times a specific label $l \in L$ was predicted correctly, normalized by the number of times that the label in fact appeared,

$$Sen_l = \frac{|T_l \cap P_l|}{|T_l|}, \tag{10}$$

and the weight v_l is defined by (8).

F-beta measure (“F-beta”)

A *weighted F-beta measure* (*F-beta score*) is defined as

$$F_{\beta,l} = (1+\beta) \sum_{l \in L} v_l \cdot \frac{Pre_l \cdot Sen_l}{\beta^2 Pre_l + Sen_l} \tag{11}$$

where the weight v_l is defined by (8).
A very popular measure: a *weighted* F_1 *measure*, i.e., by assuming $\beta = 1$.
For Pre_l and Sen_l defined by (7) and (10), respectively, we obtain

$$F_1 = \frac{2}{n} \sum_{l \in L} \frac{|T_l|}{|T_l| + |P_l|} \, |T_l \cap P_l| \, . \tag{12}$$

Examples

Example

Let us consider the following dataset:

$$D = \{(\mathbf{x}_1, 0), (\mathbf{x}_2, 1), (\mathbf{x}_3, 2), (\mathbf{x}_4, 2), (\mathbf{x}_5, 0)\} \subset \mathbb{R}^N \times \{0, 1, 2\},$$

Thus,

- $n = 5$,
- $L = \{0, 1, 2\}$.

Assume the vector of true labels

$$\mathbf{t} = [0, 1, 2, 2, 0],$$

and the vector of predicted labels

$$\mathbf{p} = [0, 0, 2, 1, 0].$$

Examples

Example

- $\mathbf{t} = [0, 1, 2, 2, 0] \Rightarrow T_0 = \{1, 5\}$, $T_1 = \{2\}$, $T_2 = \{3, 4\}$,
- $\mathbf{p} = [0, 0, 2, 1, 0] \Rightarrow P_0 = \{1, 2, 5\}$, $P_1 = \{4\}$, $P_2 = \{3\}$.
- $|T_0 \cap P_0| = 2$, $|T_1 \cap P_1| = 0$, $|T_2 \cap P_2| = 1$.
- Confusion matrix (4) is given by

$$CM = \begin{array}{c|c|c|c|} & P_0 & P_1 & P_2 \\ \hline T_0 & 2 & 0 & 0 \\ \hline T_1 & 1 & 0 & 0 \\ \hline T_2 & 0 & 1 & 1 \\ \hline \end{array} .$$

From (4), we obtain the accuracy

$$ACC = \frac{2 + 0 + 1}{5} = \frac{3}{5}.$$

Examples

Example

- $T_0 = \{1,5\}\,,\ \ T_1 = \{2\}\,,\ \ T_2 = \{3,4\}\,,$
  $P_0 = \{1,2,5\}\,, P_1 = \{4\}\,,\ \ P_2 = \{3\} \Rightarrow$
- The weights (8)
$$v_0 = \frac{2}{5},\ \ v_1 = \frac{1}{5},\ \ v_2 = \frac{2}{5},$$
- and
$$\frac{|T_0 \cap P_0|}{|P_0|} = \frac{2}{3},\quad \frac{|T_1 \cap P_1|}{|P_1|} = 0,\quad \frac{|T_2 \cap P_2|}{|P_2|} = 1.$$

Thus, the weighted precision (6) equals
$$Pre = \frac{2}{5}\cdot\frac{2}{3} + \frac{1}{5}\cdot 0 + \frac{2}{5}\cdot 1 = \frac{2}{3}.$$

Examples

Example

- $T_0 = \{1, 5\}$, $T_1 = \{2\}$, $T_2 = \{3, 4\}$,
  $P_0 = \{1, 2, 5\}$, $P_1 = \{4\}$, $P_2 = \{3\} \Rightarrow$
- $$\frac{|T_0 \cap P_0|}{|T_0|} = 1, \quad \frac{|T_1 \cap P_1|}{|T_1|} = 0, \quad \frac{|T_2 \cap P_2|}{|T_2|} = \frac{1}{2}.$$

According to (9), the weighted sensitivity is equal to

$$Sen = \frac{2}{5} \cdot \frac{2}{2} + \frac{1}{5} \cdot \frac{0}{1} + \frac{2}{5} \cdot \frac{1}{2} = \frac{3}{5}.$$

Finally, using (12), we compute the weighted F_1 measure

$$F_1 = 2\left(\frac{2}{5} \cdot \frac{2}{3+2} + \frac{1}{5} \cdot \frac{0}{1+1} + \frac{2}{5} \cdot \frac{1}{2+1}\right) = \frac{44}{75}.$$

One can easily verify that the same results can be obtained using the scikit-learn library scikit-learn (Python).

Area under the Receiver Operating Characteristic Curve (AUC)

- In many data mining applications, however, accuracy, precision, sensitivity and F_1 score do not suffice.
- One of the most important measures of a classifier's performance – and according to some studies, even the most important measure – is the AUC measure:

$$AUC = \frac{1}{r\,(r-1)} \sum_{j \in L}^{r} \sum_{\substack{k \in L \\ k \neq j}}^{r} \left(AUC_{j,k} + AUC_{k,j} \right), \tag{13}$$

where

- r is the number of classes, $(r = |L|)$.
- $AUC_{p,q}$ is the AUC with class p as the positive class and q as the negative class $(AUC_{p,q} \neq AUC_{q,p})$.
- We can compute the average AUC of all possible pairwise combinations of classes; i.e., after using one-versus-one (AVA) method.

AUC

Example

- $AUC_{p,q}$ is the AUC with class p as the positive class and q as the negative class.
- Simple dataset: $D = D_1 \cup D_0$, where
- $D_1 = \{2, 9, 0, 1, 9, 9, 5, 0, 8, 5, 2, 9, 6\}$ – class "1"
- $[2, 9, 0, 1, 9, 9, 5, 0, 8, 5, 2, 9, 6]$,
- $E(D_1) = 5.0$, $\sigma(D_1) = \sqrt{E[(D_1 - E(D_1))]^2} = 3.4862$,
- $D_0 = \{-5, 3, 0, -2, -8, -2, -4, 5, -1, 6, -2, 1, -4\}$ – class "0"
- $E(D_0) = -1.0$, $\sigma(D_0) = \sqrt{E[(D_0 - E(D_0))]^2} = 3.8431$
- Let $f_1(x)$ approximates D_1 and $f_0(x)$ approximates D_0
- Define

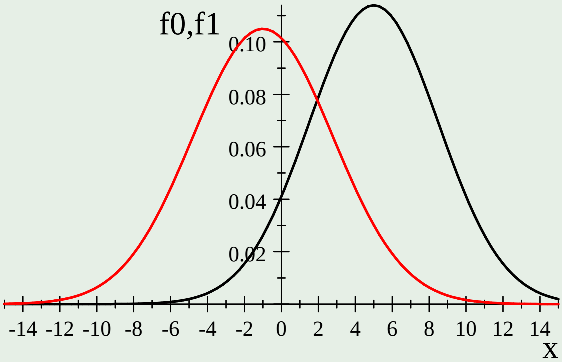
$$FPR(T) = 1 - Spe(T) = \int_T^{\infty} f_0(x)dx$$

$$TPR(T) = Sen(T) = \int_T^{\infty} f_1(x)dx$$

AUC – cont.

Example

Plot: $f_1(x) = \dfrac{1}{3.5\sqrt{2\pi}} \exp(-(x-5)^2/(2*3.5^2))$,
$f_0(x) = \dfrac{1}{3.8\sqrt{2\pi}} \exp(-(x+1)^2/(2*3.8^2))$



AUC – cont.

Example

$T = 0 \Rightarrow FPR(T) = \int_0^\infty f_0(x)dx = 0.3962,$
$TPR(T) = \int_0^\infty f_1(x)dx = 0.9234$
$T = 4 \Rightarrow FPR(T) = \int_4^\infty f_0(x)dx = 0.0941,$
$TPR(T) = \int_4^\infty f_1(x)dx = 0.6124$