Classifier evaluation

Victor Kitov

v.v.kitov@yandex.ru

Confusion matrix

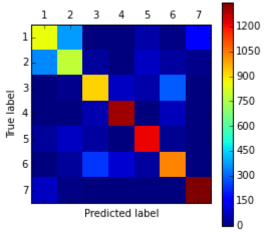
Confusion matrix $M = \{m_{ij}\}_{i,j=1}^{C}$ shows the number of ω_i class objects predicted as belonging to class ω_j.

$$
\begin{array}{ccc}
 & & \text{Forecasted classes} \\
 & & \begin{array}{cccc} 1 & 2 & \cdots & C \end{array} \\
\rotatebox{90}{\text{True classes}} & \begin{array}{c} 1 \\ 2 \\ \vdots \\ C \end{array} & \begin{bmatrix} n_{11} & n_{12} & & \\ n_{21} & n_{22} & & \\ & & \ddots & \\ & & & n_{CC} \end{bmatrix}
\end{array}
$$

Diagonal elements correspond to correct classifications and off-diagonal elements - to incorrect classifications.

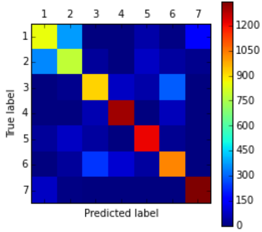
Example of confusion matrix visualization

Example of confusion matrix visualization



Example of confusion matrix visualization

Example of confusion matrix visualization



- We see here that errors here are concentrated at distinguishing between classes 1 and 2.
- We can
  - unite classes 1 and 2 into new class «1+2»
  - then solve 6-class classification problem
  - separate classes 1 and 2 for all objects assigned to class «1+2» with a separate classifier.

2 class case

Confusion matrix:

Prediction

True class		+	-
	+	TP (true positives)	FN (false negatives)
	-	FP (false positives)	TN (true negatives)

P and N - number of observations of positive and negative class.

$$P = TP + FN, \quad N = TN + FP$$

2 class case

Confusion matrix:

True class \ Prediction	+	-
+	TP (true positives)	FN (false negatives)
-	FP (false positives)	TN (true negatives)

P and N - number of observations of positive and negative class.

$$P = TP + FN, \quad N = TN + FP$$

Accuracy:	$\frac{TP+TN}{P+N}$
Error rate:	1-accuracy$=\frac{FP+FN}{P+N}$

2 class case

Confusion matrix:

		Prediction	
		+	-
True class	+	TP (true positives)	FN (false negatives)
	-	FP (false positives)	TN (true negatives)

P and N - number of observations of positive and negative class.

$$P = TP + FN, \quad N = TN + FP$$

Accuracy:	$\frac{TP+TN}{P+N}$
Error rate:	1-accuracy=$\frac{FP+FN}{P+N}$

Not informative for skewed classes and one class of interest!

“Positive class” quality metrics

FPR (error rate on negatives):	$\frac{FP}{N}$
TPR (correct rate on positives):	$\frac{TP}{P}$
Precision:	$\frac{TP}{TP+FP}$
Recall:	$\frac{TP}{P}$
F-measure:	$\frac{2}{\frac{1}{Precision}+\frac{1}{Recall}}$
Weighted F-measure:	$\frac{1}{\frac{\beta^2}{1+\beta^2}\frac{1}{Precision}+\frac{1}{1+\beta^2}\frac{1}{Recall}}$

Class label versus class probability evaluation[1]

- **Discriminability quality measures** evaluate class label prediction.
    - examples: error rate, precision, recall, etc..

[1] Give example when class labels are predicted optimally, but class probabilities - not.

Class label versus class probability evaluation[1]

- **Discriminability quality measures** evaluate class label prediction.
  - examples: error rate, precision, recall, etc..
- **Reliability quality measures** evaluate class probability prediction.
  - Example: probability likelihood:

$$\prod_{i=1}^{N} \widehat{p}(y_i|x_i)$$

  - Brier score:

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C}\left(\mathbb{I}[y_n = c] - \widehat{p}(y = c|x_n)\right)^2$$

[1] Give example when class labels are predicted optimally, but class probabilities - not.

Table of Contents

Bayes decision rule

- Loss matrix:

forecasted class

true class

	f=1	f=2
y=1	0	λ_1
y=2	λ_2	0

Bayes decision rule

- Expected loss $f=1$:
  $L(f=1)=\lambda_2 p(y=2|x)=\lambda_2 p(y=2)p(x|y=2)/p(x)$
- Expected loss $f=2$:
  $L(f=2)=\lambda_1 p(y=1|x)=\lambda_1 p(y=1)p(x|y=1)/p(x)$
- *Bayes decision rule* minimizes expected loss:

$$\widehat{y}=\arg\min_f L(f)$$

- This is equivalent to:
  $\widehat{y}=1 \Leftrightarrow \lambda_2 p(y=2)p(x|y=2)<\lambda_1 p(y=1)p(x|y=1) \Leftrightarrow$

$$\frac{p(x|y=1)}{p(x|y=2)}>\frac{\lambda_2 p(y=2)}{\lambda_1 p(y=1)}=\mu$$

Discriminant decision rules

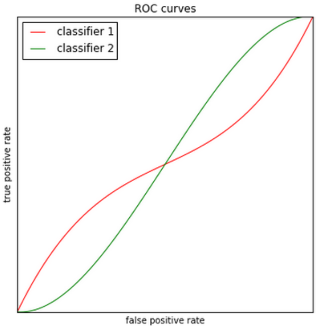
- Decision rule based on discriminant functions:
  - predict $\omega_1 \iff g_1(x) - g_2(x) > \mu$
  - predict $\omega_1 \iff g_1(x)/g_2(x) > \mu \quad$ (for $g_1(x) > 0$, $g_2(x) > 0$)
- Decision rule based on probabilities:
  - predict $\omega_1 \iff P(\omega_1|x) > \mu$

ROC curve[2]

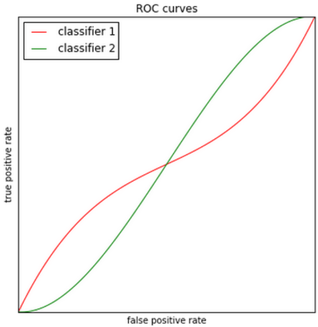
- ROC curve - is a function TPR(FPR).
- It shows how the probability of correct classification on positive classes (“recognition rate”) changes with probability of incorrect classification on negative classes (“false alarm”).
- It is build as a set of points TPR(μ), FPR(μ).
- If $\mu \downarrow$, the algorithm predicts ω_1 more often and
    - TPR=$1 - \varepsilon_1 \uparrow$
    - FPR=$\varepsilon_2 \uparrow$
- Characterizes classification accuracy for different μ.
    - more concave ROC curves are better

[2] Prove that diagonal ROC corresponds to random assignment of ω_1 and ω_2 with probabilities p and $1-p$.

Comparison of classifiers using ROC curves



Comparison of classifiers using ROC curves



How to compare different classifiers?

Area under the curve

- AUC - area under the ROC curve:
  - global quality characteristic for different μ
  - AUC$\in [0,1]$
    - AUC=0.5 - equivalent to random guessing
    - AUC=1 - no errors classification.
  - AUC property: it is equal to probability that for 2 random objects $x_1 \in \omega_1$ and $x_2 \in \omega_2$ it will hold that: $\widehat{p}(\omega_1|x_1) > \widehat{p}(\omega_2|x)$