Classifier evaluation

Victor Kitov

Skoltech
Skolkovo Institute of Science and Technology

November-December 2015.

Confusion matrix

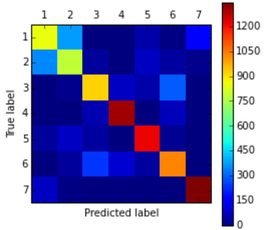
Confusion matrix $M = \{m_{ij}\}_{i,j=1}^{C}$ shows the number of ω_i class objects predicted as belonging to class ω_j.

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

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

Example of confusion matrix visualization

Example of confusion matrix visualization



- 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, and finally 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.

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

Quality metrics

Accuracy:	$\frac{TP+TN}{P+N}$
Error rate:	1-accuracy=$\frac{FP+FN}{P+N}$
FPR (error rate on negatives):	$\frac{FP}{N}$
TPR (error 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

- **Discriminability quality measures** evaluate class label prediction.
  - examples: previously mentioned measures: 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:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{C}\left(\mathbb{I}[x_i \in \omega_c] - \widehat{p}(\omega_c|x_i)\right)^2$$

- Example when class labels are predicted accurately, but class probabilities - not.

Table of Contents

Bayes decision rule

- Definition: $\widehat{\omega}_i$ means, that «prediction is equal to ω_i»
- Loss matrix:

predicted class

true class		$\widehat{\omega}_1$	$\widehat{\omega}_2$
	ω_1	0	λ_1
	ω_2	λ_2	0

- λ_1, λ_2 - costs of incorrect classification of objects, belonging to classes ω_1 and ω_2 respectively.

Bayes decision rule

- Expected loss of prediction $\widehat{\omega}_1$:
  $L(\widehat{\omega}_1) = \lambda_2 p(\omega_2|x) = \lambda_2 p(\omega_2)p(x|\omega_2)/p(x)$
- Expected loss of prediction $\widehat{\omega}_2$:
  $L(\widehat{\omega}_2) = \lambda_1 p(\omega_1|x) = \lambda_1 p(\omega_1)p(x|\omega_1)/p(x)$
- *Bayes decision rule* minimizes expected loss:

$$\widehat{\omega}^* = \arg\min_{\widehat{\omega}} L(\widehat{\omega})$$

- This is equivalent to:
  $\widehat{\omega}* = \widehat{\omega}_1 \Leftrightarrow \lambda_2 p(\omega_2)p(x|\omega_2) < \lambda_1 p(\omega_1)p(x|\omega_1) \Leftrightarrow$

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{\lambda_2 p(\omega_2)}{\lambda_1 p(\omega_1)} = \mu$$

Discriminant decision rules

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

ROC curve

- 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=$\varepsilon_2 \uparrow$
- Diagonal points correspond to random assignment of ω_1 and ω_2 with probabilities p and $1-p$.
- Characterizes classification accuracy for different μ.
    - more concave ROC curves are better

Iso-loss lines

- Define $\varepsilon_1, \varepsilon_2$ - probabilities of error on objects of class ω_1 and ω_2 respectively.
- $1 - \varepsilon_1 = TPR$, $\varepsilon_2 = FPR$
- Expected loss:

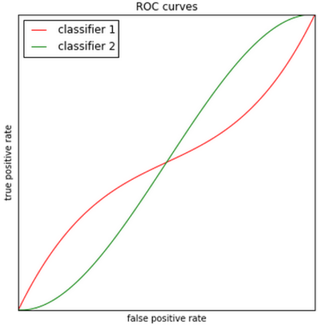
$$L = \lambda_2 p(\omega_2)\varepsilon_2 + \lambda_1 p(\omega_1)\varepsilon_1 = \lambda_2 p(\omega_2)\varepsilon_2 - \lambda_1 p(\omega_1)(1-\varepsilon_1) + \lambda_1 p(\omega_1)$$

- Iso-loss line:

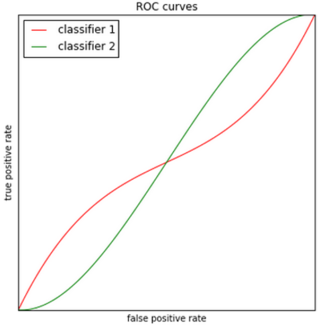
$$(1-\varepsilon_1) = \frac{\lambda_2 p(\omega_2)}{\lambda_1 p(\omega_1)}\varepsilon_2 + \frac{\lambda_1 p(\omega_1) - L}{\lambda_1 p(\omega_1)}$$

- In the optimal point iso-loss line is tangent to the ROC curve with slope of the curve equal to $\frac{\lambda_2 p(\omega_2)}{\lambda_1 p(\omega_1)}$

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: $\widehat{p}(\omega_1|x_1) > \widehat{p}(\omega_2|x)$

Bayes decision rule with uncertainty about λ_1 and λ_2

- Predefined λ_1, λ_2: too specific.
  - estimate losses associated with yield point estimates of classifiers
- Undefined λ_1, λ_2: too broad
  - compare AUC of different classifiers

LC index

- LC index - classifier comparison in intermediary case:
  1. Scale λ_1 and λ_2 so that $\lambda_1 + \lambda_2 = 1$
  2. define $\lambda_1 = \lambda$, $\lambda_2 = 1 - \lambda$
  3. for each $\lambda \in [0, 1]$ calculate
     $$L(\lambda) = \begin{cases} +1 & \text{if 1st classifier is better} \\ -1 & \text{if 2nd classifier is better} \end{cases}$$
  4. define probability density distribution of λ: $p(\lambda)$ (for example, from "triangular" class)
  5. select classifier 1 if $\int_0^1 L(\lambda)p(\lambda)d\lambda > 0$ and classifier 2 otherwise.

Error rate distribution

- Define e - probability of error on the new object.
- What is the distribution of e?
  - we know that on held-out sample of size n there were k mistakes.

Probability to make k mistakes on sample of size n:

$$p(k|e,n) = \begin{pmatrix} n \\ k \end{pmatrix} e^k (1-e)^{n-k}$$

Then

$$p(e|k,n) = \frac{p(e,k|n)}{p(k|n)} = \frac{p(k|e,n)p(e|n)}{\int p(k|n)p(e|n)de}$$

Assuming no prior knowledge about the error rate $p(e|n) \equiv const$, we obtain

$$p(e|k,n) = \frac{p(k|e,n)}{\int p(k|n)de} \propto e^k(1-e)^{n-k}$$

Error rate distribution

Since the beta distribution has the form
$Be(x|\alpha,\beta)=[\Gamma(\alpha+\beta)/(\Gamma(\alpha)\Gamma(\beta))]x^{\alpha-1}(1-x)^{\beta-1}$, so

$$p(e|k,n)\sim Be(k+1,n-k+1)$$

Beta distribution:
$\xi\sim Be(\alpha,\beta)\Rightarrow\mathbb{E}[\xi]=\frac{\alpha}{\alpha+\beta},\ Var[\xi]=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$