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.

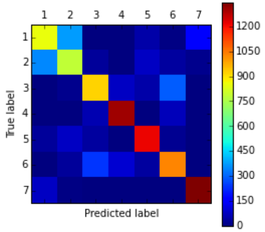
Forecasted classes

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

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

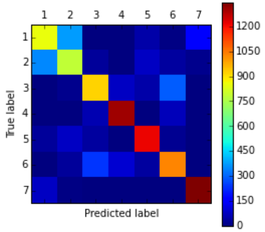
Example of confusion matrix visualization

Example of confusion matrix visualization



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		
		+	-	total
True class	+	TP (true positives)	FN (false negatives)	P
	-	FP (false positives)	TN (true negatives)	N
	total	$\widehat{P}$	$\widehat{N}$	

2 class case

Confusion matrix:

		Prediction		
		+	-	total
True class	+	TP (true positives)	FN (false negatives)	P
	-	FP (false positives)	TN (true negatives)	N
	total	$\widehat{P}$	$\widehat{N}$	

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

2 class case

Confusion matrix:

Prediction

True class		+	-	total
	+	TP (true positives)	FN (false negatives)	P
	-	FP (false positives)	TN (true negatives)	N
	total	$\widehat{P}$	$\widehat{N}$	

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

Precision	$\frac{TP}{\widehat{P}}$
Recall (=TPR)	$\frac{TP}{P}$
F-measure	$\frac{2}{\frac{1}{Precision}+\frac{1}{Recall}}$
Weighted F-measure	$\frac{1}{\frac{\beta^2}{1+\beta^2}\frac{1}{Precision}+\frac{1}{1+\beta^2}\frac{1}{Recall}}$

TPR (=recall)	$\frac{TP}{P}$
FPR	$\frac{FP}{N}$

- TPR = correct rate on positives, recognition rate
- FRP = error rate on negatives, false alarm.

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:

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

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

Table of Contents

Discriminant decision rules

- Define $g(x) = g_{+1}(x) - g_{-1}(x)$.
- Standard classification $\widehat{y}(x) = \text{sign}(g(x))$.
- Classification with variable eagerness to assign $\widehat{y} = +1$:

$$\widehat{y}(x) = \text{sign}(g(x) - \alpha)$$

  - small α: more $\widehat{y} = +1$
  - large α: less $\widehat{y} = -1$.
- Use case: unequal costs: $\lambda_{+1} \neq \lambda_{-1}$
  - $\lambda_{+1} = \text{cost}(\widehat{y} = -1 | \, y = +1)$
  - $\lambda_{-1} = \text{cost}(\widehat{y} = +1 | \, y = -1)$
- Costs may vary depending on regime:
  - target detection in peace/war time.
  - credit scoring during economic growth/recession.

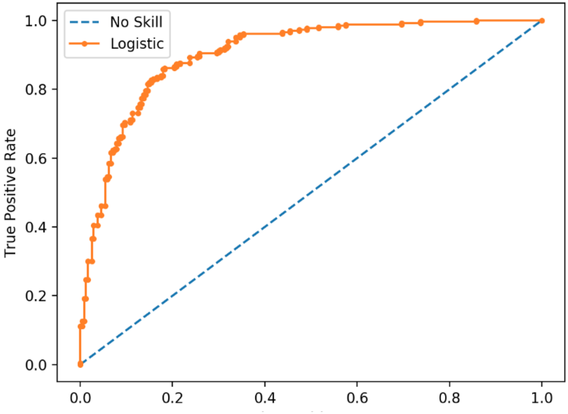
ROC curve[2]

- $TPR = TPR(\alpha)$, $FPR = FPR(\alpha)$.
- ROC curve is a function TPR(FPR).
- Questions:
  - How TPR and FPR change with α?
  - Higher ROC corresponds to better of worse classifier?
  - ROC curve for random guessing.
  - How to improve classifier with concave ROC curve?
  - How ROC curve will change for inverted classifier $(g(x) := -g(x))$?
- AUC - general quality measure.

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

ROC curve: TPR(FPR)

$$TPR = \frac{TP}{P} \qquad FPR = \frac{FP}{N}$$



Properties of ROC curve

- **Invariance to monotone transform**
    - $\widehat{y}(x) = \text{sign}(g(x) > \alpha) \iff \text{sign}(f(g(x)) > f(\alpha))$, for any $\uparrow f(\cdot)$
    - $TPR(\alpha), FPR(\alpha)$ corresponds to $TPR(f(\alpha)), FPR(f(\alpha))$.
- **AUC=proportion of correctly ordered object pairs**
    - take random $(x_1, y_1 = -1)$
    - take random $(x_2, y_2 = +1)$
    - AUC=probability that ordering by $g(x)$ is correct:

$$AUC = p\left(g(x_1) < g(x_2)\right)$$

Order objects $x_{(1)}, ...x_{(N)}$ by $g\left(x_{(1)}\right) < g\left(x_{(2)}\right) < ... < g\left(x_{(N)}\right)$.

Proof that AUC=proportion of correctly ordered pairs

$$\mathrm{TPR}(\tau) = \frac{\sum_{i=1}^{N}[y_i = +1][A(x_i) \ge \tau]}{N_+} \quad \text{and} \quad \mathrm{FPR}(\tau) = \frac{\sum_{i=1}^{N}[y_i = -1][A(x_i) \ge \tau]}{N_-}$$

(where [boolean expression] is 1 if expression is true, and 0 otherwise). Then, ROC curve is built from points of the form $(\mathrm{FPR}(\tau), \mathrm{TPR}(\tau))$ for different values of τ. Moreover, it's easy to see that if we order our samples $x_{(i)}$ (note the parentheses) according to the algorithm's output $A(x_i)$, then neither TPR nor FPR changes for τ between consecutive samples $A(x_{(i)}) < \tau < A(x_{(i+1)})$. So it's enough to evaluate FPR and TPR only for $\tau \in \{A(x_{(1)}), \dots, A(x_{(N)})\}$. For k^{th} point we have

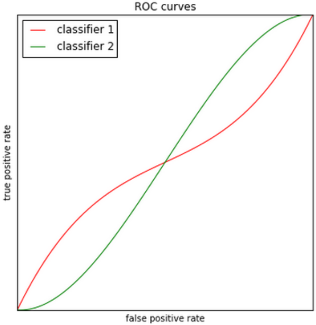
$$\mathrm{TPR}_k = \frac{\sum_{i=k}^{N}[y_{(i)} = +1]}{N_+} \quad \text{and} \quad \mathrm{FPR}_k = \frac{\sum_{i=k}^{N}[y_{(i)} = -1]}{N_-}$$

(Note both sequences are non-increasing in k). These sequences define `x` and `y` coordinates of points on the ROC curve. Next, we linearly interpolate these points to get the curve itself and calculate area under the curve (Using a formula for area of a trapezoid):

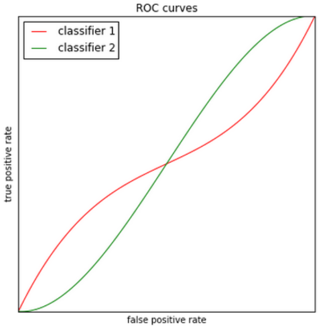
$$\begin{aligned}
\mathrm{AUC} &= \sum_{k=1}^{N-1} \frac{\mathrm{TPR}_{k+1} + \mathrm{TPR}_k}{2}(\mathrm{FPR}_k - \mathrm{FPR}_{k+1}) \\
&= \sum_{k=1}^{N-1} \frac{\sum_{i=k+1}^{N}[y_{(i)} = +1] + \frac{1}{2}[y_{(k)} = +1]}{N_+} \frac{[y_{(k)} = -1]}{N_-} \\
&= \frac{1}{N_+N_-} \sum_{k=1}^{N-1} \sum_{i=k+1}^{N} [y_{(i)} = +1][y_{(k)} = -1] = \frac{1}{N_+N_-} \sum_{k<i} [y_{(k)} < y_{(i)}]
\end{aligned}$$

Here I used the fact that $[y = -1][y = +1] = 0$ for any y.

Comparison of classifiers using ROC curves



Comparison of classifiers using ROC curves



How to compare different classifiers?

- Fixed $\lambda_{+1}, \lambda_{-1}$: point on ROC curve
- Unknown $\lambda_{+1}, \lambda_{-1}$: by AUC
- Approximate $\lambda_{+1}, \lambda_{-1}$: by LC index

Iso-loss lines

- Mistake probabilities: $p(\widehat{y}=-1|y=+1)=1-TPR$, $p(\widehat{y}=+1|y=-1)=FPR$
- Iso-loss line (loss$\equiv L$):

$$L=p(y=+1)(1-TPR)\lambda_{+1}+p(y=-1)FPR\lambda_{-1}$$
$$(TPR-1)\,p(y=+1)\lambda_{+1}=-L+\lambda_{-1}p(y=-1)FPR$$
$$TPR=1+\frac{\lambda_{-1}p(y=-1)FPR-L}{\lambda_{+1}p(y=+1)}$$

- Optimal point: iso-loss line tangent to ROC, so ROC slope $=\frac{\lambda_{-1}p(y=-1)}{\lambda_{+1}p(y=+1)}$

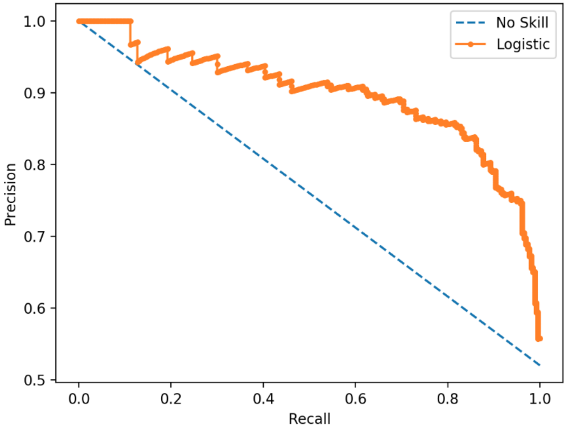
Area under the curve

AUC - area under the ROC curve:

- global quality characteristic for different α
- AUC$\in [0, 1]$
- AUC=0.5 - equivalent to random guessing
- AUC=1 - no errors classification.
- AUC p

Precision-recall curve: precision(recall)

$$Precision = \frac{TP}{\widehat{P}} \qquad Recall = \frac{TP}{P}$$



Conclusion

- Confusion matrix: localization of hardly separable classes
- Precision, recall or (TPR, FPR) for imbalanced classes
- Label prediction vs. class probability prediction.
- Unequal costs:
  - F-measure, Precision, Recall
  - ROC curve best point
  - AUC-ROC