Classification Evaluation

Albert Bifet (@abifet)

Paris, 27 September 2016

albert.bifet@telecom-paristech.fr

Evaluation

1. Error estimation: *Hold-out or Cross-Validation*
2. Evaluation performance measures: *Accuracy or κ-statistic*
3. Statistical significance validation: *MacNemar or Nemenyi test*

Evaluation Framework

Error Estimation

Data available for testing

- Holdout an independent test set
- Apply the current decision model to the test set
- The loss estimated in the holdout is an unbiased estimator

Holdout Evaluation

1. Error Estimation

Not enough data available for testing

- Divide dataset in 10 folds
- Repeat 10 times: use one fold for testing and the rest for training

k-fold Cross-validation

2. Evaluation performance measures

	Predicted Class+	Predicted Class-	Total
Correct Class+	75	8	83
Correct Class-	7	10	17
Total	82	18	100

Table: Simple confusion matrix example

2. Evaluation performance measures

	Predicted Class+	Predicted Class-	Total
Correct Class+	tp	fn	tp+fn
Correct Class-	fp	tn	fp+tn
Total	tp+fp	fn+tn	N

Table: Simple confusion matrix example

- Precision = $\frac{tp}{tp+fp}$
- Recall = $\frac{tp}{tp+fn}$
- $F_1 = 2 \cdot \frac{\text{precision}\cdot\text{recall}}{\text{precision}+\text{recall}}$

2. Evaluation performance measures

	Predicted Class+	Predicted Class-	Total
Correct Class+	75	8	83
Correct Class-	7	10	17
Total	82	18	100

Table: Simple confusion matrix example

- Accuracy = $\frac{75}{100} + \frac{10}{100} = \frac{75}{83}\frac{83}{100} + \frac{10}{17}\frac{17}{100} = 85\%$
- Arithmetic mean = $(\frac{75}{83} + \frac{10}{17})/2 = 74.59\%$
- Geometric mean = $\sqrt{\frac{75}{83}\frac{10}{17}} = 72.90\%$

2. Performance Measures with Unbalanced Classes

	Predicted Class+	Predicted Class-	Total
Correct Class+	75	8	83
Correct Class-	7	10	17
Total	82	18	100

Table: Simple confusion matrix example

	Predicted Class+	Predicted Class-	Total
Correct Class+	68.06	14.94	83
Correct Class-	13.94	3.06	17
Total	82	18	100

Table: Confusion matrix for chance predictor

2. Performance Measures with Unbalanced Classes

Kappa Statistic

- p_0: classifier's prequential accuracy
- p_c: probability that a chance classifier makes a correct prediction.
- κ statistic
$$\kappa = \frac{p_0 - p_c}{1 - p_c}$$
- $\kappa = 1$ if the classifier is always correct
- $\kappa = 0$ if the predictions coincide with the correct ones as often as those of the chance classifier

Matthews correlation coefficient (MCC)

$$\frac{tp \times tn - fp \times fn}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}}$$

2. Evaluation performance measures

	Predicted Class+	Predicted Class-	Total
Correct Class+	tp	fn	tp+fn
Correct Class-	fp	tn	fp+tn
Total	tp+fp	fn+tn	N

Table: Simple confusion matrix example

AUC Area under the curve

A ROC space is defined by FPR and TPR (recall)

- FPR = $\frac{fp}{fp+tp}$
- TPR = $\frac{tp}{tp+fn}$

3. Statistical significance validation (2 Classifiers)

	Classifier A Class+	Classifier A Class-	Total
Classifier B Class+	c	a	c+a
Classifier B Class-	b	d	b+d
Total	c+b	a+d	a+b+c+d

$$M = |a - b - 1|^2/(a + b)$$

The test follows the χ^2 distribution. At 0.99 confidence it rejects the null hypothesis (the performances are equal) if $M > 6.635$.

McNemar test

3. Statistical significance validation ($>$ 2 Classifiers)

Two classifiers are performing differently if the corresponding average ranks differ by at least the critical difference

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}$$

- k is the number of learners, N is the number of datasets,
- critical values q_α are based on the Studentized range statistic divided by $\sqrt{2}$.

Nemenyi test

3. Statistical significance validation ($>$ 2 Classifiers)

Two classifiers are performing differently if the corresponding average ranks differ by at least the critical difference

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}$$

- k is the number of learners, N is the number of datasets,
- critical values q_α are based on the Studentized range statistic divided by $\sqrt{2}$.

# classifiers	2	3	4	5	6	7
$q_{0.05}$	1.960	2.343	2.569	2.728	2.850	2.949
$q_{0.10}$	1.645	2.052	2.291	2.459	2.589	2.693

Table: Critical values for the Nemenyi test