Binary classification performances measure cheat sheet

Damien François – v1.1 - 2009 (damien.francois@uclouvain.be)

Confusion matrix for two possible outcomes *p* (positive) and *n* (negative)

		Actual		
		p	n	Total
Predicted	p'	**true positive**	**false postive**	P
	n'	**false negative**	**true negative**	N
	total	P'	N'	

Classification accuracy
(TP + TN) / (TP + TN + FP + FN)
Error rate
(FP + FN) / (TP + TN + FP + FN)

Paired criteria

Precision: (or Positive predictive value) proportion of predicted positives which are actual positive
TP / (TP + FP)

Recall: proportion of actual positives which are predicted positive
TP / (TP + FN)

Sensitivity: proportion of actual positives which are predicted positive
TP / (TP + FN)

Specificity: proportion of actual negative which are predicted negative
TN / (TN + FP)

True positive rate: proportion of actual positives which are predicted positive
TP / (TP + FN)

True negative rate: proportion of actual negative which are predicted negative
TN / (TN + FP)

Positive likelihood: likelihood that a predicted positive is an actual positive
sensitivity / (1 - specificity)

Negative likelihood: likelihood that a predicted negative is an actual negative
(1 - sensitivity) / specificity

Combined criteria

BCR: Balanced Classification Rate
½ (TP / (TP + FN) + TN / (TN + FP))
BER: Balanced Error Rate, or **HTER**:
Half Total Error Rate: 1 - BCR

F-measure harmonic mean between precision and recall
2 (precision . recall) /
(precision + recall)
F_β-measure weighted harmonic mean between precision and recall
$(1+\beta)^2$ TP / ($(1+\beta)^2$ TP + β^2 FN + FP)

The harmonic mean between specificity and sensitivity is also often used and sometimes referred to as F-measure.

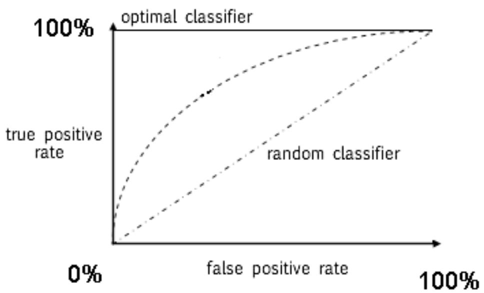
Youden's index: arithmetic mean between sensitivity and specificity
sensitivity - (1 - specificity)

Matthews correlation correlation between the actual and predicted
(TP . TN – FP . FN) /
$((TP+FP)(TP+FN)(TP+FP)(TN+FN))^{1/2}$
comprised between -1 and 1

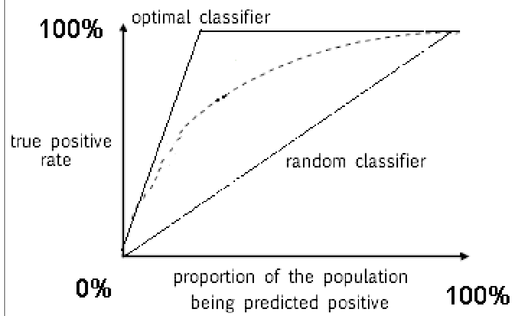
Discriminant power normalised likelihood index
sqrt(3) / ·J .
(log (sensitivity / (1 – specificity)) +
log (specificity / (1 - sensitivity)))
<1 = poor, >3 = good, fair otherwise

Graphical tools

ROC curve receiver operating characteristic curve : 2-D curve parametrized by one parameter of the classification algorithm, e.g. some threshold in the « true postivie rate / false positive rate » space
AUC The area under the ROC is between 0 and 1



(Cumlative) Lift chart plot of the true positive rate as a function of the proportion of the population being predicted positive, controlled by some classifier parameter (e.g. a threshold)



Relationships

sensitivity = recall = true positive rate
specificity = true negative rate
BCR = ½ . (sensitivity + specificity)
BCR = 2 . Youden's index - 1
F-measure = F_1measure
Accuracy = 1 – error rate

References

Sokolova, M. and Lapalme, G. 2009. A systematic analysis of performance measures for classification tasks. Inf. Process. Manage. 45, 4 (Jul. 2009), 427-437.
Demsar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research **7** (2006) 1–30

Regression performances measure cheat sheet

Damien François – v0.9 - 2009 (damien.francois@uclouvain.be)

Let $D = \{(x_i, y_i)\}$ be a set of input/output pairs and f a function such that for $i = 1..n$,

$$y_i = f(x_i) + \epsilon_i$$

Squared error

SSE Sum of Squared Errors, or
RSS Residual Sum of Squares

$$\sum_i \epsilon_i^2$$

MSE Mean Squared Error

$$\frac{1}{n} \sum_i \epsilon_i^2$$

RMSE Root Mean Squared Error

$$\sqrt{\frac{1}{n} \sum_i \epsilon_i^2}$$

NMSE Normalised Mean Squared Error

$$\frac{SSE}{var(\{y_i\})}$$

where var is the empirical variance in the sample.

R-squared

$$1 - \frac{SSE}{var(y_i)}$$

where var is the empirical variance in the sample

Absolute error

MAD Mean Absolute Deviation

$$\frac{1}{n} \sum |\epsilon_i|$$

MAPE Mean Absolute Percentage Error

$$\frac{1}{n} \sum_i \frac{|\epsilon_i|}{y_i}$$

Predicted error

PRESS Predicted REsidual Sums of Squares

$$\frac{1}{n} \|diag(XX^T)(XX^T - I)Y\|_2^2$$

where X is a matrix built by stacking the x_i in rows. Y is the vector of y_i

GCV Generalised Cross Validation

$$\frac{\frac{1}{n}\|(I - X(X^T X + nI)^{-1} X^T)Y\|^2}{(\frac{1}{n} Trace(I - X(X^T X + nI)^{-1} X^T)^2}$$

where X is a matrix built by stacking the x_i in rows. Y is the vector of y_i

Information criteria

AIC Akaike Information Criterion

$$n \log MSE + 2k$$

where k is the number of parameters in the model

BIC Bayesian Information Criterion

$$n \log MSE + k \cdot \log n$$

where k is the number of parameters in the model

Robust error measures

Median Squared error

$$median(\epsilon_i^2)$$

α-trimmed MSE

$$\frac{1}{\#I} \sum_{i \in I} \epsilon_i^2$$

where I is the set of residuals ϵ_i where α percents of the largest values are discarded.

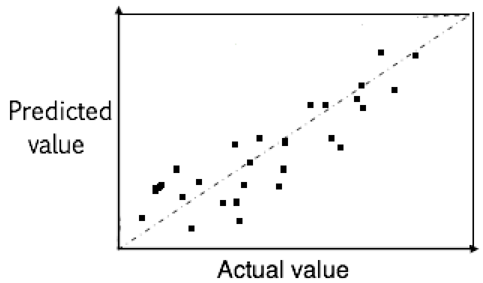
M-estimators

$$\frac{1}{n} \sum_i \rho(\epsilon_i)$$

where \rho is a non-negative function with a mininmum in 0, like the parabola, the Hubber function, or the bisquare function.

Graphical tool

Plot of predicted value against actual value. A perfect model places all dots on the diagonal.



Resampling methods

LOO – Leave-one-out: build the model on $n-1$ data elements and test on the remaining one. Iterate n times to collect all ϵ_i and compute mean error.

X-Val – Cross validation. Randomly split the data in two parts, use the first one to build the model and the second one to test it. Iterate to get a distribution of the test error of the model.

K-Fold – Cut the data into K parts. Build the model on the K-1 first parts and test on the Kth one. Iterate from 1 to K to get a distribution of the test error of the model.

Bootstrap – Draw a random subsample of the data with replacement. Compute the error on the whole dataset minus the training error of the model and Iterate to get a distribution of such values. The mean of the distribution is the optimism. The bootstrap error estimate is the training error on the whole dataset plus the optimism.