CLASSIFICATION VALIDATION METHODS

Ardo Allik

DIAGNOSTICS OF CLASSIFIERS

There are several methods and tools to evaluate how well the classifier performs.

A **confusion matrix** (or **contingency table**) is a specific table layout that allows the visualization of the performance of a classifier.

		Predicted class	
		Positive	Negative
Actual class	Positive	True Positives (TP)	False Negatives (FN)
	Negative	False Positives (FP)	True Negatives (TN)

TAL TECH

DIAGNOSTICS OF CLASSIFIERS

- **True positives (TP)** are the number of positive instances the classifier correctly identified as positive.
- **False positives (FP)** are the number of instances in which the classifier identified as positive but in reality are negative.
- **True negatives (TN)** are the number of negative instances the classifier correctly identified as negative.
- **False negatives (FN)** are the number of instances classified as negative but in reality are positive.

TP and TN are the correct guesses – a good classifier should have large TP and TN and small (ideally zero) number for FP and FN.

		Predicted class	
		Positive	Negative
Actual class	Positive	True Positives (TP)	False Negatives (FN)
	Negative	False Positives (FP)	True Negatives (TN)

TAL TECH

TRUE POSITIVE RATE

True Positive Rate (TPR), shows what percent of positive instances the classifier correctly identified. TPR is also called **sensitivity** or **recall**.

It is defined as TP divided by all the positive instances:

$$TPR = \frac{TP}{TP + FN}$$

A well-performed model should have a high TPR that is ideally 1.

TAL TECH

TRUE NEGATIVE RATE

True Negative Rate (TNR), shows what percent of negative instances the classifier correctly marked as negative. TNR is also called **specifity** or **selectivity**.

It is defined as TN divided by all the negative instances:

$$TNR = \frac{TN}{TN + FP}$$

A well-performed model should have a high TNR that is ideally 1.

TAL TECH

FALSE POSITIVE RATE

False Positive Rate (FPR), shows what percent of negatives the classifier marked as positive. FPR is also called the **false alarm rate**, **type I error rate** or **fall-out**.

It is defined as FP divided by all the negative instances:

$$FPR = \frac{FP}{FP + TN}$$

A well-performed model should have a low FPR that is ideally 0.

TAL TECH

FALSE NEGATIVE RATE

False Negative Rate (FNR), shows what percent of positives the classifier marked as negatives**.** FNR is also called the **miss rate** or **type II error rate**.

It is defined as FN divided by all the positive instances:

$$FNR = \frac{FN}{TP + FN}$$

A well-performed model should have a low FNR that is ideally 0**.**

TAL TECH

ACCURACY

Accuracy (ACC), or the overall success rate, is a metric defining the rate at which a model has classified the records correctly.

It is defined as the sum of TP and TN divided by the total number of instances:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \cdot 100\%$$

A good model should have a high accuracy score, but having a high accuracy score alone does not guarantee the model is well established.

TAL TECH

PRECISION AND F_1 SCORE

Precision is the percentage of instances marked positive that, really are positive. Precision is also called **confidence**.

It is defined as TP divided by all the instances classified as positive:

$$Precision = \frac{TP}{TP + FP}$$

F_1 score is a measure of performance that considers both precision and recall (TPR). F_1 score is also called the **F-score** or **F-measure**.

It is defined by harmonic mean of precision and recall:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

TAL TECH

		True condition			
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$	F_1 score = $\frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$

TAL TECH

IN-CLASS EXERCISE

Calculate the performance characteristics based on the given confusion matrix:

$$TPR = \frac{TP}{TP + FN}$$

$$TNR = \frac{TN}{TN + FP}$$

$$FPR = \frac{FP}{FP + TN}$$

$$FNR = \frac{FN}{TP + FN}$$

		Predicted class	
		Positive	Negative
Actual class	Positive	3	8
	Negative	2	87

$$Precision = \frac{TP}{TP + FP}$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

TAL TECH

IN-CLASS EXERCISE

Calculate the performance characteristics based on the given confusion matrix:

		Predicted class	
		Positive	Negative
Actual class	Positive	3	8
	Negative	2	87

$$TPR = \frac{TP}{TP + FN} = \frac{3}{3 + 8} \approx 0.273$$

$$TNR = \frac{TN}{TN + FP} = \frac{87}{87 + 2} \approx 0.978$$

$$FPR = \frac{FP}{FP + TN} = \frac{2}{2 + 87} \approx 0.023$$

$$FNR = \frac{FN}{TP + FN} = \frac{8}{3 + 8} \approx 0.727$$

$$Precision = \frac{TP}{TP + FP} = \frac{3}{3 + 2} = 0.6$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} = 2 \cdot \frac{0.6 \cdot 0.273}{0.6 + 0.273} = 0.375$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{3 + 87}{3 + 87 + 2 + 8} = 0.9$$

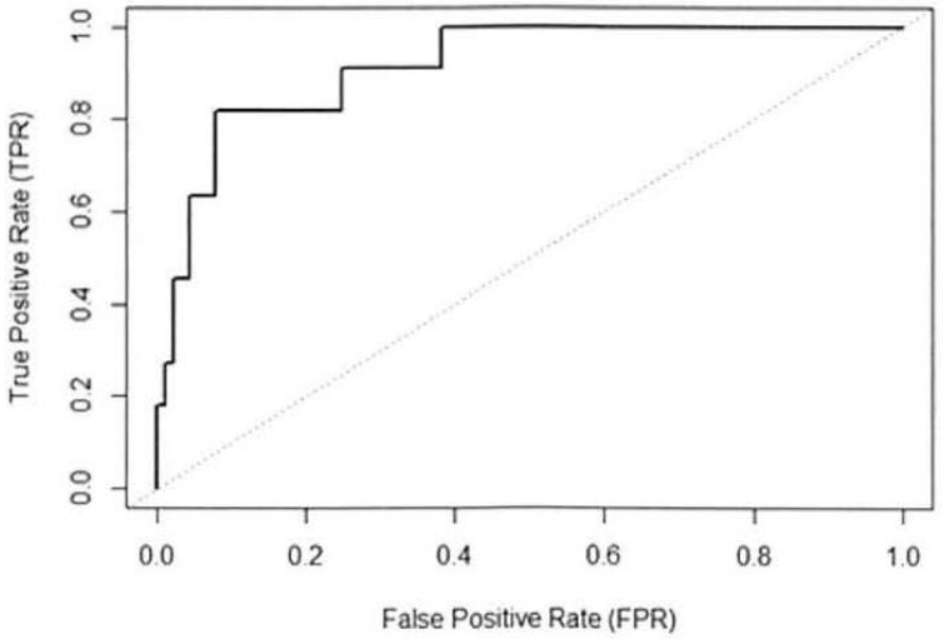
TAL TECH

RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE

ROC curve is a common tool to evaluate classifiers. The abbreviation stands for **receiver operating characteristic**, a term used in signal detection to characterize the trade-off between hit rate and false-alarm rate over a noisy channel.

A ROC curve evaluates the performance of a classifier based on the TP and FP, regardless of other factors such as class distribution and error costs.

The vertical axis is the True Positive Rate (TPR), and the horizontal axis is the False Positive Rate (FPR).



TAL TECH

EMC Education Services, Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, 2015 John Wiley @ Sons

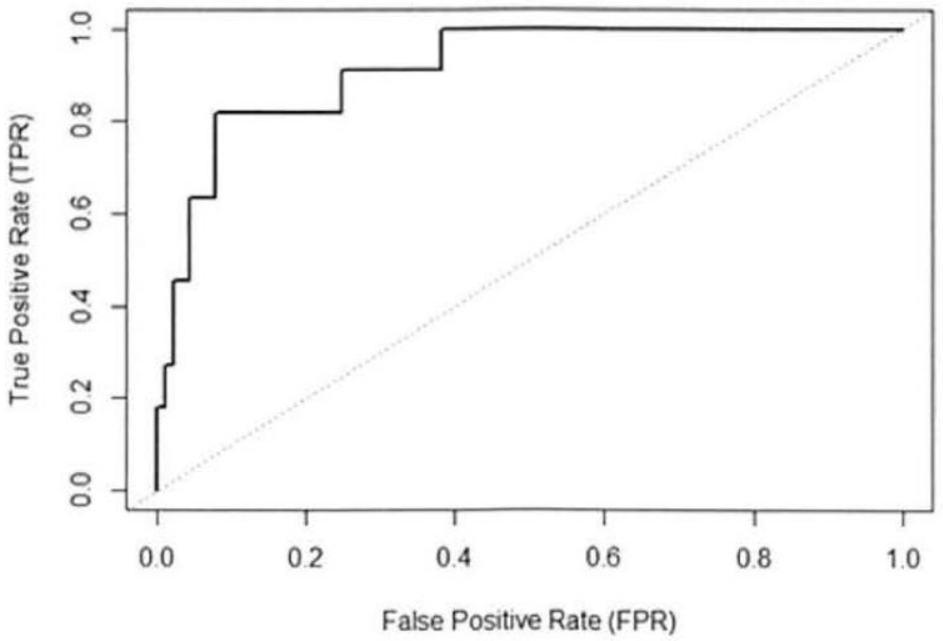
RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE

Any classifier can achieve the bottom left of the graph where TPR=FPR=0 by classifying everything as negative. Similarly, any classifier can achieve the top right of the graph where TPR=FPR=1 by classifying everything as positive.

If a classifier performs "at chance" by random guessing the results, it can achieve any point on the diagonal line TPR=FPR.

The ROC curve of ideal classifiers goes straight up from TPR=FPR=0 to the top-left corner (TPR=1, FPR=0) and moves straight right to the top-right corner (TPR=1, FPR=1)

In reality, it can be difficult to achieve the top-left corner. But a better classifier should be closer to the top left, separating it from ohter classifiers that are closer to the diagonal line.

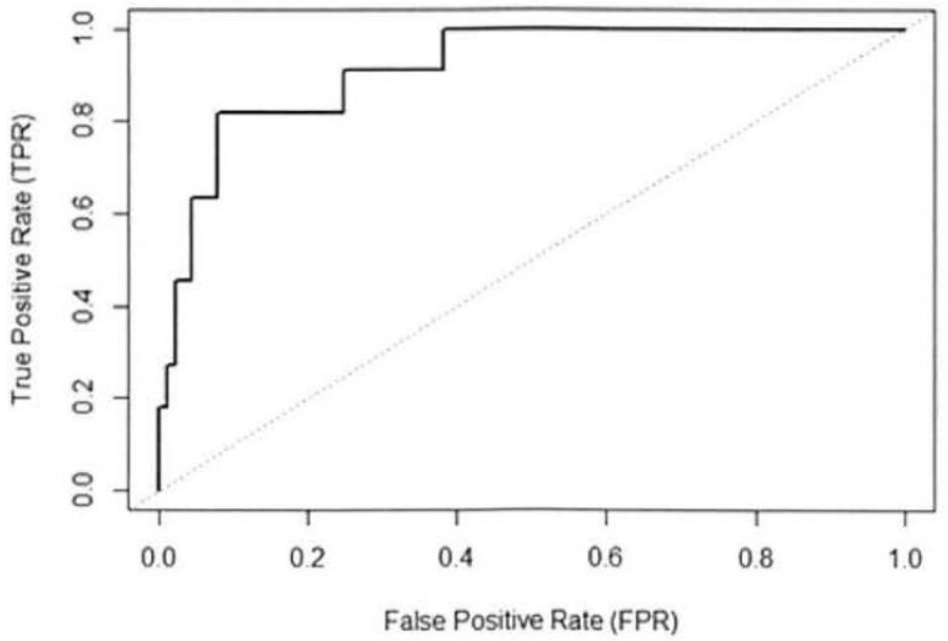


EMC Education Services, Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, 2015 John Wiley @ Sons

AREA UNDER THE CURVE (AUC)

Related to the ROC curve is the area under the curve (AUC). The AUC is calculated by measuring the area under the ROC curve. Higher AUC scores mean the classifier performs better.

The score can range from 0.5 (for the diagonal line TPR=FPR) to 1.0 (with ROC passing through the top-left corner).



TAL TECH

EMC Education Services, Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, 2015 John Wiley @ Sons

CROSS-VALIDATION METHODS

Cross-validation methods are used to assess how the results of the classification will generalize to an independent data set.

During validation dataset is usually divided into two datasets:

- **Training data** – dataset that is used for classifier training
- **Testing data** – dataset the classifier is tested against

TAL TECH

CROSS-VALIDATION METHOD TYPES

Cross-validation can be distinguished by two types:

Exhaustive cross-validation methods learn and test on all possible ways to divide the original sample into training and a validation set.

- **Leave-one-out cross-validation**
- **Leave-p-out cross-validation**

Non-exhaustive cross-validation methods do not compute all the ways of splitting the original sample.

- **k-fold cross-validation**
- **Holdout method**

TAL TECH

LEAVE-ONE-OUT CROSS-VALIDATION

Example: n=4

	1	2	3	4
#1	TEST	Train	Train	Train
#2	Train	TEST	Train	Train
#3	Train	Train	TEST	Train
#4	Train	Train	Train	TEST

TAL TECH

LEAVE-P-OUT CROSS-VALIDATION

Example: n=4 ; p=2

	1	2	3	4
#1	TEST	TEST	Train	Train
#2	TEST	Train	TEST	Train
#3	TEST	Train	Train	TEST
#4	Train	TEST	TEST	Train
#5	Train	TEST	Train	TEST
#6	Train	Train	TEST	TEST

TAL TECH

K-FOLD CROSS-VALIDATION

Example: n=6 ; k=3

	2	1	4	5	6	3
#1	TEST	TEST	Train	Train	Train	Train
#2	Train	Train	TEST	TEST	Train	Train
#3	Train	Train	Train	Train	TEST	TEST

If n=k then same as leave-one-out cross-validation

TAL TECH

HOLDOUT METHOD

Training 75% & Testing 25%

	2	8	4	5	6	3	1	7
#1	TEST	TEST	Train	Train	Train	Train	Train	Train

TAL TECH

TALLINNA TEHNIKAÜLIKOOL

Ehitajate tee 5, 19086 Tallinn, Tel 620 2002 (E-R 8.30–17.00)

taltech.ee