Validation / Evaluation

Christian Salvatore
Scuola Universitaria Superiore IUSS Pavia

VALIDATION & PERFORMANCE EVALUATION

How to test a machine-learning classifier?

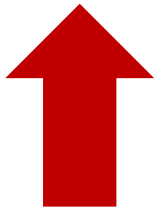
A good validation process allows to obtain a minimally biased estimate of the true diagnostic performance of the classifier

- Correct quantification of the discriminatory power of a given model (model evaluation)
- Possibility to compare classification techniques based on different approaches (model selection)

How to test a machine-learning classifier?

For example, if parameter selection, training of the predictive model and validation are performed using the same dataset, the generated classifier will show limited generalization ability when classifying unseen samples

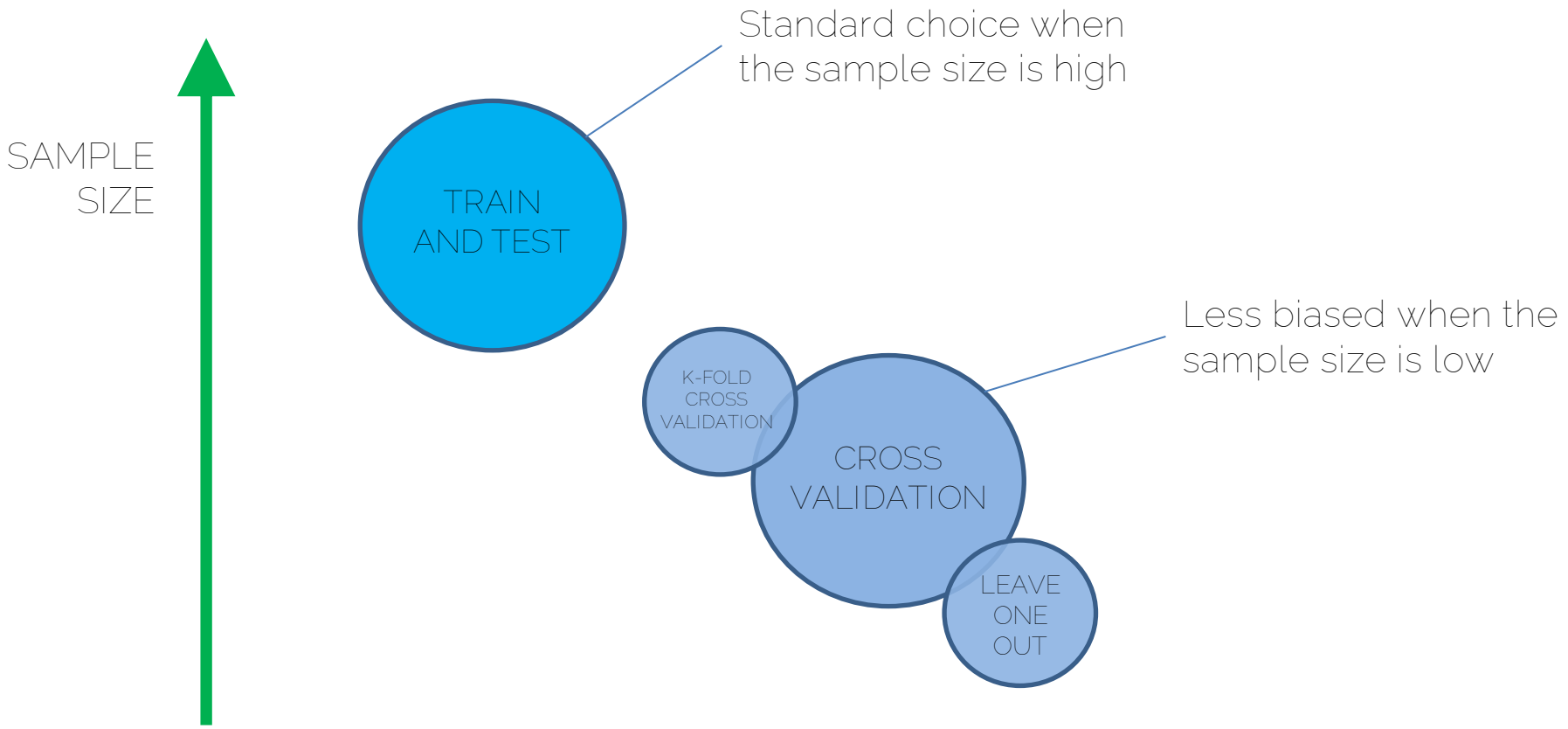
LOW
TRAINING ERROR



HIGH
TESTING ERROR
(low generalization ability)

OVERFITTING

Which validation approach?

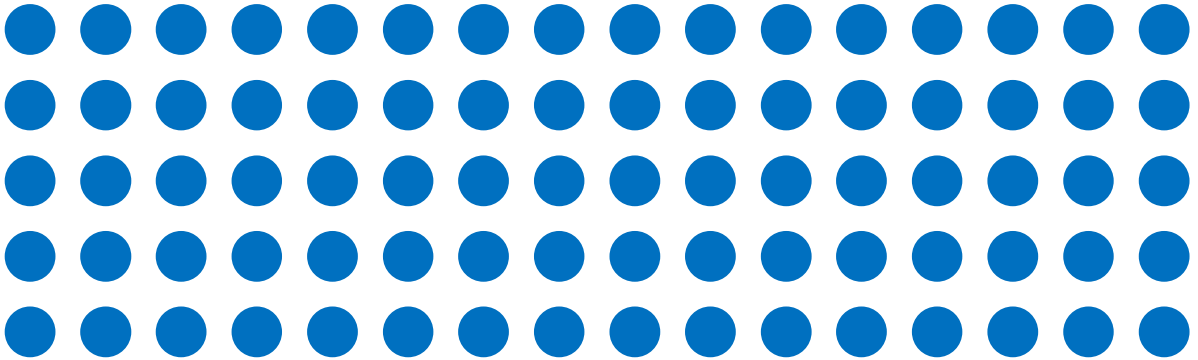


Train-and-test

This kind of procedure is used when the number of samples in the original dataset is high enough to allow its splitting into two subsets including different samples, which can be used to train and test the classifier.

Train-and-test

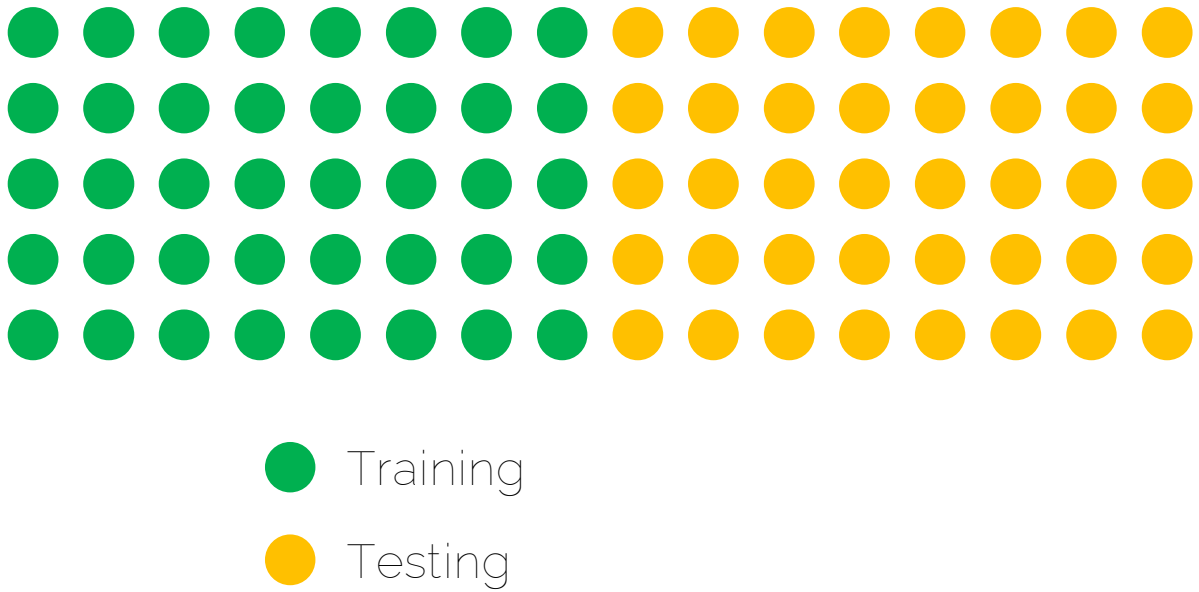
This kind of procedure is used when the number of samples in the original dataset is high enough to allow its splitting into two subsets including different samples, which can be used to train and test the classifier.



The original dataset is partitioned into 2 complementary subsets, the TRAINING set and the TESTING set.

The TRAINING set is used to train the classifier
The TESTING set is used for validation

Train-and-test



Train-and-test

Advantages:

- Over-training problems are reduced, because the training and testing sets are completely independent

Drawbacks:

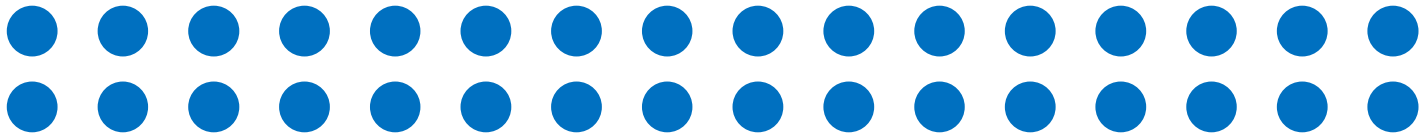
- Results could be related to the particular choice of the partition subsets

Leave-one-out cross validation

Leave-One-Out (LOO) CV can be considered a particular form of k-fold CV in which

k = number of samples in the original dataset

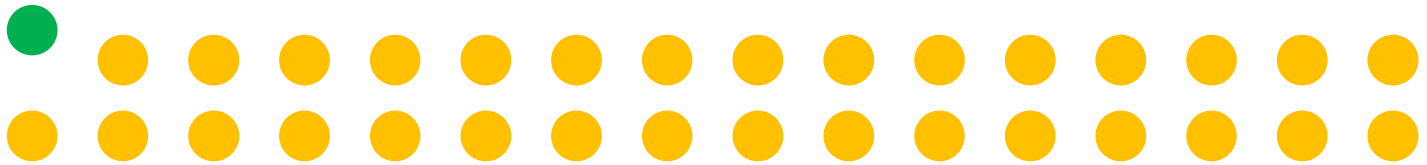
Leave-one-out cross validation



TRAINING of the classifier is performed using n-1 samples of the original dataset

TESTING is performed using the remaining sample (n being the total number of samples in the original dataset)

Leave-one-out cross validation

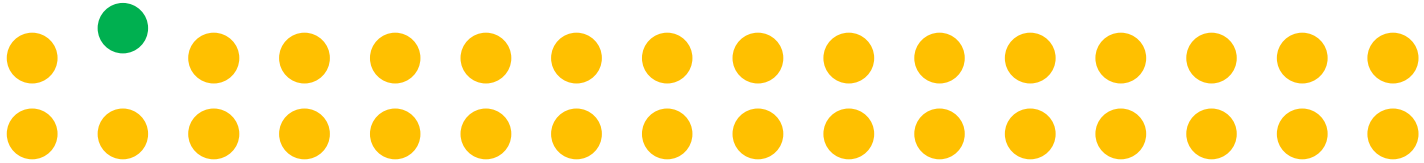


Training

Testing

The procedure is then repeated n times, until all samples are used once for validation.

Leave-one-out cross validation

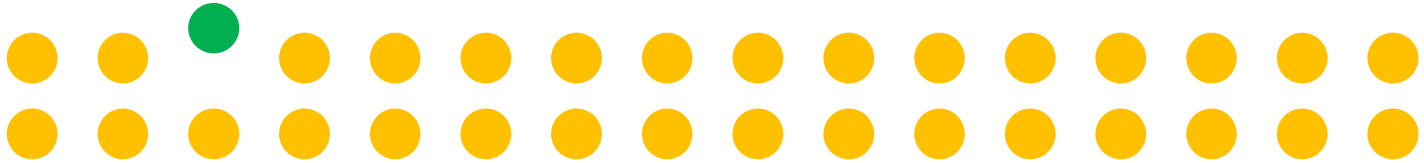


Training

Testing

The procedure is then repeated n times, until all samples are used once for validation.

Leave-one-out cross validation

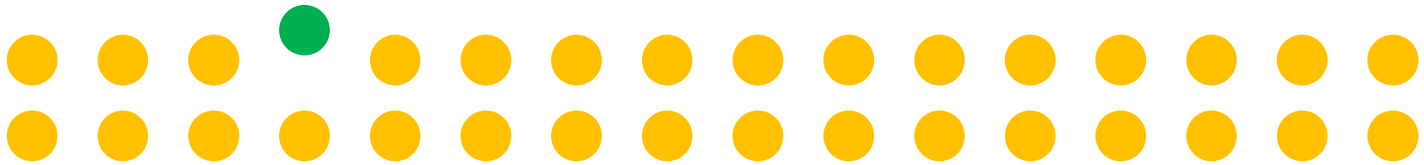


Training

Testing

The procedure is then repeated n times, until all samples are used once for validation.

Leave-one-out cross validation



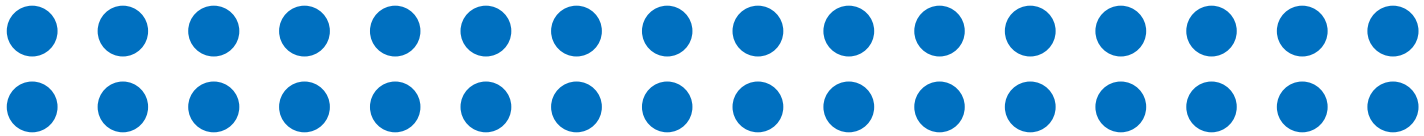
Training

Testing

and so on…

Cross validation

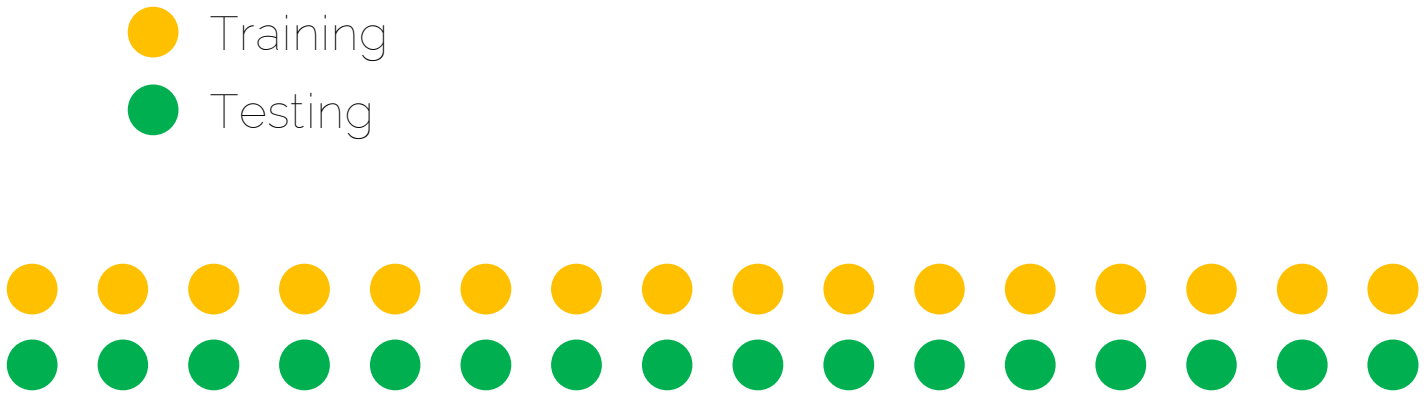
Quantification of the discriminatory power of a predictive model even if the size of the dataset is small



CV involves partitioning the original dataset into complementary subsets, the training set and the testing set

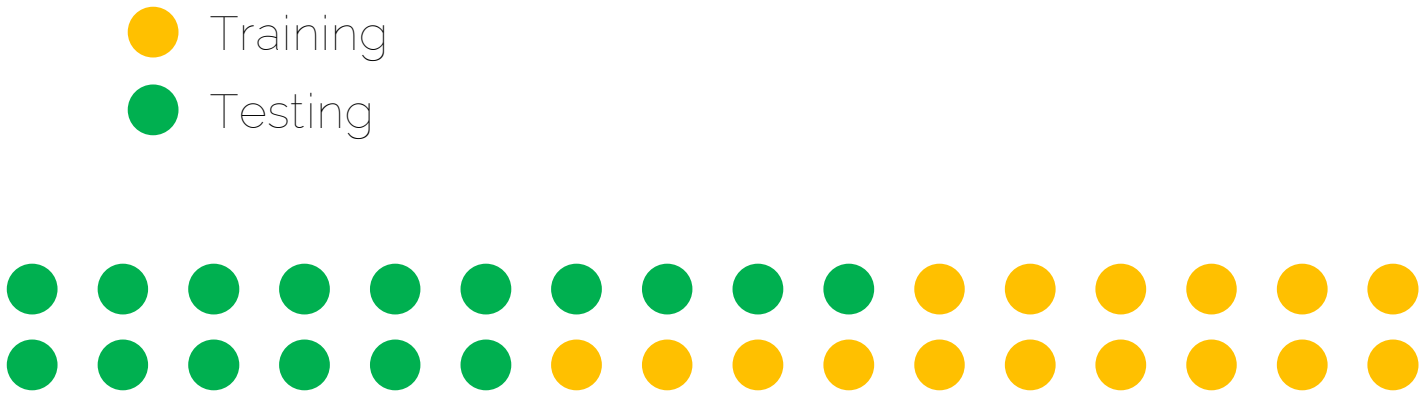
The TRAINING set is used to train the classifier
The TESTING set is used to validate the generated predictive model

Cross validation



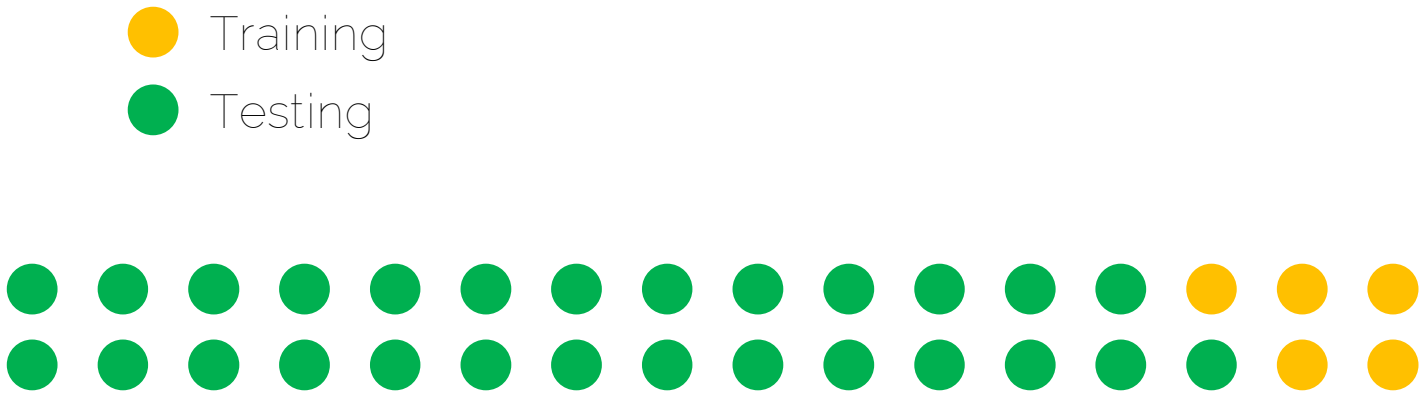
By using different partitions of the original dataset, multiple rounds of CV can be performed, which can aid reducing the variability of the partitioned subsets.

Cross validation



By using different partitions of the original dataset, multiple rounds of CV can be performed, which can aid reducing the variability of the partitioned subsets.

Cross validation

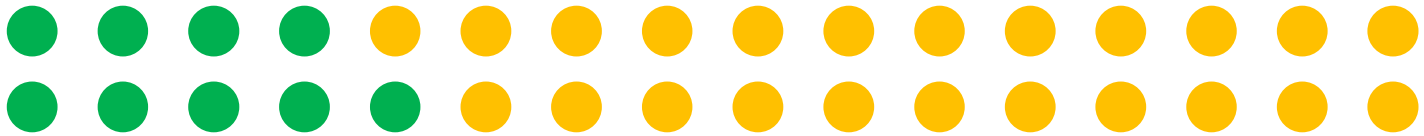


By using different partitions of the original dataset, multiple rounds of CV can be performed, which can aid reducing the variability of the partitioned subsets.

Cross validation

Training

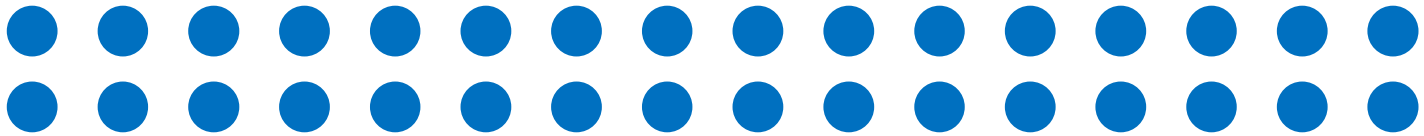
Testing



and so on…

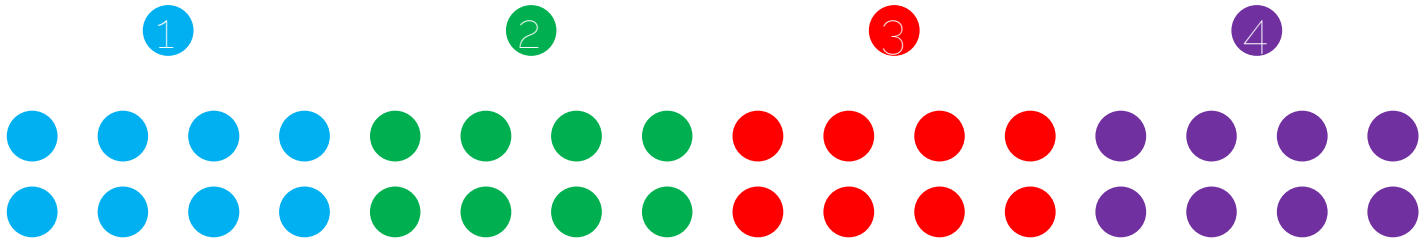
Results obtained from multiple rounds can be averaged in order to obtain a quantification of the performance of the classifier.

K-fold cross validation



The original dataset is randomly
partitioned into k subsets of equal size.

K-fold cross validation

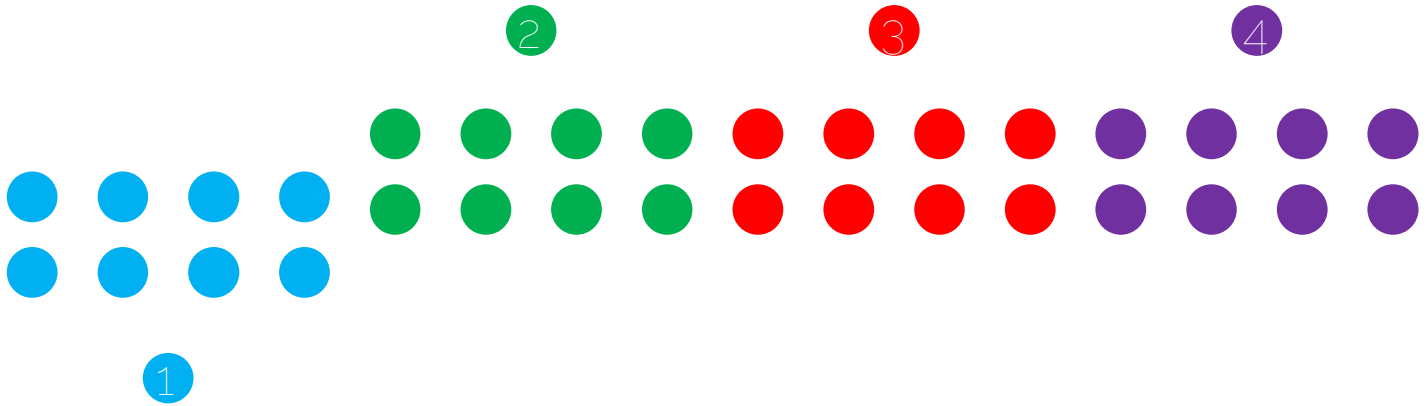


The original dataset is randomly partitioned into k subsets of equal size

TRAINING of the classifier is performed using k-1 subsets

TESTING is performed using the remaining subset

K-fold cross validation

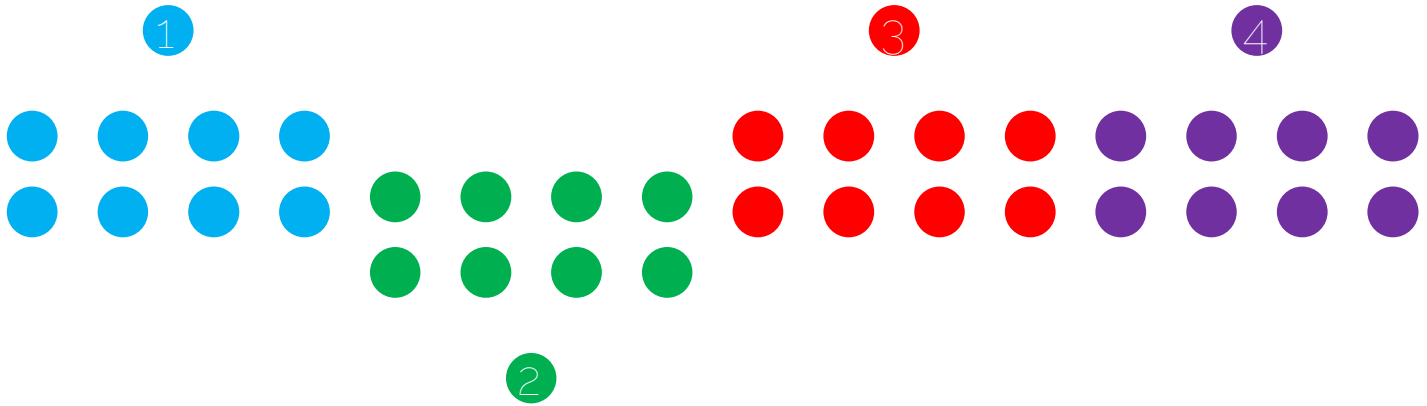


2 3 4 Training

1 Testing

The procedure is then repeated k times, until all subsets are used once as testing set.

K-fold cross validation

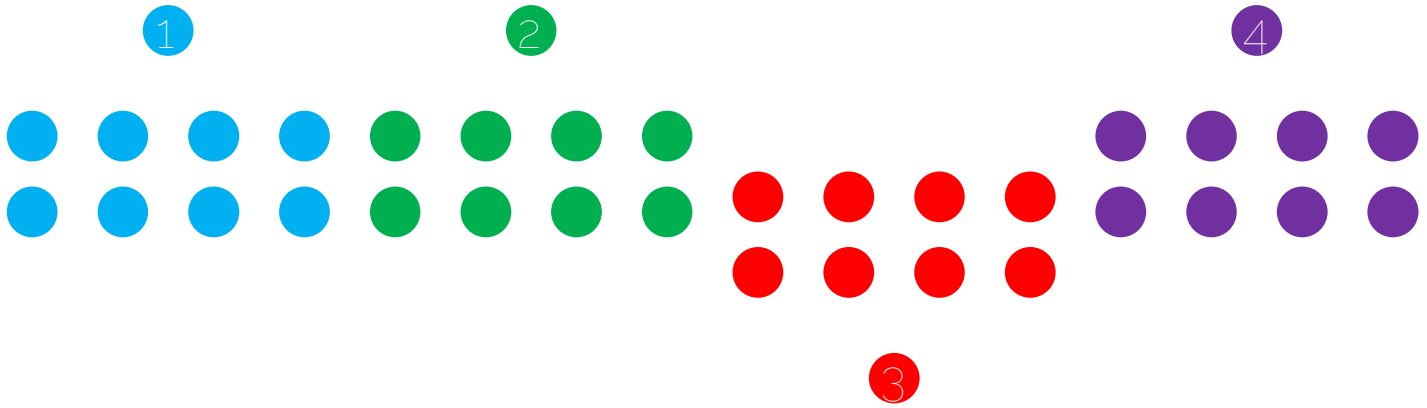


1 3 4 Training

2 Testing

The procedure is then repeated k times, until all subsets are used once as testing set.

K-fold cross validation

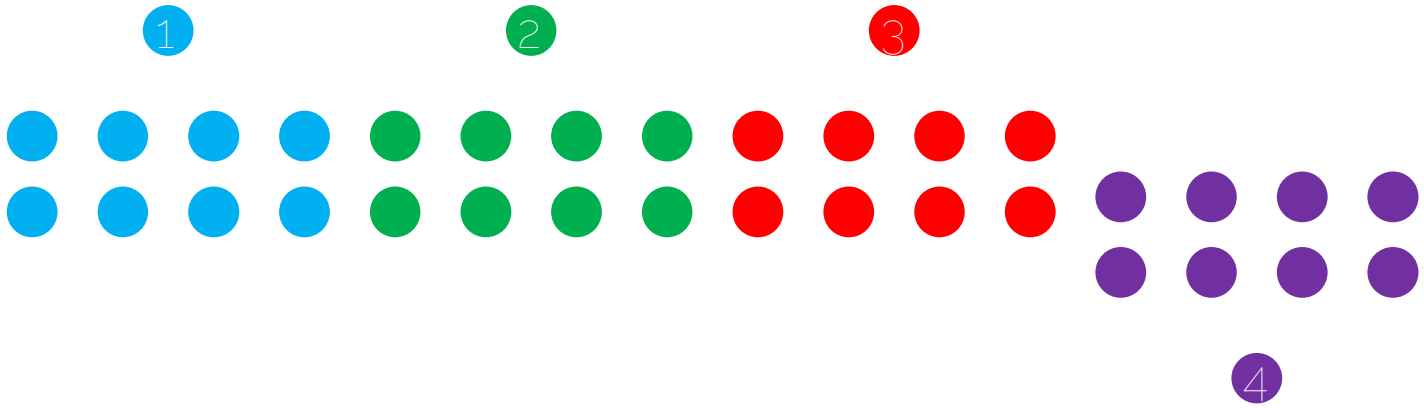


1 2 4 Training

3 Testing

The procedure is then repeated k times, until all subsets are used once as testing set.

K-fold cross validation



1 2 3 Training

4 Testing

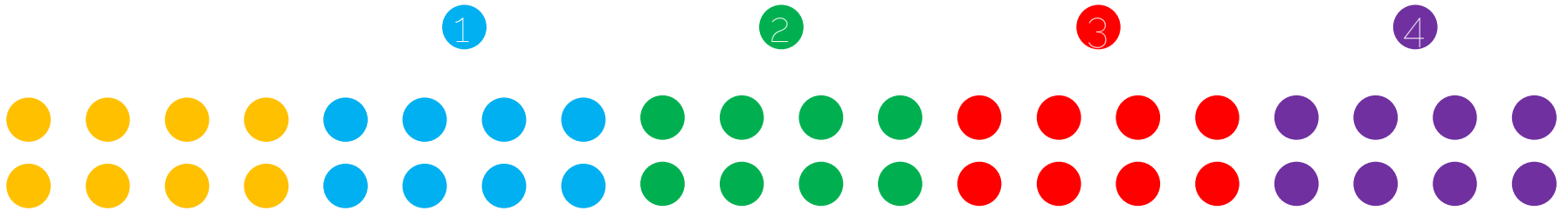
The procedure is then repeated k times, until all subsets are used once as testing set.

K-fold cross validation

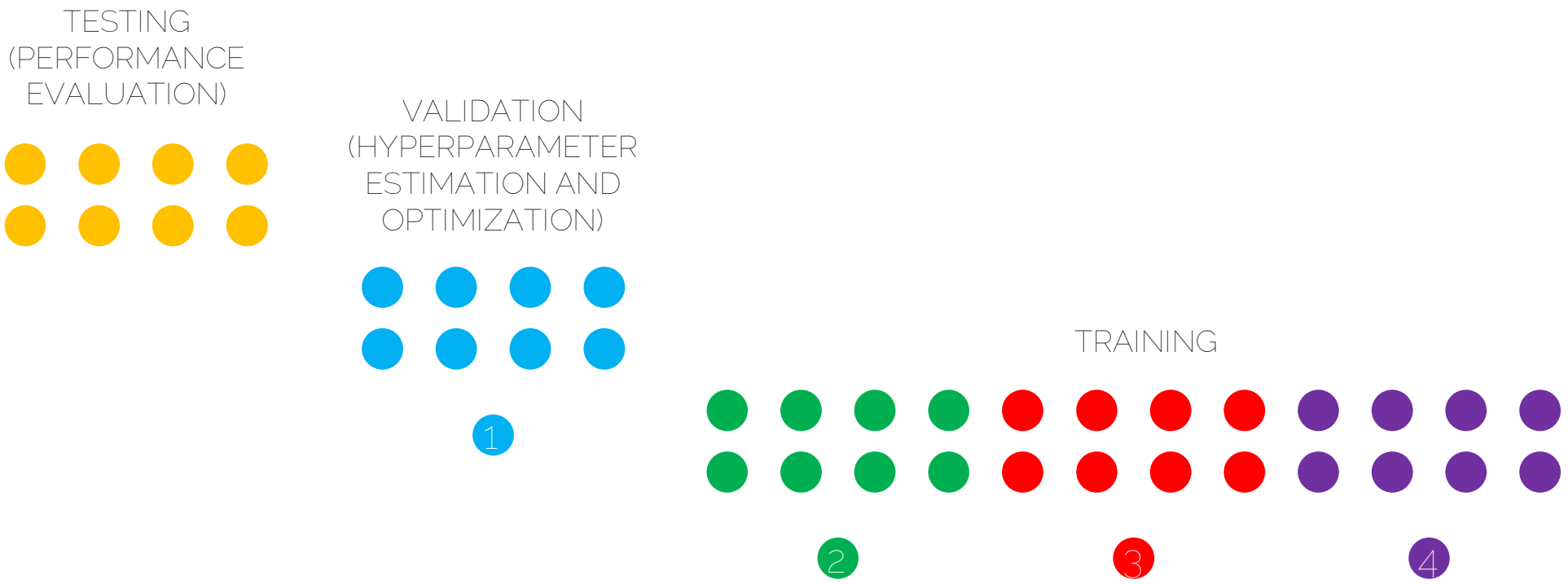
Advantages:

- each sample of the original dataset is used once for validation
- all samples being used for both training and testing phases

Nested K-fold cross validation



Nested K-fold cross validation



Nested K-fold cross validation

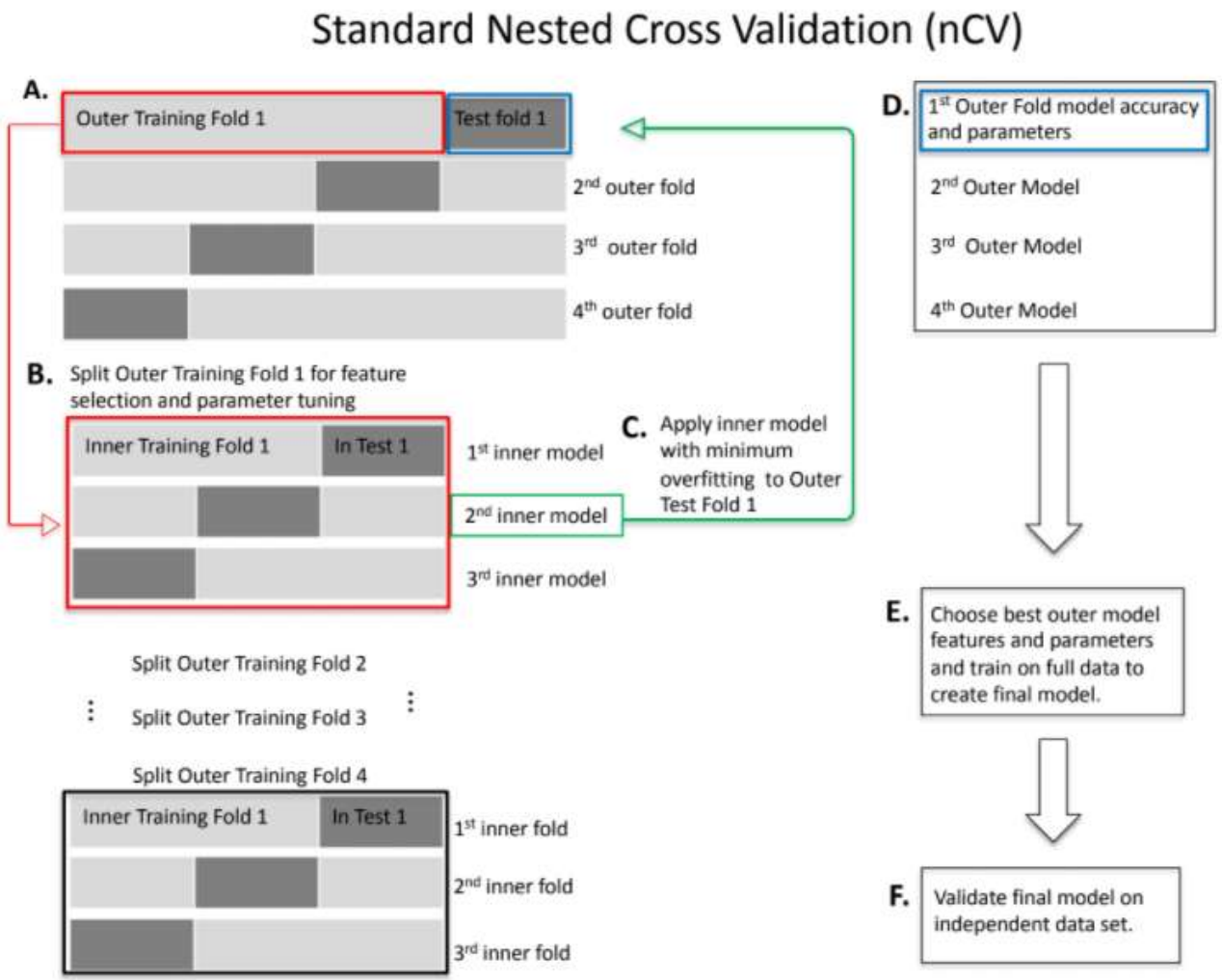


Fig. 1. Standard nested Cross-Validation (nCV). **A.** Split the data into outer folds of training and testing data pairs (4 outer folds in this illustration). Then do the following for each outer training fold (illustration starting with Outer Training Fold 1 (red box, A)). **B.** Split outer training fold into inner folds for feature selection and possible hyperparameter tuning by grid search. **C.** Use the best inner training model including features and parameters (2nd inner model, green box, for illustration) based on minimum overfitting (difference between training and test accuracies) in the inner folds to test on the outer test fold (green arrow to blue box, Test Fold 1). **D.** Save the best model for this outer fold including the features and test accuracies. Repeat B-D for the remaining outer folds. **E.** Choose the best outer model with its features based on minimum overfitting. Train on the full data to create the final model. **F.** Validate the final model on independent data.

PARVANDEH ET AL. 2020

Nested K-fold cross validation

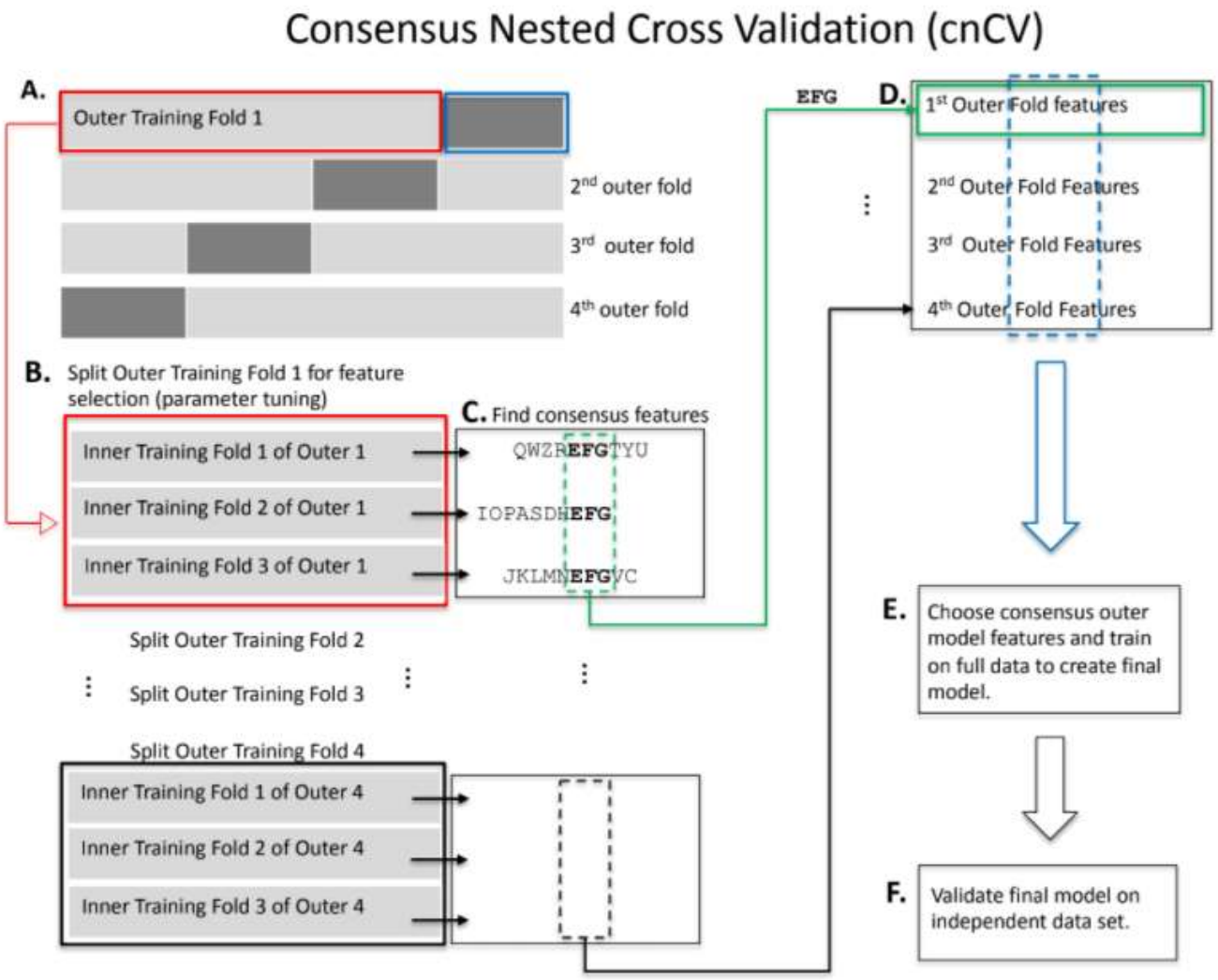


Fig. 2. Consensus Nested Cross-Validation (cnCV). **A.** Split the data into outer folds (4 outer folds in this illustration). Then do the following for each outer training fold (illustration starting with Outer Training Fold 1 (red box, A)). **B.** Split outer training fold into inner folds for feature selection and optional hyperparameter tuning by grid search. **C.** Find consensus features. For each fold, features with positive Relief scores are collected (e.g., "QWZREFGTYU" for fold 1). Negative Relief scores have high probability of being irrelevant to classification. The implementation allows for different feature importance methods and tuning the number of input features. Consensus features (found in all folds) are used as the best features in the corresponding outer fold. For example, features "EFG" are shared across the three inner folds. This procedure is used in the inner and outer folds of cnCV. Classification is not needed to select consensus features. **D.** The best outer fold features (green arrow to green box) are found for each fold (i.e., Repeat B-D for all outer folds). **E.** Choose the consensus features across all the outer folds to train the final model on full data. Consensus features are selected based on training data only. Classification is not performed until the outer consensus features are selected (A-D). **F.** Validate the final model on independent data.

PARVANDEH ET AL. 2020

Performance metrics

A brief overview. . .

1. Accuracy
2. Sensitivity and Specificity
3. Precision and Recall
4. Positive and Negative Predictive Value
5. False/True Positive/Negative Rates
6. Imbalanced datasets
7. ROC-AUC

Accuracy

Accuracy is the most used metric in classification problems

Accuracy of classification =

correctly classified samples (for both classes)

/

classified samples

If the error rate is defined as the number of misclassified samples (both classes) divided by the total number of classified samples, it is evident that accuracy and error rate are complementary measures.

Sensitivity and Specificity

Two metrics of great importance in medicine are sensitivity and specificity, as they measure the rate of correctly classified samples in the positive (pathological) and negative (normal) class, respectively.

Sensitivity (also known as True Positive Rate) is given by the number of correctly classified samples belonging to the positive class (true positives) divided by the total number of samples belonging to the positive class (true positives plus false negatives).

$$\text{Sensitivity} = \frac{\#\text{ correctly classified samples in the positive class}}{\#\text{ positive samples}}$$

Sensitivity and Specificity

Specificity (also known as True Negative Rate) is given by the number of correctly classified samples belonging to the negative class (true negatives) divided by the total number of samples belonging to the negative class (true negatives plus false positives).

$$\text{Specificity} = \frac{\#\text{ correctly classified samples in the negative class}}{\#\text{ negative samples}}$$

Here, true positive (negative) gives the number of correctly classified samples belonging to the positive (negative) class, while false positive (negative) gives the number of misclassified samples belonging to the negative (positive) class.

Precision and Recall

Precision is the number of correctly classified samples belonging to the positive class (true positives) divided by the total number of samples predicted as positive by the classifier (true positives plus false positives).

$$\text{Precision} = \frac{\#\text{ correctly classified samples in the positive class}}{\#\text{ predicted-as-positive samples}}$$

Precision and Recall

Recall is the number of correctly classified samples belonging to the positive class (true positives) divided by the total number of samples in the positive class (true positives plus false negatives = positive samples).

Recall =

correctly classified samples in the positive class / # actually-positive samples

Positive and Negative Predictive Value

Positive Predictive Value is the number of correctly classified samples belonging to the positive class (true positives) divided by the total number of samples predicted as positive by the classifier (true positives plus false positives).

PPV =

correctly classified samples in the positive class / # predicted-as-positive samples

Positive and Negative Predictive Value

Negative Predictive Value is the number of correctly classified samples belonging to the negative class (true negatives) divided by the total number of samples predicted as negative by the classifier (true negatives plus false negatives).

NPV =

$$\frac{\text{\# correctly classified samples in the negative class}}{\text{\# predicted-as-negative samples}}$$

False/True Positive/Negative Rates

TPR =

correctly classified samples in the positive class / # positive samples

FPR =

incorrectly classified samples in the positive class / # negative samples

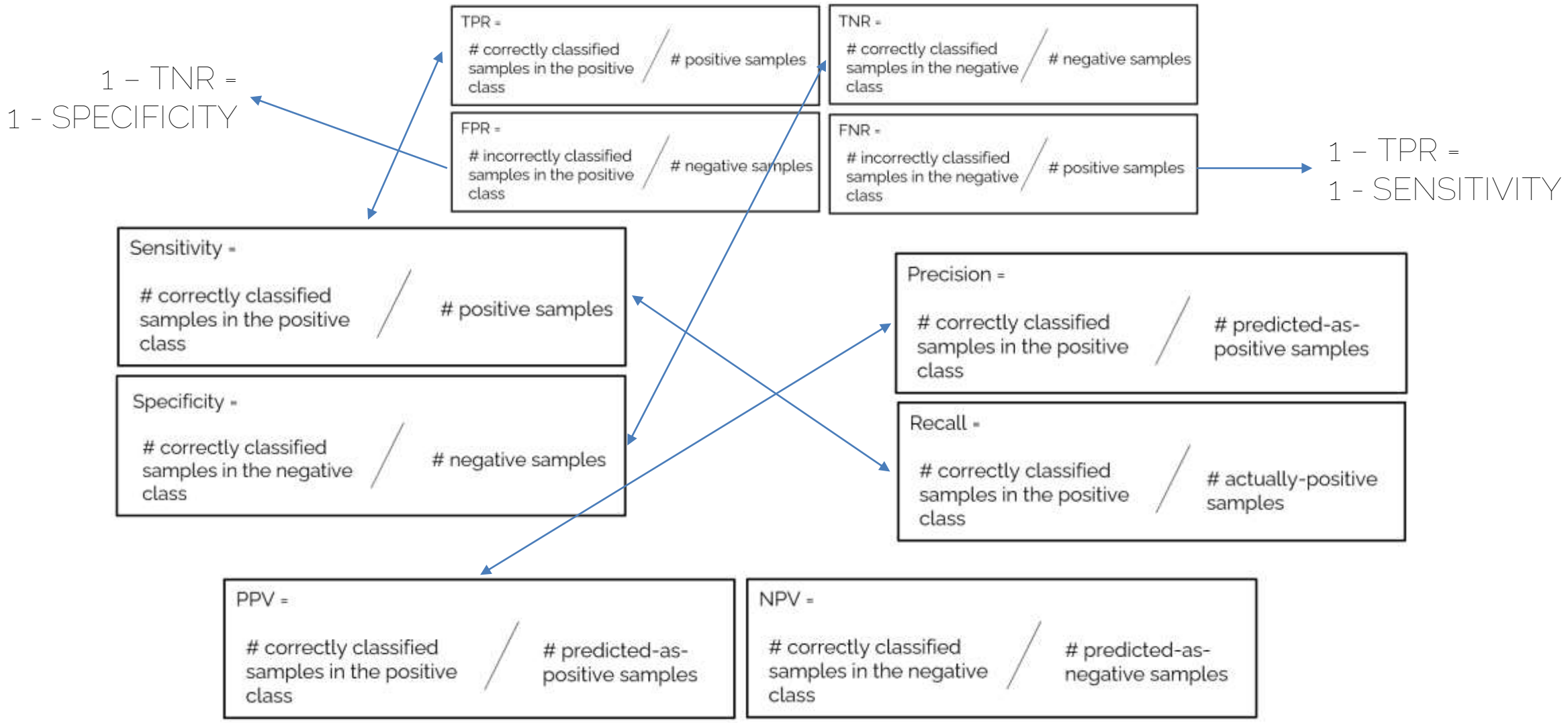
TNR =

correctly classified samples in the negative class / # negative samples

FNR =

incorrectly classified samples in the negative class / # positive samples

Precision and Recall



Confusion matrix

		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

Positive	95/100
Negative	75/100

	P PREDICTED	N PREDICTED
P ACTUAL	95	5
N ACTUAL	25	75

Confusion matrix

1. Accuracy
1. Sensitivity and Specificity
1. Precision and Recall
1. Positive and Negative Predictive Value
1. False/True Positive Rates
1. False/True Negative Rates

(95+75) / (95+25+5+75) = 85%

95 / (95+5) = 95%

75/ (25+75) = 75%

95 / (95+25) = 79%

95 / (95+5) = 95%

95 / (95+25) = 79%

75 / (75+5) = 94%

5 / (25+75) = 5%

95 / (95+5) = 95%

25 / (95+5) = 25%

75 / (25+75) = 75%

Positive	95/100
Negative	75/100

	P PREDICTED	N PREDICTED
P ACTUAL	95	5
N ACTUAL	25	75

Confusion matrix

P 1	95/100
P 2	55/100
P 3	90/100
N	75/100

CLASS-X ACCURACY

	P 1 PREDICTED	P 2 PREDICTED	P 3 PREDICTED	N PREDICTED
P 1 ACTUAL	95	0	1	4
P 2 ACTUAL	5	55	5	35
P 3 ACTUAL	2	1	90	7
N ACTUAL	5	15	5	75

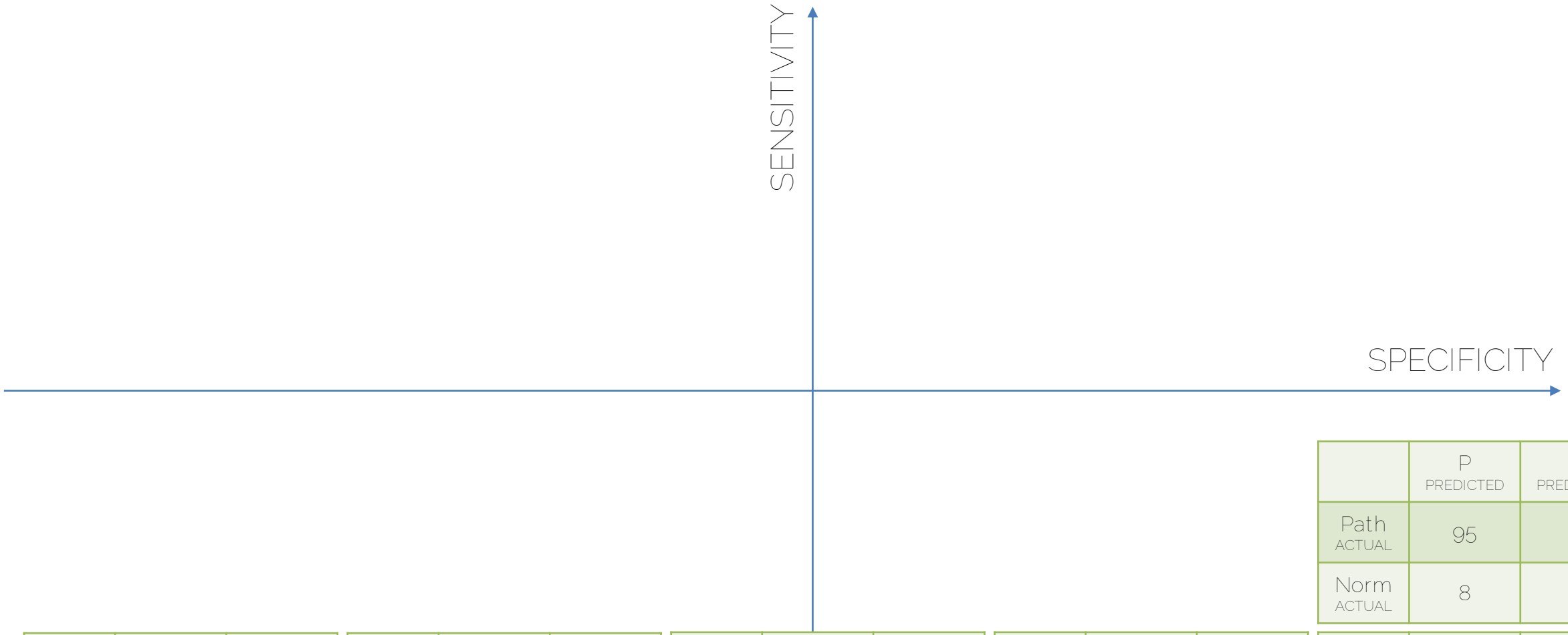
95%

55%

90%

75%

Confusion matrix



	P PREDICTED	N PREDICTED
P ACTUAL	95	5
N ACTUAL	25	75

	P PREDICTED	N PREDICTED
P ACTUAL	25	75
N ACTUAL	65	35

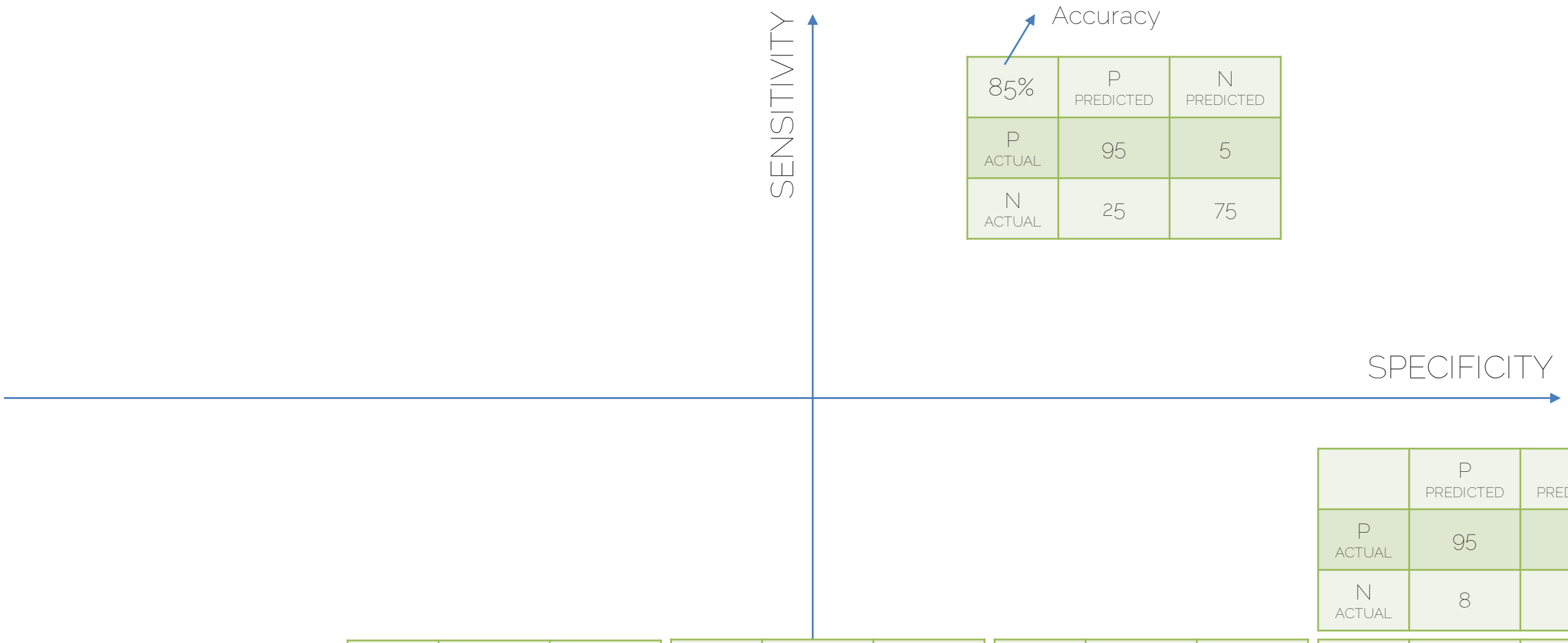
	P PREDICTED	N PREDICTED
P ACTUAL	95	5
N ACTUAL	75	25

	P PREDICTED	N PREDICTED
P ACTUAL	25	75
N ACTUAL	15	85

	P PREDICTED	N PREDICTED
P ACTUAL	5	10
N ACTUAL	15	85

	P PREDICTED	N PREDICTED
Path ACTUAL	95	5
Norm ACTUAL	8	2

Confusion matrix



Accuracy

85%	P PREDICTED	N PREDICTED
P ACTUAL	95	5
N ACTUAL	25	75

	P PREDICTED	N PREDICTED
P ACTUAL	95	5
N ACTUAL	8	2

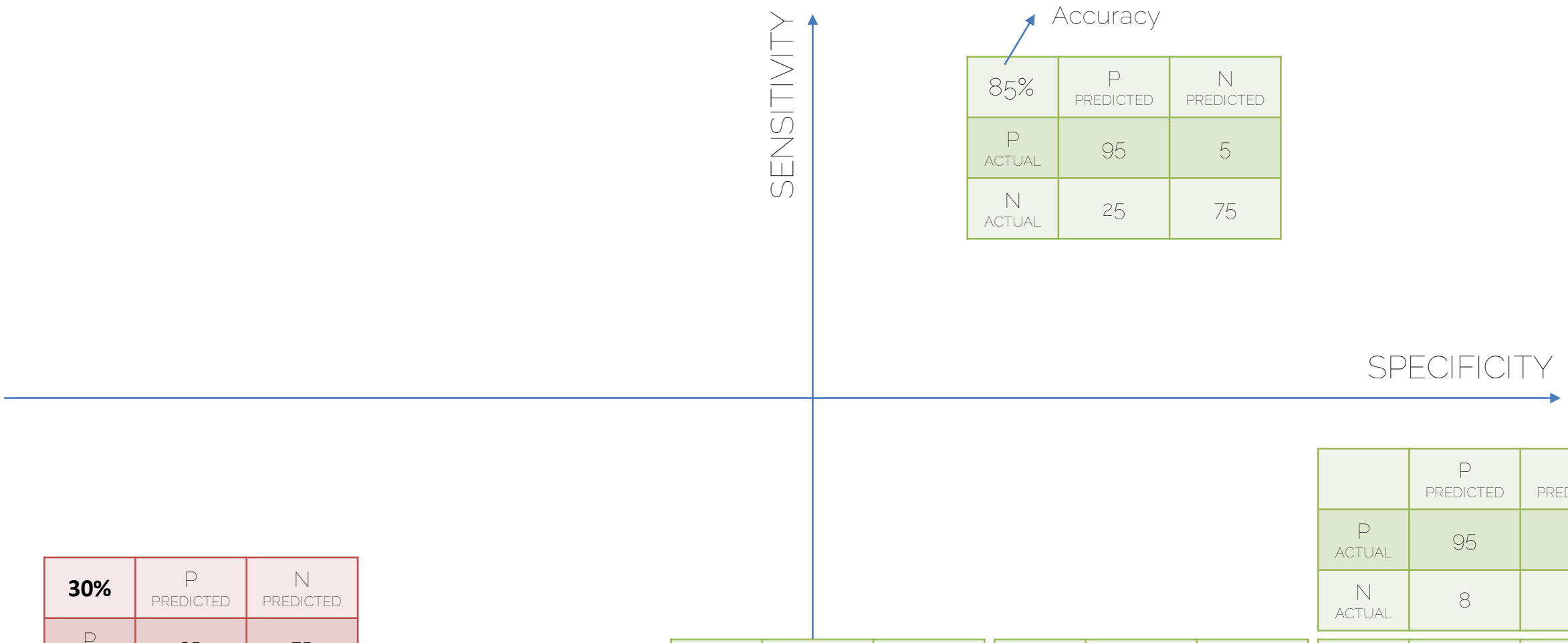
	P PREDICTED	N PREDICTED
P ACTUAL	25	75
N ACTUAL	65	35

	P PREDICTED	N PREDICTED
P ACTUAL	95	5
N ACTUAL	75	25

	P PREDICTED	N PREDICTED
P ACTUAL	25	75
N ACTUAL	15	85

	P PREDICTED	N PREDICTED
P ACTUAL	5	10
N ACTUAL	15	85

Confusion matrix



85%	P PREDICTED	N PREDICTED
P ACTUAL	95	5
N ACTUAL	25	75

30%	P PREDICTED	N PREDICTED
P ACTUAL	**25**	**75**
N ACTUAL	**65**	**35**

	P PREDICTED	N PREDICTED
P ACTUAL	95	5
N ACTUAL	75	25

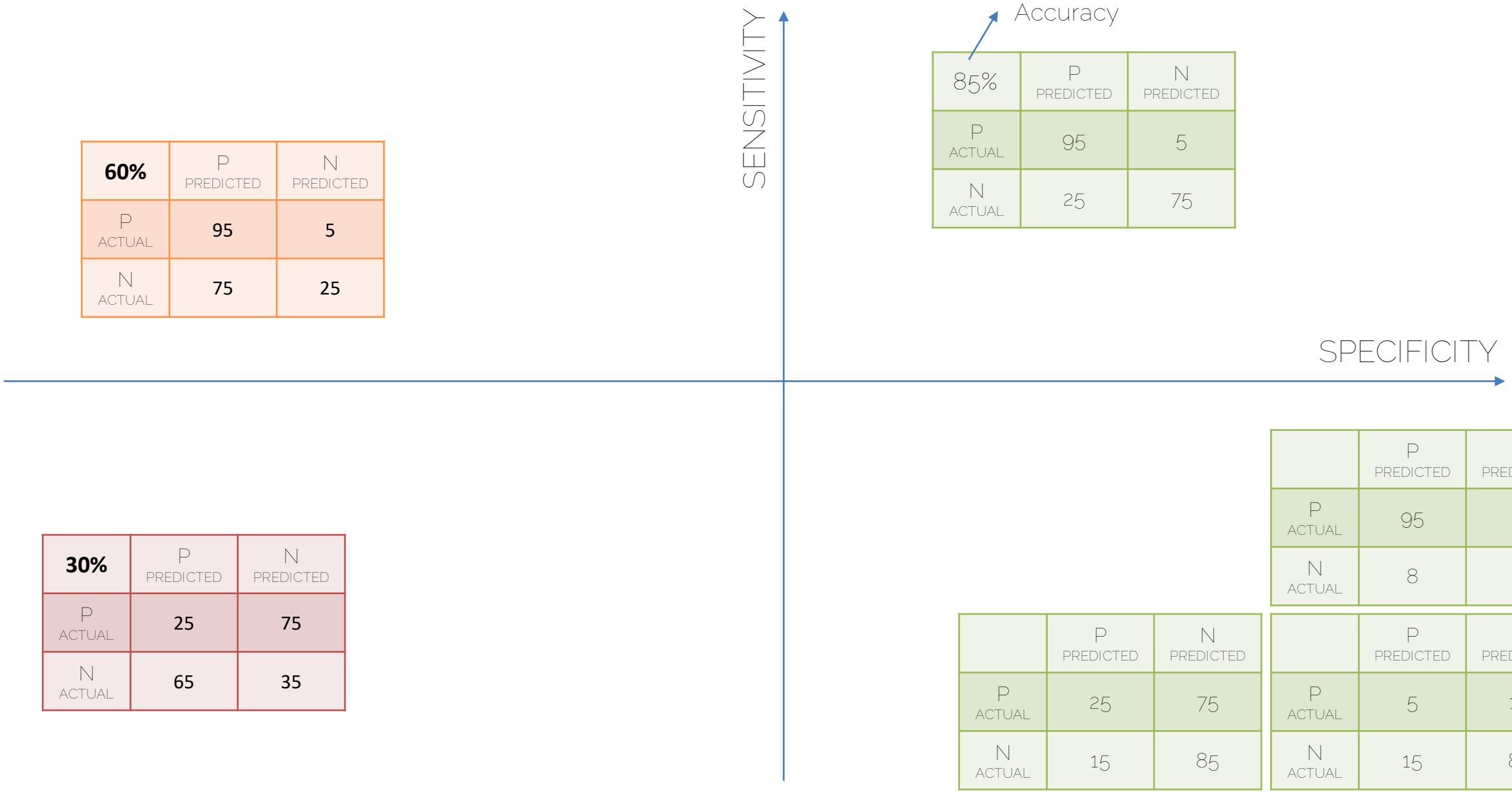
	P PREDICTED	N PREDICTED
P ACTUAL	25	75
N ACTUAL	15	85

	P PREDICTED	N PREDICTED
P ACTUAL	95	5
N ACTUAL	8	2

	P PREDICTED	N PREDICTED
P ACTUAL	5	10
N ACTUAL	15	85

Confusion matrix

60%	P PREDICTED	N PREDICTED
P ACTUAL	95	5
N ACTUAL	75	25



Accuracy

85%	P PREDICTED	N PREDICTED
P ACTUAL	95	5
N ACTUAL	25	75

30%	P PREDICTED	N PREDICTED
P ACTUAL	25	75
N ACTUAL	65	35

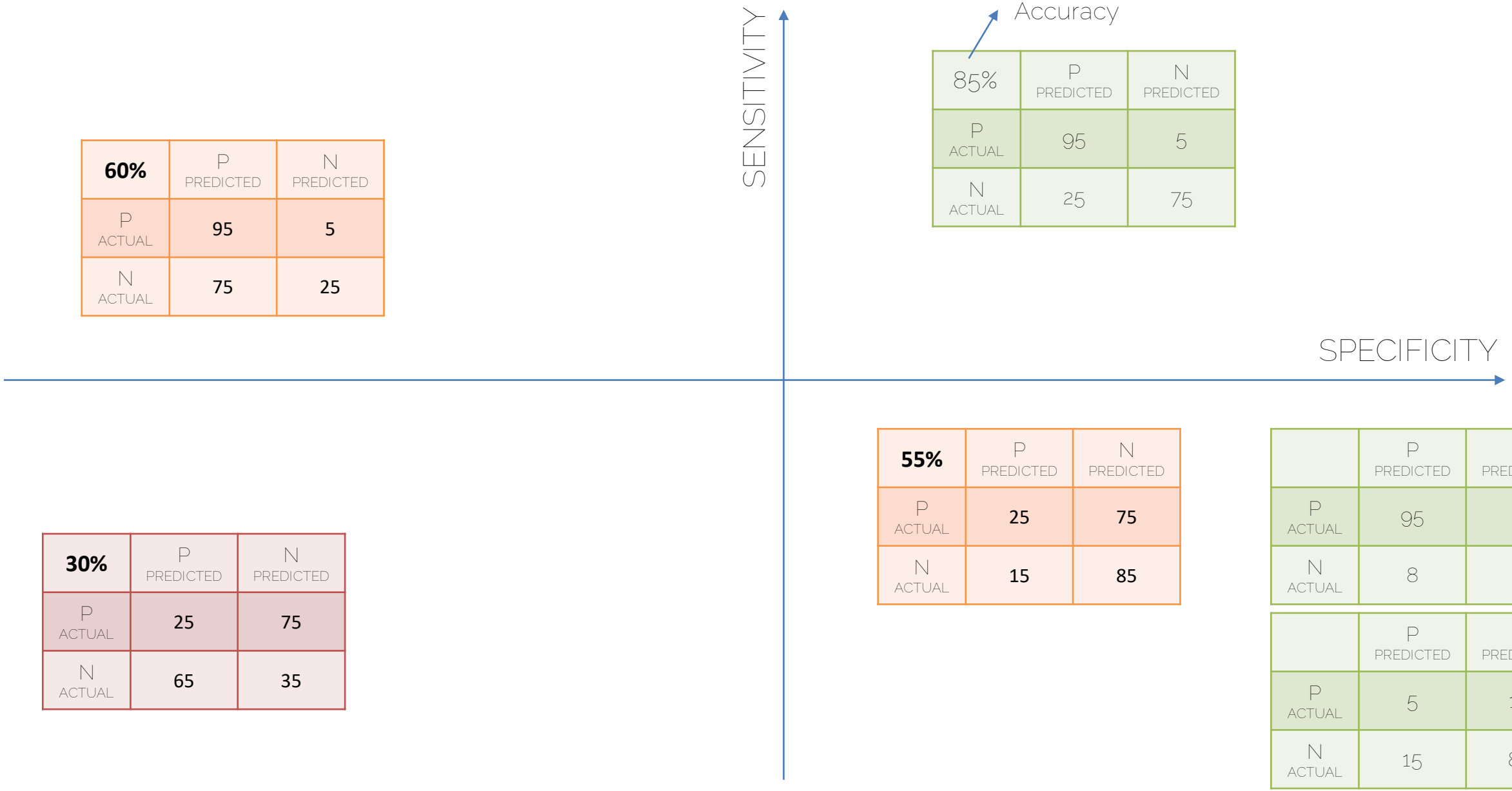
	P PREDICTED	N PREDICTED
P ACTUAL	95	5
N ACTUAL	8	2

	P PREDICTED	N PREDICTED
P ACTUAL	25	75
N ACTUAL	15	85

	P PREDICTED	N PREDICTED
P ACTUAL	5	10
N ACTUAL	15	85

Confusion matrix

60%	P PREDICTED	N PREDICTED
P ACTUAL	95	5
N ACTUAL	75	25



Accuracy

85%	P PREDICTED	N PREDICTED
P ACTUAL	95	5
N ACTUAL	25	75

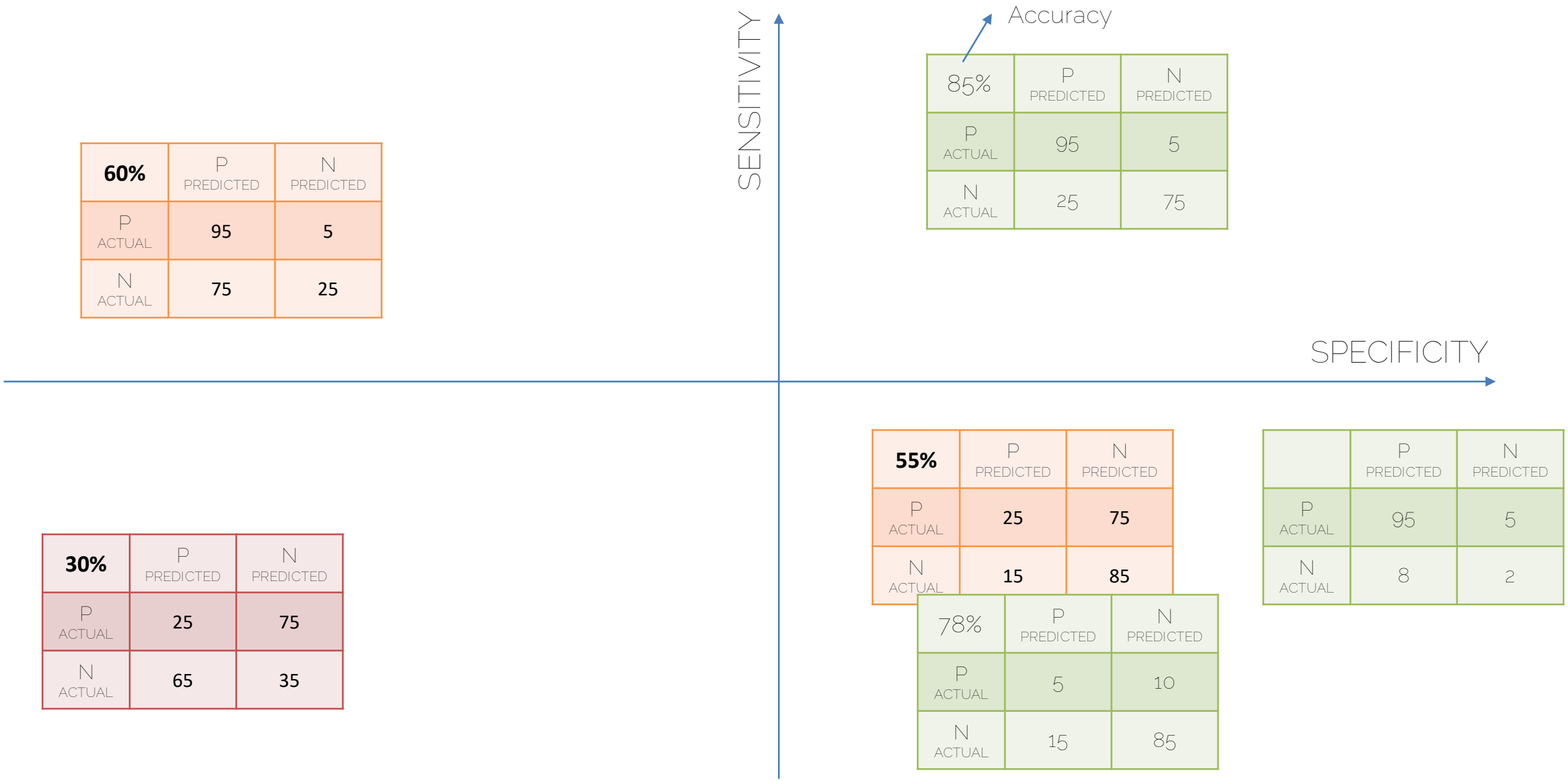
30%	P PREDICTED	N PREDICTED
P ACTUAL	25	75
N ACTUAL	65	35

55%	P PREDICTED	N PREDICTED
P ACTUAL	25	75
N ACTUAL	15	85

	P PREDICTED	N PREDICTED
P ACTUAL	95	5
N ACTUAL	8	2

	P PREDICTED	N PREDICTED
P ACTUAL	5	10
N ACTUAL	15	85

Confusion matrix



60%	P PREDICTED	N PREDICTED
P ACTUAL	95	5
N ACTUAL	75	25

85%	P PREDICTED	N PREDICTED
P ACTUAL	95	5
N ACTUAL	25	75

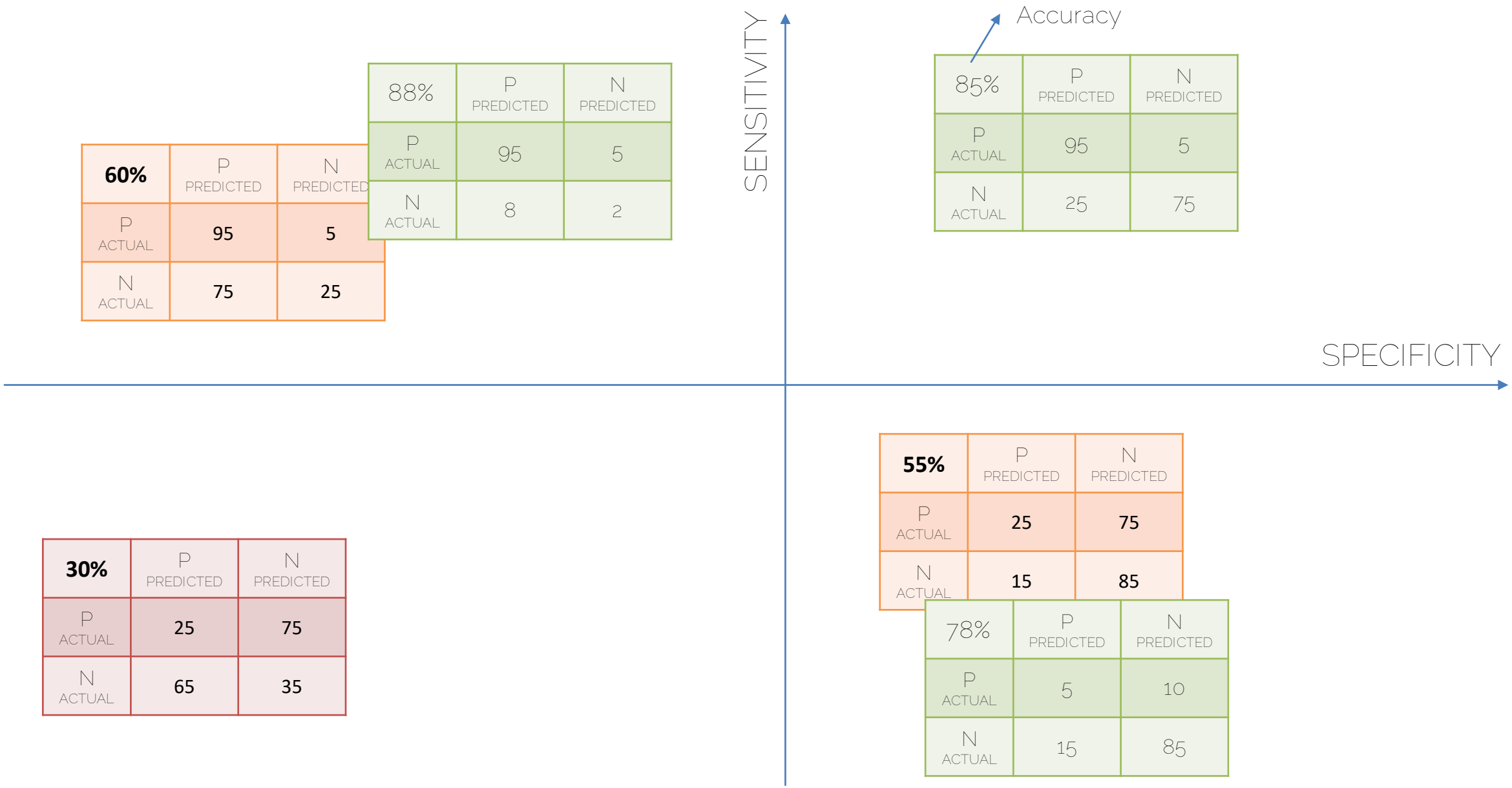
30%	P PREDICTED	N PREDICTED
P ACTUAL	25	75
N ACTUAL	65	35

55%	P PREDICTED	N PREDICTED
P ACTUAL	25	75
N ACTUAL	15	85

78%	P PREDICTED	N PREDICTED
P ACTUAL	5	10
N ACTUAL	15	85

	P PREDICTED	N PREDICTED
P ACTUAL	95	5
N ACTUAL	8	2

Confusion matrix



60%	P PREDICTED	N PREDICTED
P ACTUAL	95	5
N ACTUAL	75	25

88%	P PREDICTED	N PREDICTED
P ACTUAL	95	5
N ACTUAL	8	2

85%	P PREDICTED	N PREDICTED
P ACTUAL	95	5
N ACTUAL	25	75

30%	P PREDICTED	N PREDICTED
P ACTUAL	25	75
N ACTUAL	65	35

55%	P PREDICTED	N PREDICTED
P ACTUAL	25	75
N ACTUAL	15	85

78%	P PREDICTED	N PREDICTED
P ACTUAL	5	10
N ACTUAL	15	85

Confusion matrix

88%	P PREDICTED	N PREDICTED	
P ACTUAL	95	5	95%
N ACTUAL	8	2	20%

60%	P PREDICTED	N PREDICTED	
P ACTUAL	**95**	**5**	95%
N ACTUAL	**75**	**25**	25%

Accuracy

85%	P PREDICTED	N PREDICTED	
P ACTUAL	95	5	95%
N ACTUAL	25	75	75%

SENSITIVITY

SPECIFICITY

30%	P PREDICTED	N PREDICTED	
P ACTUAL	**25**	**75**	25%
N ACTUAL	**65**	**35**	35%

55%	P PREDICTED	N PREDICTED	
P ACTUAL	**25**	**75**	25%
N ACTUAL	**15**	**85**	85%

78%	P PREDICTED	N PREDICTED	
P ACTUAL	5	10	33%
N ACTUAL	15	85	85%

Performance metrics for imbalanced datasets

Prevalence is "the proportion of a particular population with a given condition".
In our case, it is the ratio between the number of actually-positive samples and the entire-sample size

$$\text{Prevalence} = \frac{\text{\# samples in the positive class (actual)}}{\text{\# samples (actual positives + negatives)}}$$

Performance metrics for imbalanced datasets

Geometric Mean of the true rates is the geometric mean between sensitivity and specificity

$$GM = (Sensitivity \times Specificity)^{(1/2)}$$

Which values does it span?

Performance metrics for imbalanced datasets

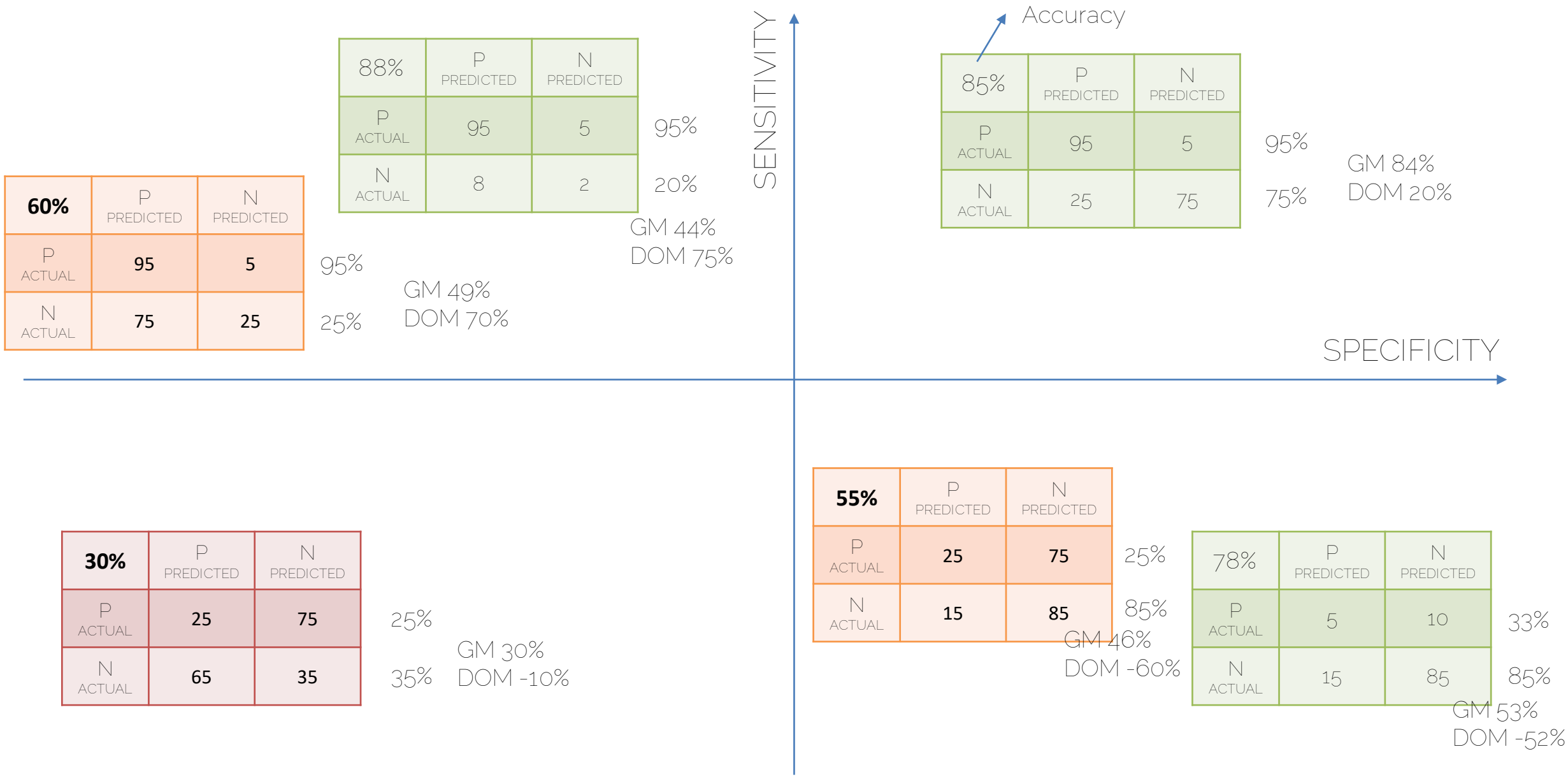
Dominance is the difference between sensitivity and specificity

Dominance =

Sensitivity - Specificity

Which values does it span?

Confusion matrix



F1 Score

F1 score is an armonic mean of precision (PPV) and recall (= sensitivity)

$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2\frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} = \frac{2\mathrm{tp}}{2\mathrm{tp} + \mathrm{fp} + \mathrm{fn}}$$

Why useful?

As above, it is convenient to have one single metric to compare classification performance (not both sensitivity/specificity or precision/recall)

Standard average is not always a good measure $\frac{P + R}{2}$

F1 score behaves better (what happens when precision = 0? when recall = 0? when is F1 score = 1?)

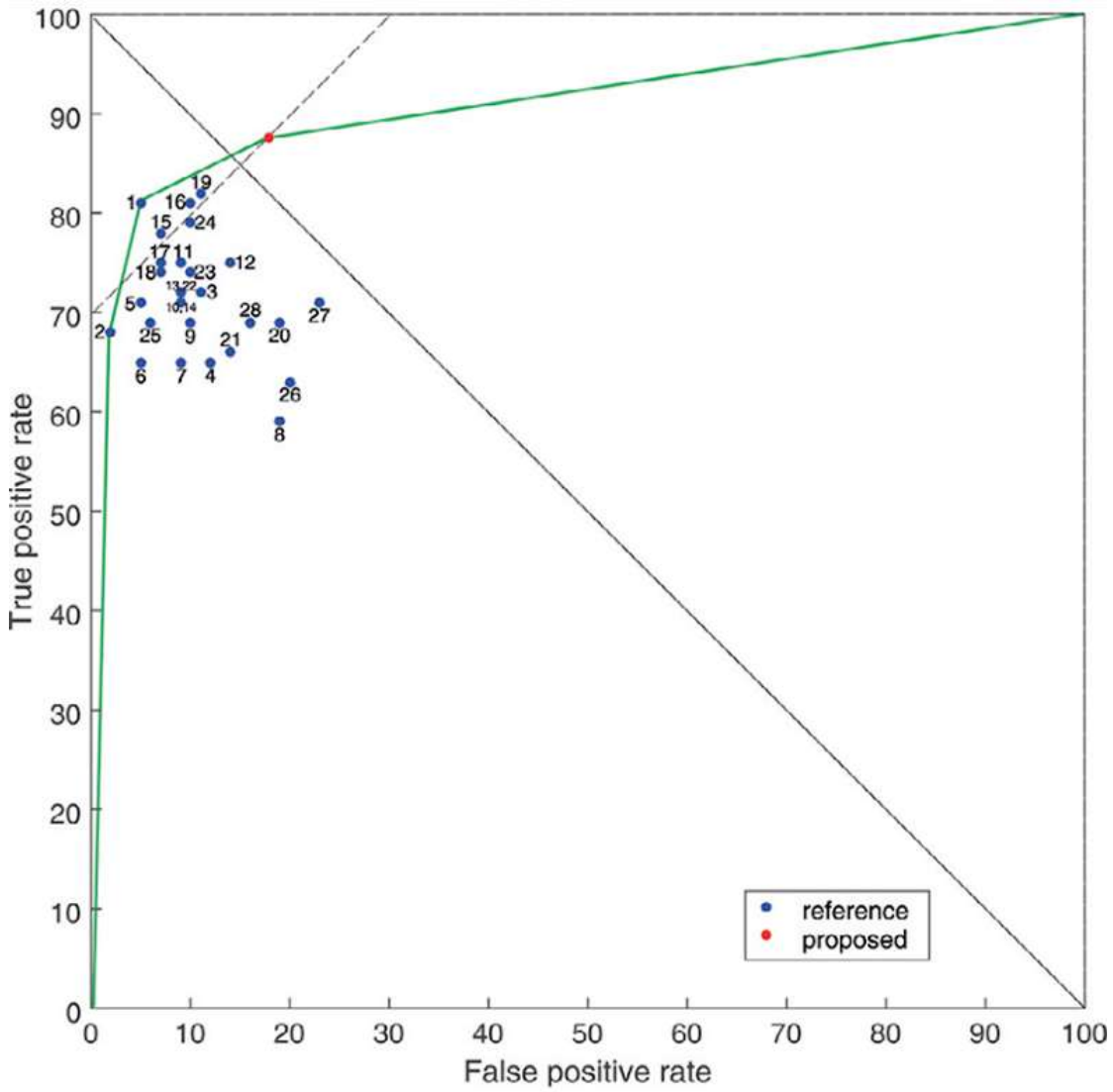
ROC analysis and Area Under the (ROC) Curve

Another important metric in classification problems is given by the study of the Receiver Operating Characteristic (ROC) curve.

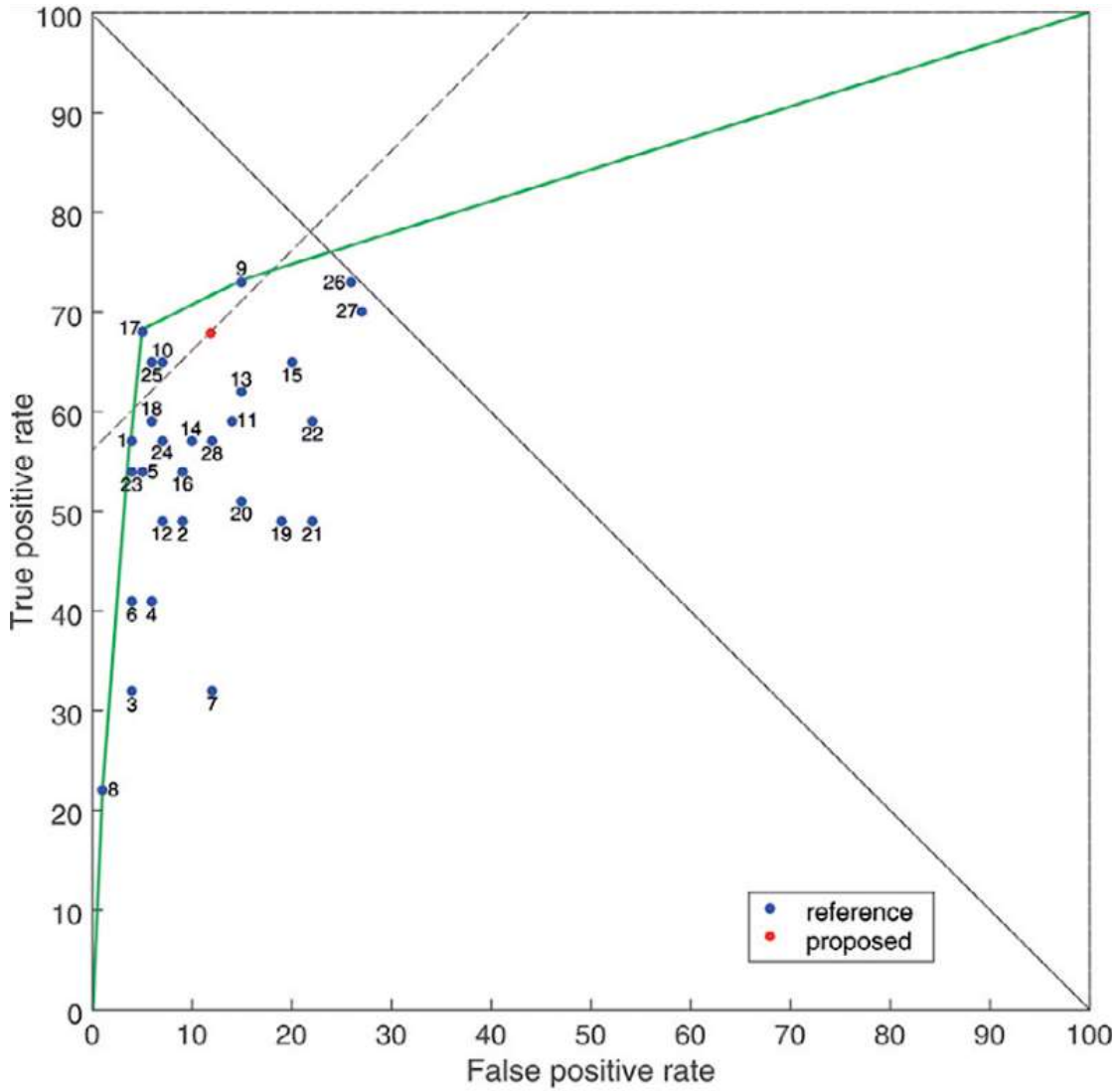
For a binary classifier, A ROC curve is a plot of the TPR (sensitivity) against the FPR (1 – specificity), which can be obtained at different setting thresholds.

The Area Under the ROC Curve (AUC) gives a quantification of the classifier performance, with a higher statistical consistency than accuracy.

ROC analysis and Area Under the (ROC) Curve

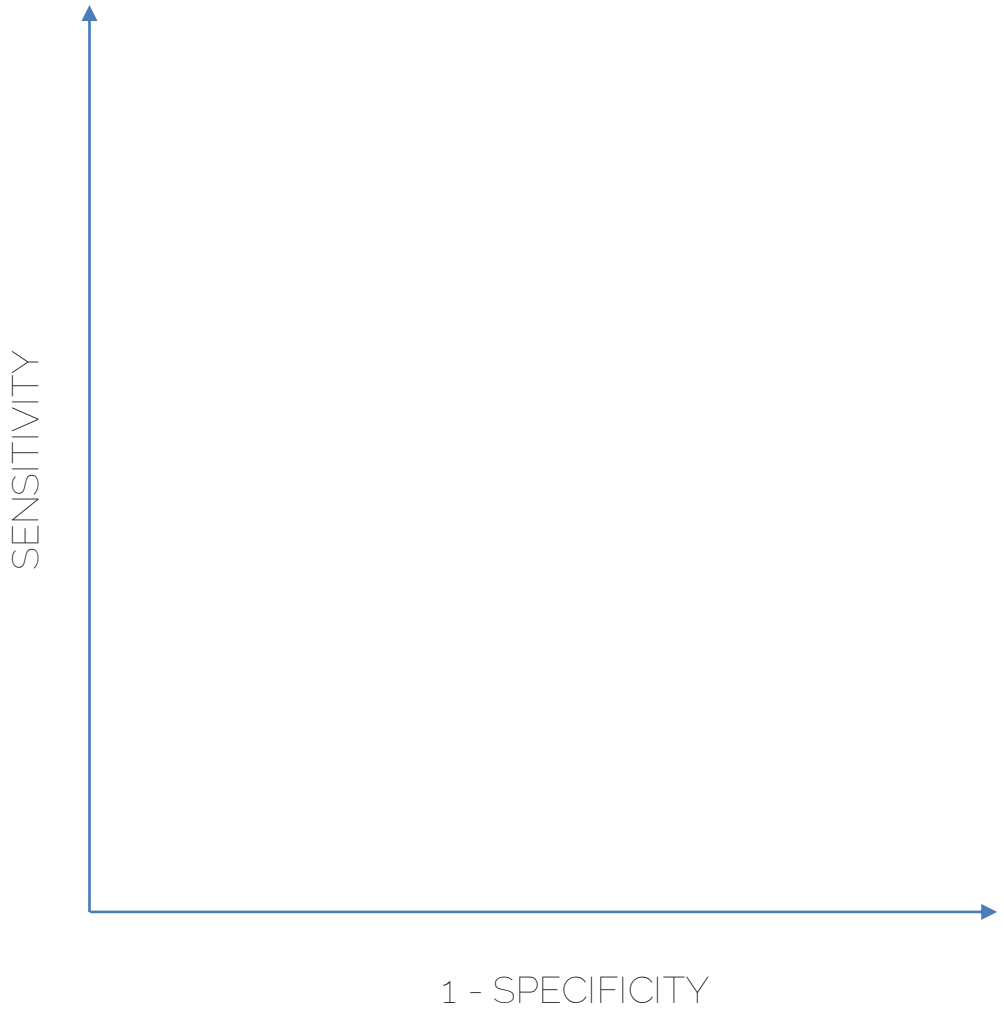


ROC analysis and Area Under the (ROC) Curve

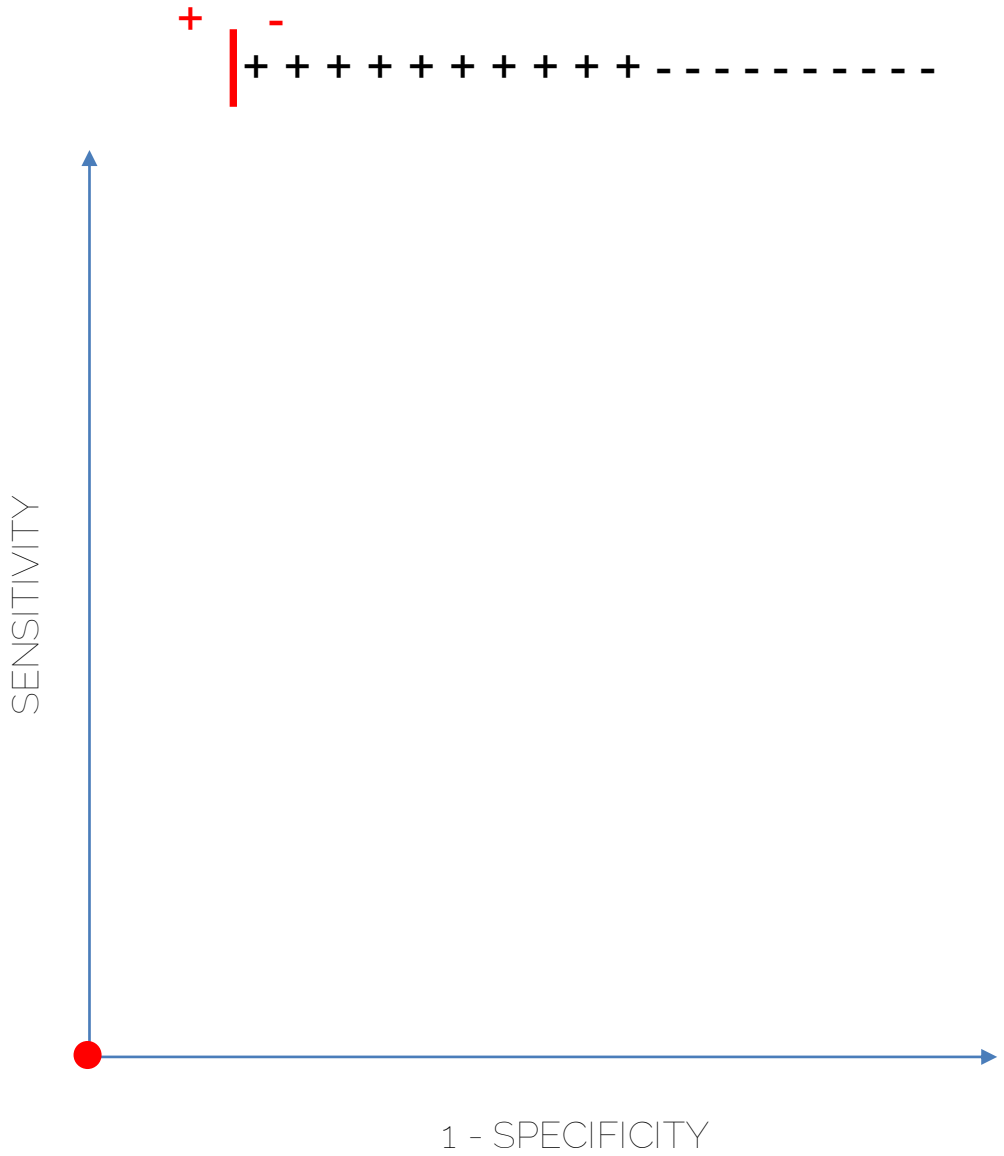


ROC analysis and Area Under the (ROC) Curve

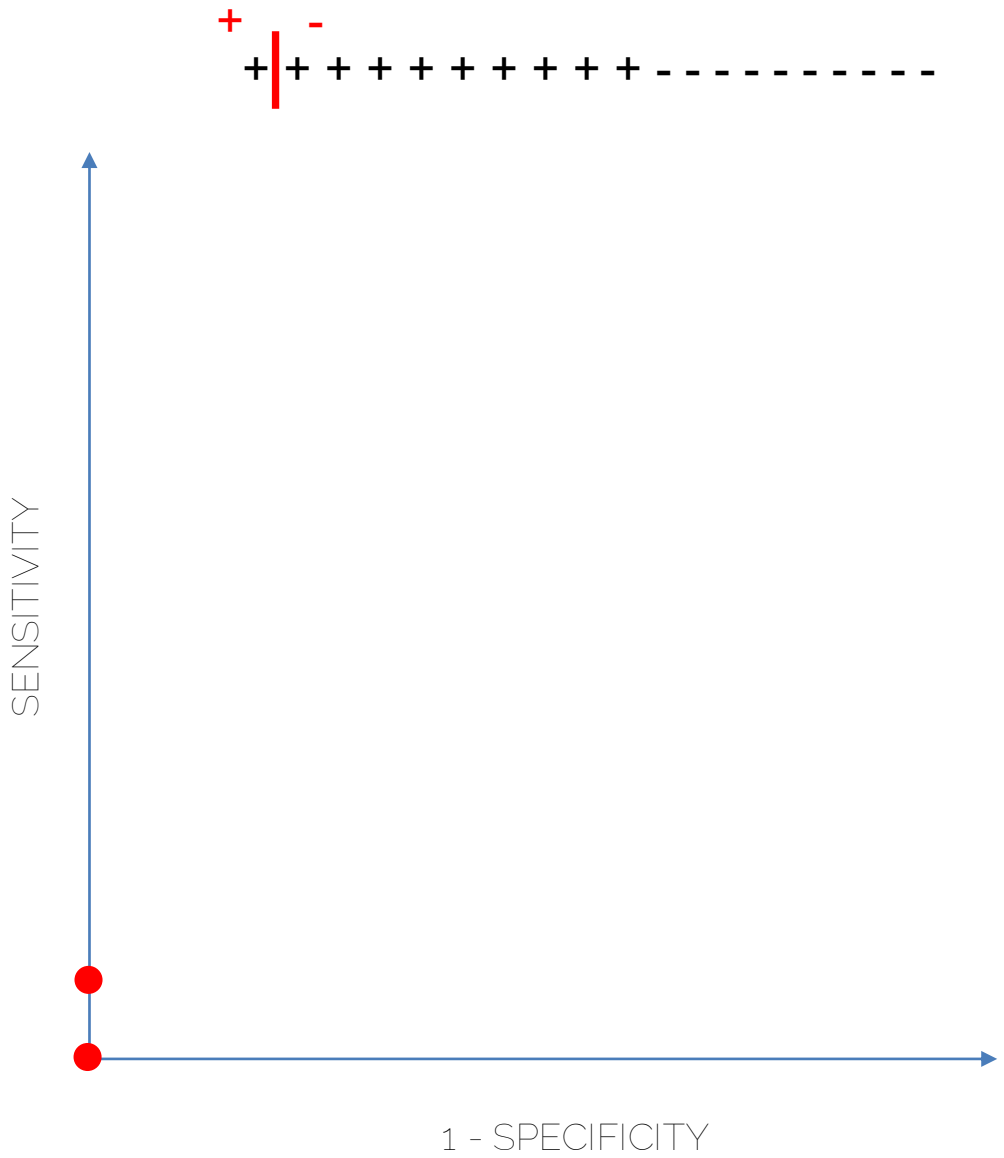
+ + + + + + + + + + - - - - - - - - - -



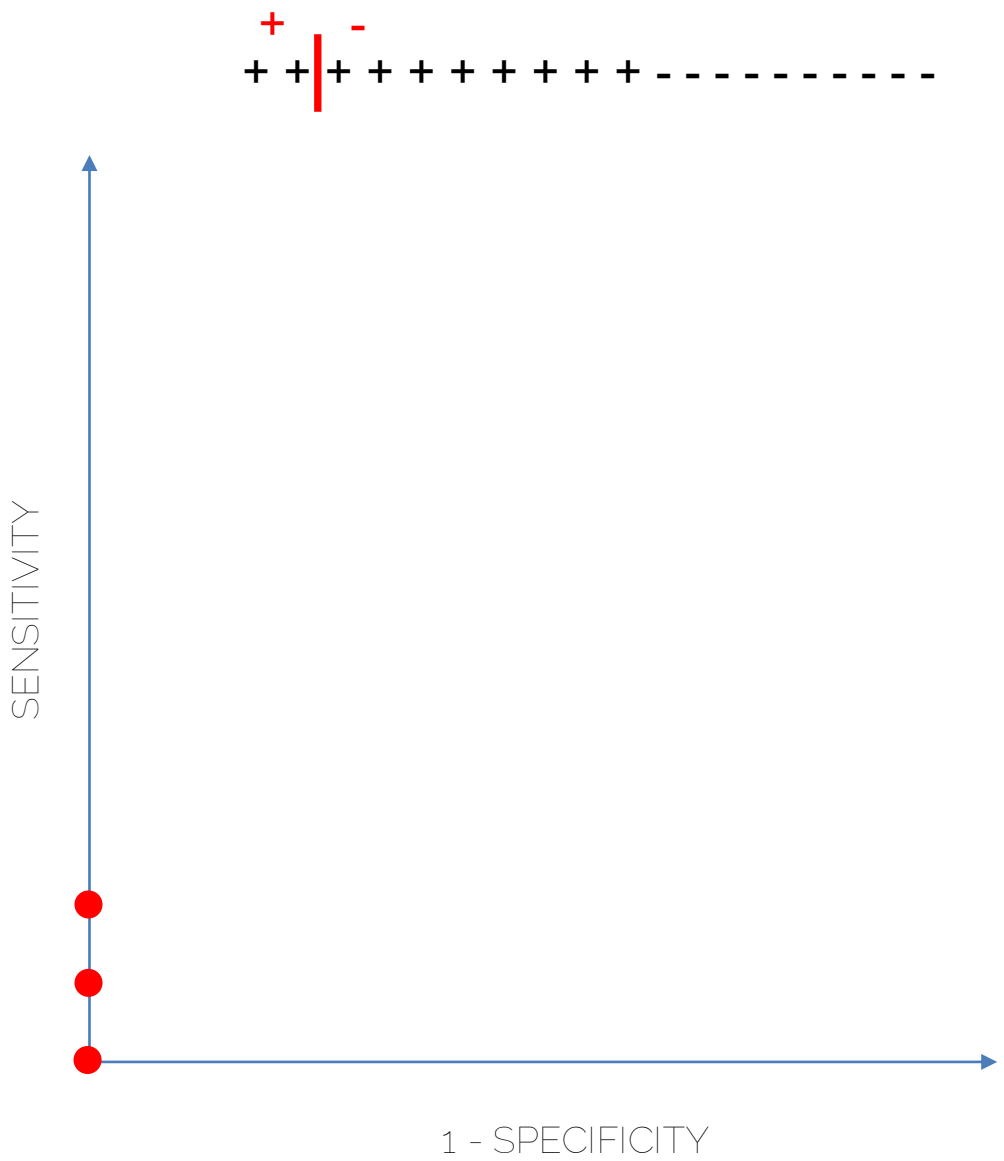
ROC analysis and Area Under the (ROC) Curve



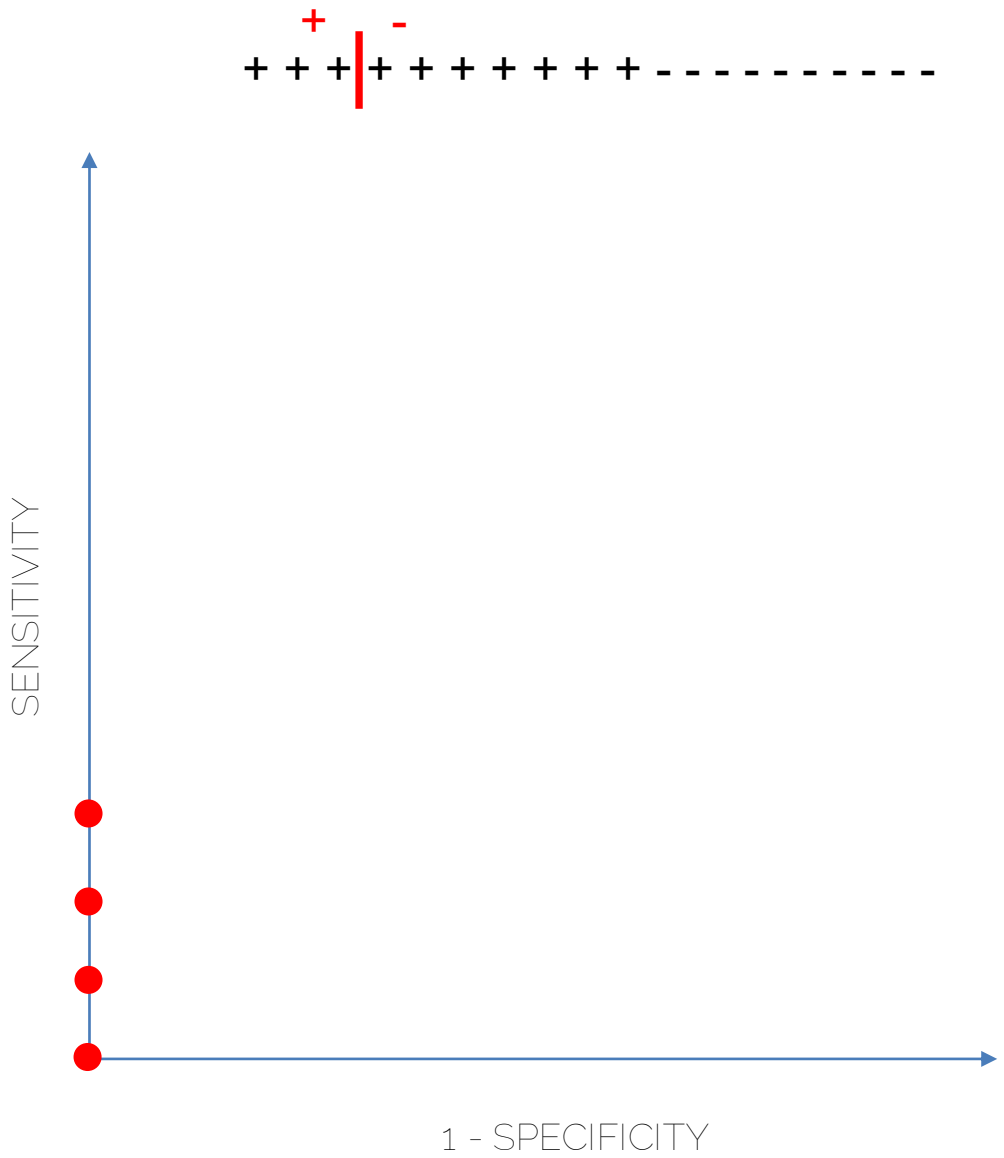
ROC analysis and Area Under the (ROC) Curve



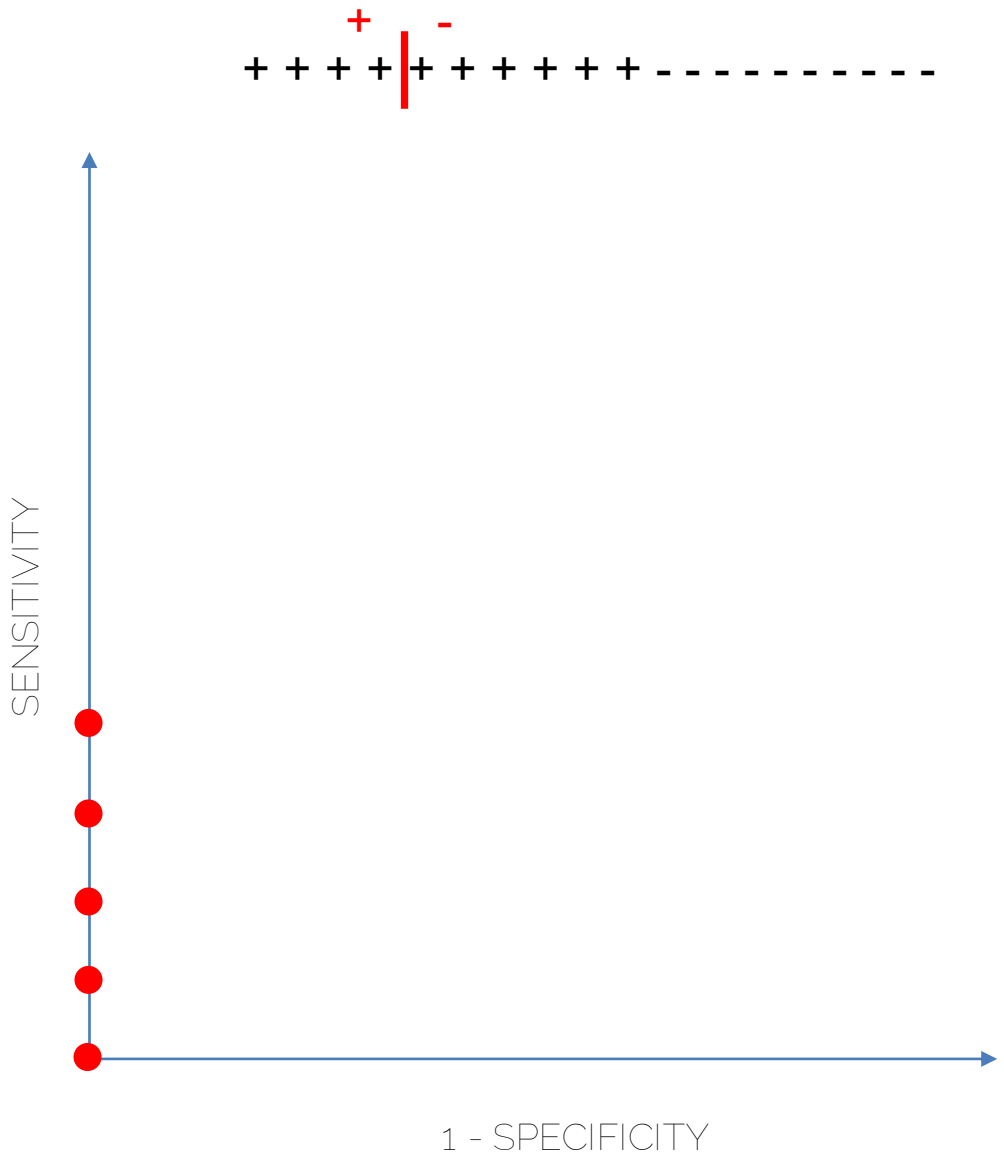
ROC analysis and Area Under the (ROC) Curve



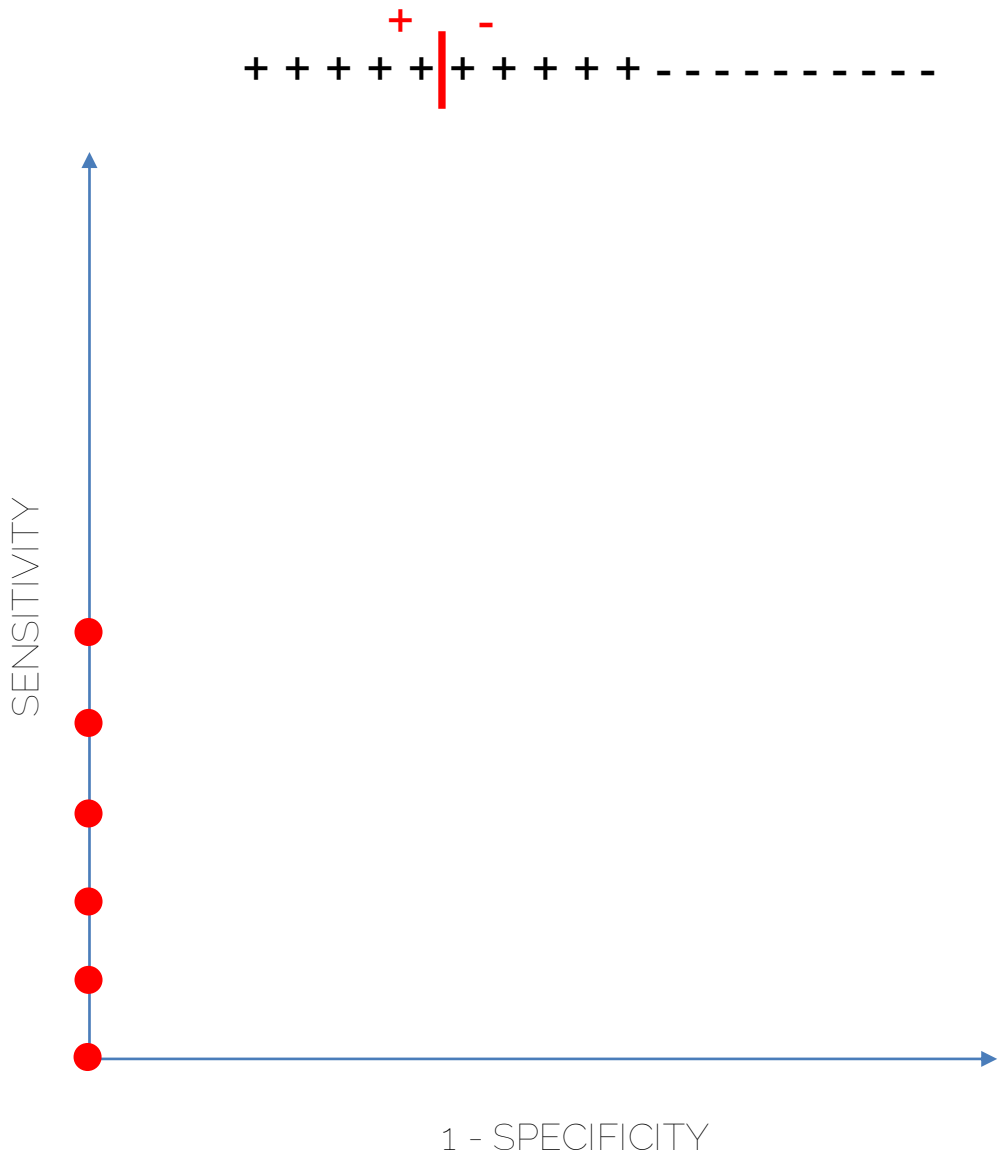
ROC analysis and Area Under the (ROC) Curve



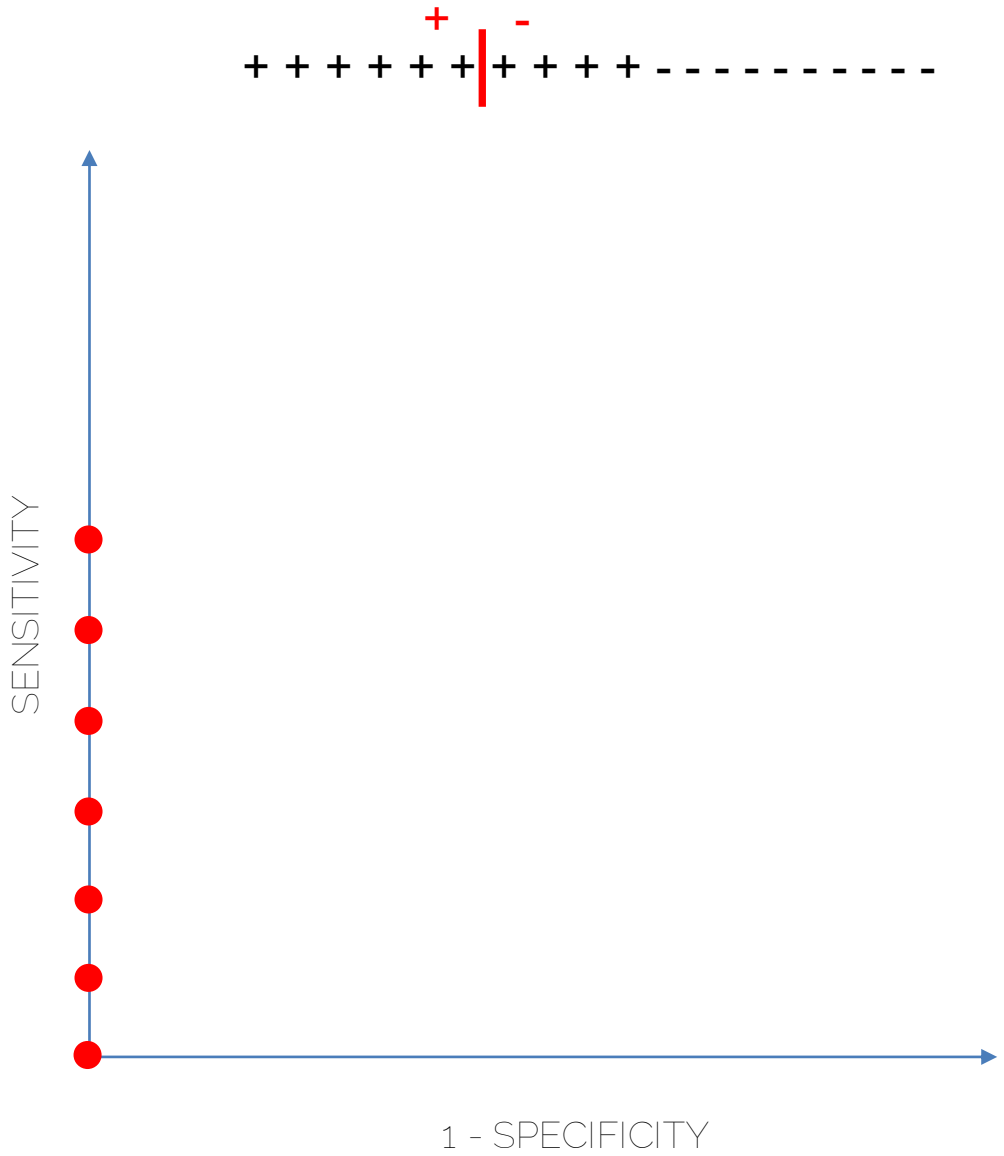
ROC analysis and Area Under the (ROC) Curve



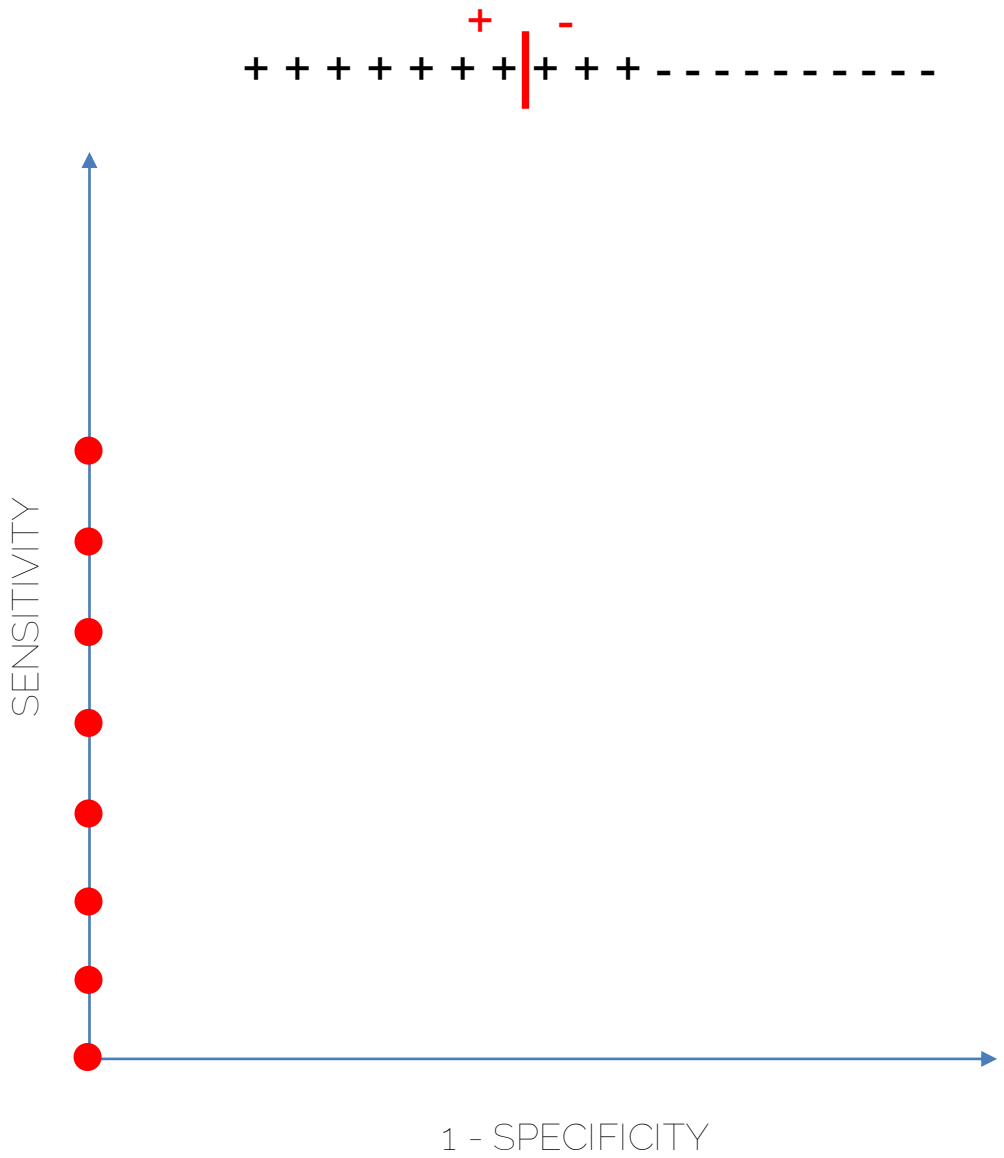
ROC analysis and Area Under the (ROC) Curve



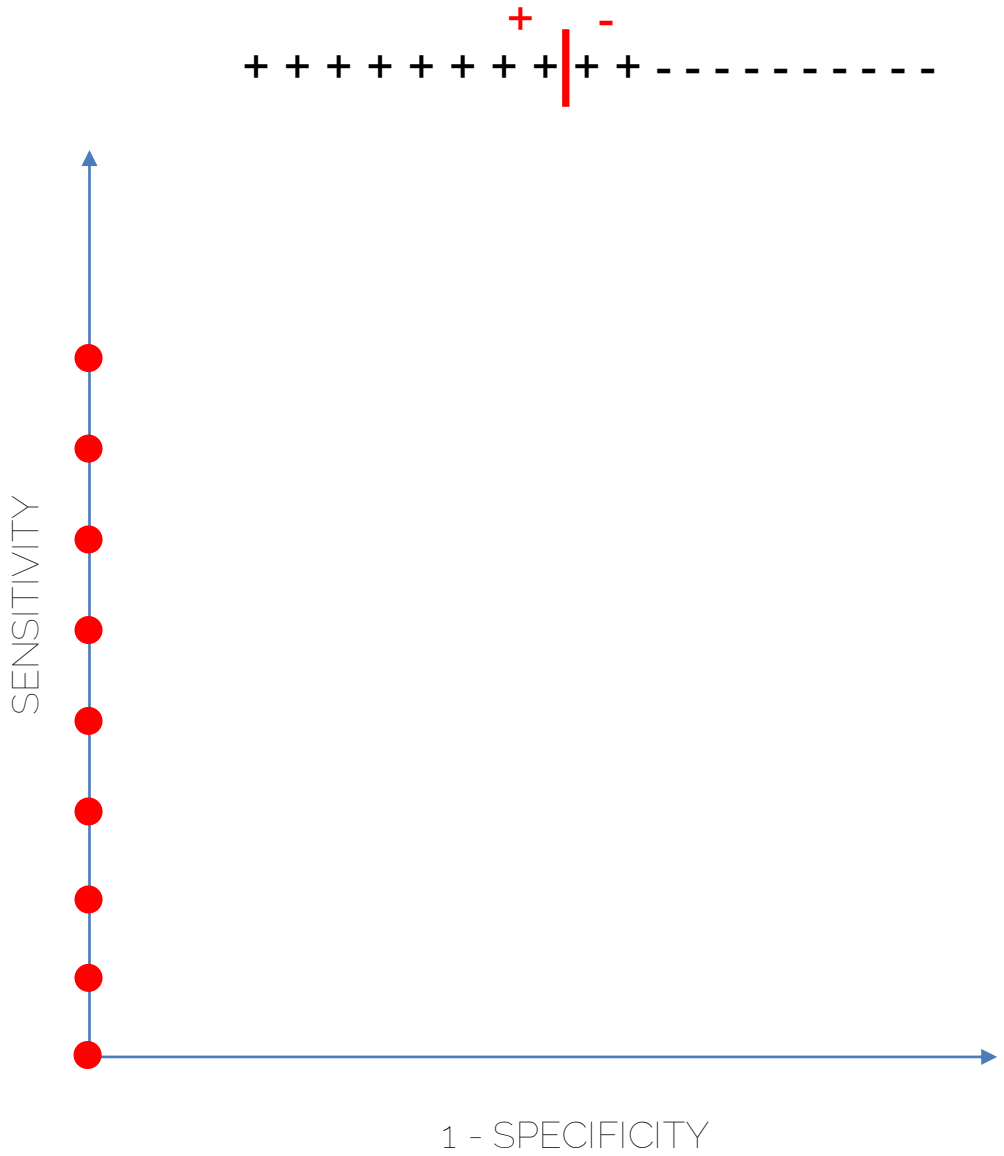
ROC analysis and Area Under the (ROC) Curve



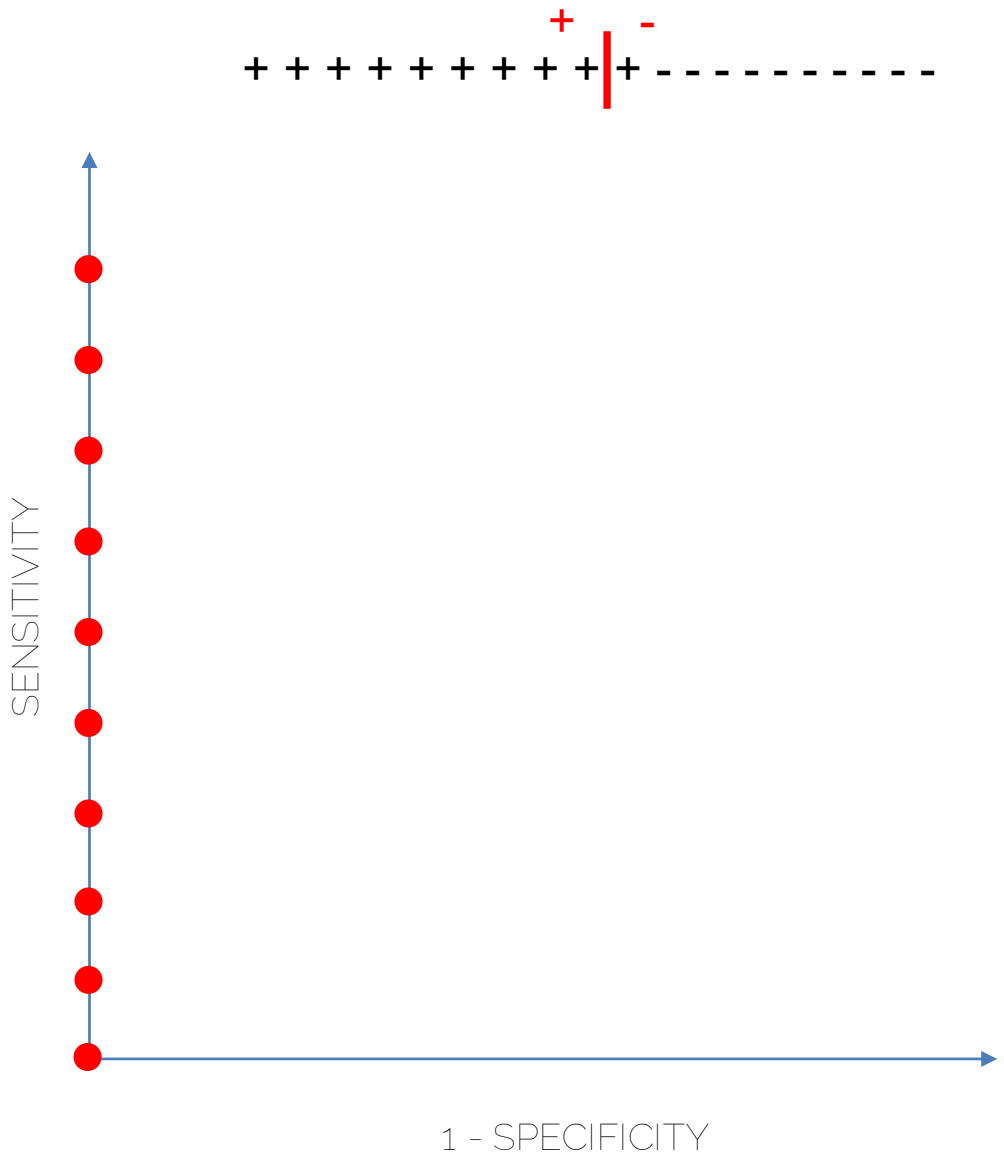
ROC analysis and Area Under the (ROC) Curve



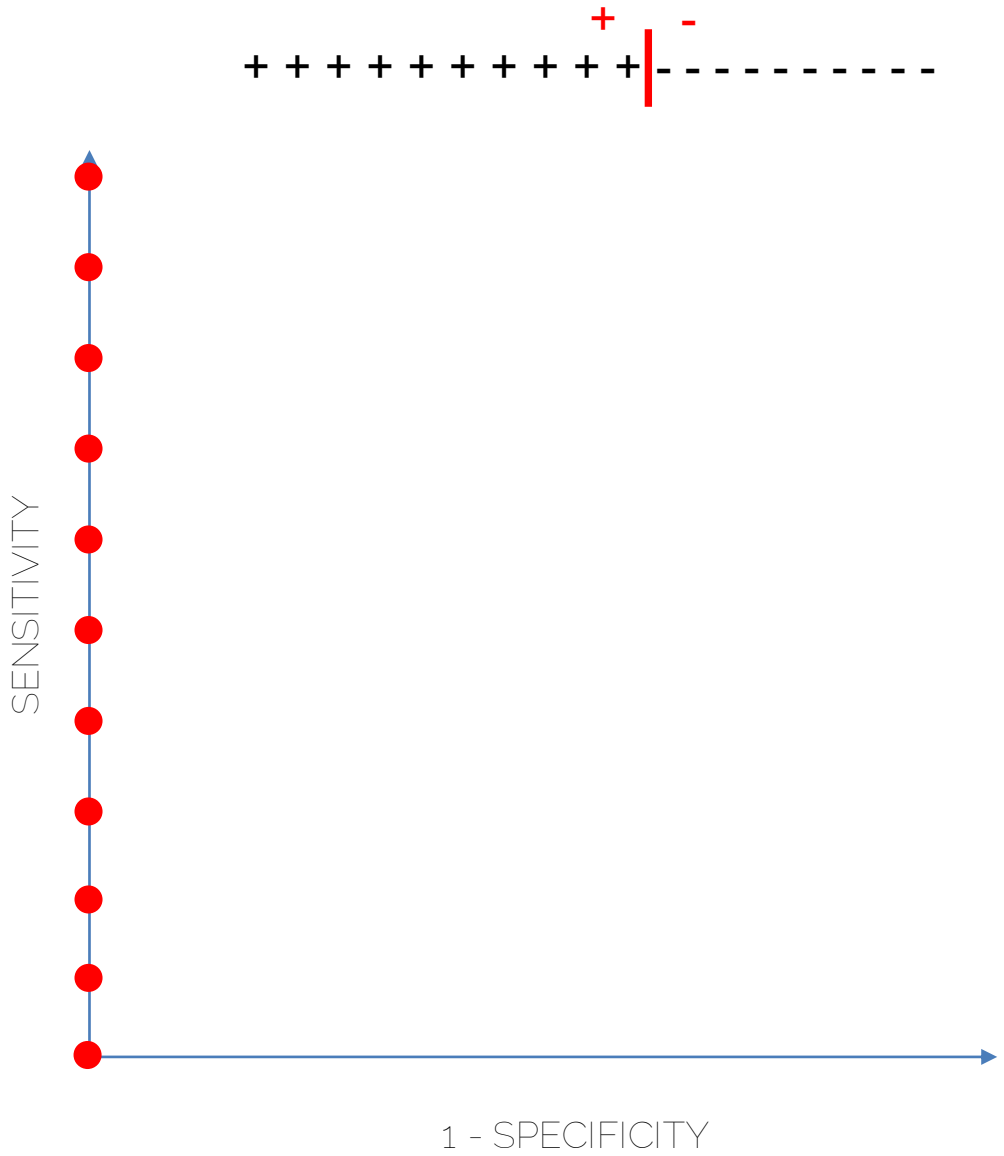
ROC analysis and Area Under the (ROC) Curve



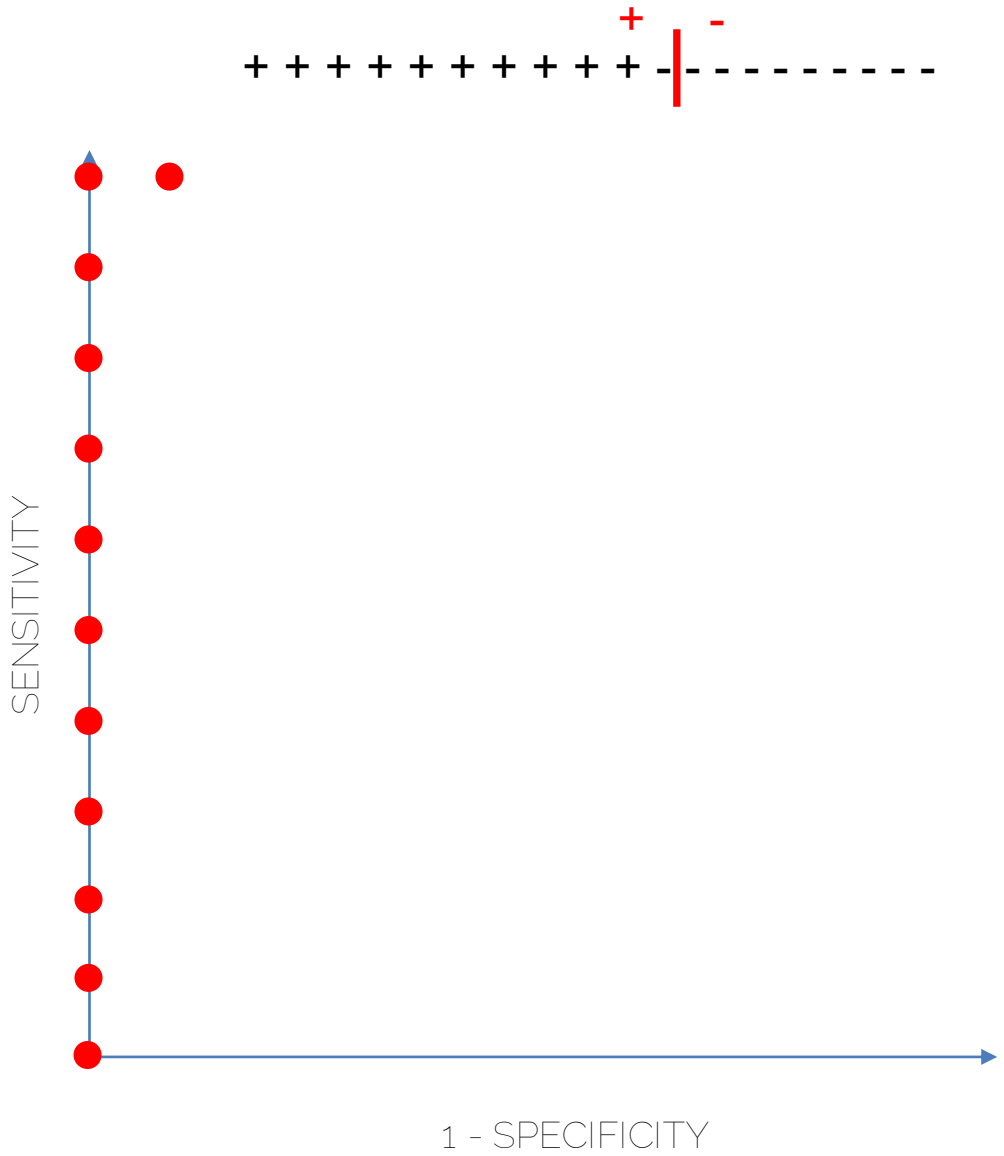
ROC analysis and Area Under the (ROC) Curve



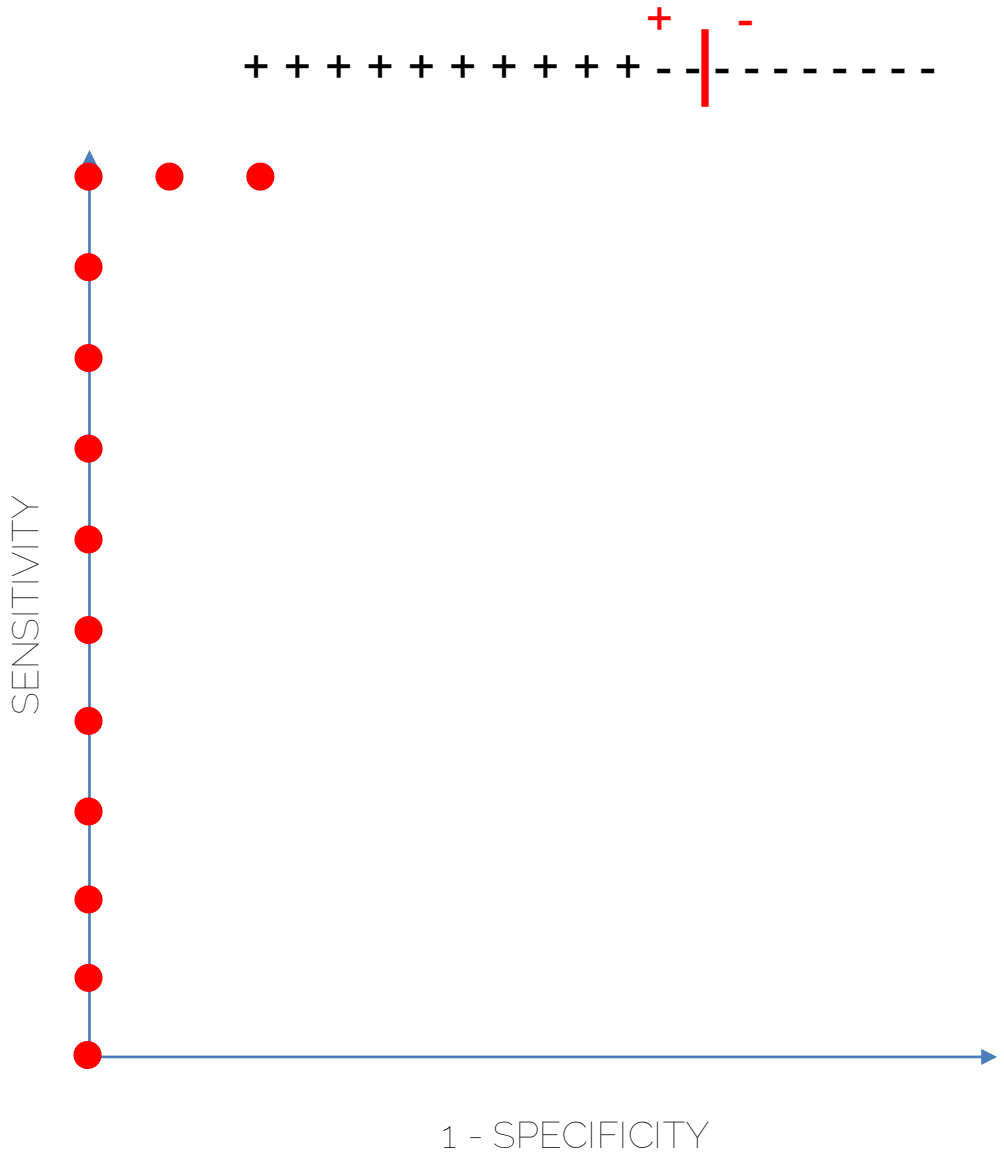
ROC analysis and Area Under the (ROC) Curve



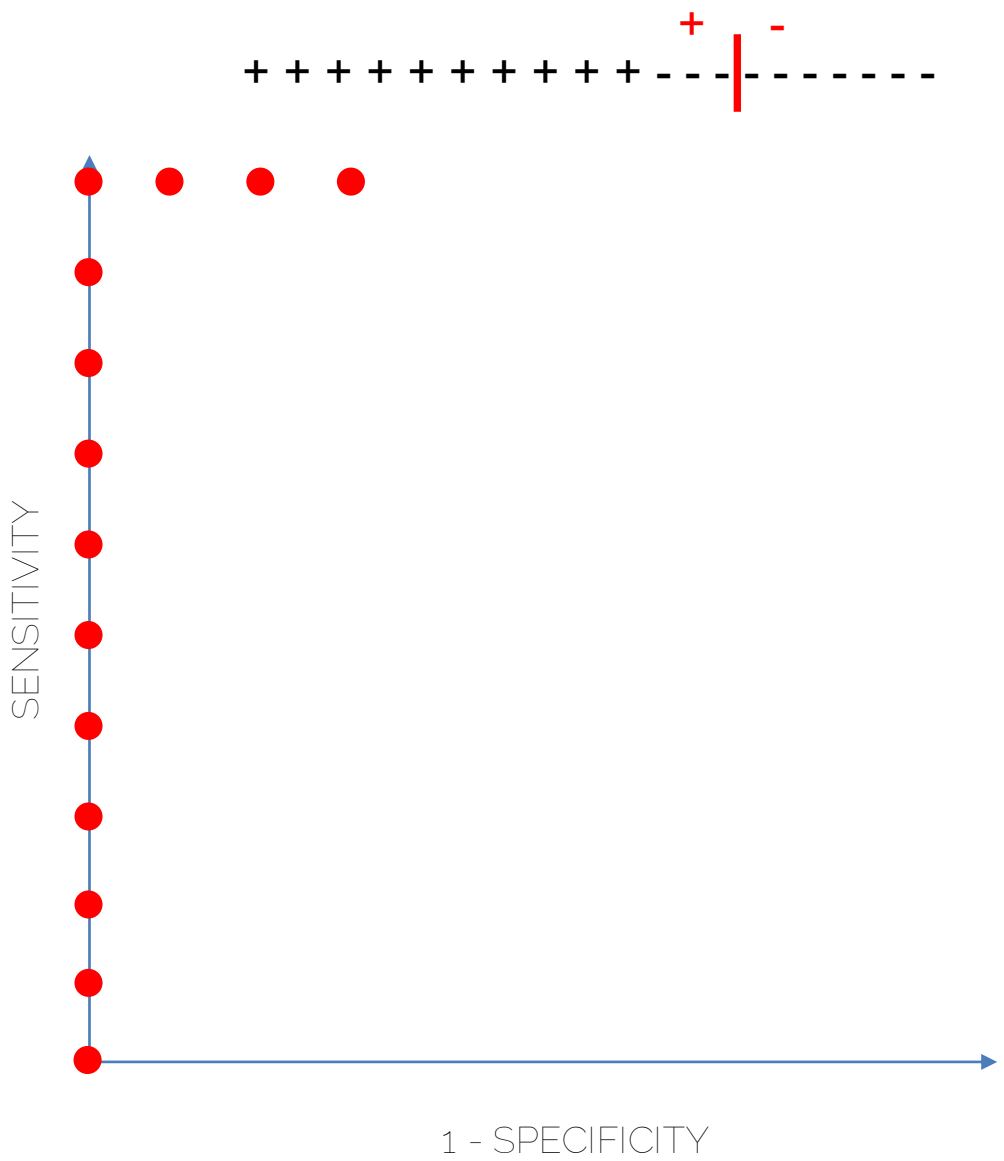
ROC analysis and Area Under the (ROC) Curve



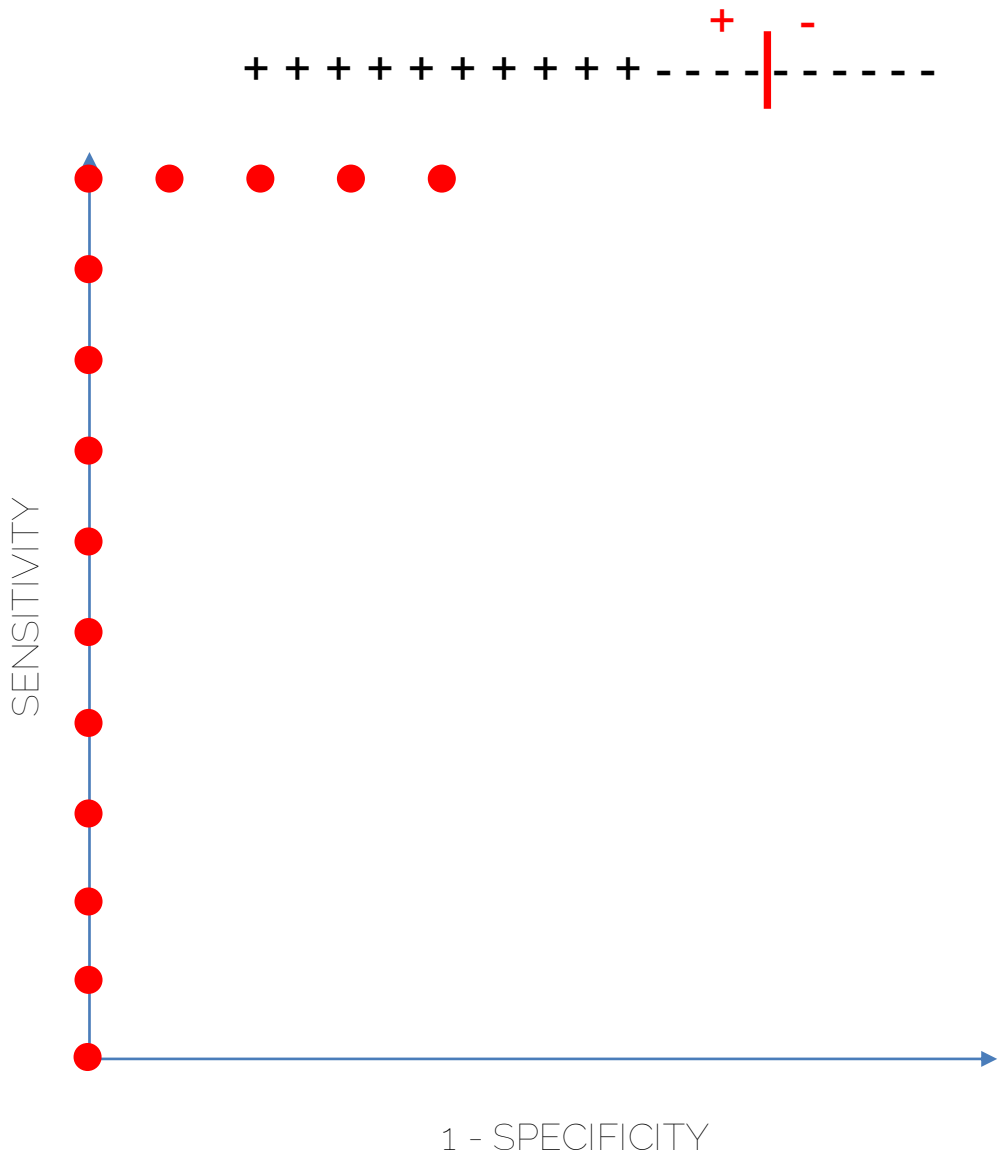
ROC analysis and Area Under the (ROC) Curve



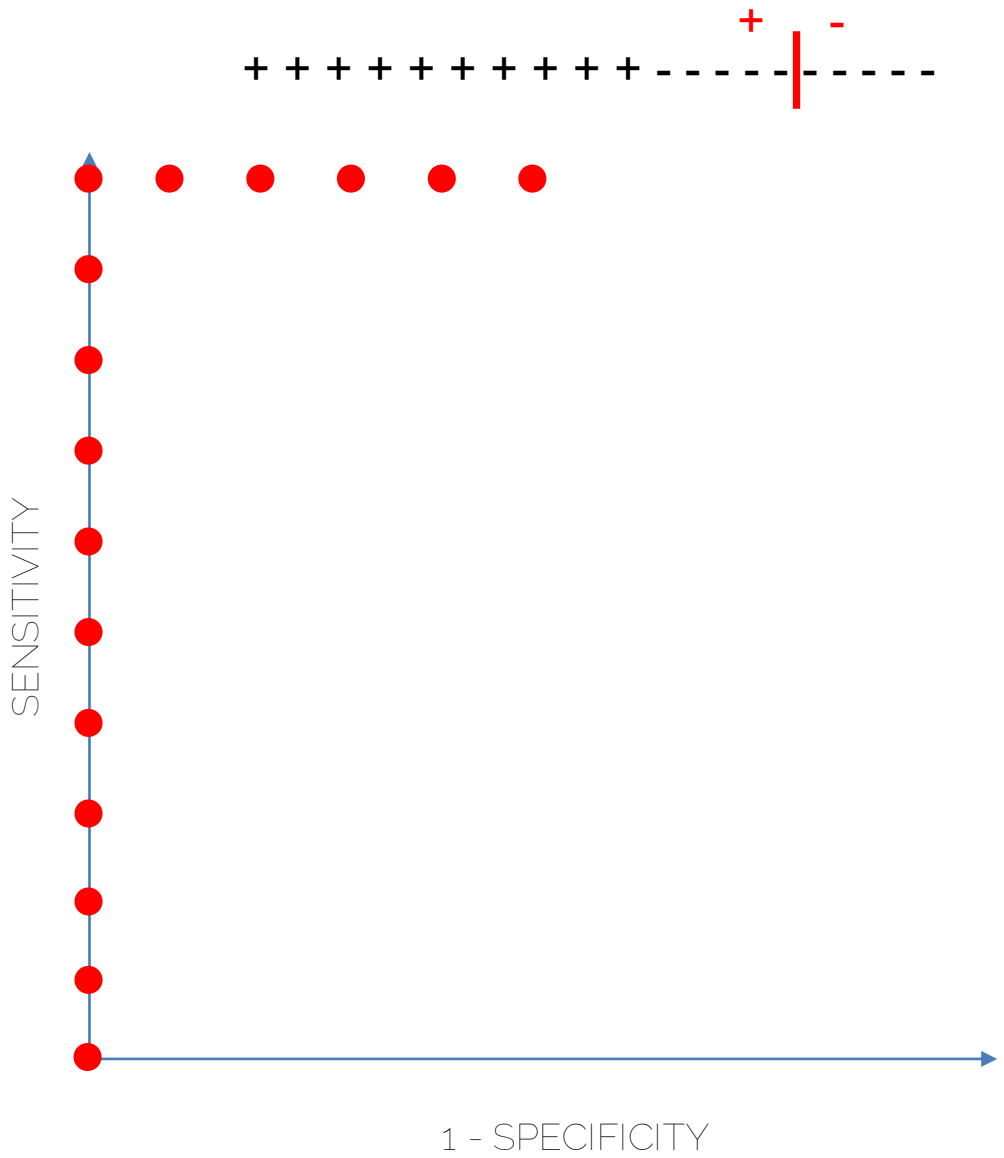
ROC analysis and Area Under the (ROC) Curve



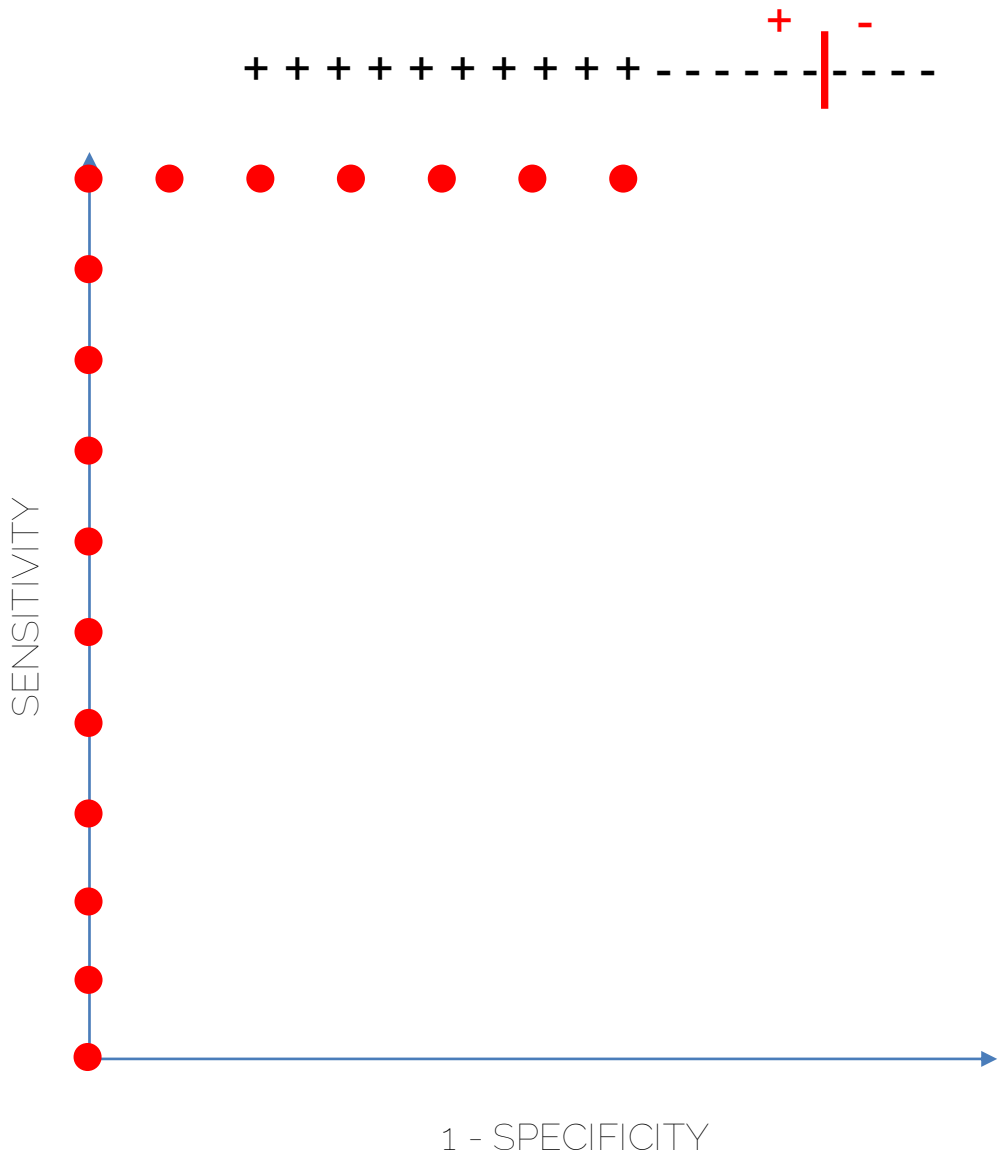
ROC analysis and Area Under the (ROC) Curve



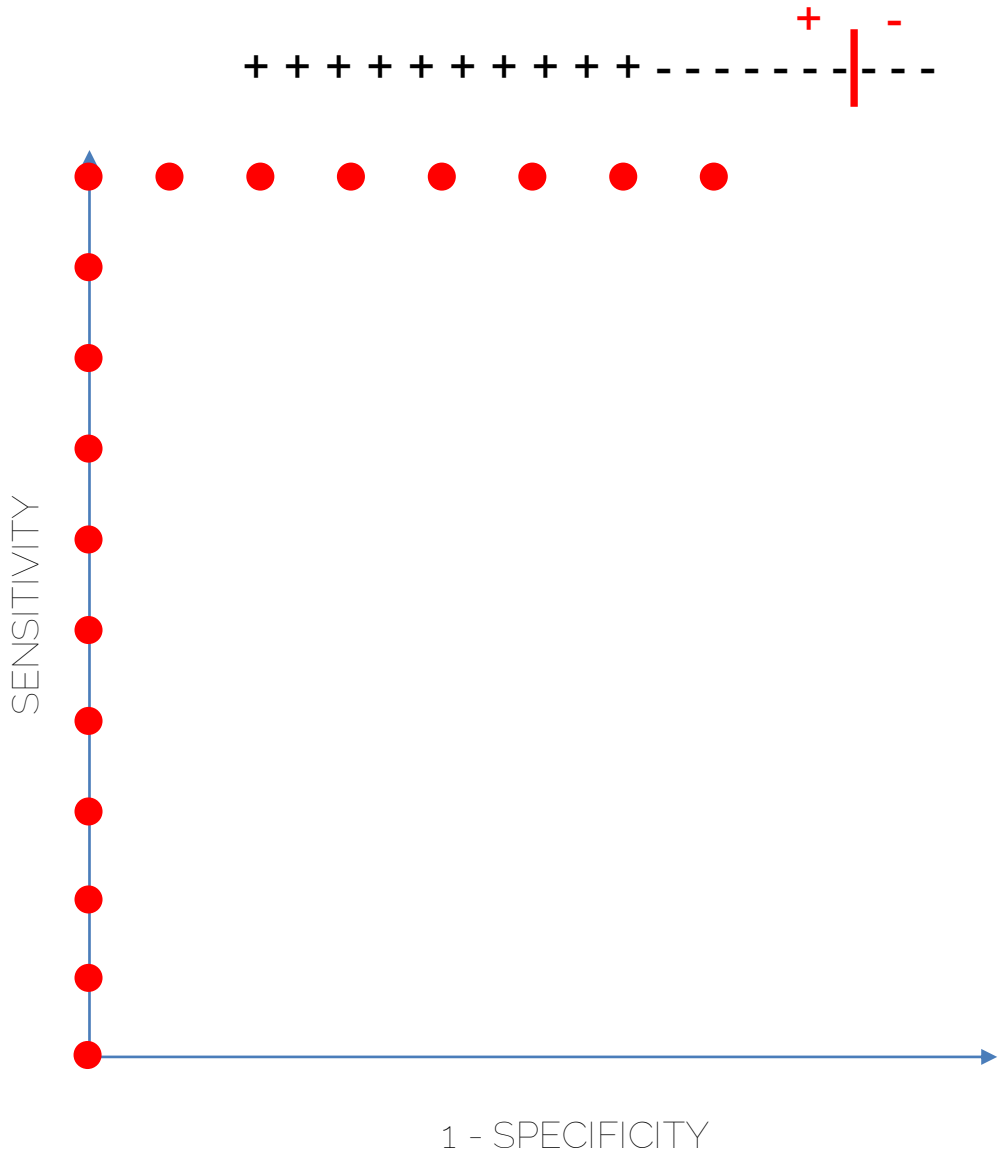
ROC analysis and Area Under the (ROC) Curve



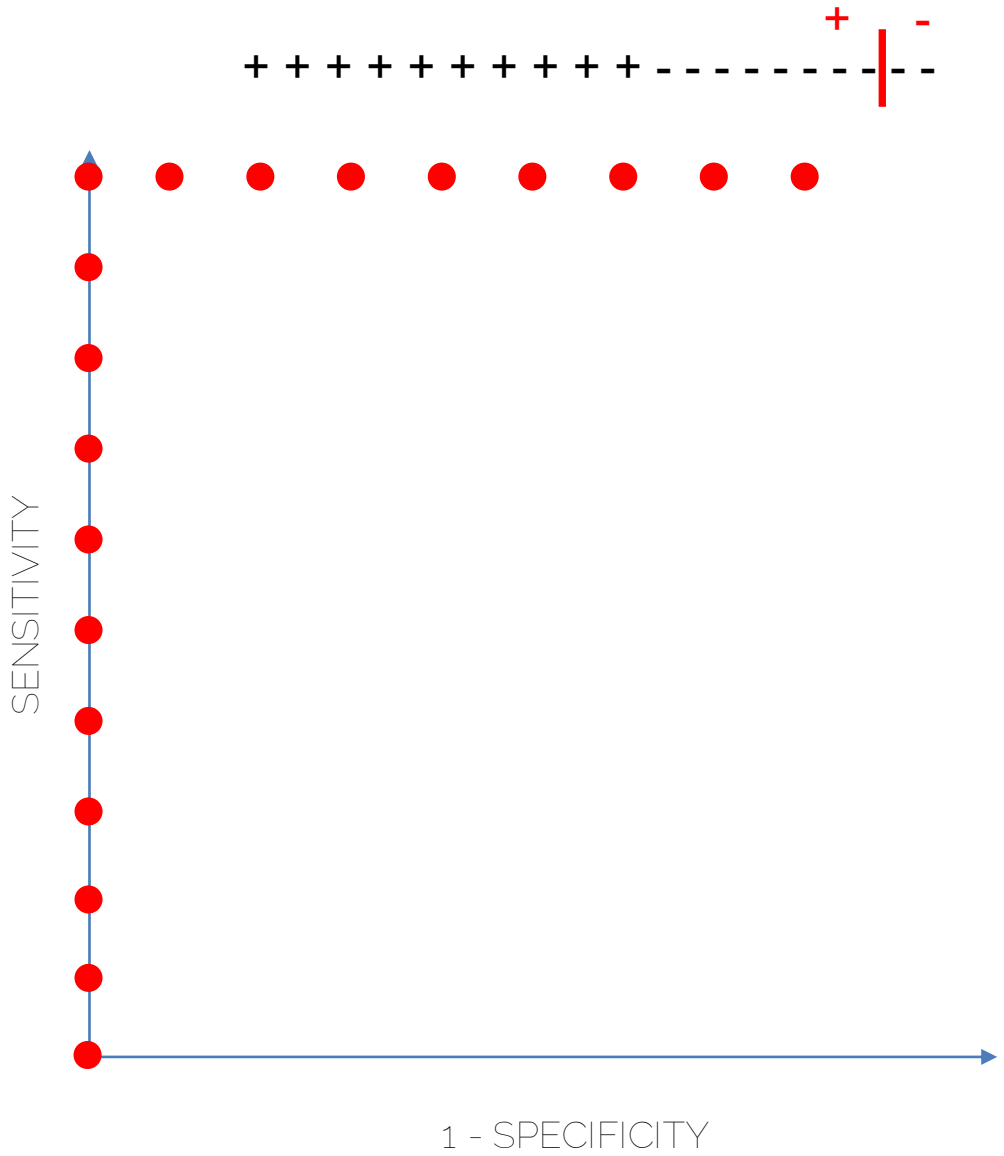
ROC analysis and Area Under the (ROC) Curve



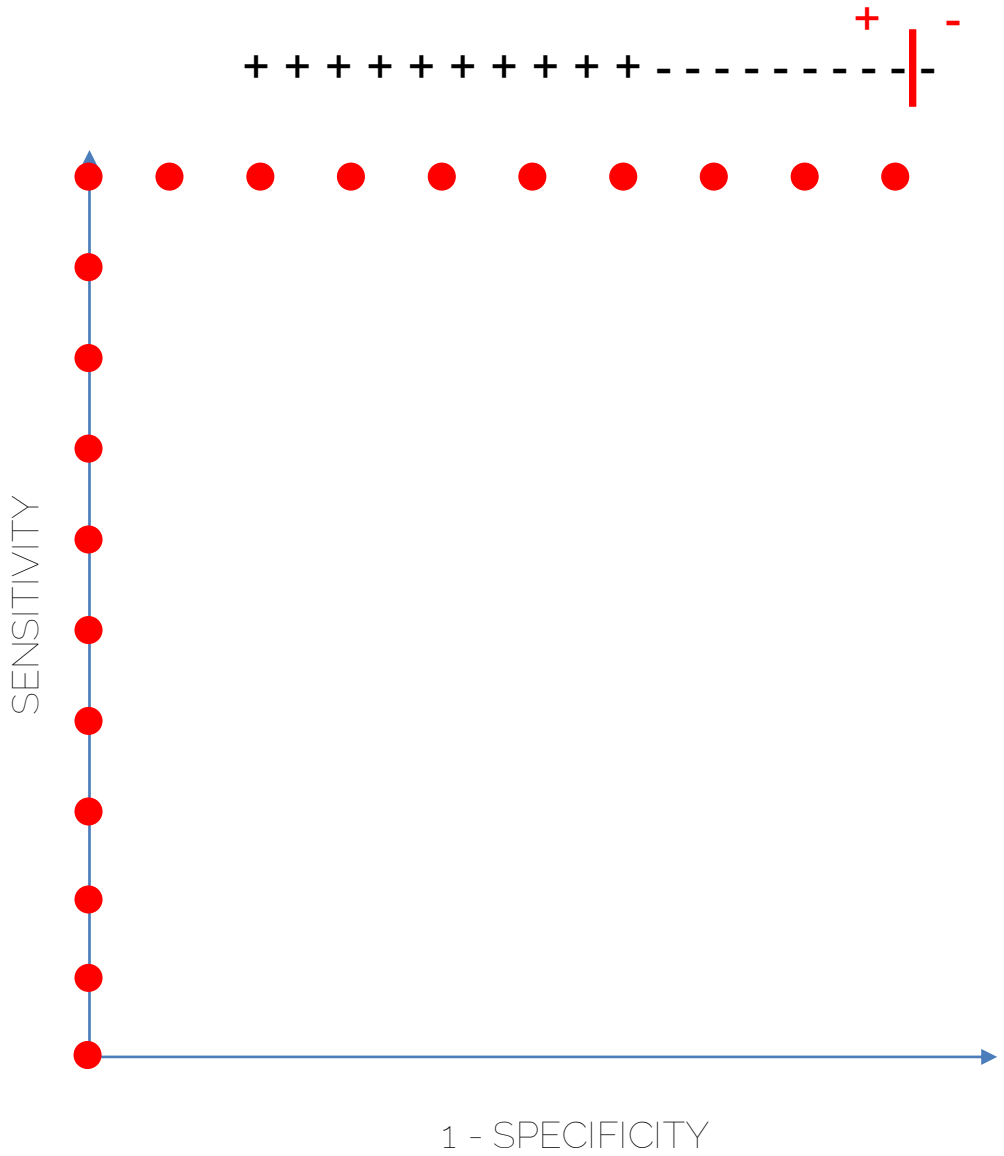
ROC analysis and Area Under the (ROC) Curve



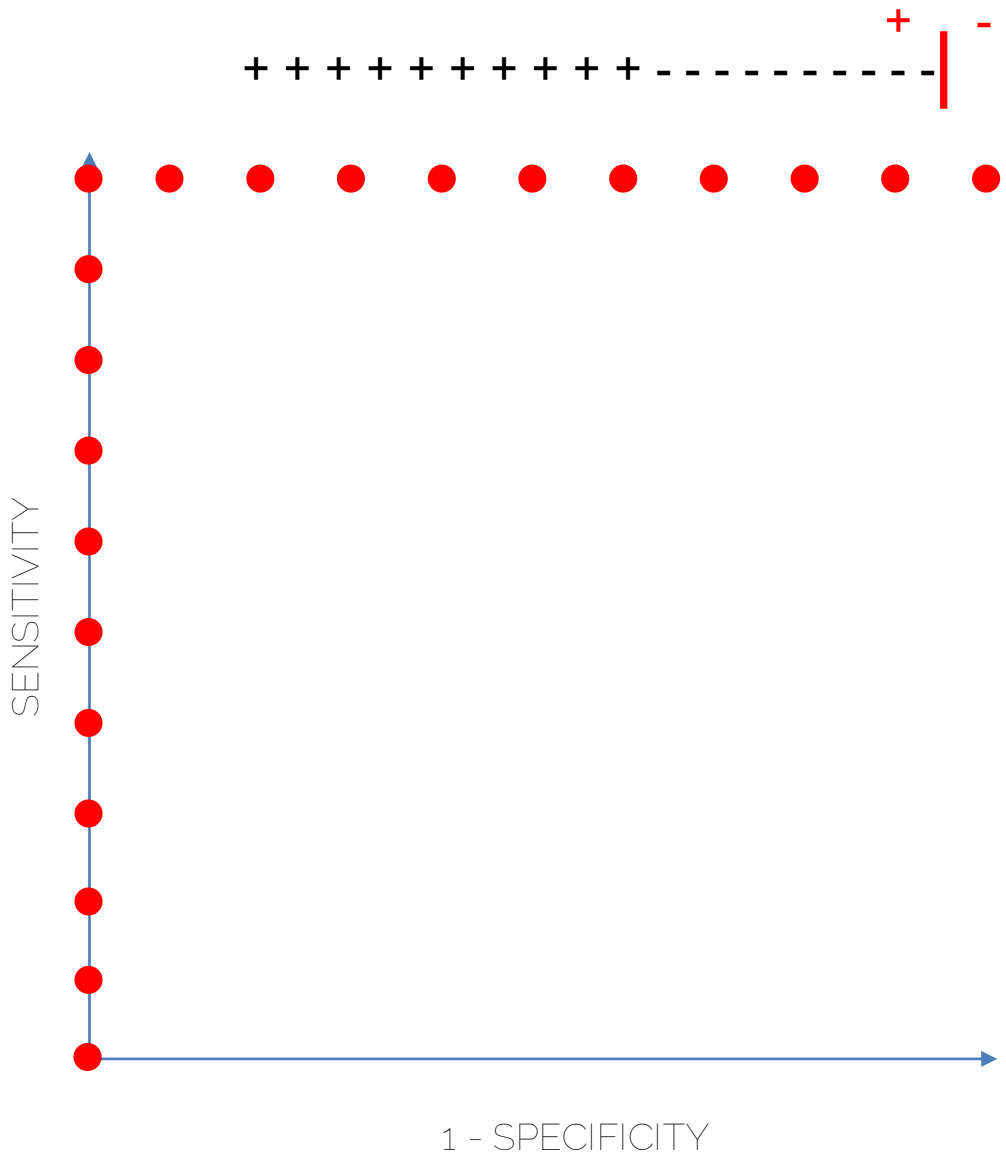
ROC analysis and Area Under the (ROC) Curve



ROC analysis and Area Under the (ROC) Curve

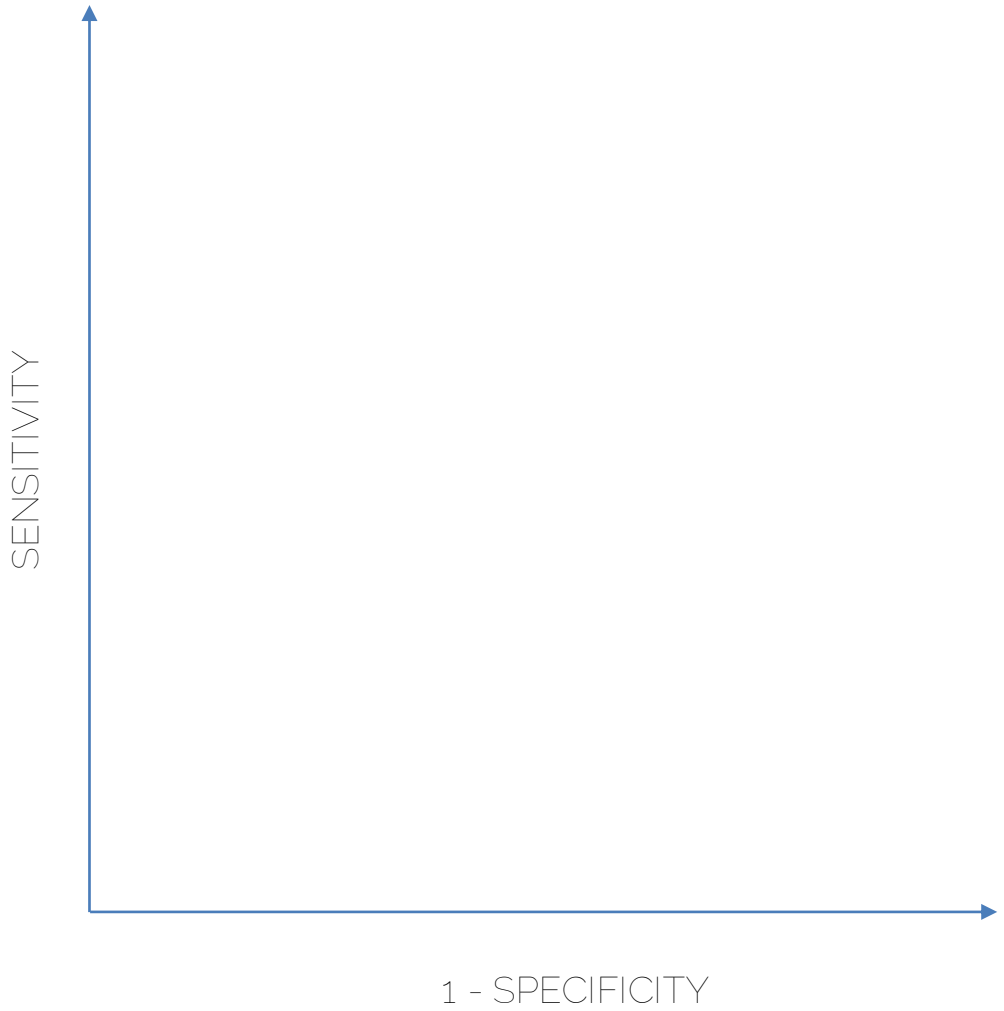


ROC analysis and Area Under the (ROC) Curve

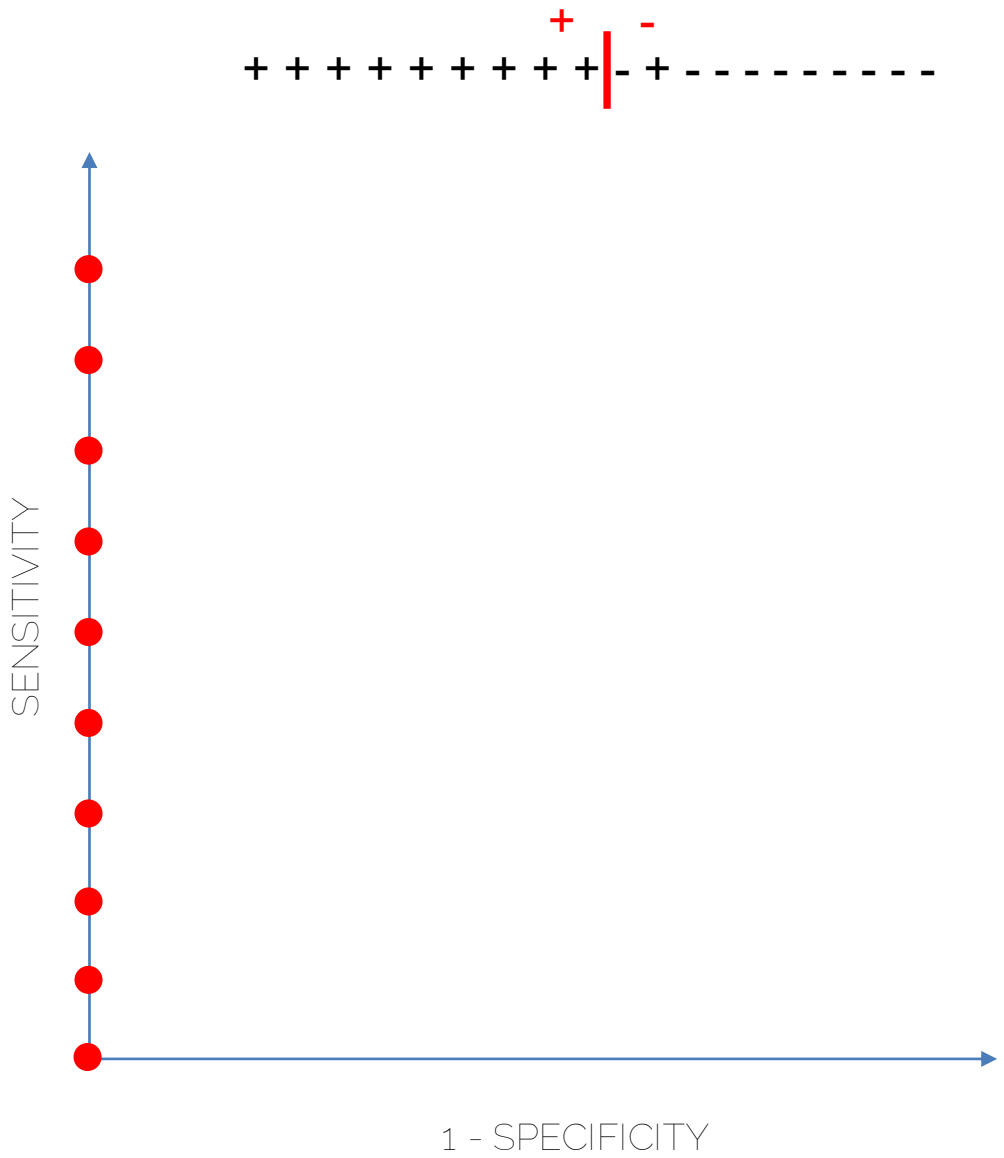


ROC analysis and Area Under the (ROC) Curve

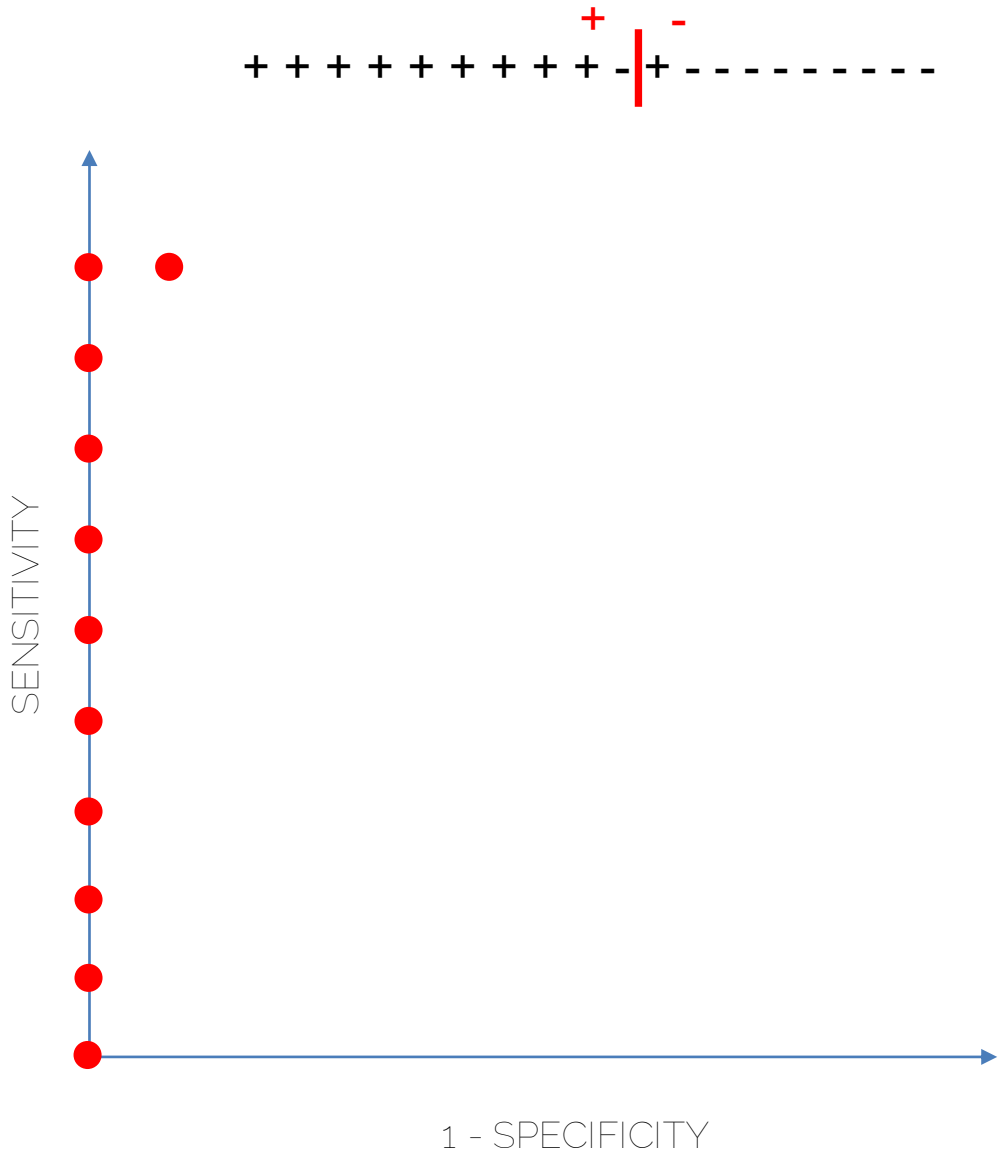
+ + + + + + + + + + - + - - - - - - - - -



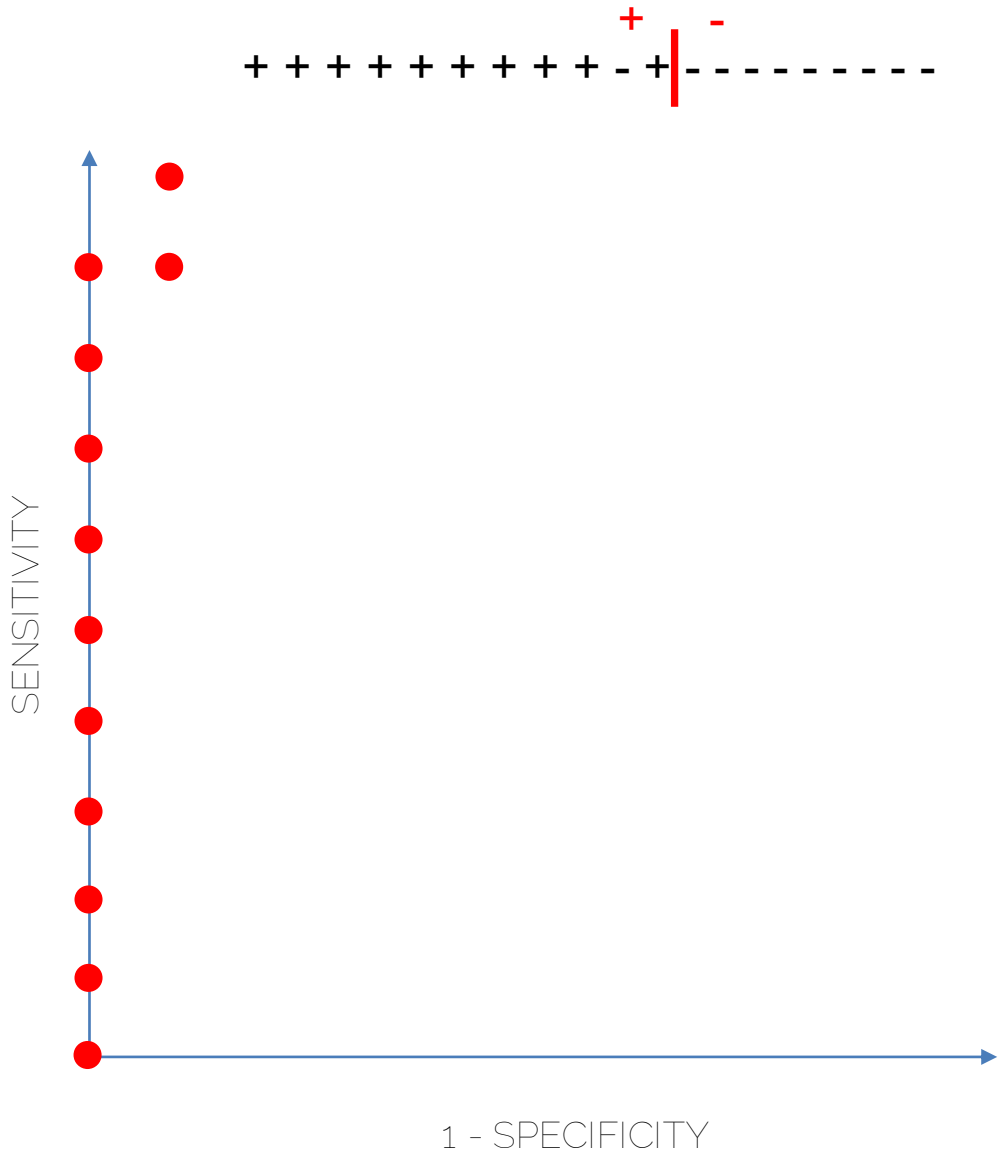
ROC analysis and Area Under the (ROC) Curve



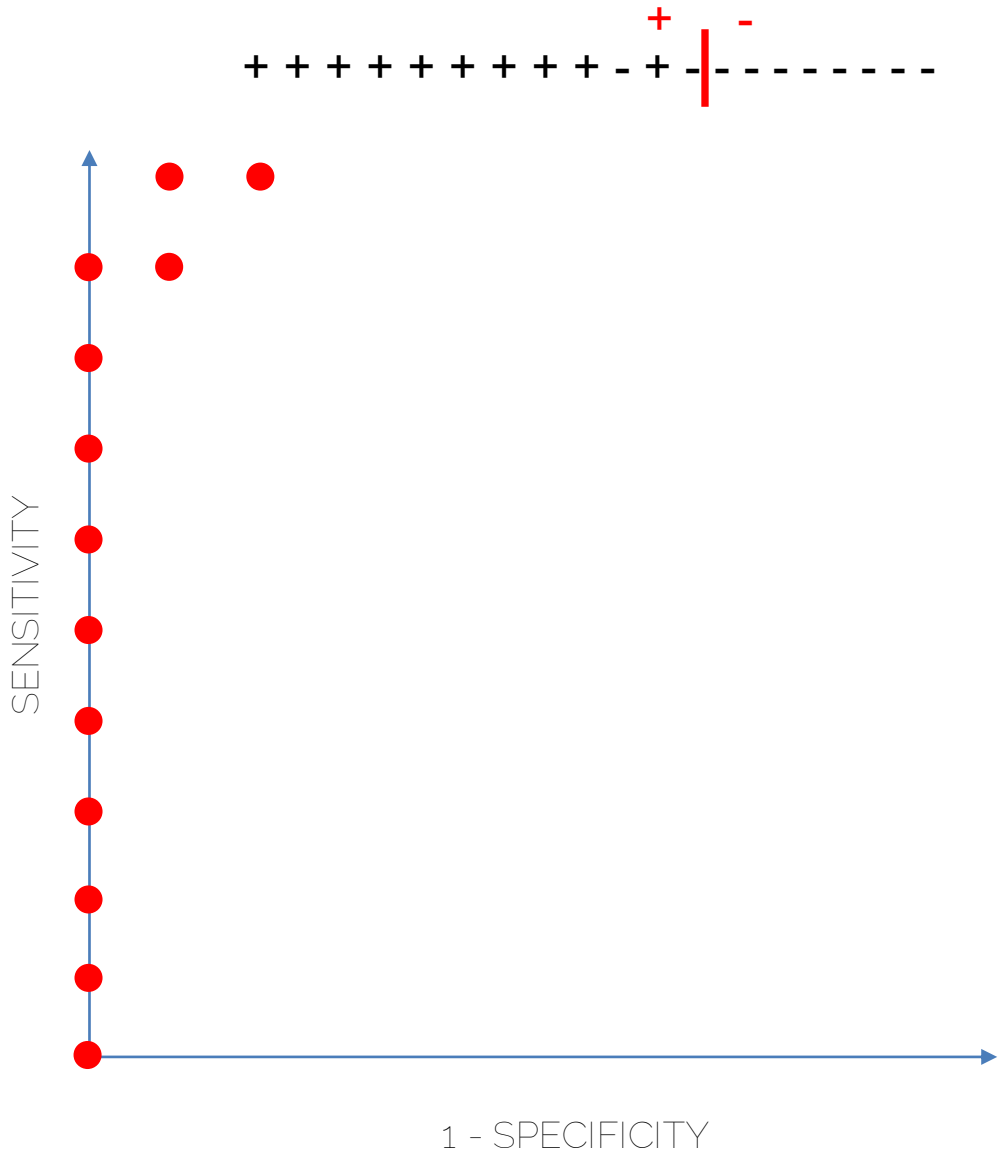
ROC analysis and Area Under the (ROC) Curve



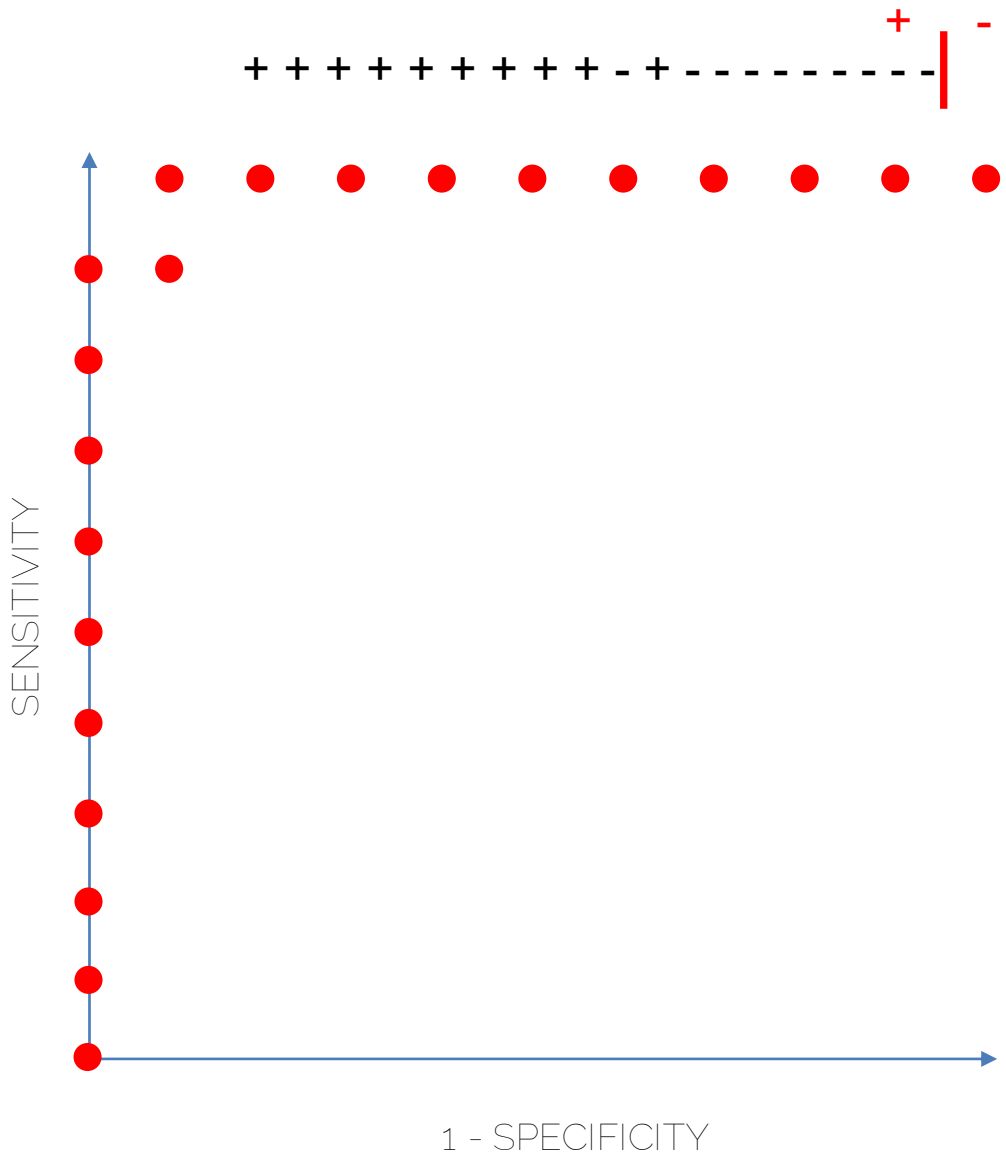
ROC analysis and Area Under the (ROC) Curve



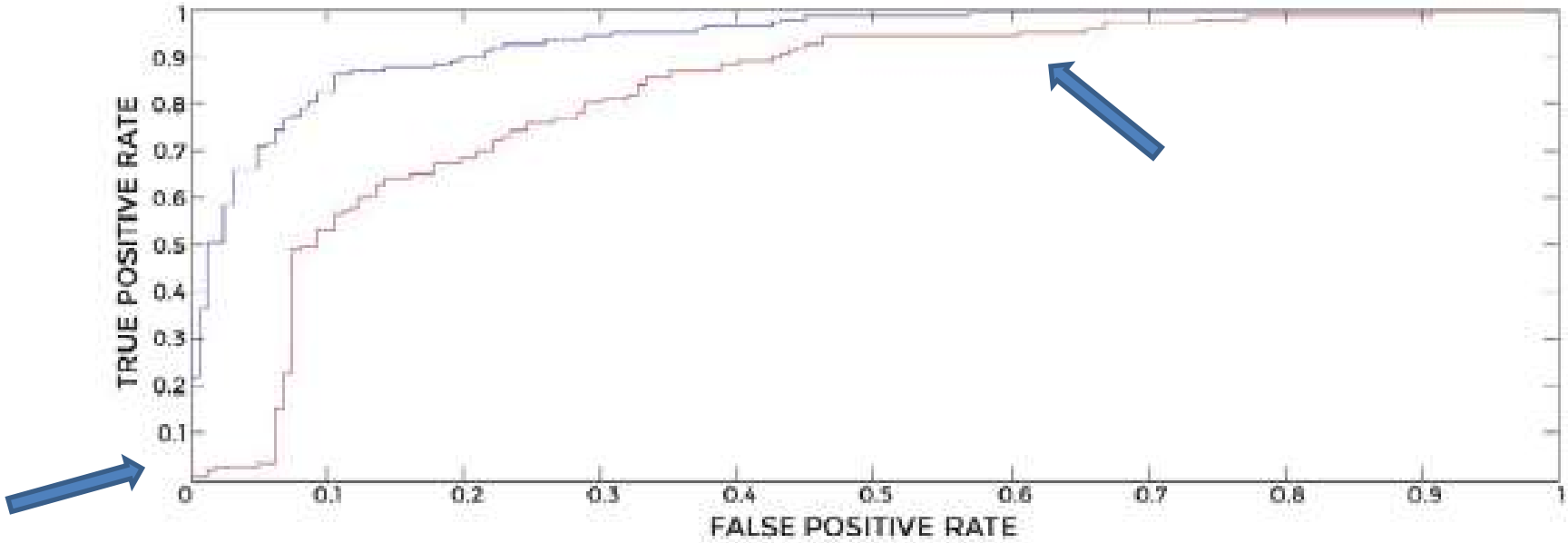
ROC analysis and Area Under the (ROC) Curve

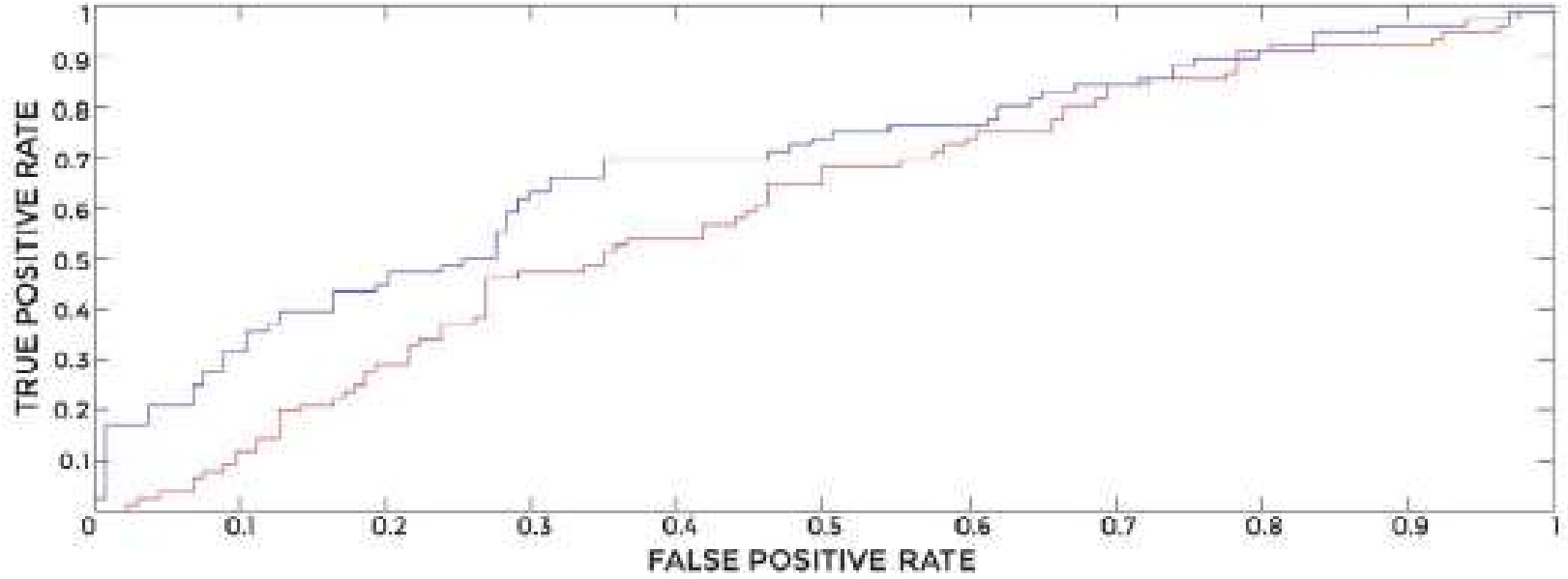


ROC analysis and Area Under the (ROC) Curve



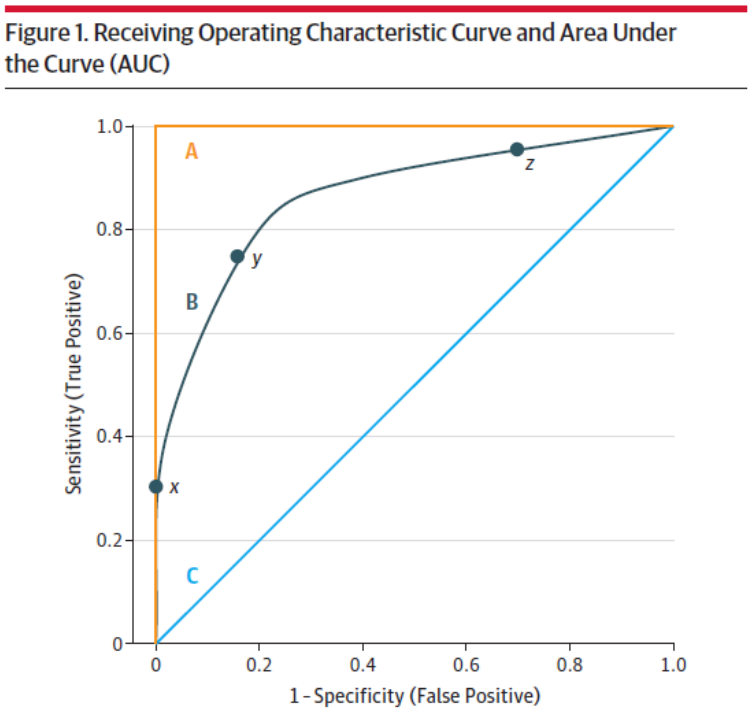
ROC analysis and Area Under the (ROC) Curve

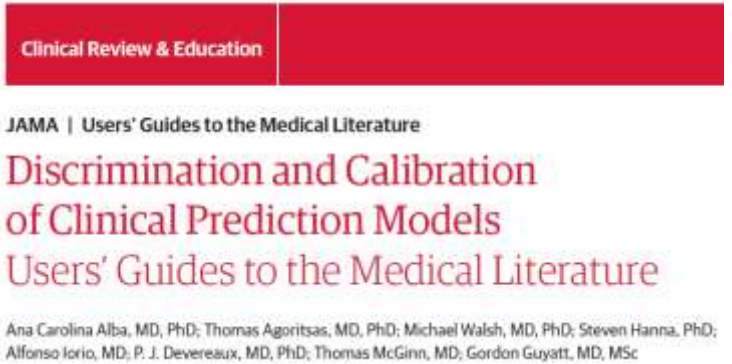




Discrimination and Calibration of Clinical Prediction Models

Discrimination refers to how well the model differentiates those at higher risk of having an event from those at lower risk. Discrimination depends on the distribution of patient characteristics in the population in which the model is being used. A model could well discriminate patients with events from those without events in a heterogenous population with widely different values of predictors included in the model (e.g., age, sex, laboratory values); however, the same model could fail to discriminate patients in a more homogenous population.





But…

A model can have excellent discrimination (and thus a high C statistic) and still provide misleading absolute risks.

Discrimination and Calibration of Clinical Prediction Models

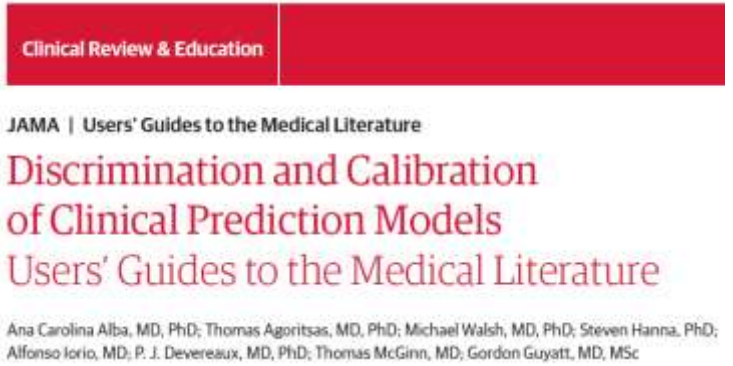
Calibration or goodness of fit is often considered the most important property of a model, and **reflects the extent to which a model correctly estimates the absolute risk** (i.e., if the values predicted by the model agree with the observed values). Poorly calibrated models will underestimate or overestimate the outcome of interest.

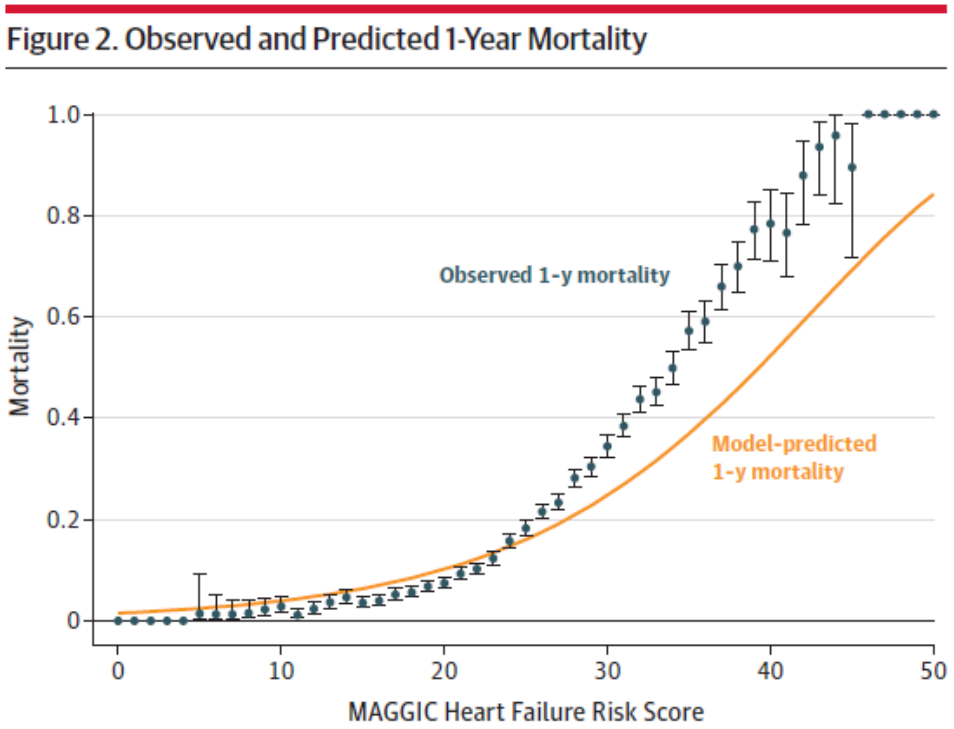
Assessing calibration involves comparing predicted and observed risk at different levels in (1) the whole population (mean calibration); (2) different groups of patients based on predicted risk; or (3) different groups of patients based on combinations of predictors (covariates).

An excellent model will show strong calibration for different groups of patients with different characteristics; however, simulation exercises suggest that accurate calibration for different populations may be unrealistic.

Models that are well calibrated to predict average population risk could have poor discrimination and thus no clinical value.

A model with adequate calibration by predicted risk strata will provide useful information for clinical decision making.





+ statistical tests, e.g. Hosmer-Lemeshow test

Comparison of two models: net reclassification

Net reclassification involves classifying patients in risk categories and **determining how a new model reclassifies patients into various risk categories compared with a previous model**. Risk differences are classified based on the actual outcome patients experienced

Clinical Review & Education

JAMA | Users' Guides to the Medical Literature

Discrimination and Calibration of Clinical Prediction Models
Users' Guides to the Medical Literature

Ana Carolina Alba, MD, PhD; Thomas Agoritsas, MD, PhD; Michael Walsh, MD, PhD; Steven Hanna, PhD; Alfonso Iorio, MD; P. J. Devereaux, MD, PhD; Thomas McGinn, MD; Gordon Guyatt, MD, MSc

Patients with events (n=10 000, 50% of total)

OLD MODEL	NEW MODEL: Low risk	Moderate risk	High risk	Totals, old model
Low risk	2500	1200 a^+	400 b^+	4100
Moderate risk	300 d^+	2500	600 c^+	3400
High risk	50 e^+	350 f^+	2100	2500
Totals, new model	2850	4050	3100	10 000

Patients without events (n=10 000, 50% of total)

OLD MODEL	NEW MODEL: Low risk	Moderate risk	High risk	Totals, old model
Low risk	4000	300 a^-	100 b^-	4400
Moderate risk	100 d^-	4000	150 c^-	4250
High risk	50 e^-	100 f^-	1200	1350
Totals, new model	4150	4400	1450	10 000

	Patients with events, No.	Patients without events, No.		
Correct reclassification	2200	250	Additive NRI	12
Incorrect reclassification	700	550	Absolute NRI	6%
Net reclassification	1500	-300		

Comparison of two models: net reclassification

Net reclassification involves classifying patients in risk categories and **determining how a new model reclassifies patients into various risk categories compared with a previous model**. Risk differences are classified based on the actual outcome patients experienced

Clinical Review & Education

JAMA | Users' Guides to the Medical Literature

Discrimination and Calibration of Clinical Prediction Models Users' Guides to the Medical Literature

Ana Carolina Alba, MD, PhD; Thomas Agoritsas, MD, PhD; Michael Walsh, MD, PhD; Steven Hanna, PhD; Alfonso Iorio, MD; P. J. Devereaux, MD, PhD; Thomas McGinn, MD; Gordon Guyatt, MD, MSc

	Patients with events	Patients without events
Correct reclassification	$x^+ = a^+ + b^+ + c^+$	$x^- = d^- + e^- + f^-$
Incorrect reclassification	$y^+ = d^+ + e^+ + f^+$	$y^- = a^- + b^- + c^-$
Net reclassification	$z^+ = x^+ - y^+$	$z^- = x^- - y^-$

- No reclassification
- Correct reclassification
- Incorrect reclassification

$$\text{Additive NRI} = \left(\frac{z^+}{\text{Total No. of patients with event}} \times 100\right) + \left(\frac{z^-}{\text{Total No. of patients without event}} \times 100\right)$$

$$\text{Absolute NRI} = \frac{(z^+ + z^-)}{\text{Total No. of patients}} \times 100$$