Machine learning: basi e sue applicazioni

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?

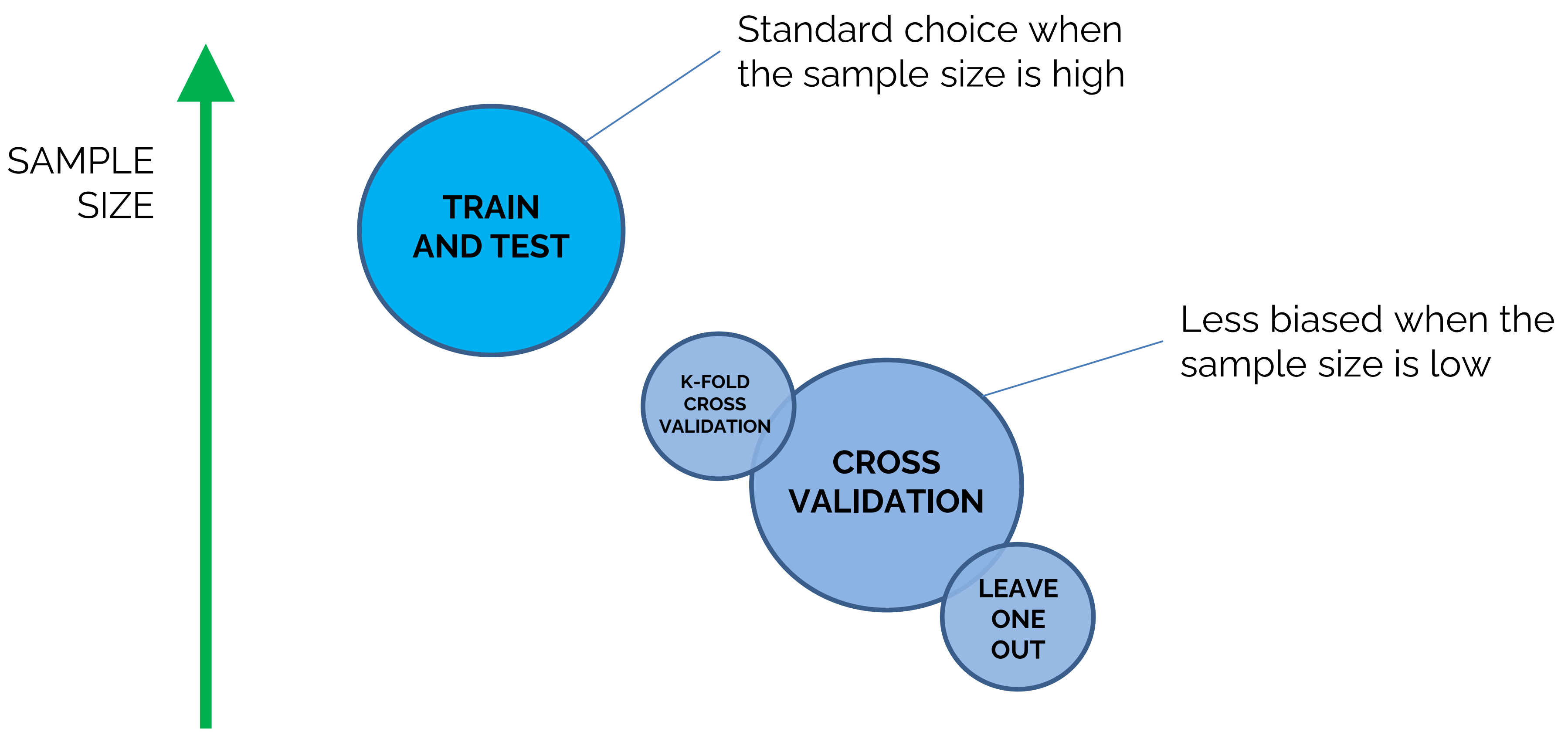
For example, if parameter selection, training of the predictive model and validation are performed using the same dataset, the generated classifier will show limited generalization ability when classifying unseen samples

LOW
TRAINING ERROR

HIGH
TESTING ERROR
(low generalization ability)

OVERFITTING

Which validation approach?

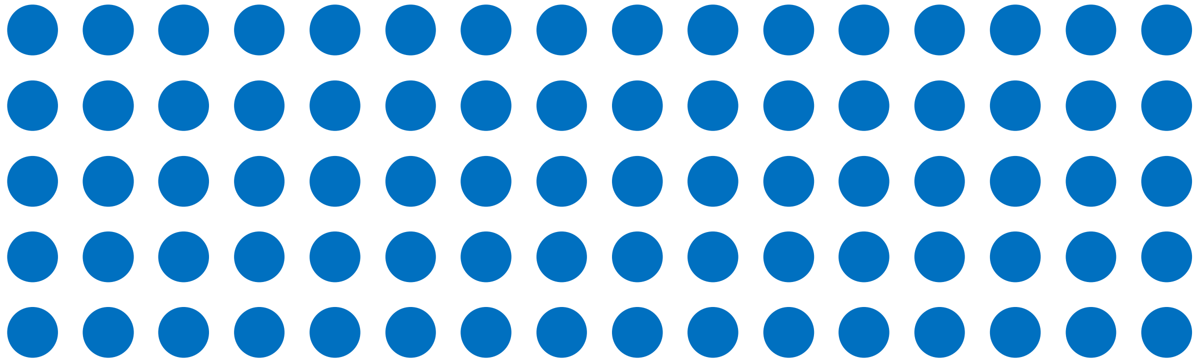


Train-and-test

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:

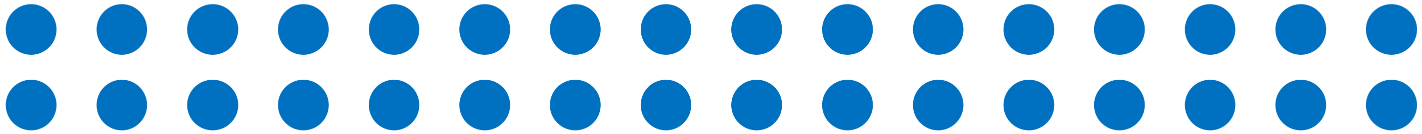
- Results could be related to the particular choice of the partition subsets

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

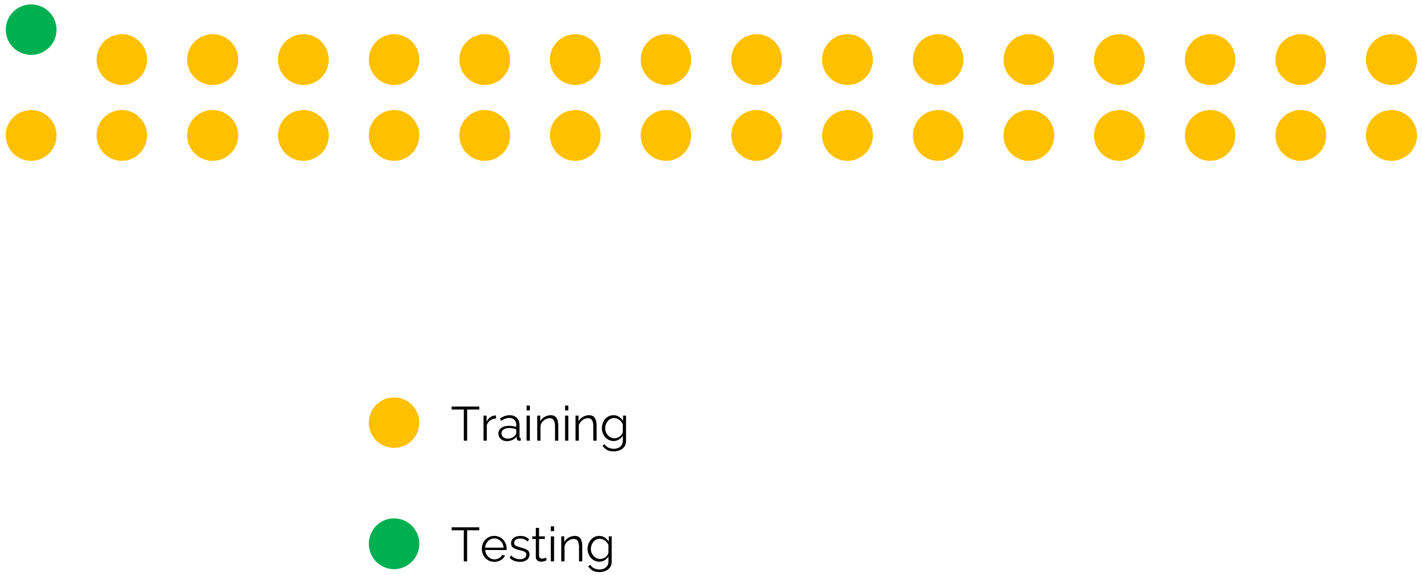
Leave-one-out cross validation



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

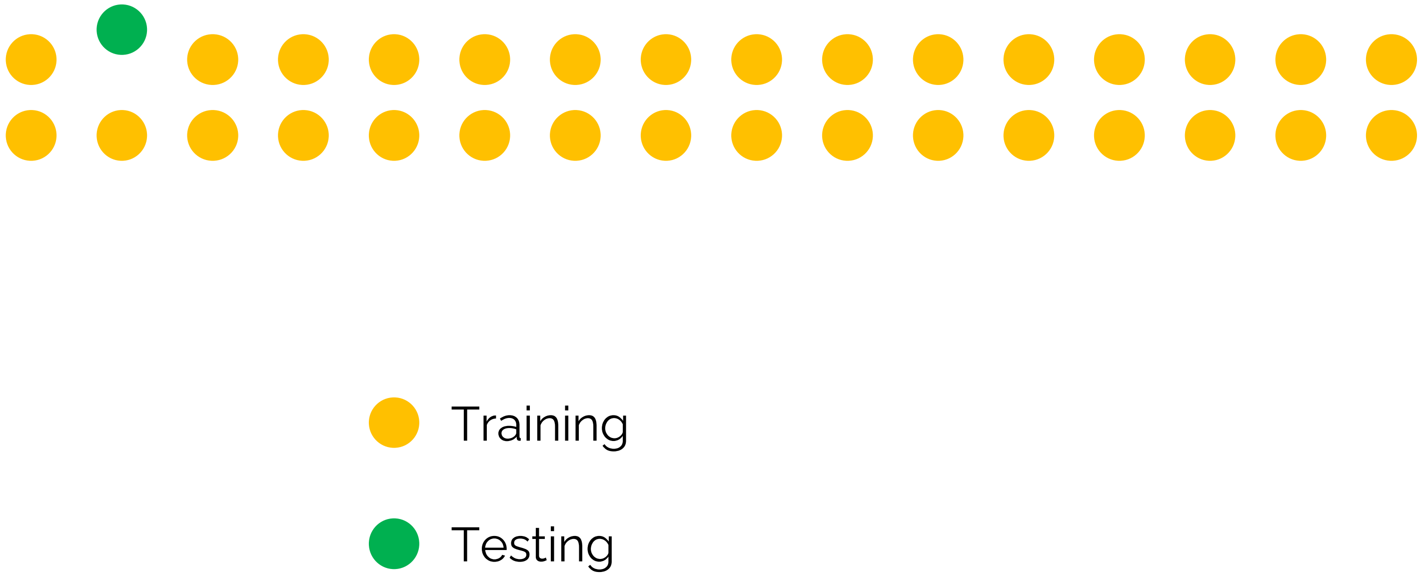
TESTING is performed using the remaining sample (n being the total number of samples in the original dataset)

Leave-one-out cross validation



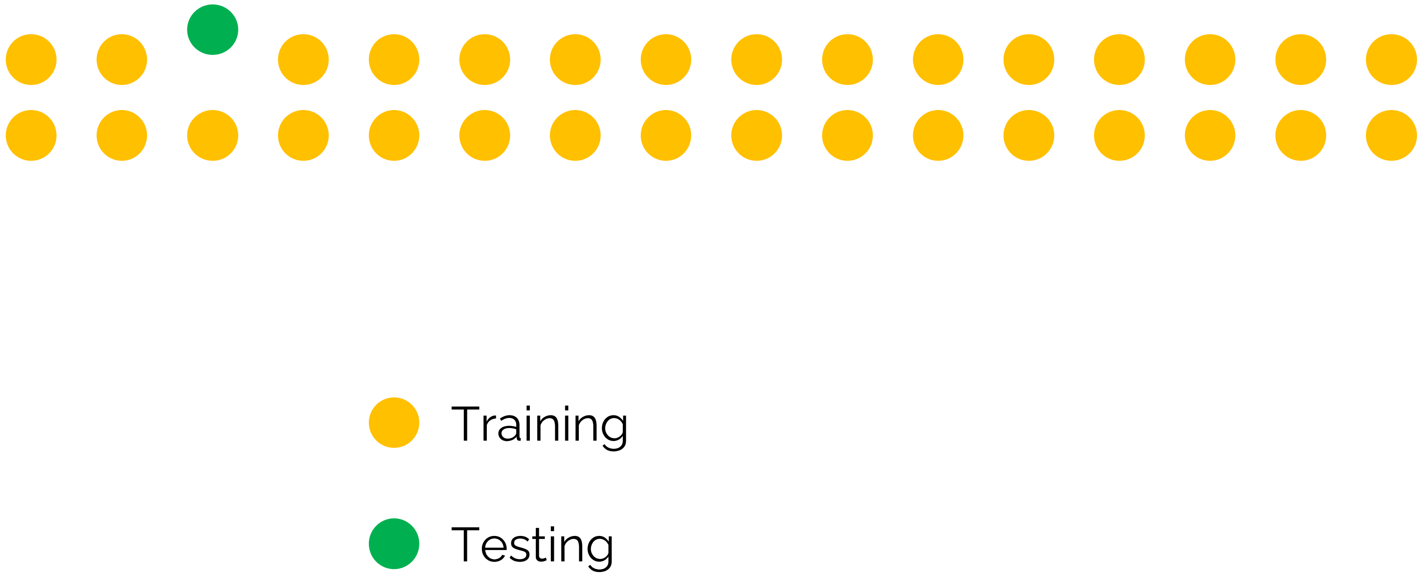
The procedure is then repeated n times, until all samples are used once for validation.

Leave-one-out cross validation



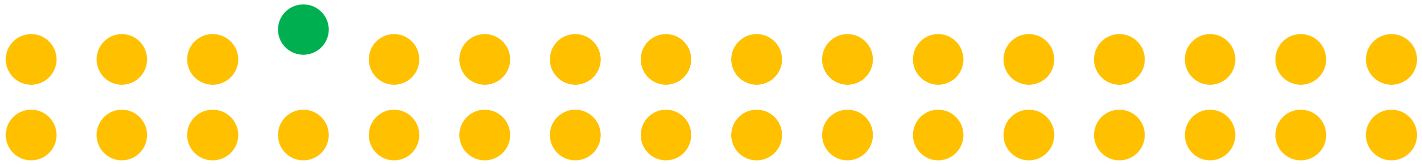
The procedure is then repeated n times, until all samples are used once for validation.

Leave-one-out cross validation



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

Leave-one-out cross validation



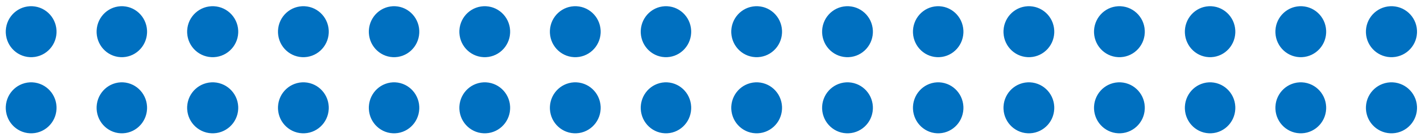
Training

Testing

and so on…

Cross validation

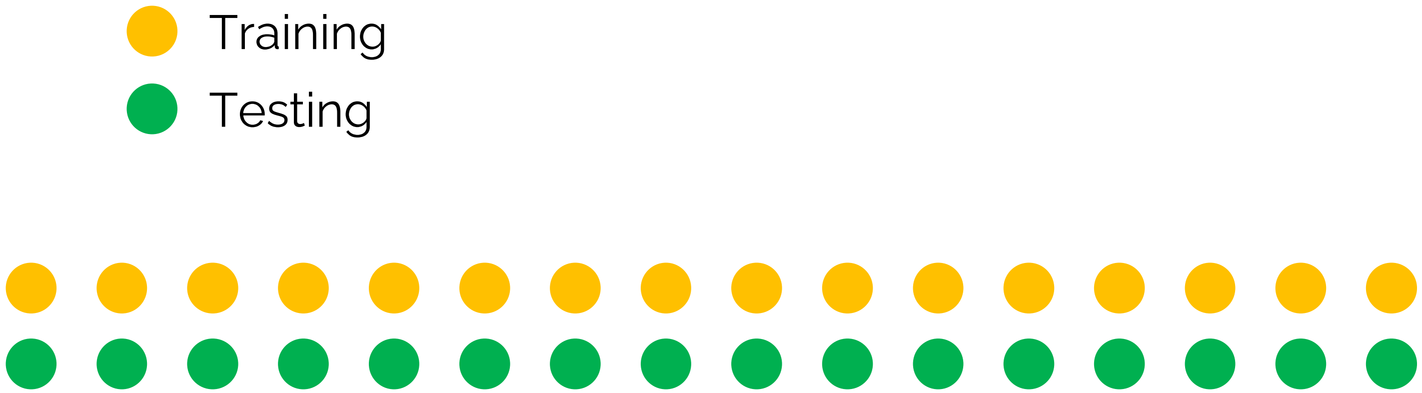
Quantification of the discriminatory power of a predictive model even if the size of the dataset is small



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

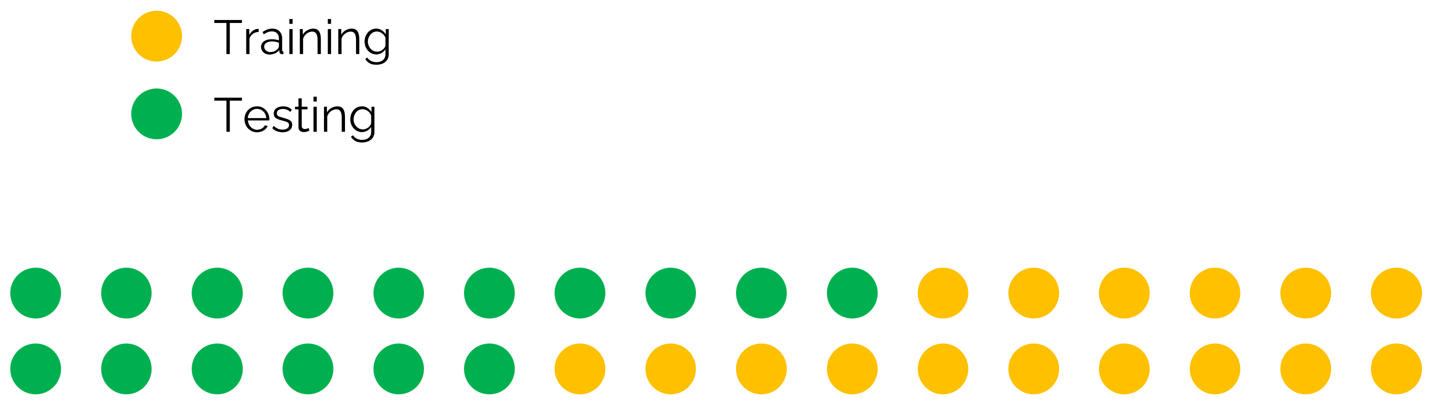
The TRAINING set is used to train the classifier
The TESTING set is used to validate the generated predictive model

Cross validation



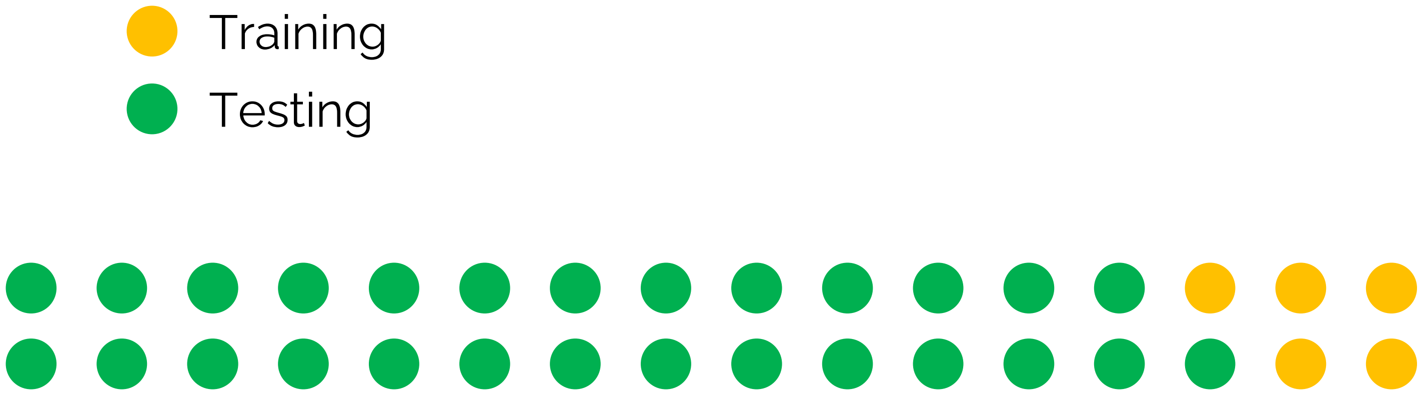
By using different partitions of the original dataset, multiple rounds of CV can be performed, which can aid reducing the variability of the partitioned subsets.

Cross validation



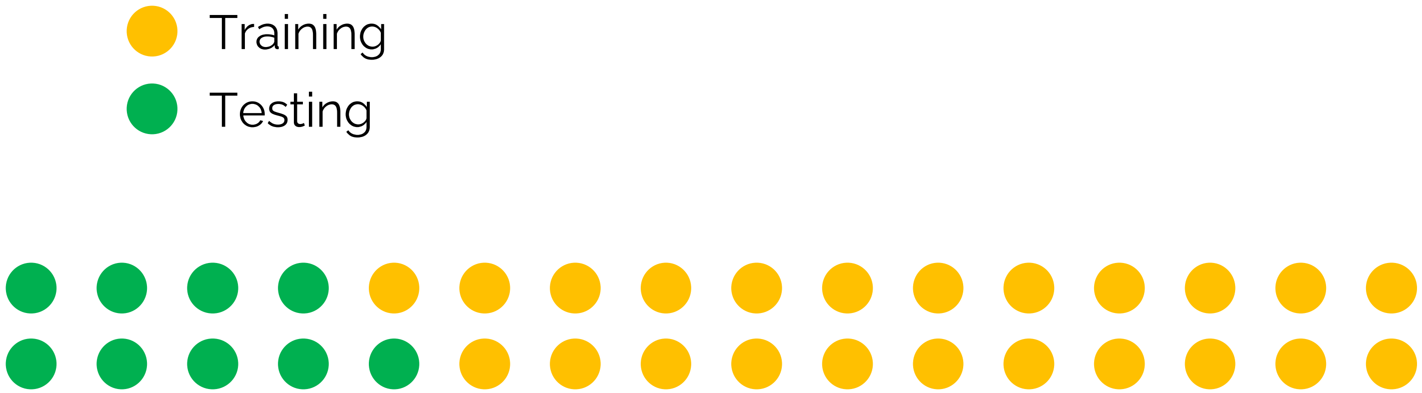
By using different partitions of the original dataset, multiple rounds of CV can be performed, which can aid reducing the variability of the partitioned subsets.

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.

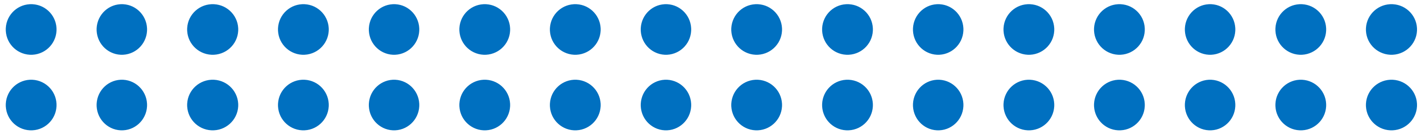
Cross validation



and so on…

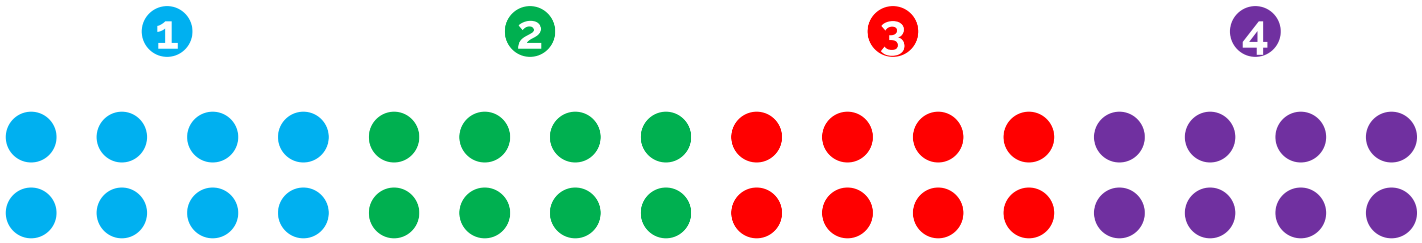
Results obtained from multiple rounds can be averaged in order to obtain a quantification of the performance of the classifier.

K-fold cross validation



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

K-fold cross validation

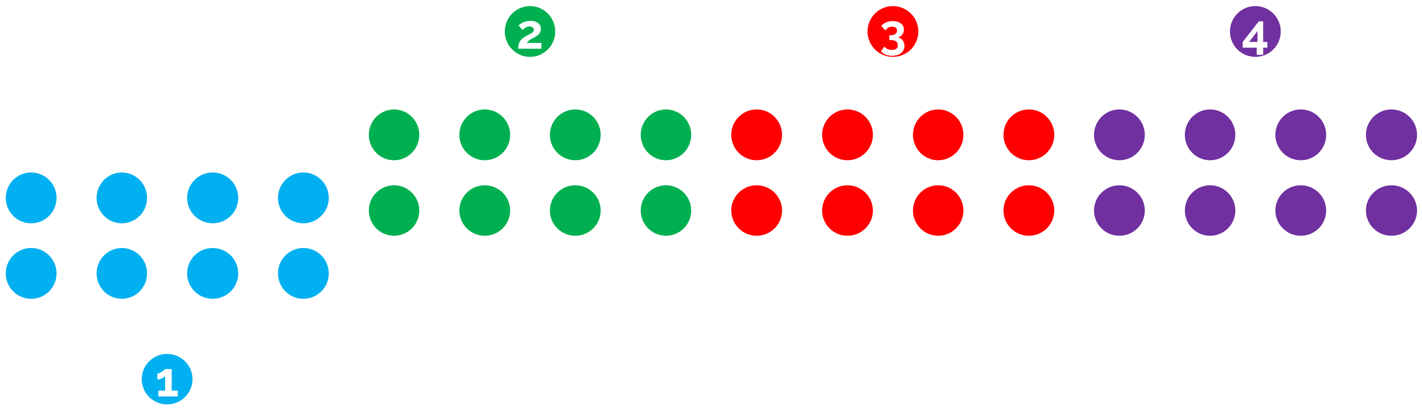


The original dataset is randomly partitioned into k subsets of equal size

TRAINING of the classifier is performed using k-1 subsets

TESTING is performed using the remaining subset

K-fold cross validation

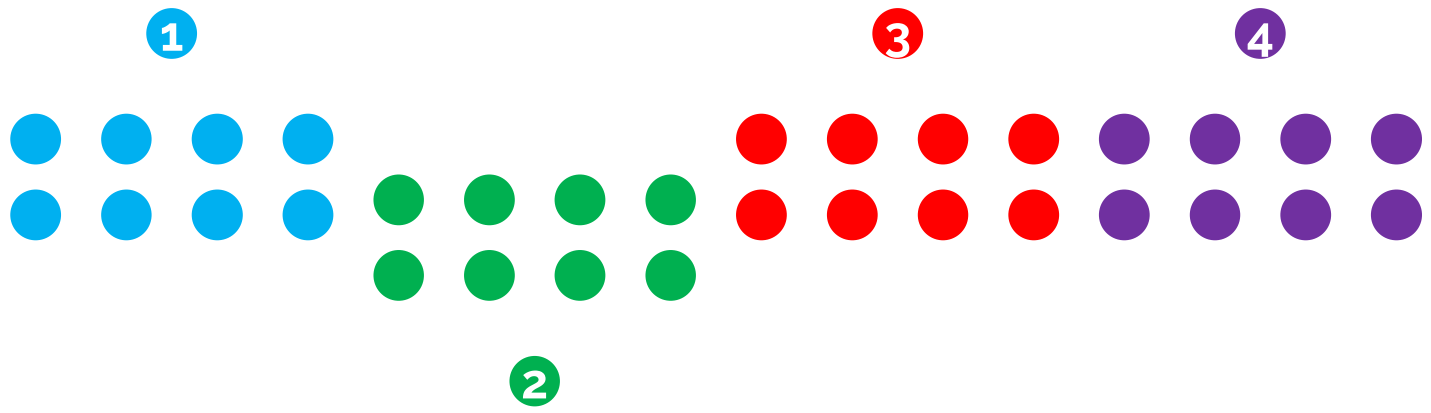


2 3 4 Training

1 Testing

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

K-fold cross validation

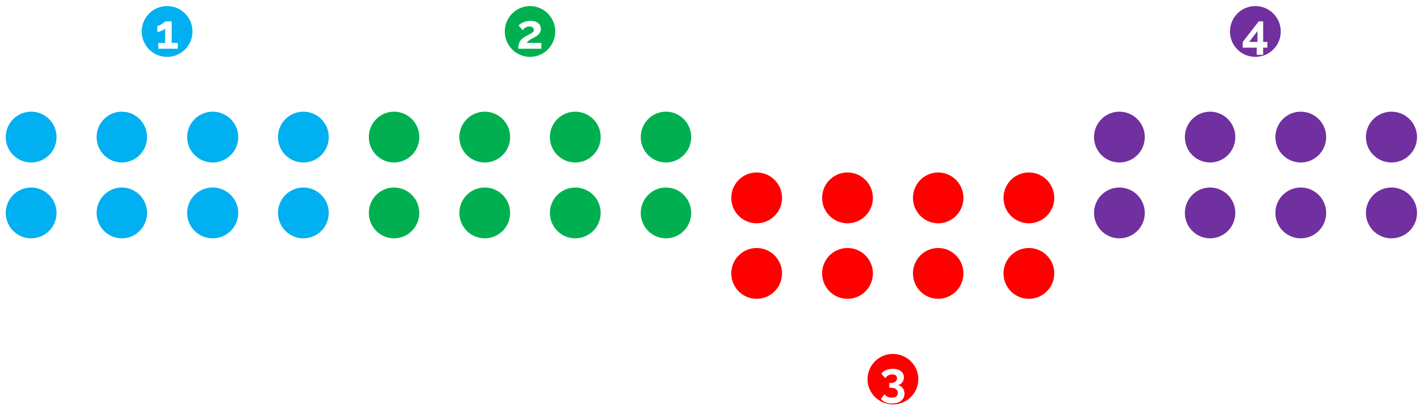


1 3 4 Training

2 Testing

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

K-fold cross validation

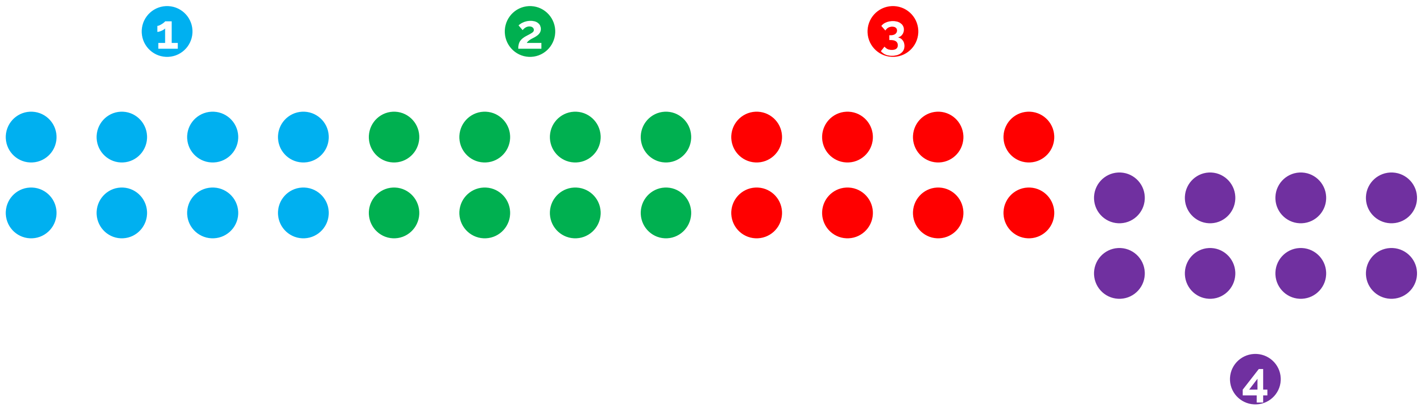


1 2 4 Training

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

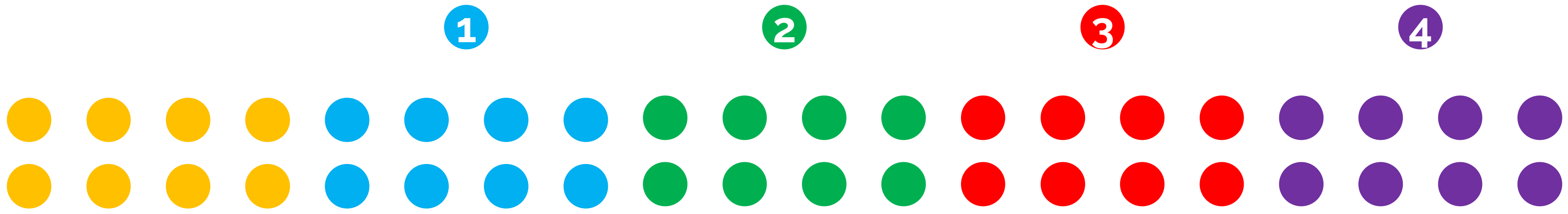
The procedure is then repeated k times, until all subsets are used once as testing set.

K-fold cross validation

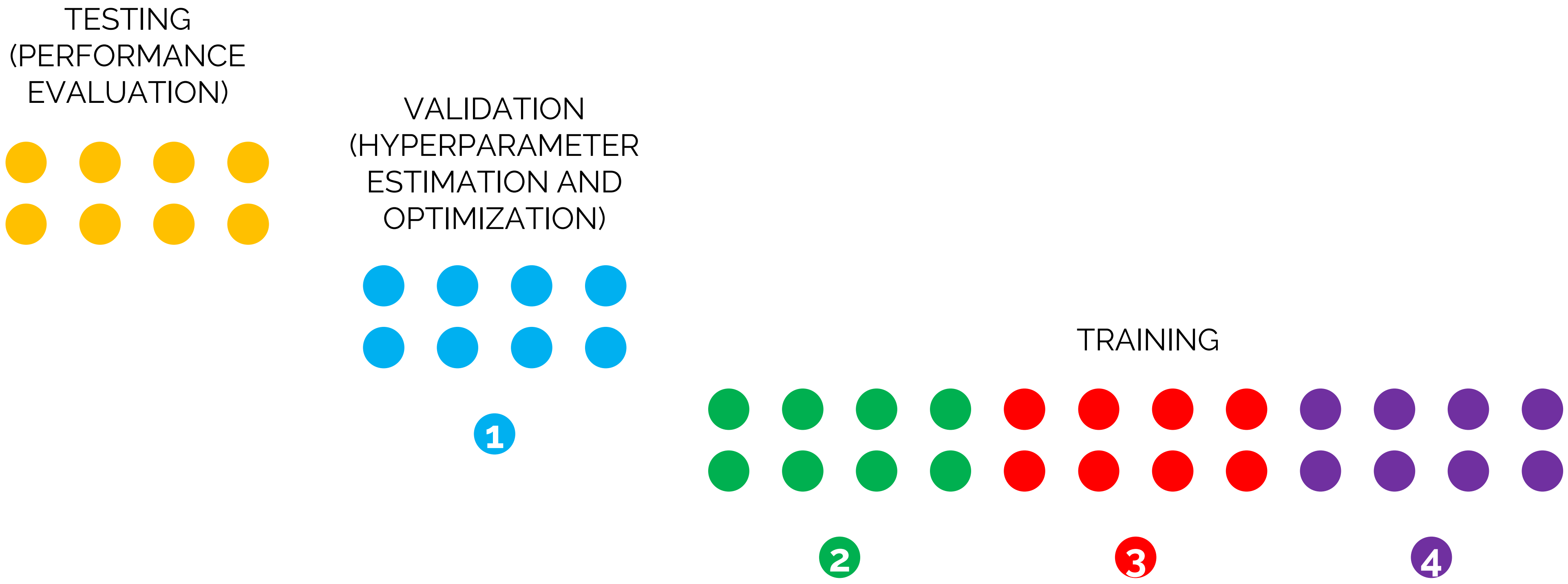
Advantages:

- 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

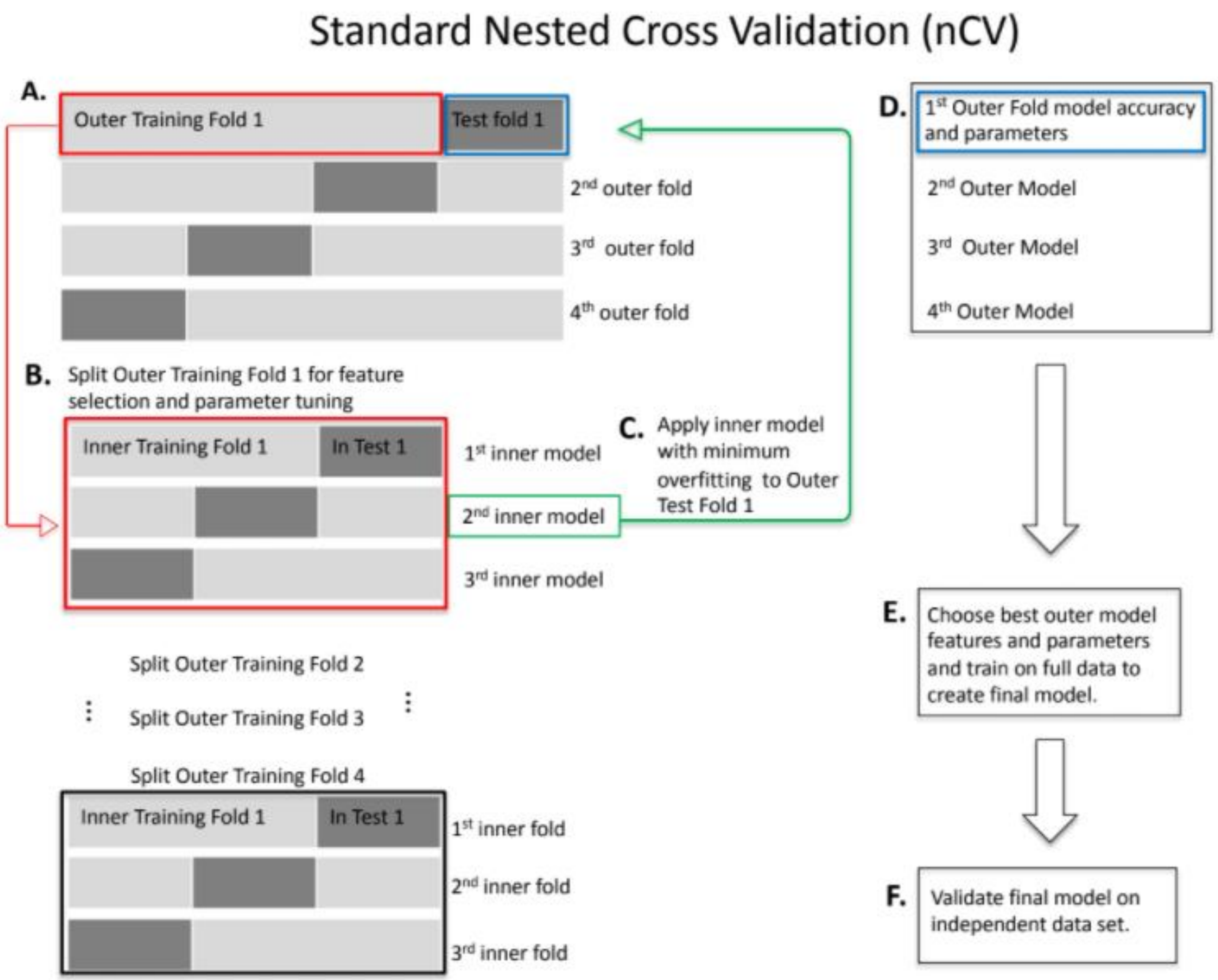


Fig. 1. Standard nested Cross-Validation (nCV). A. Split the data into outer folds of training and testing data pairs (4 outer folds in this illustration). Then do the following for each outer training fold (illustration starting with Outer Training Fold 1 (red box, A)). **B.** Split outer training fold 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

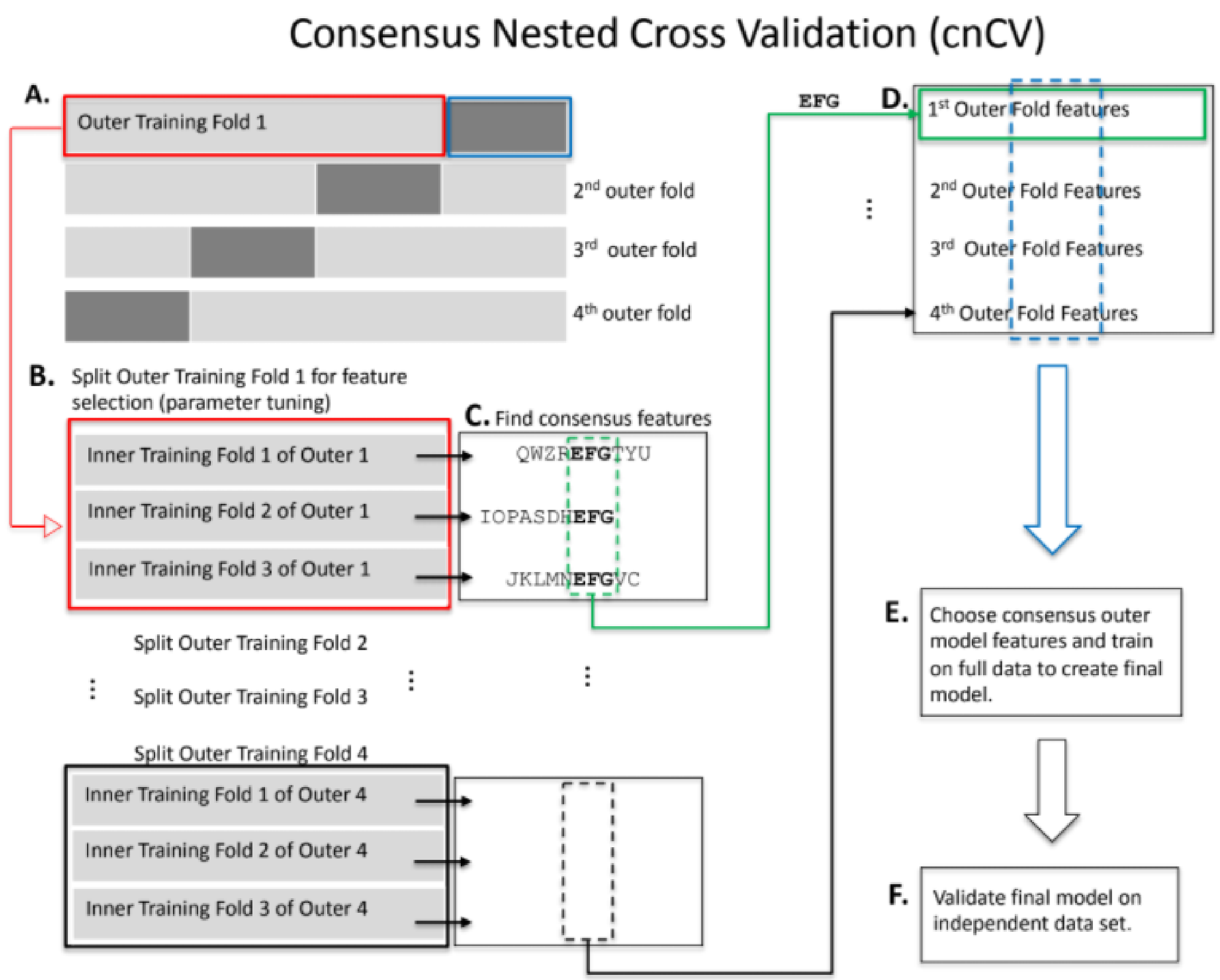


Fig. 2. Consensus Nested Cross-Validation (cnCV). A. Split the data into outer folds (4 outer folds in this illustration). Then do the following for each outer training fold (illustration starting with Outer Training Fold 1 (red box, A)). **B.** Split outer training fold into inner folds for feature 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).

Sensitivity =

correctly classified samples in the positive class / # 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).

Specificity =

correctly classified samples in the negative class / # 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).

Precision =

correctly classified samples in the positive class / # 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 =

correctly classified samples in the negative class / # 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

TPR = # correctly classified samples in the positive class / # positive samples

TNR = # correctly classified samples in the negative class / # negative samples

FPR = # incorrectly classified samples in the positive class / # negative samples

1 – TNR =
1 - SPECIFICITY

FNR = # incorrectly classified samples in the negative class / # positive samples

1 – TPR =
1 - SENSITIVITY

Sensitivity = # correctly classified samples in the positive class / # positive samples

Specificity = # correctly classified samples in the negative class / # negative samples

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

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

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

NPV = # correctly classified samples in the negative class / # predicted-as-negative samples

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)

Star	95/100
Other	75/100

	Star PREDICTED	Other PREDICTED
Star ACTUAL	95	5
Other ACTUAL	25	75

Confusion matrix

Star	95/100
Other	75/100

	Star PREDICTED	Other PREDICTED
Star ACTUAL	95	5
Other ACTUAL	25	75

1. Accuracy
   (95+75) / (95+25+5+75) = 85%

1. Sensitivity and Specificity
   95 / (95+5) = 95%
   75/ (25+75) = 75%

1. Precision and Recall
   95 / (95+25) = 79%
   95 / (95+5) = 95%

1. Positive and Negative Predictive Value
   95 / (95+25) = 79%
   75 / (75+5) = 94%

1. False/True Positive Rates
   5 / (25+75) = 5%
   95 / (95+5) = 95%

1. False/True Negative Rates
   25 / (95+5) = 25%
   75 / (25+75) = 75%

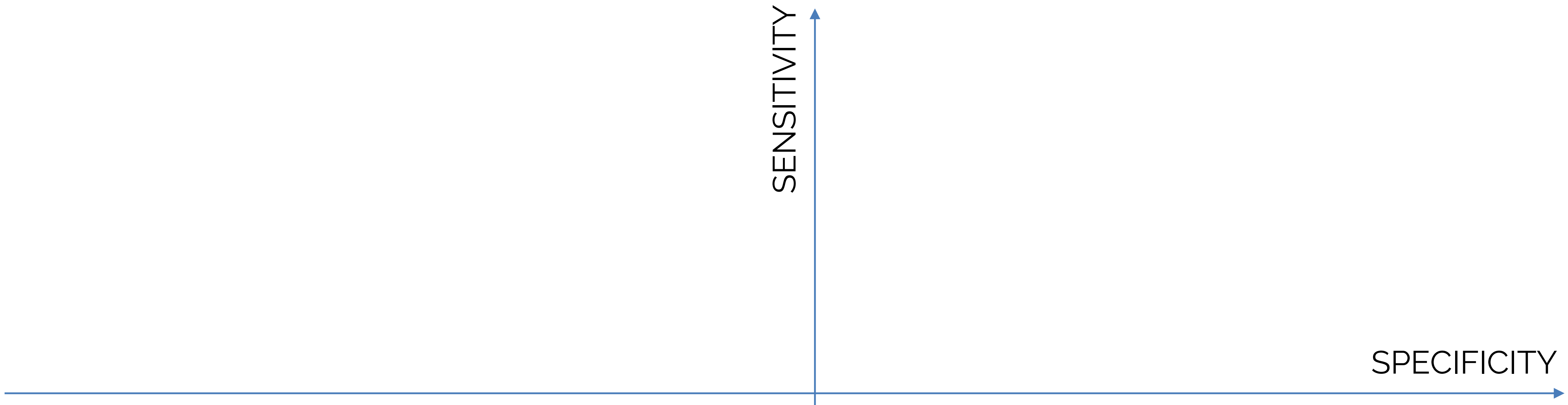
Confusion matrix

Star	**95/100**
Planet	**55/100**
Galaxy	**90/100**
Other	**75/100**

CLASS-X **ACCURACY**

	Star PREDICTED	**Planet** PREDICTED	**Galaxy** PREDICTED	**Other** PREDICTED	
Star ACTUAL	95	0	1	4	95%
Planet ACTUAL	5	55	5	35	55%
Galaxy ACTUAL	2	1	90	7	90%
Other ACTUAL	5	15	5	75	75%

Confusion matrix



	Star PREDICTED	Other PREDICTED
Star ACTUAL	95	5
Other ACTUAL	8	2

	Star PREDICTED	Other PREDICTED
Star ACTUAL	95	5
Other ACTUAL	25	75

	Star PREDICTED	Other PREDICTED
Star ACTUAL	25	75
Other ACTUAL	65	35

	Star PREDICTED	Other PREDICTED
Star ACTUAL	95	5
Other ACTUAL	75	25

	Star PREDICTED	Other PREDICTED
Star ACTUAL	25	75
Other ACTUAL	15	85

	Star PREDICTED	Other PREDICTED
Star ACTUAL	5	10
Other ACTUAL	15	85

Confusion matrix

SENSITIVITY

Accuracy

85%	Star PREDICTED	Other PREDICTED
Star ACTUAL	95	5
Other ACTUAL	25	75

SPECIFICITY

	Star PREDICTED	Other PREDICTED
Star ACTUAL	95	5
Other ACTUAL	8	2

	Star PREDICTED	Other PREDICTED
Star ACTUAL	25	75
Other ACTUAL	65	35

	Star PREDICTED	Other PREDICTED
Star ACTUAL	95	5
Other ACTUAL	75	25

	Star PREDICTED	Other PREDICTED
Star ACTUAL	25	75
Other ACTUAL	15	85

	Star PREDICTED	Other PREDICTED
Star ACTUAL	5	10
Other ACTUAL	15	85

Confusion matrix

SENSITIVITY

Accuracy

85%	Star PREDICTED	Other PREDICTED
Star ACTUAL	95	5
Other ACTUAL	25	75

SPECIFICITY

	Star PREDICTED	Other PREDICTED
Star ACTUAL	95	5
Other ACTUAL	8	2

30%	Star PREDICTED	Other PREDICTED
Star ACTUAL	25	75
Other ACTUAL	65	35

	Star PREDICTED	Other PREDICTED
Star ACTUAL	95	5
Other ACTUAL	75	25

	Star PREDICTED	Other PREDICTED
Star ACTUAL	25	75
Other ACTUAL	15	85

	Star PREDICTED	Other PREDICTED
Star ACTUAL	5	10
Other ACTUAL	15	85

Confusion matrix

SENSITIVITY

Accuracy

85%	**Star** PREDICTED	**Other** PREDICTED
Star ACTUAL	95	5
Other ACTUAL	25	75

60%	**Star** PREDICTED	**Other** PREDICTED
Star ACTUAL	95	5
Other ACTUAL	75	25

SPECIFICITY

	Star PREDICTED	**Other** PREDICTED
Star ACTUAL	95	5
Other ACTUAL	8	2

30%	**Star** PREDICTED	**Other** PREDICTED
Star ACTUAL	25	75
Other ACTUAL	65	35

	Star PREDICTED	**Other** PREDICTED
Star ACTUAL	25	75
Other ACTUAL	15	85

	Star PREDICTED	**Other** PREDICTED
Star ACTUAL	5	10
Other ACTUAL	15	85

Confusion matrix

SENSITIVITY

60%	Star PREDICTED	Other PREDICTED
Star ACTUAL	95	5
Other ACTUAL	75	25

Accuracy

85%	Star PREDICTED	Other PREDICTED
Star ACTUAL	95	5
Other ACTUAL	25	75

SPECIFICITY

30%	Star PREDICTED	Other PREDICTED
Star ACTUAL	25	75
Other ACTUAL	65	35

55%	Star PREDICTED	Other PREDICTED
Star ACTUAL	25	75
Other ACTUAL	15	85

	Star PREDICTED	Other PREDICTED
Star ACTUAL	95	5
Other ACTUAL	8	2

	Star PREDICTED	Other PREDICTED
Star ACTUAL	5	10
Other ACTUAL	15	85

Confusion matrix

SENSITIVITY

Accuracy

85%	Star PREDICTED	Other PREDICTED
Star ACTUAL	95	5
Other ACTUAL	25	75

60%	Star PREDICTED	Other PREDICTED
Star ACTUAL	95	5
Other ACTUAL	75	25

SPECIFICITY

55%	Star PREDICTED	Other PREDICTED
Star ACTUAL	25	75
Other ACTUAL	15	85

	Star PREDICTED	Other PREDICTED
Star ACTUAL	95	5
Other ACTUAL	8	2

30%	Star PREDICTED	Other PREDICTED
Star ACTUAL	25	75
Other ACTUAL	65	35

78%	Star PREDICTED	Other PREDICTED
Star ACTUAL	5	10
Other ACTUAL	15	85

Confusion matrix

SENSITIVITY

Accuracy

88%	Star PREDICTED	Other PREDICTED
Star ACTUAL	95	5
Other ACTUAL	8	2

60%	Star PREDICTED	Other PREDICTED
Star ACTUAL	95	5
Other ACTUAL	75	25

85%	Star PREDICTED	Other PREDICTED
Star ACTUAL	95	5
Other ACTUAL	25	75

SPECIFICITY

55%	Star PREDICTED	Other PREDICTED
Star ACTUAL	25	75
Other ACTUAL	15	85

78%	Star PREDICTED	Other PREDICTED
Star ACTUAL	5	10
Other ACTUAL	15	85

30%	Star PREDICTED	Other PREDICTED
Star ACTUAL	25	75
Other ACTUAL	65	35

Confusion matrix

Accuracy

SENSITIVITY

SPECIFICITY

88%	Star PREDICTED	Other PREDICTED	
Star ACTUAL	95	5	95%
Other ACTUAL	8	2	20%

60%	Star PREDICTED	Other PREDICTED	
Star ACTUAL	95	5	95%
Other ACTUAL	75	25	25%

85%	Star PREDICTED	Other PREDICTED	
Star ACTUAL	95	5	95%
Other ACTUAL	25	75	75%

30%	Star PREDICTED	Other PREDICTED	
Star ACTUAL	25	75	25%
Other ACTUAL	65	35	35%

55%	Star PREDICTED	Other PREDICTED	
Star ACTUAL	25	75	25%
Other ACTUAL	15	85	85%

78%	Star PREDICTED	Other PREDICTED	
Star ACTUAL	5	10	33%
Other 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

Prevalence =

samples in the positive class (actual) / # samples (actual positives + negatives)

Performance metrics for imbalanced datasets

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

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

Which values does it span?

Performance metrics for imbalanced datasets

Dominance is the difference between sensitivity and specificity

$$\boxed{\text{Dominance} = \text{Sensitivity} - \text{Specificity}}$$

Which values does it span?

Confusion matrix

SENSITIVITY

88%	Star PREDICTED	Other PREDICTED	
Star ACTUAL	95	5	95%
Other ACTUAL	8	2	20%

GM 44%
DOM 75%

Accuracy

85%	Star PREDICTED	Other PREDICTED	
Star ACTUAL	95	5	95%
Other ACTUAL	25	75	75%

GM 84%
DOM 20%

60%	Star PREDICTED	Other PREDICTED	
Star ACTUAL	95	5	95%
Other ACTUAL	75	25	25%

GM 49%
DOM 70%

SPECIFICITY

55%	Star PREDICTED	Other PREDICTED	
Star ACTUAL	25	75	25%
Other ACTUAL	15	85	85%

GM 46%
DOM -60%

30%	Star PREDICTED	Other PREDICTED	
Star ACTUAL	25	75	25%
Other ACTUAL	65	35	35%

GM 30%
DOM -10%

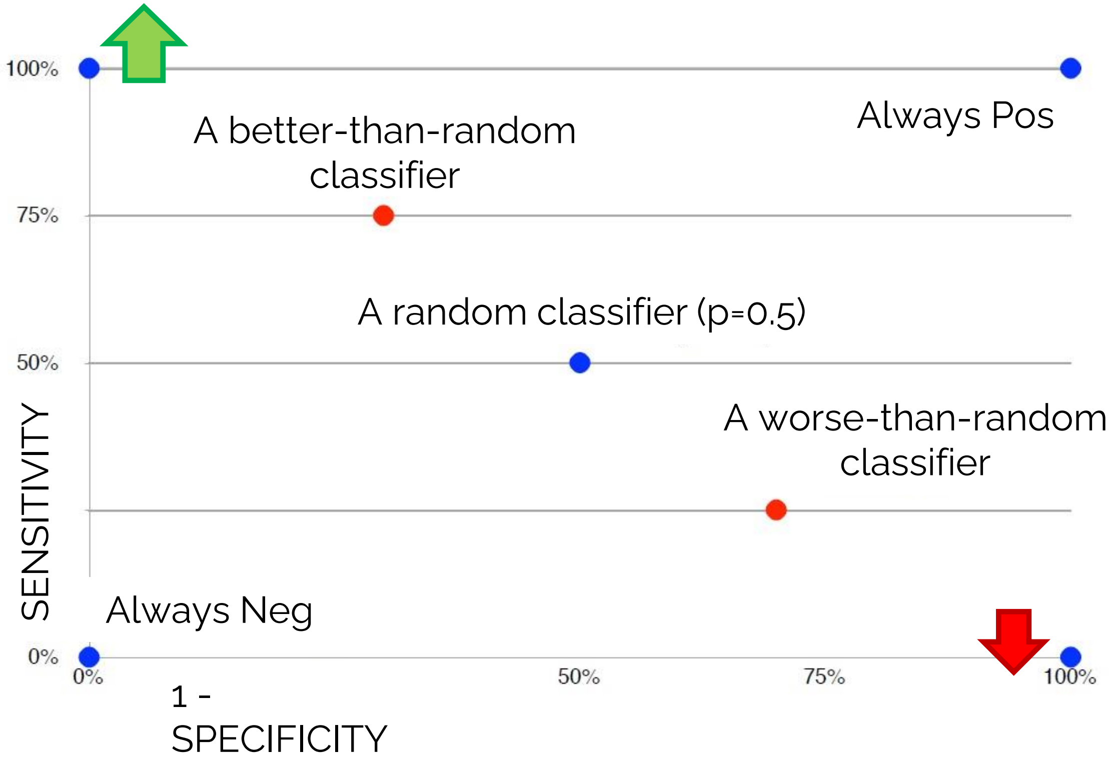
78%	Star PREDICTED	Other PREDICTED	
Star ACTUAL	5	10	33%
Other ACTUAL	15	85	85%

GM 53%
DOM -52%

ROC analysis and Area Under the (ROC) Curve

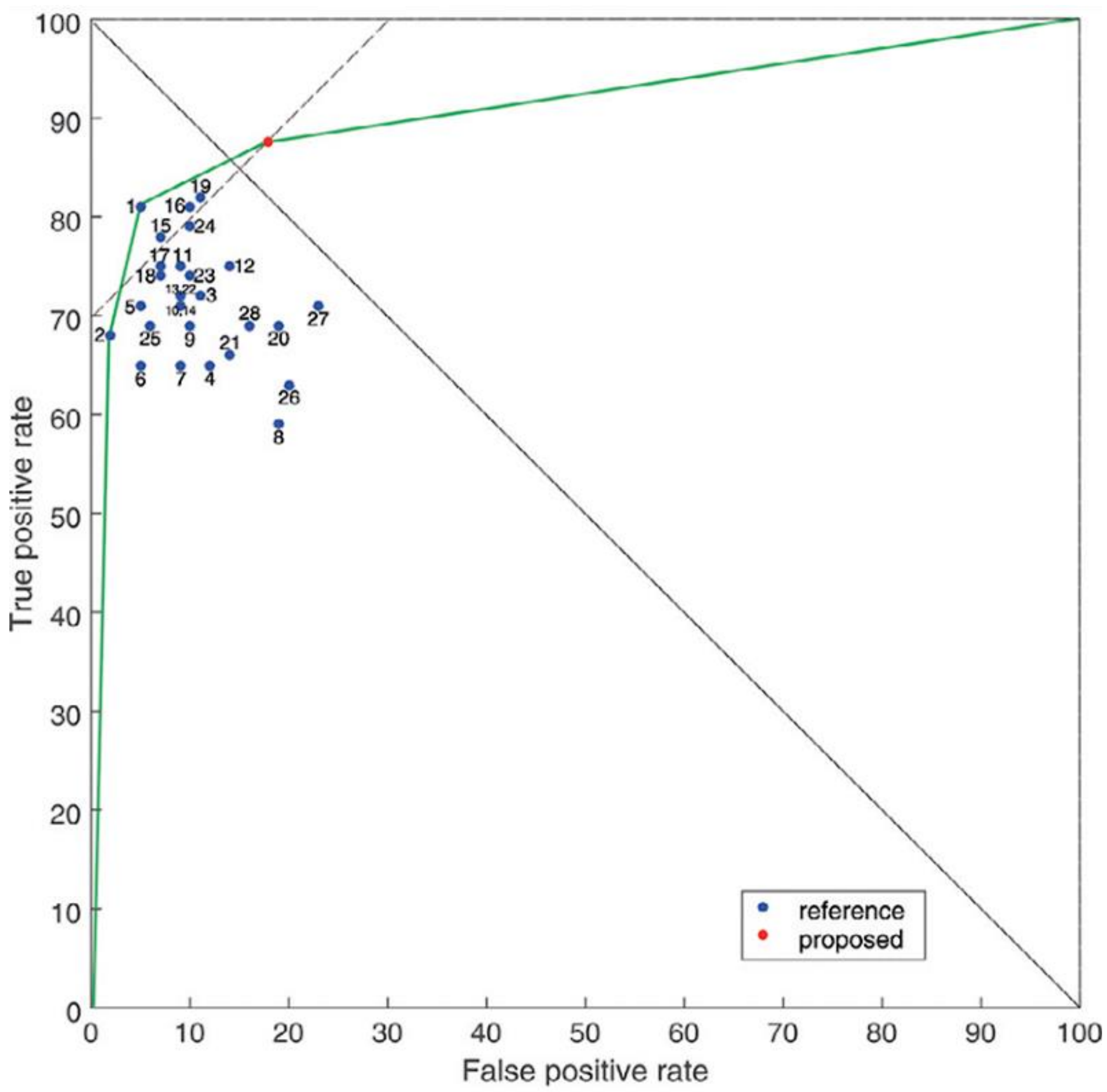
Another important metric in classification problems is given by the study of the Receiver Operating Characteristic (ROC) curve.

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.

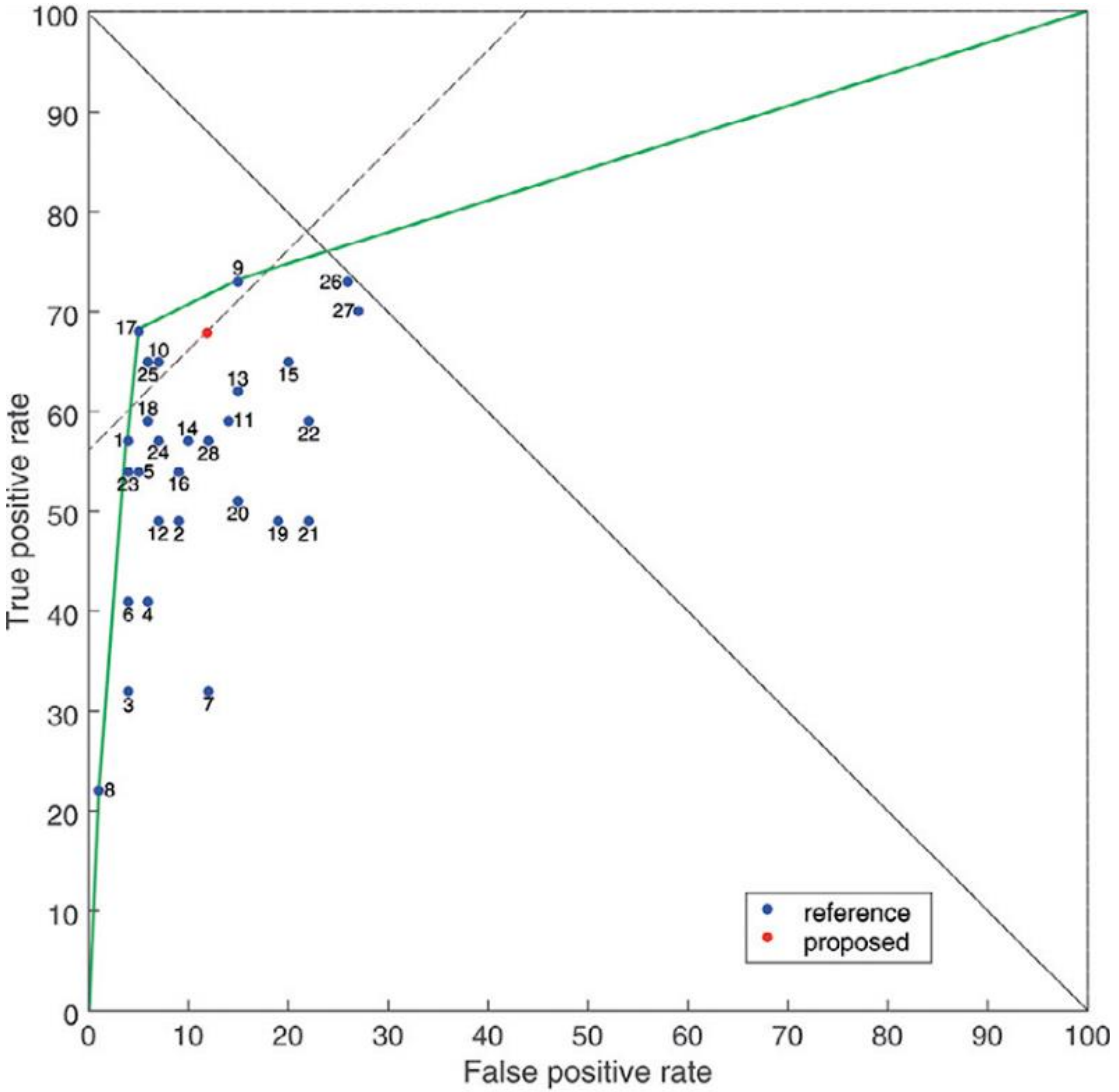


The Area Under the ROC Curve (AUC) gives a quantification of the classifier performance, with a higher statistical consistency than accuracy.

ROC analysis and Area Under the (ROC) Curve

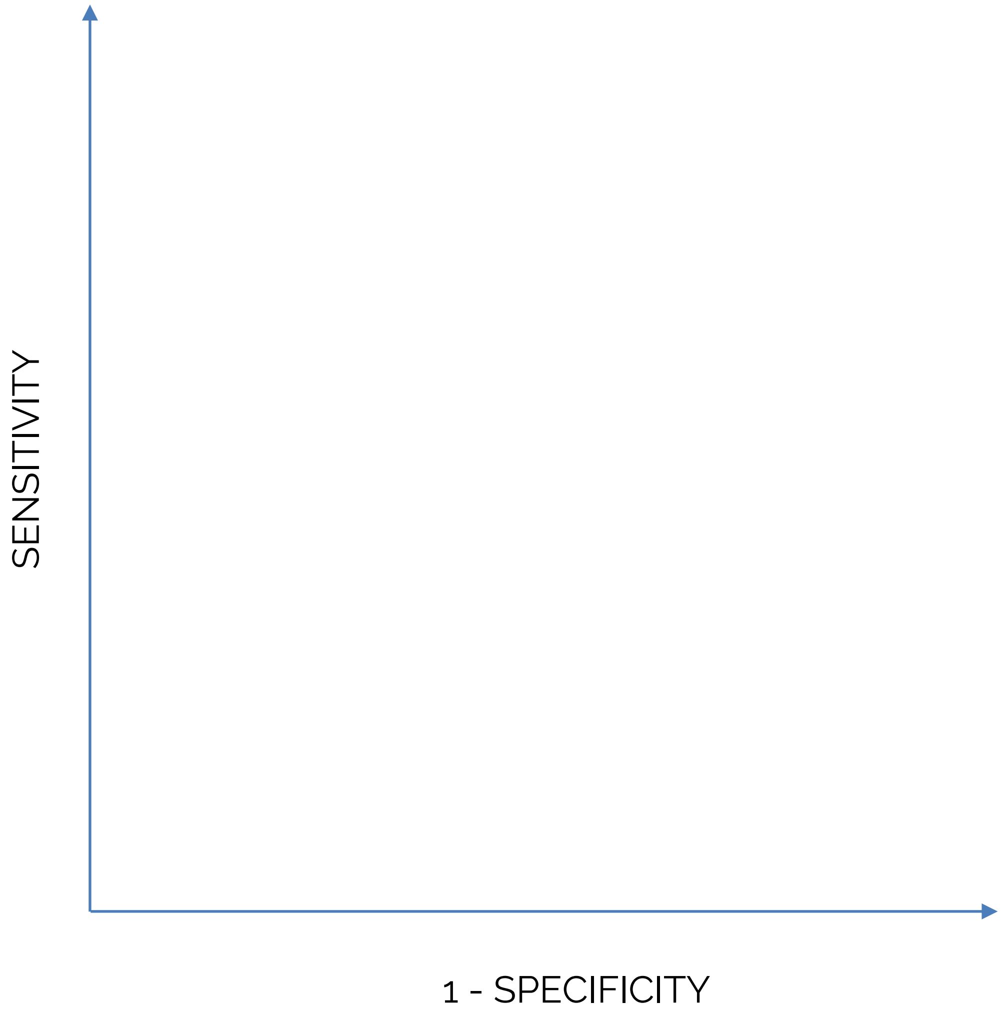


ROC analysis and Area Under the (ROC) Curve

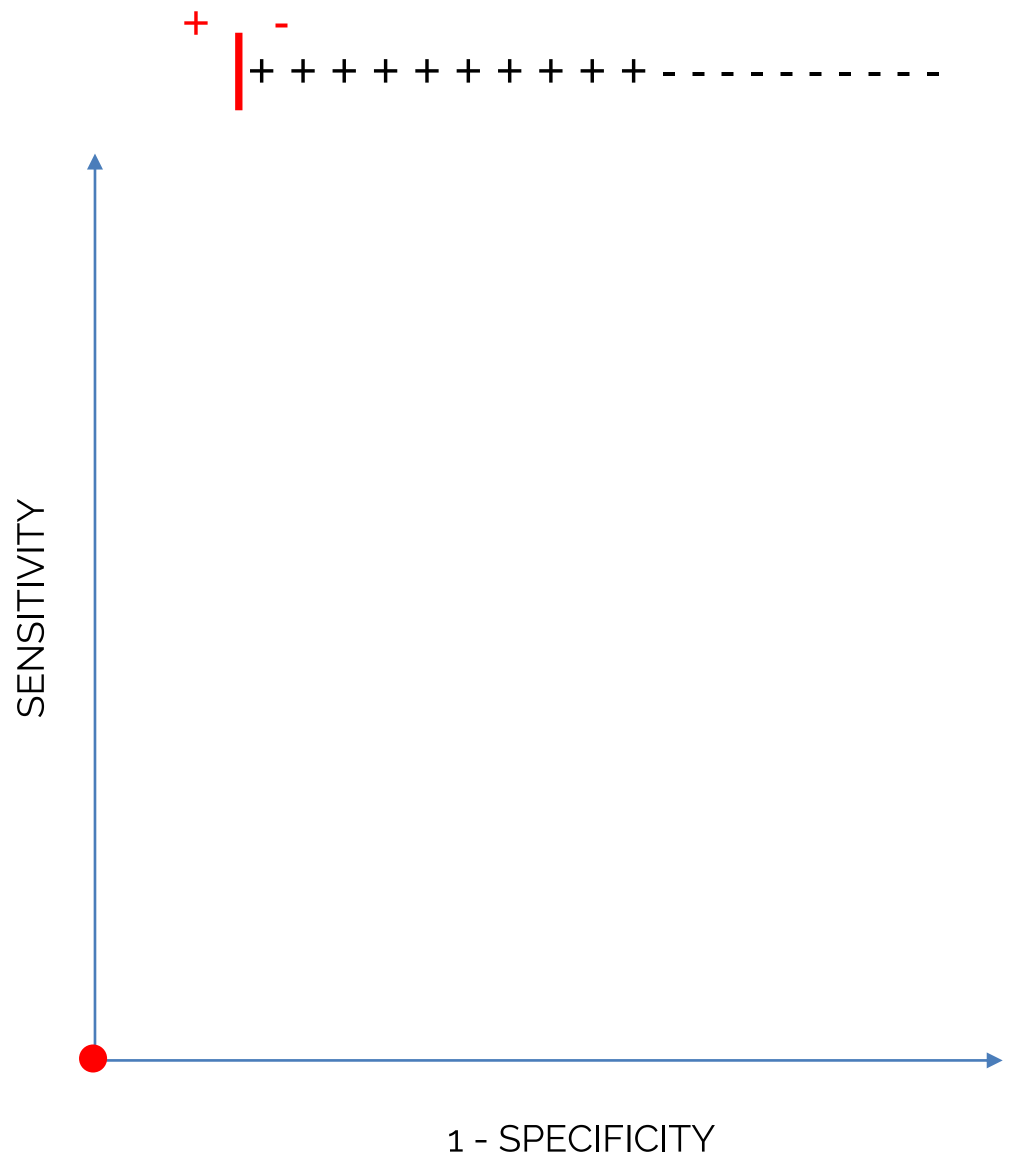


ROC analysis and Area Under the (ROC) Curve

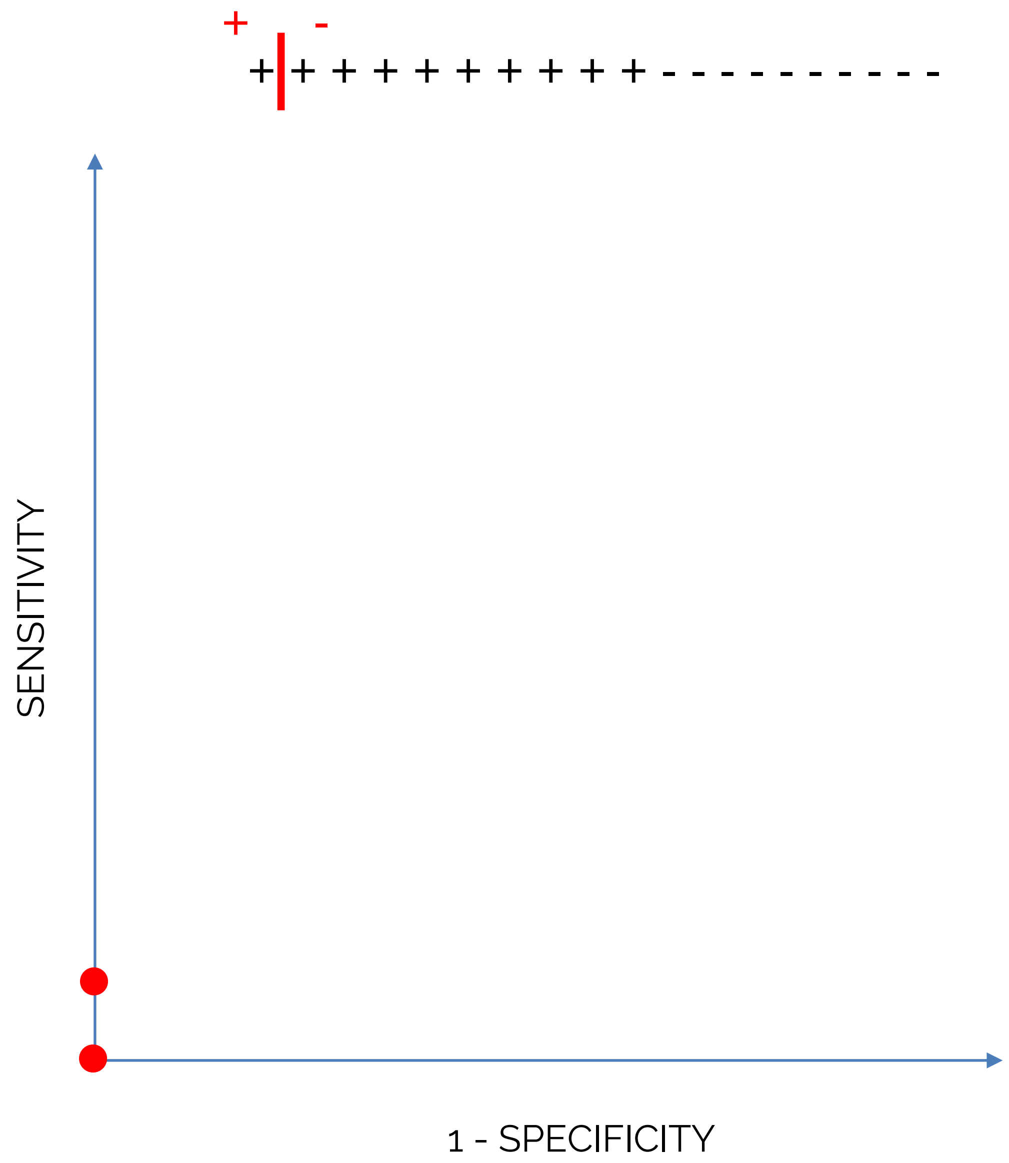
\+ + + + + + + + + + - - - - - - - - - -



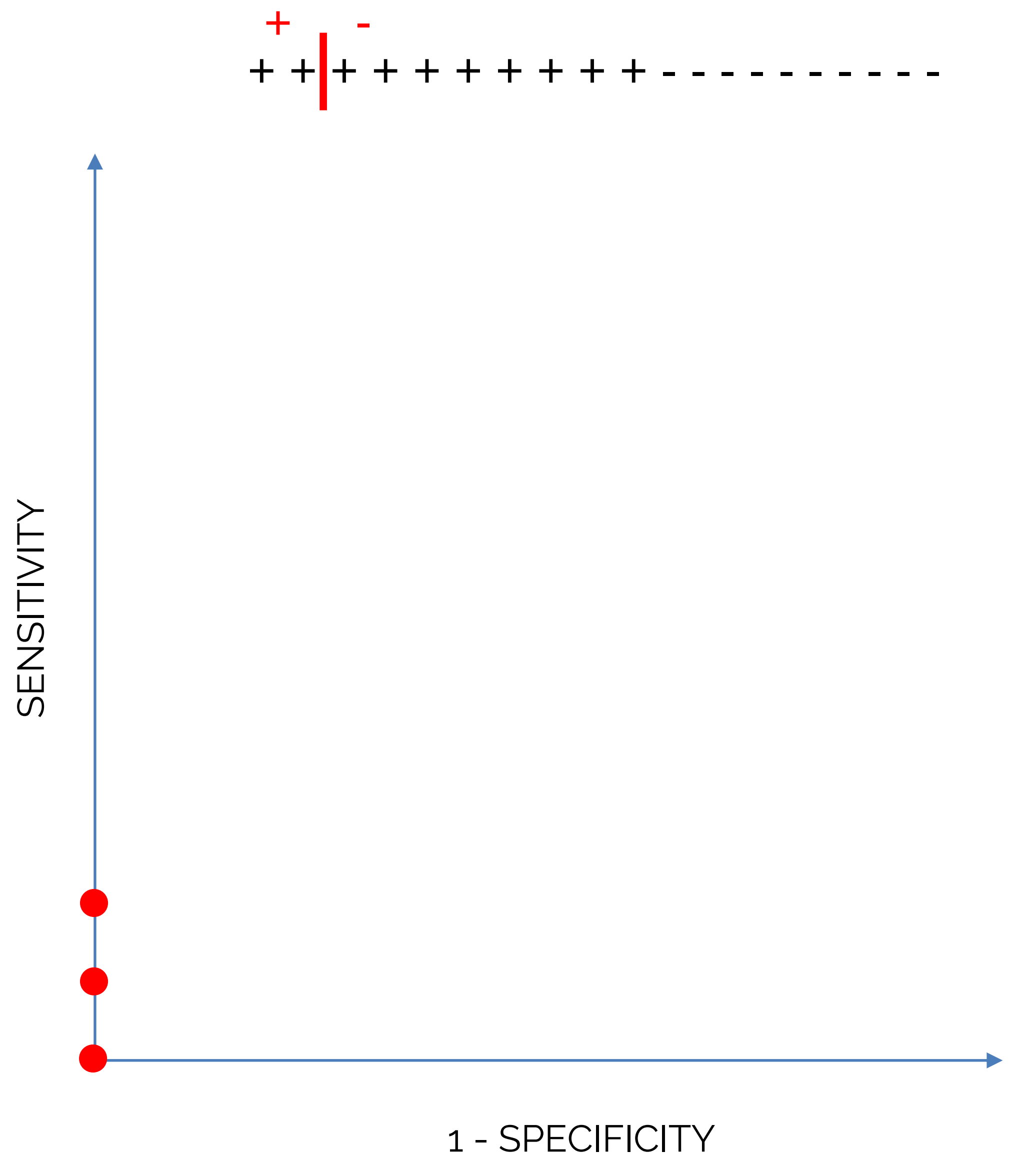
ROC analysis and Area Under the (ROC) Curve



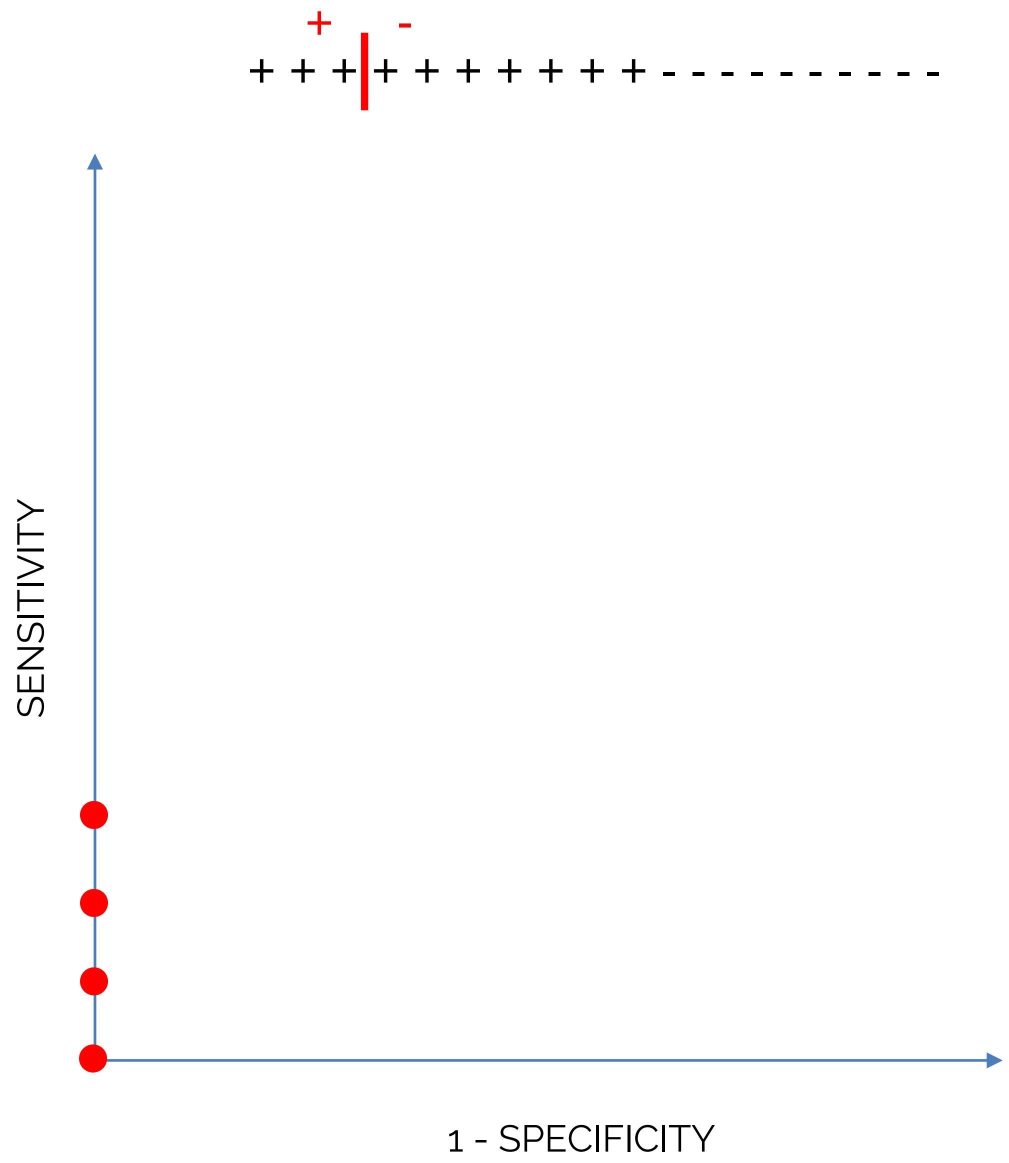
ROC analysis and Area Under the (ROC) Curve



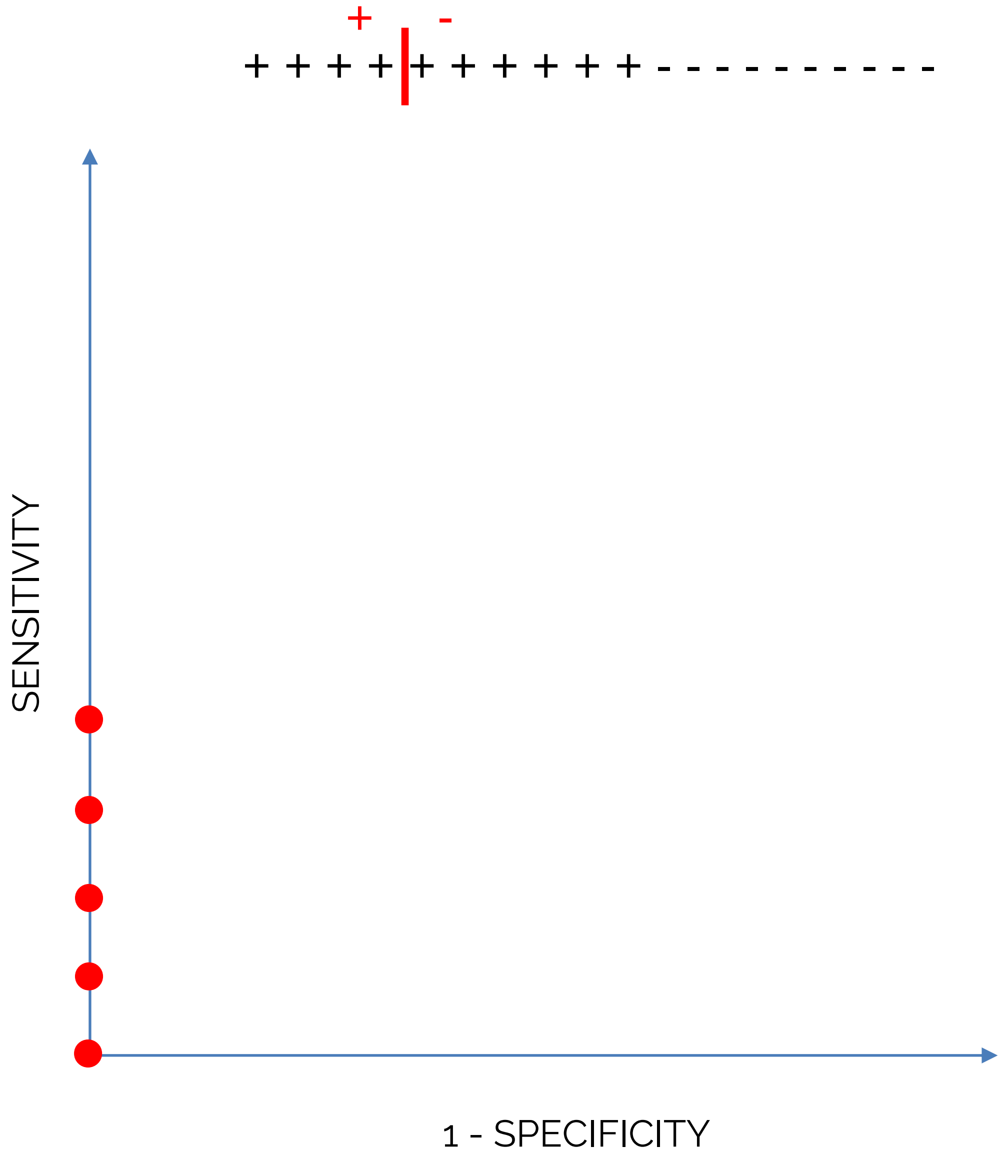
ROC analysis and Area Under the (ROC) Curve



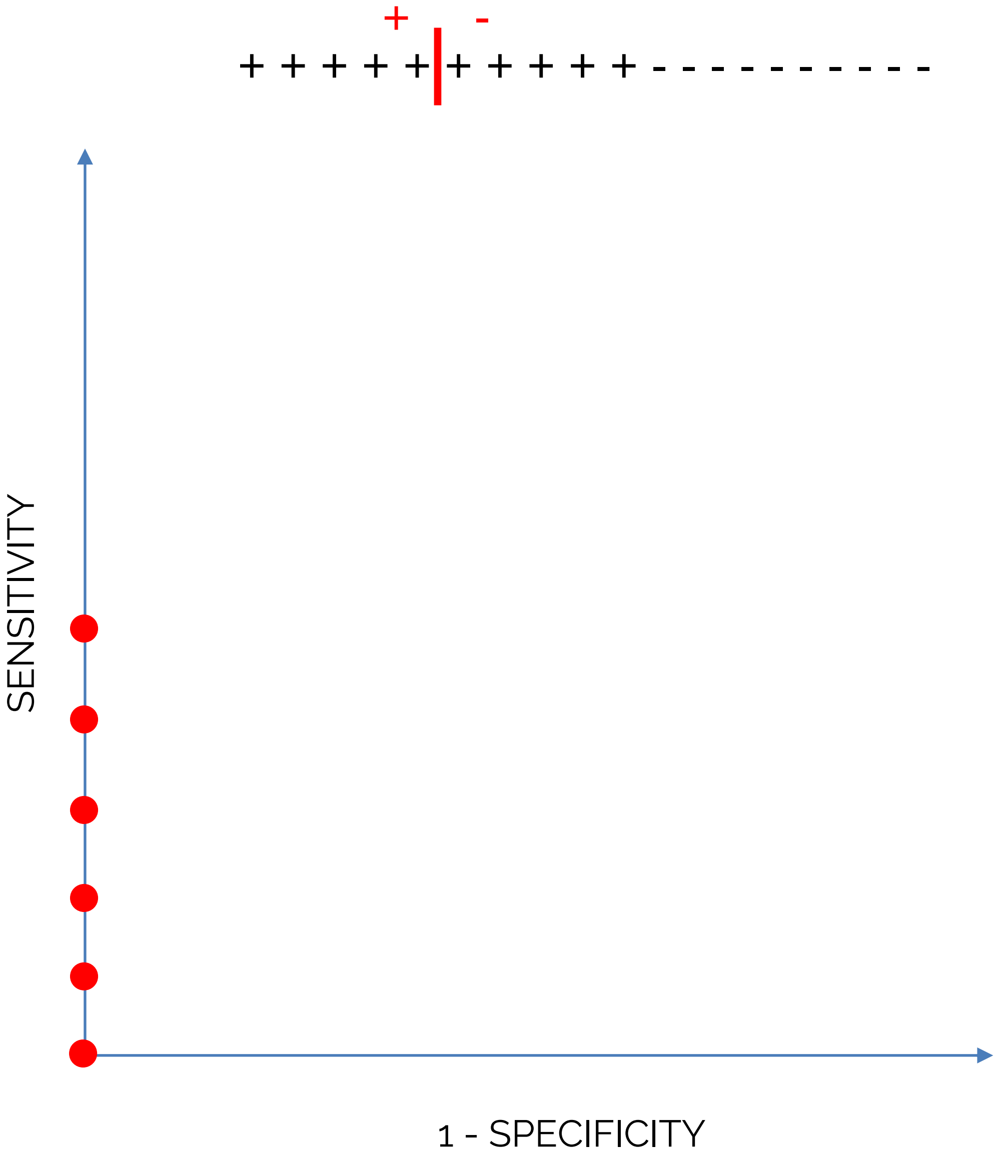
ROC analysis and Area Under the (ROC) Curve



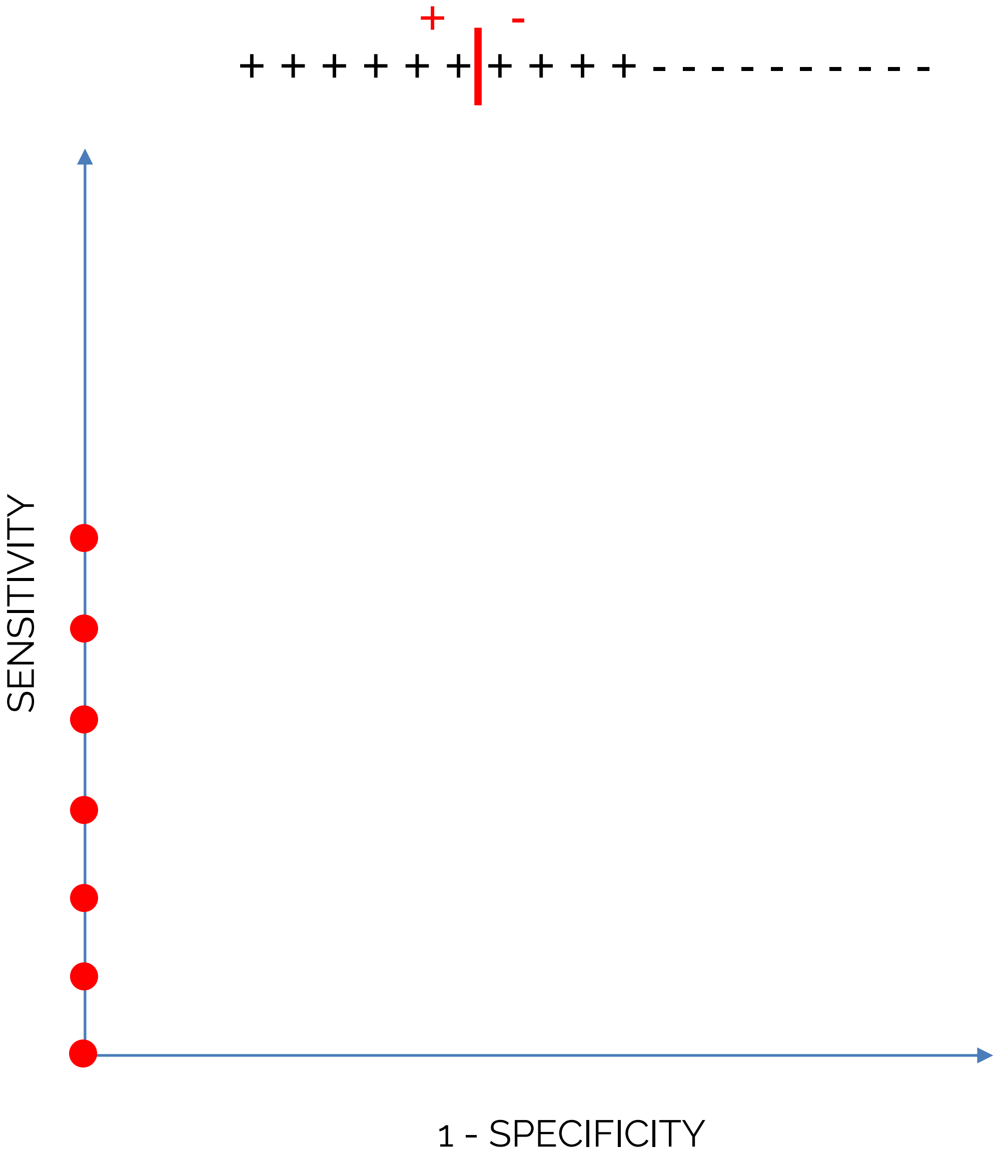
ROC analysis and Area Under the (ROC) Curve



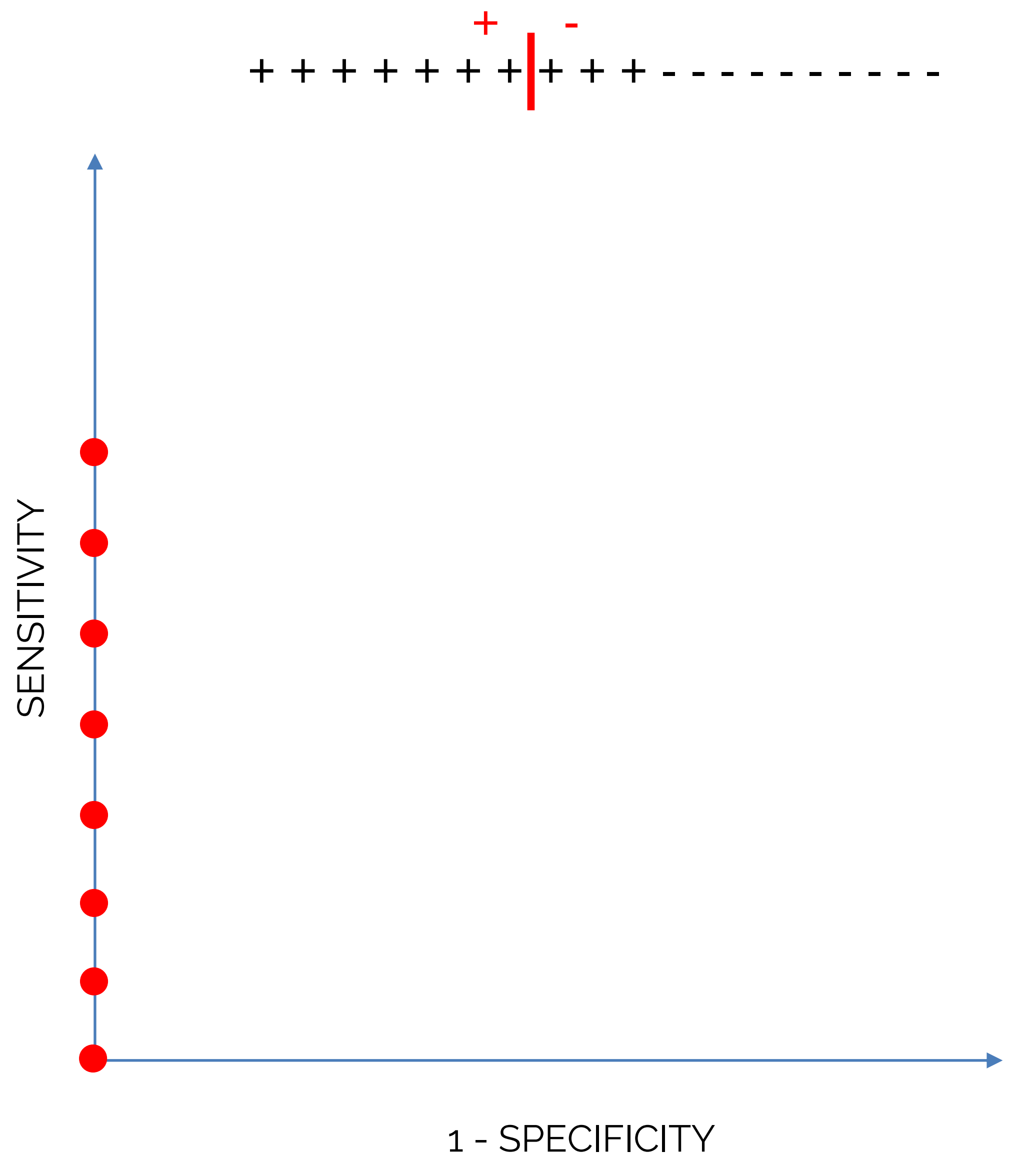
ROC analysis and Area Under the (ROC) Curve



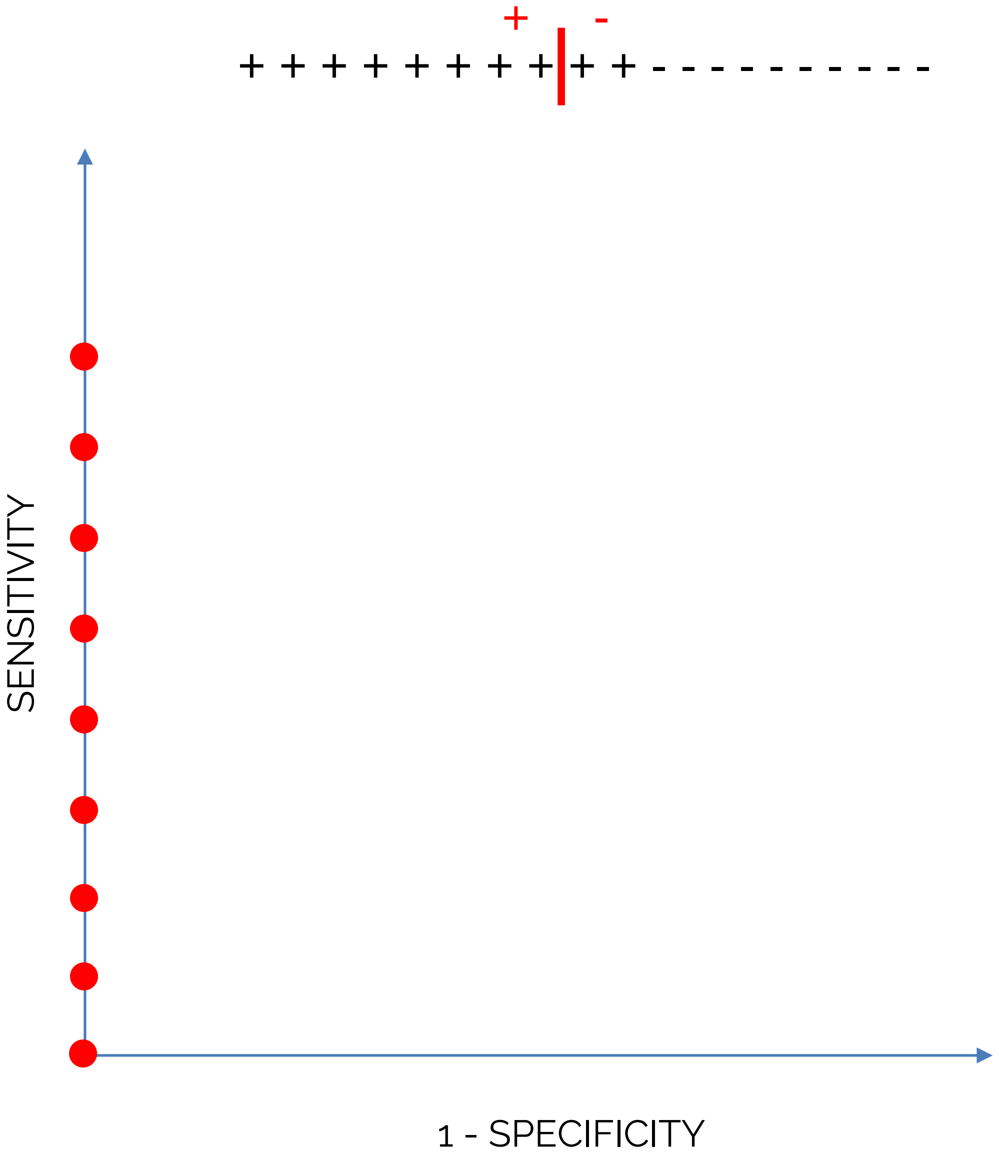
ROC analysis and Area Under the (ROC) Curve



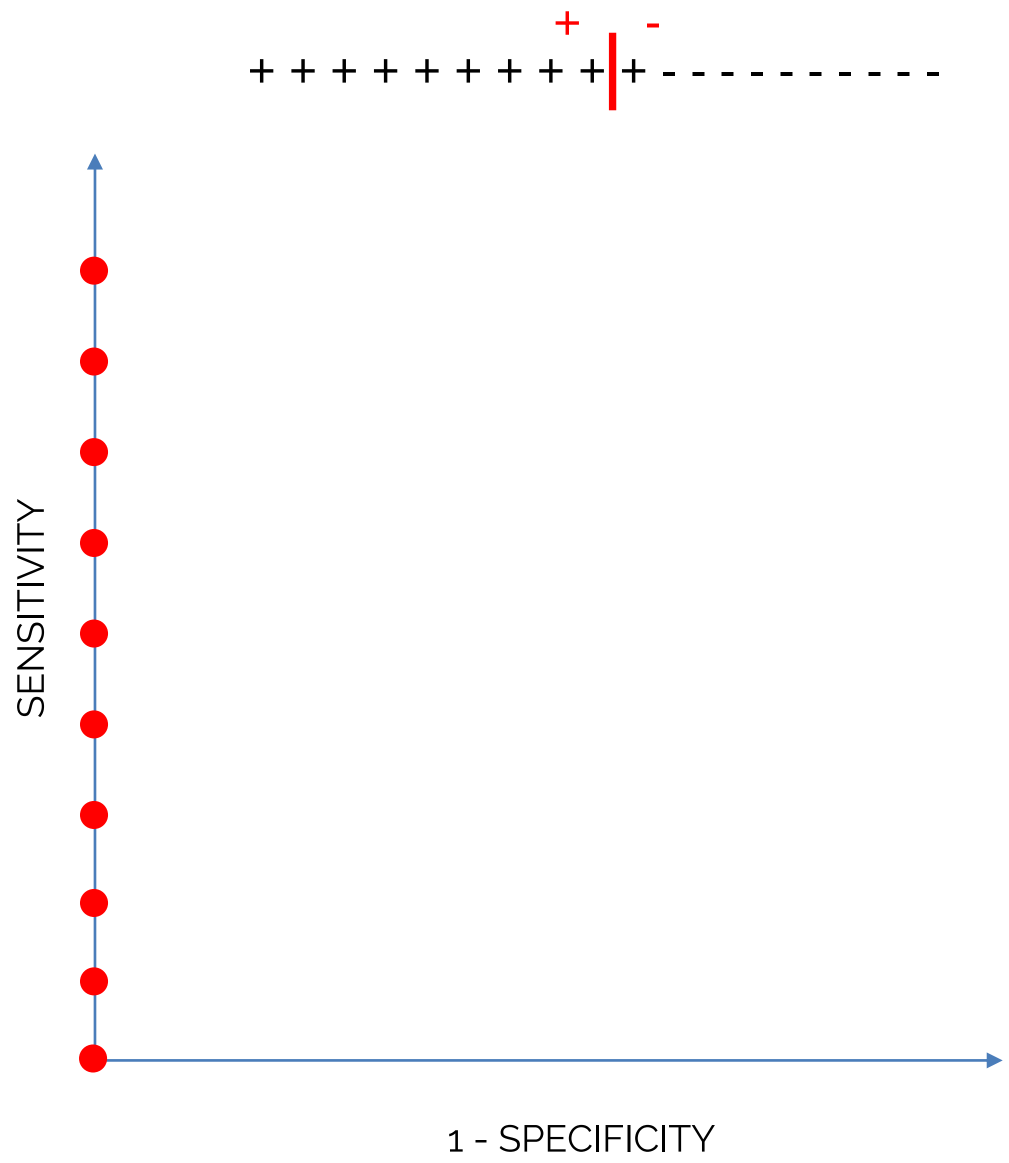
ROC analysis and Area Under the (ROC) Curve



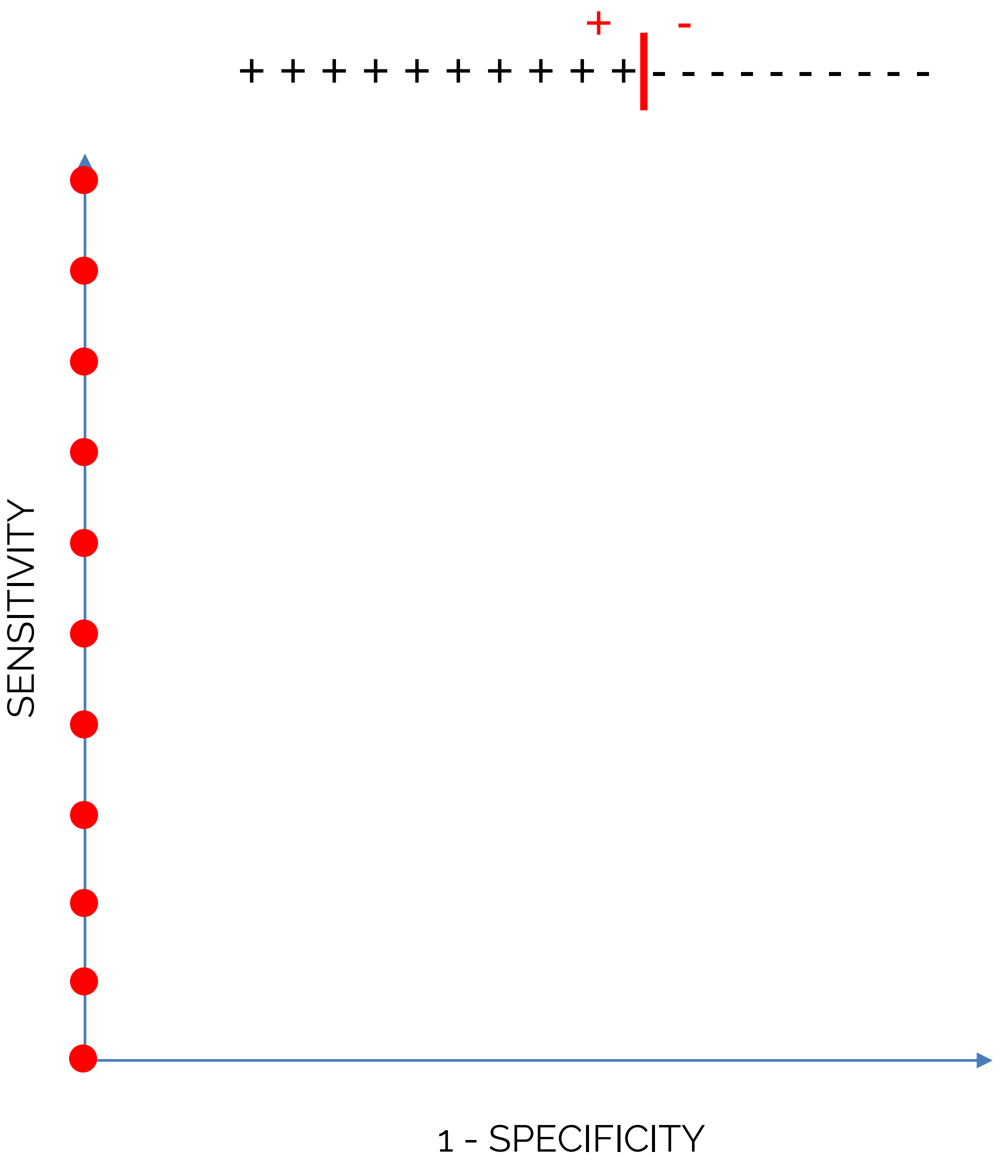
ROC analysis and Area Under the (ROC) Curve



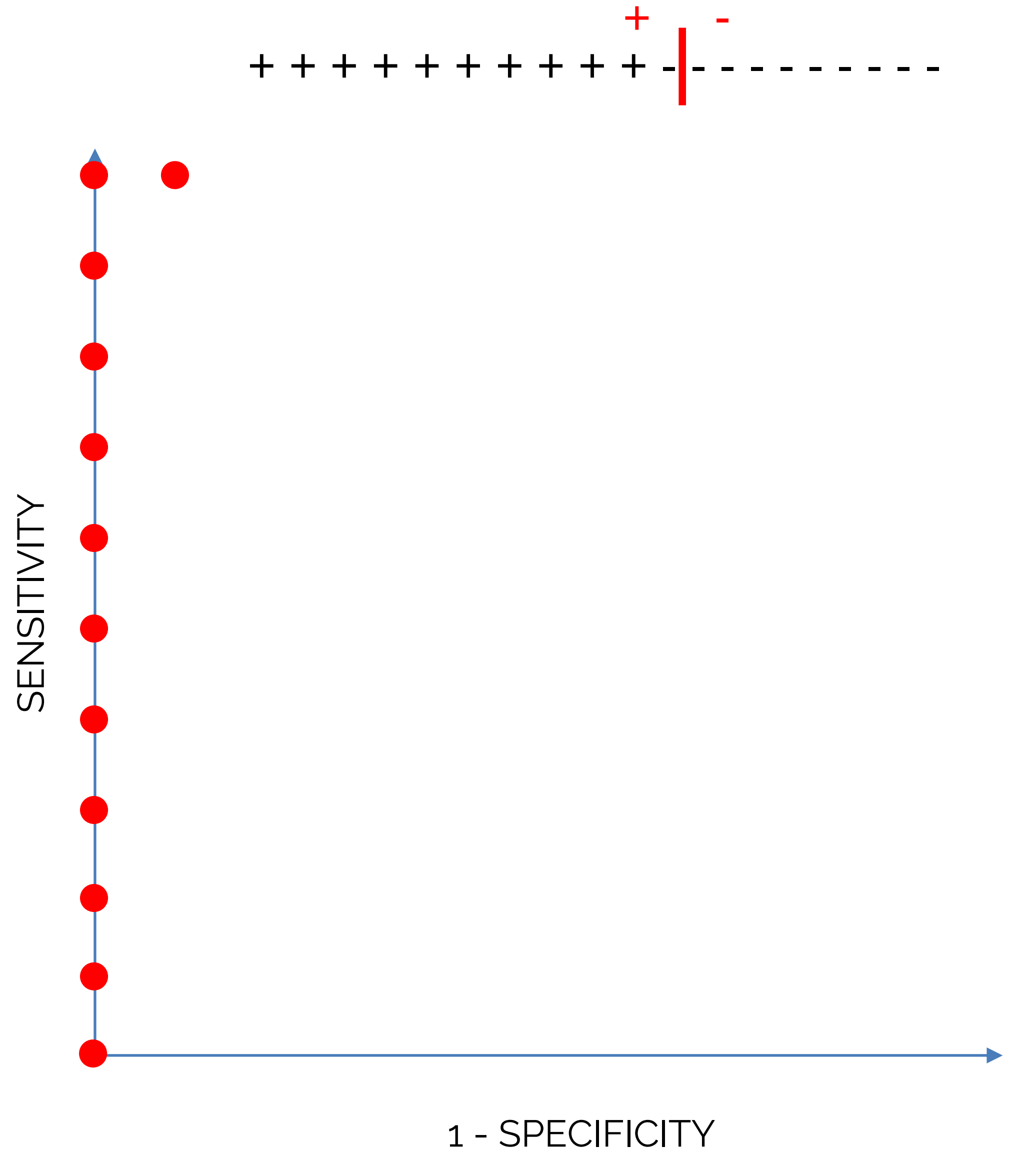
ROC analysis and Area Under the (ROC) Curve



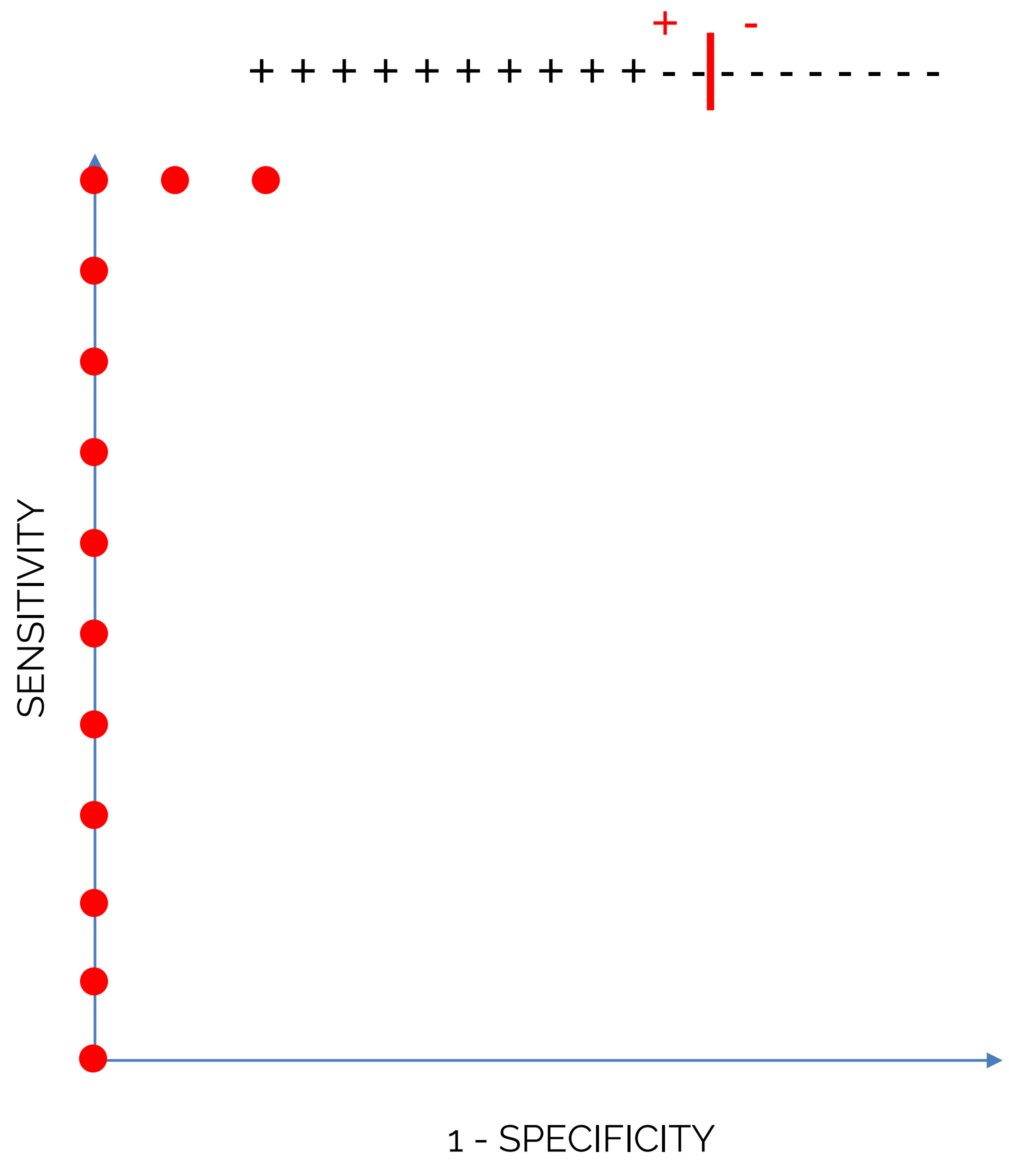
ROC analysis and Area Under the (ROC) Curve



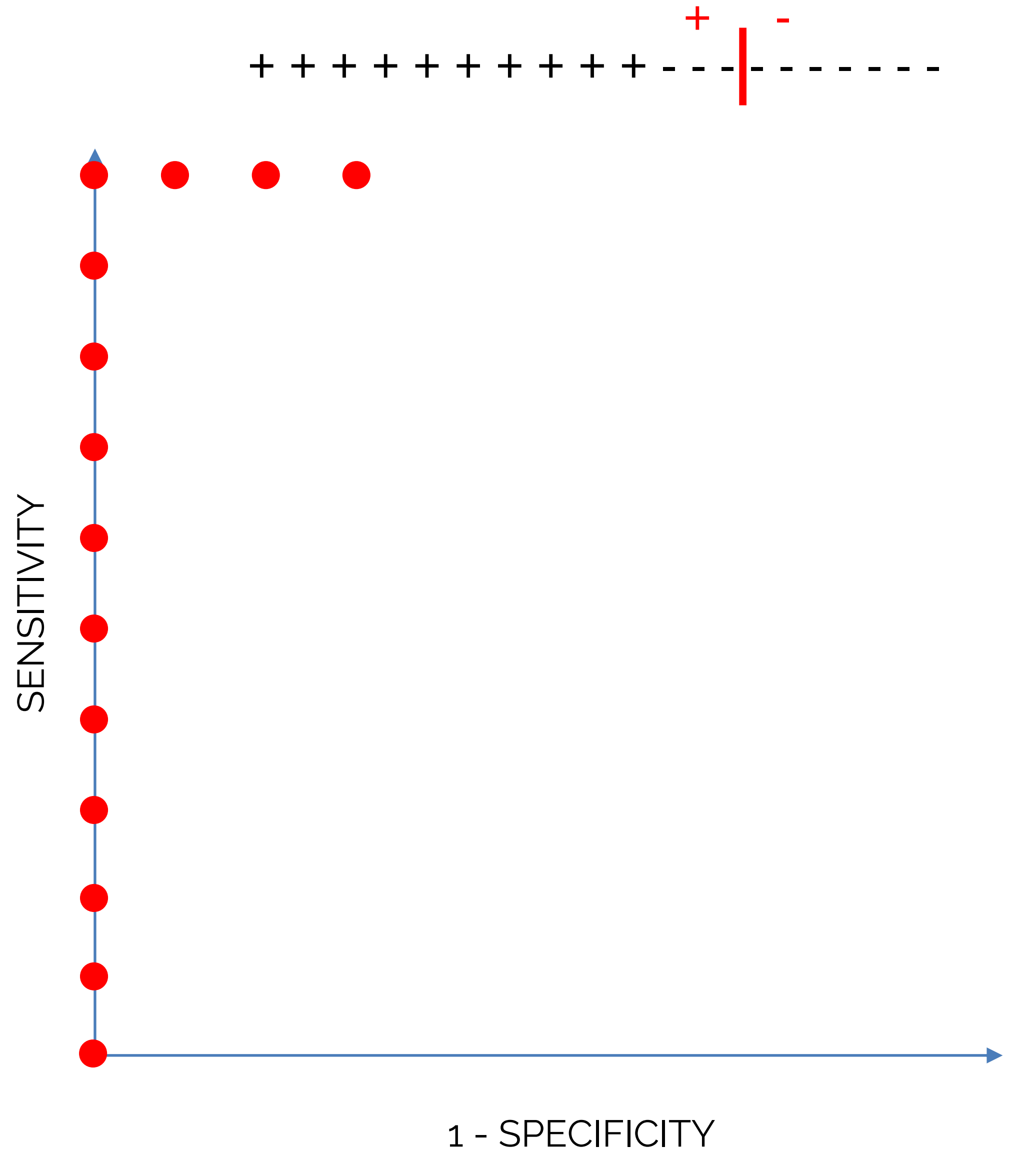
ROC analysis and Area Under the (ROC) Curve



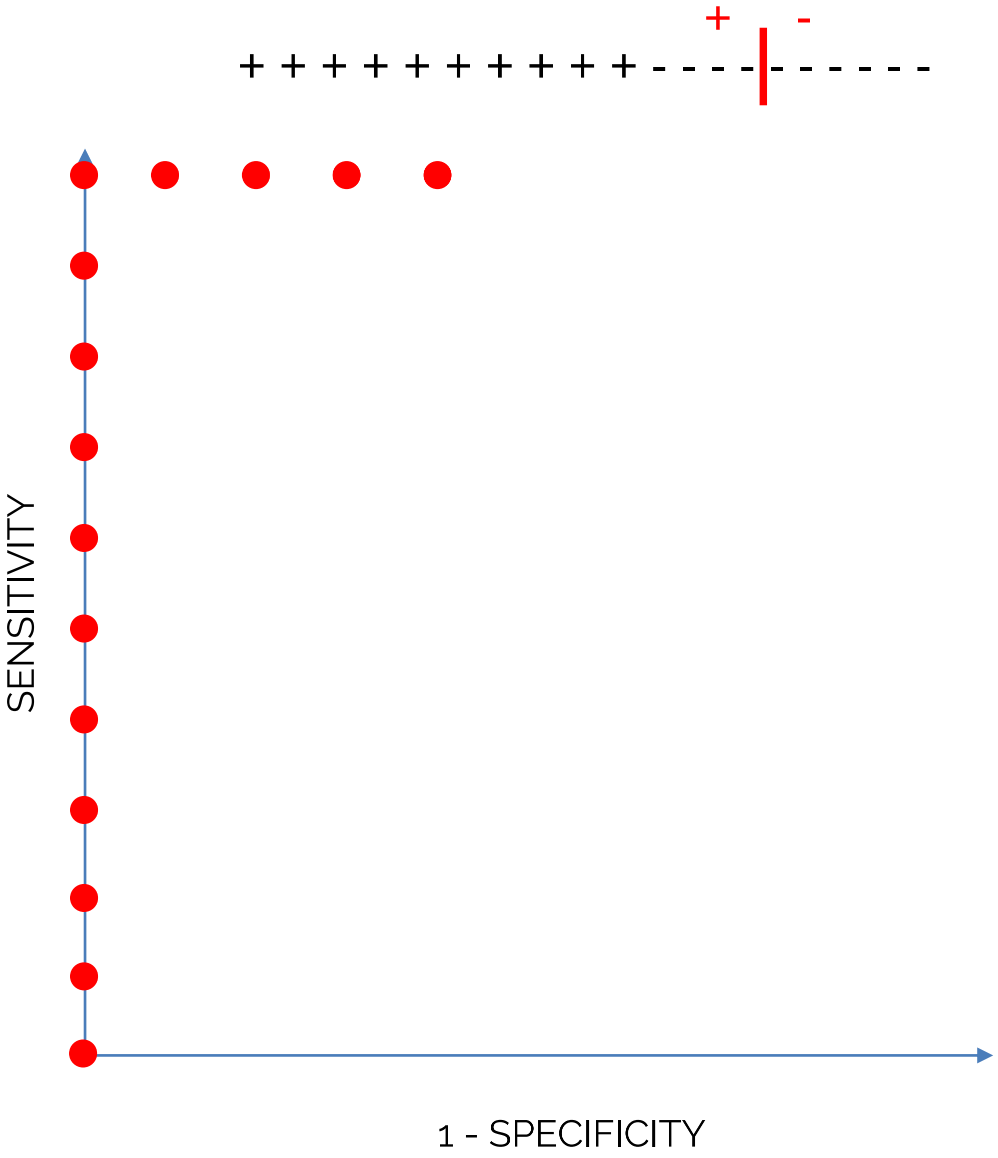
ROC analysis and Area Under the (ROC) Curve



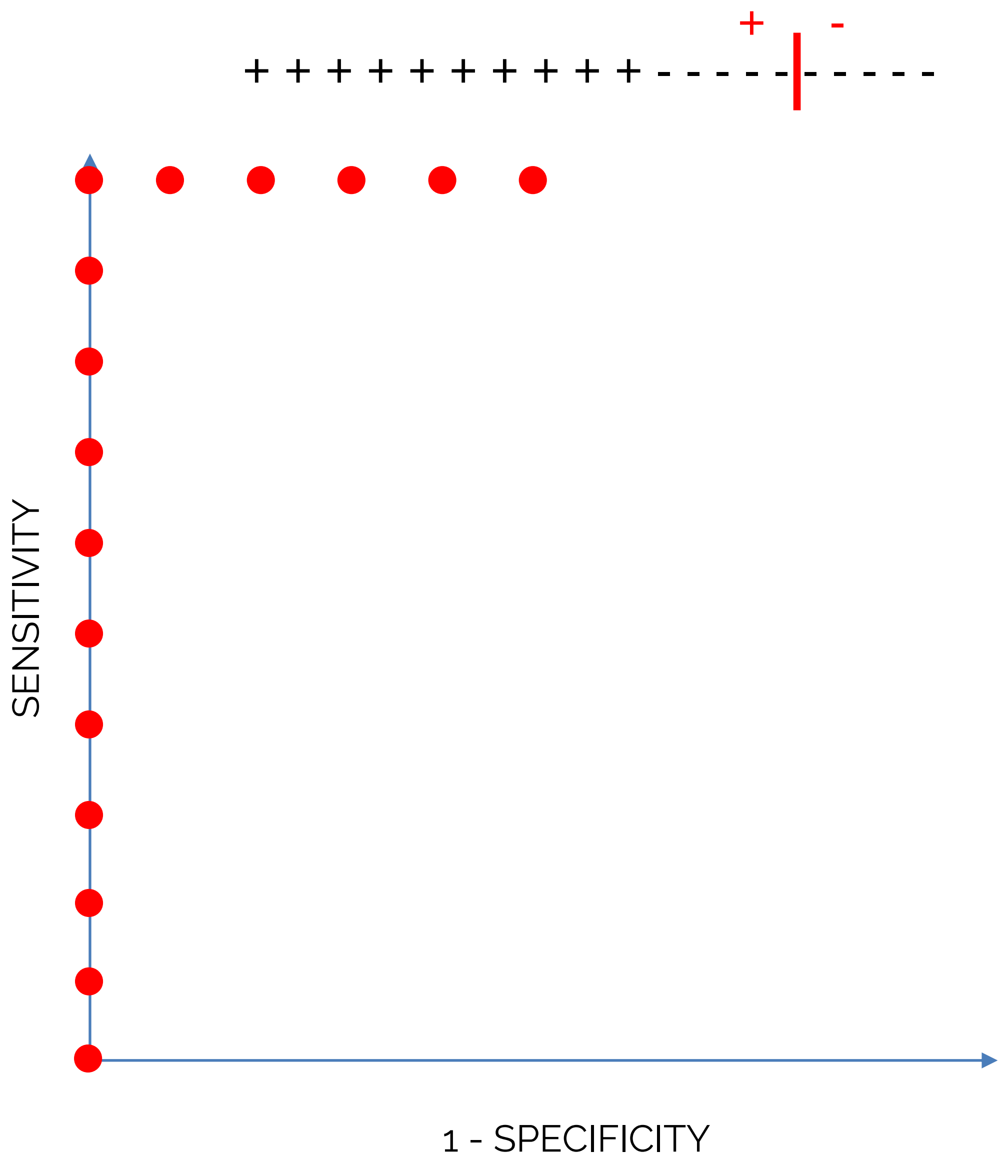
ROC analysis and Area Under the (ROC) Curve



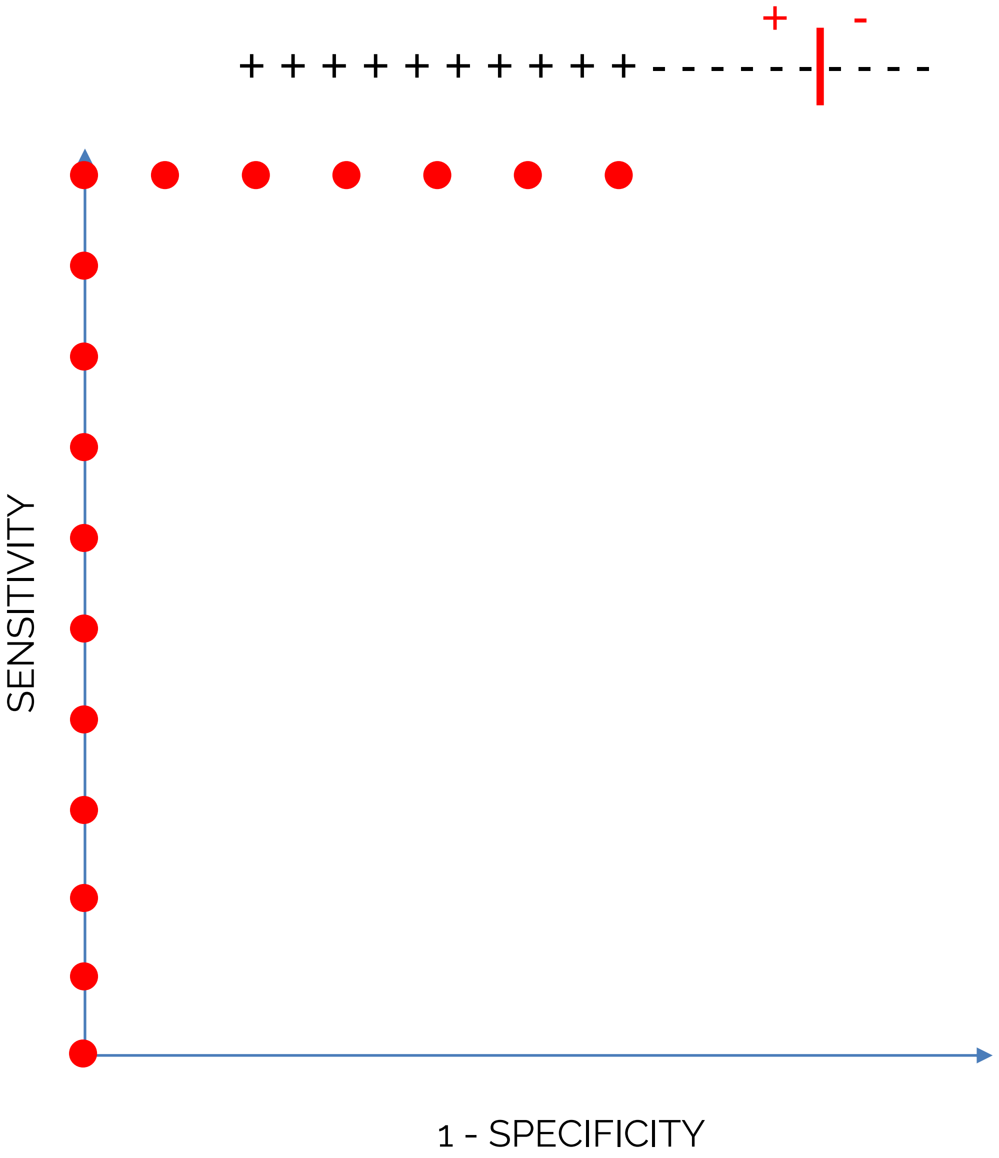
ROC analysis and Area Under the (ROC) Curve



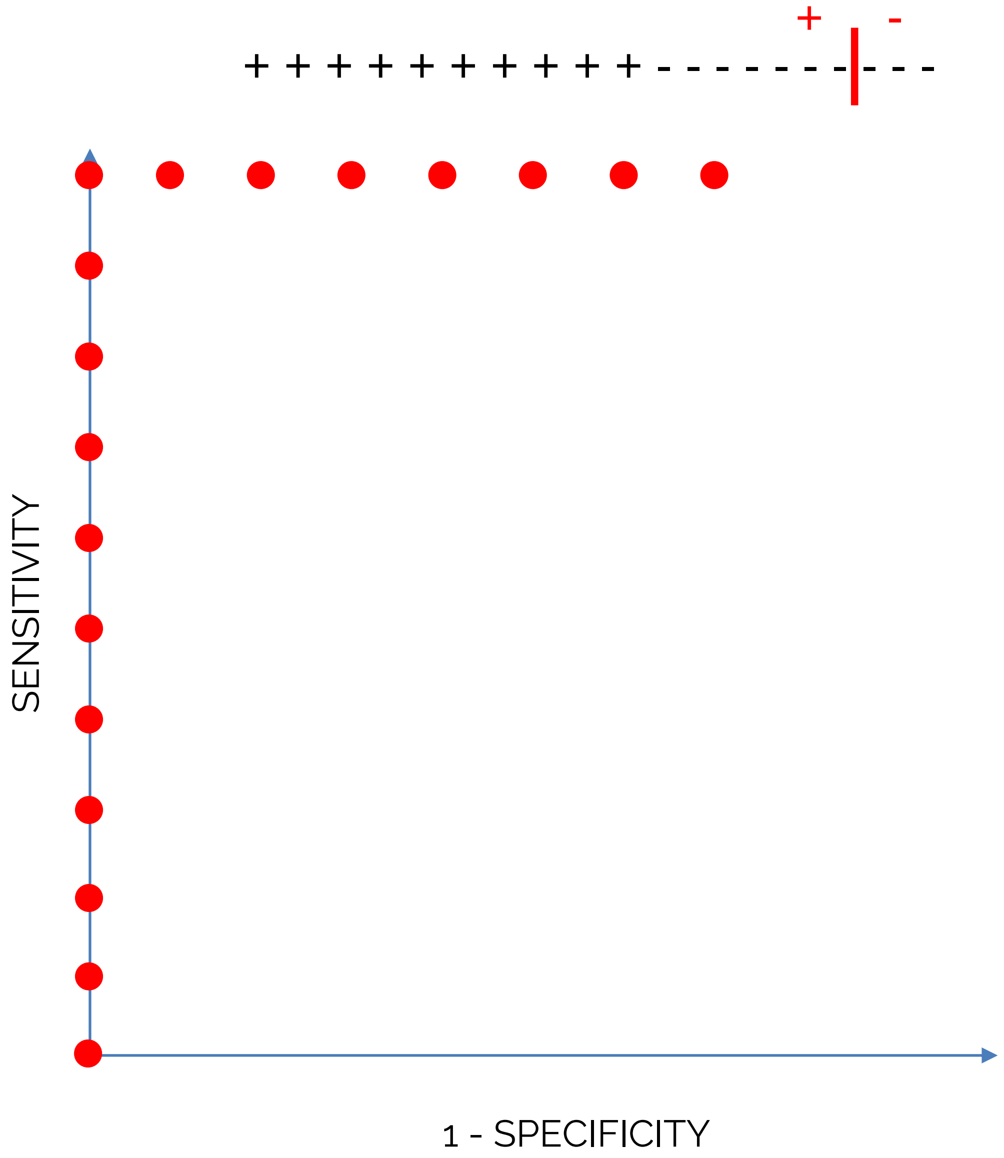
ROC analysis and Area Under the (ROC) Curve



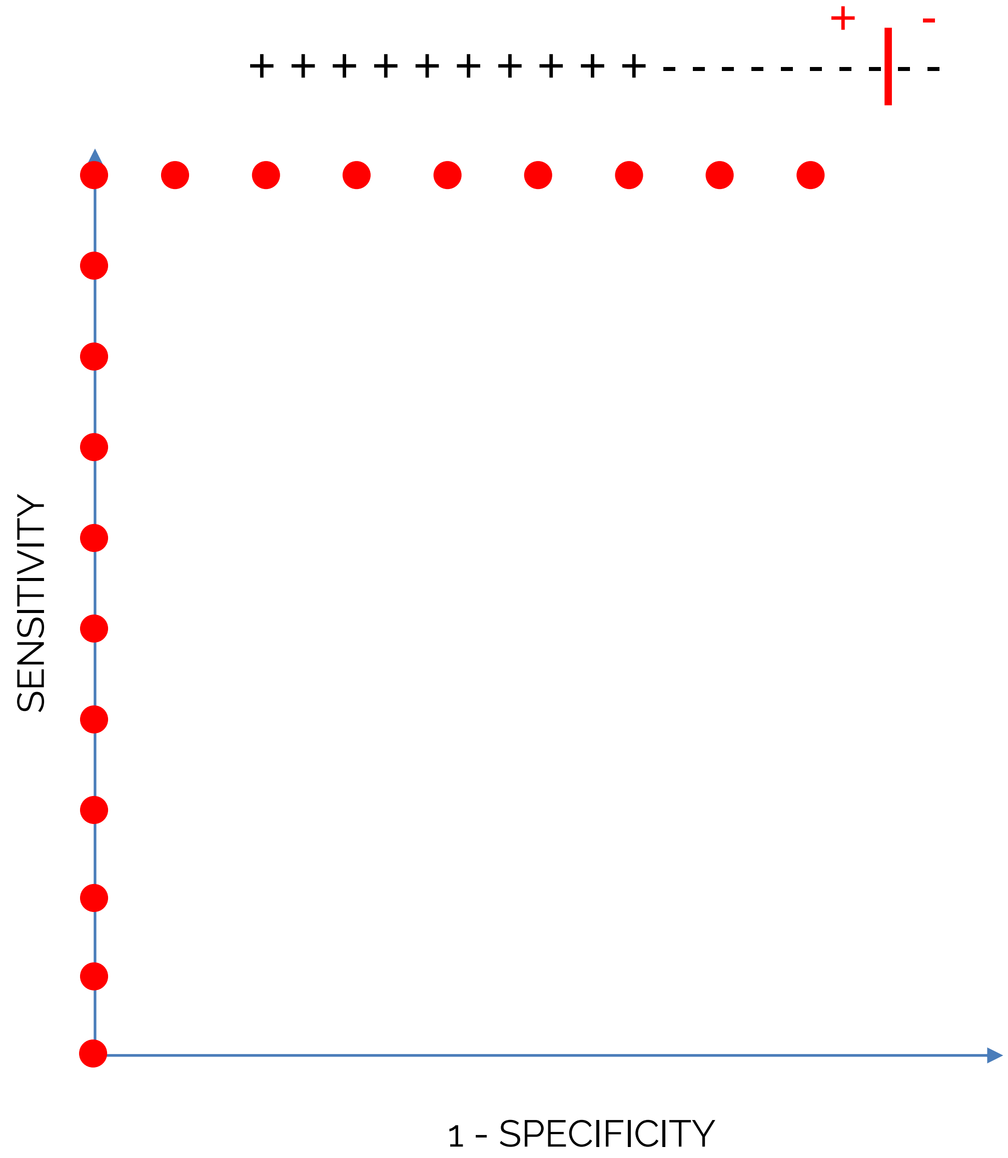
ROC analysis and Area Under the (ROC) Curve



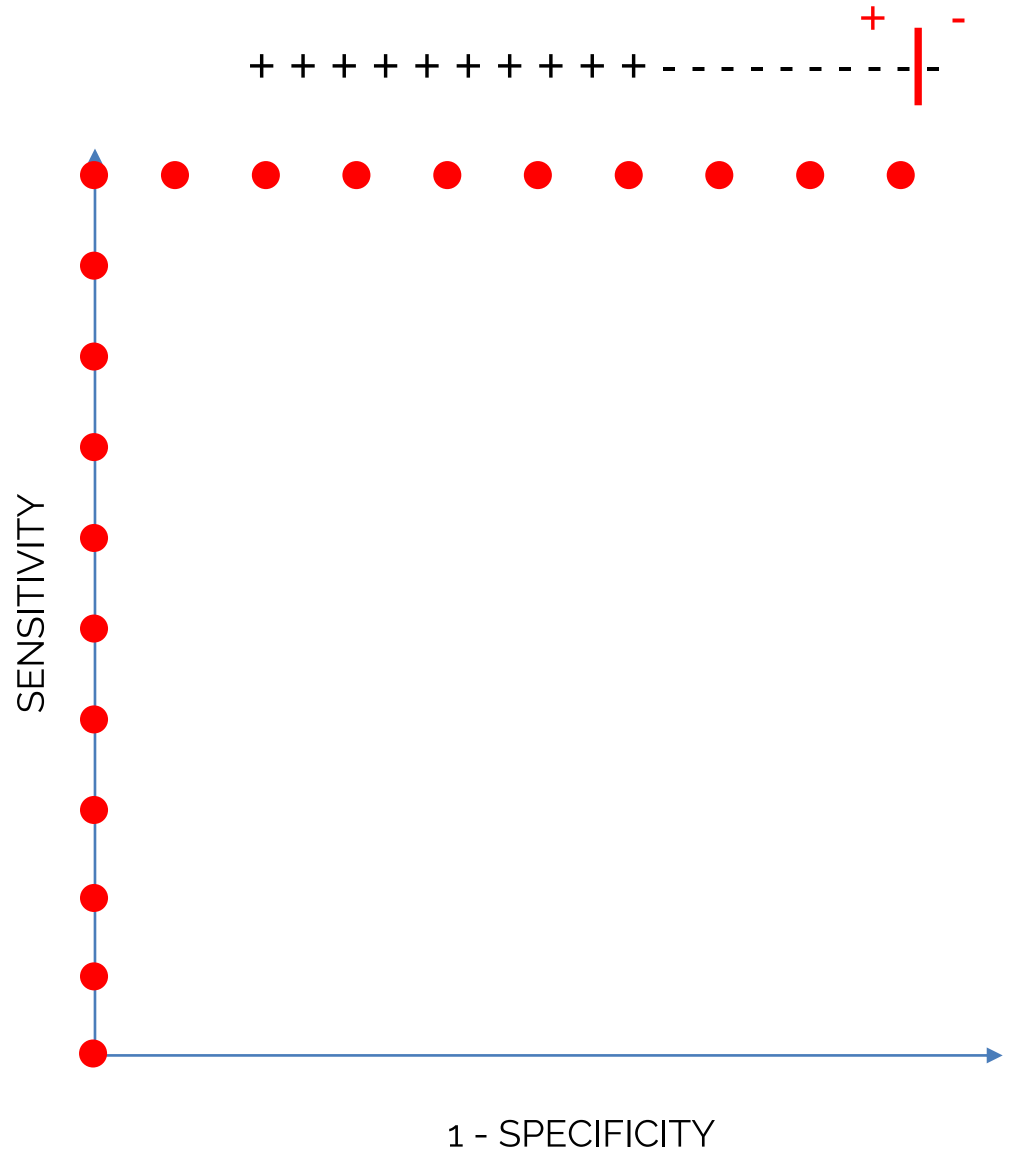
ROC analysis and Area Under the (ROC) Curve



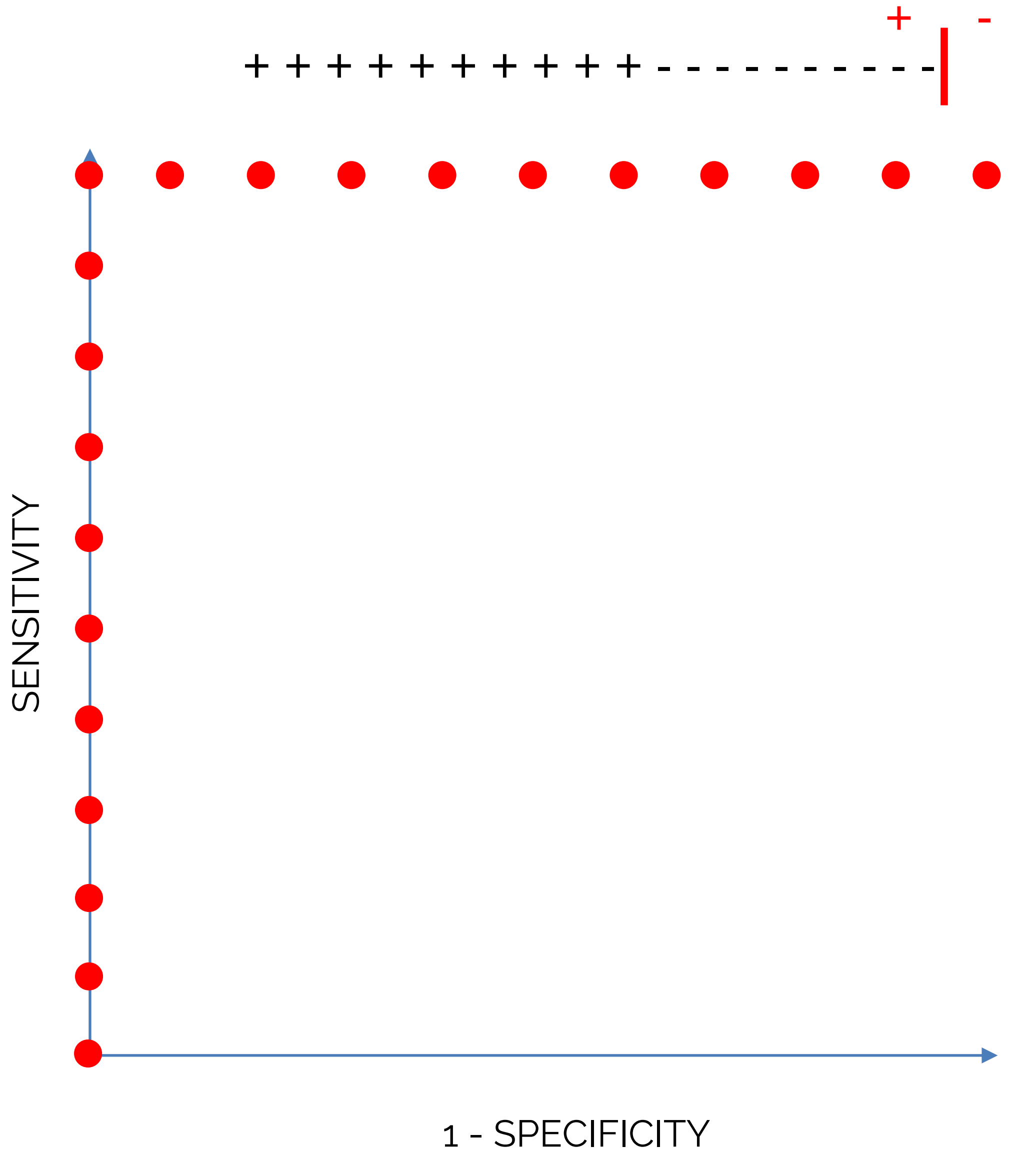
ROC analysis and Area Under the (ROC) Curve



ROC analysis and Area Under the (ROC) Curve

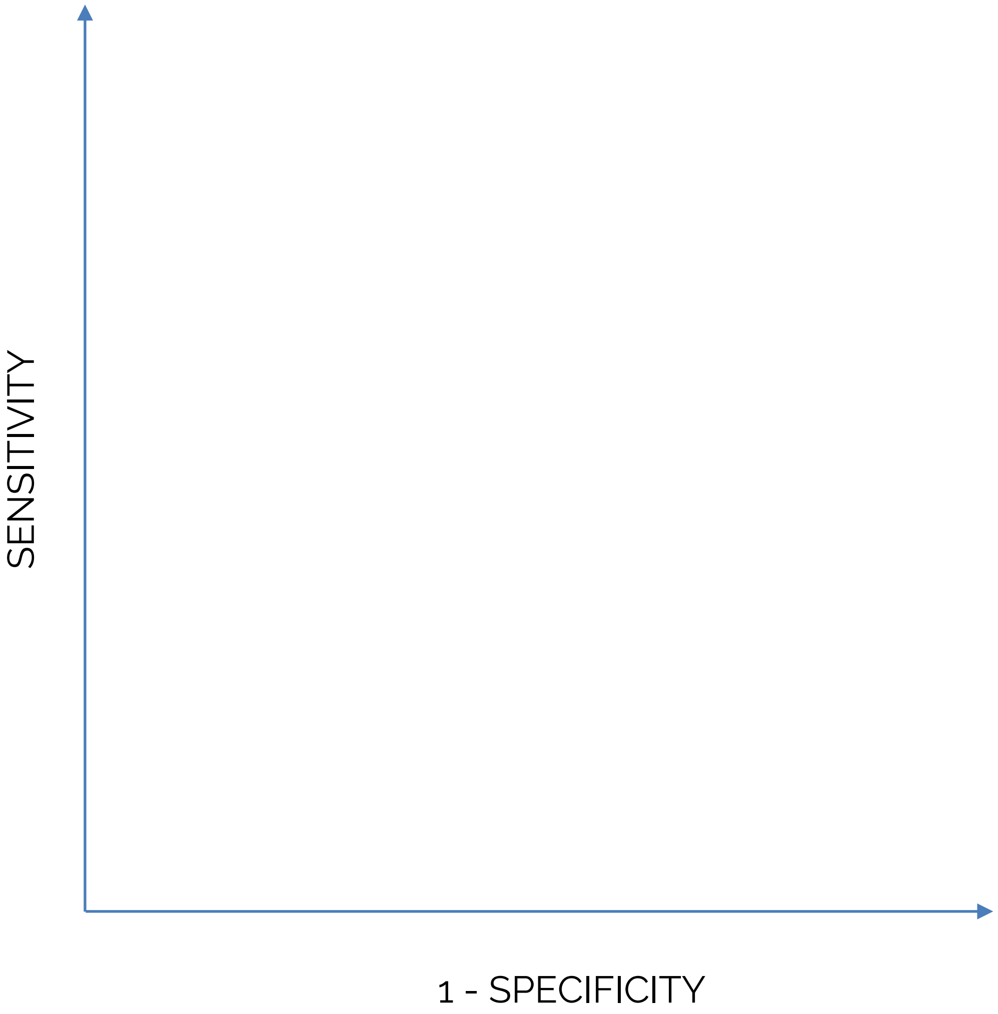


ROC analysis and Area Under the (ROC) Curve

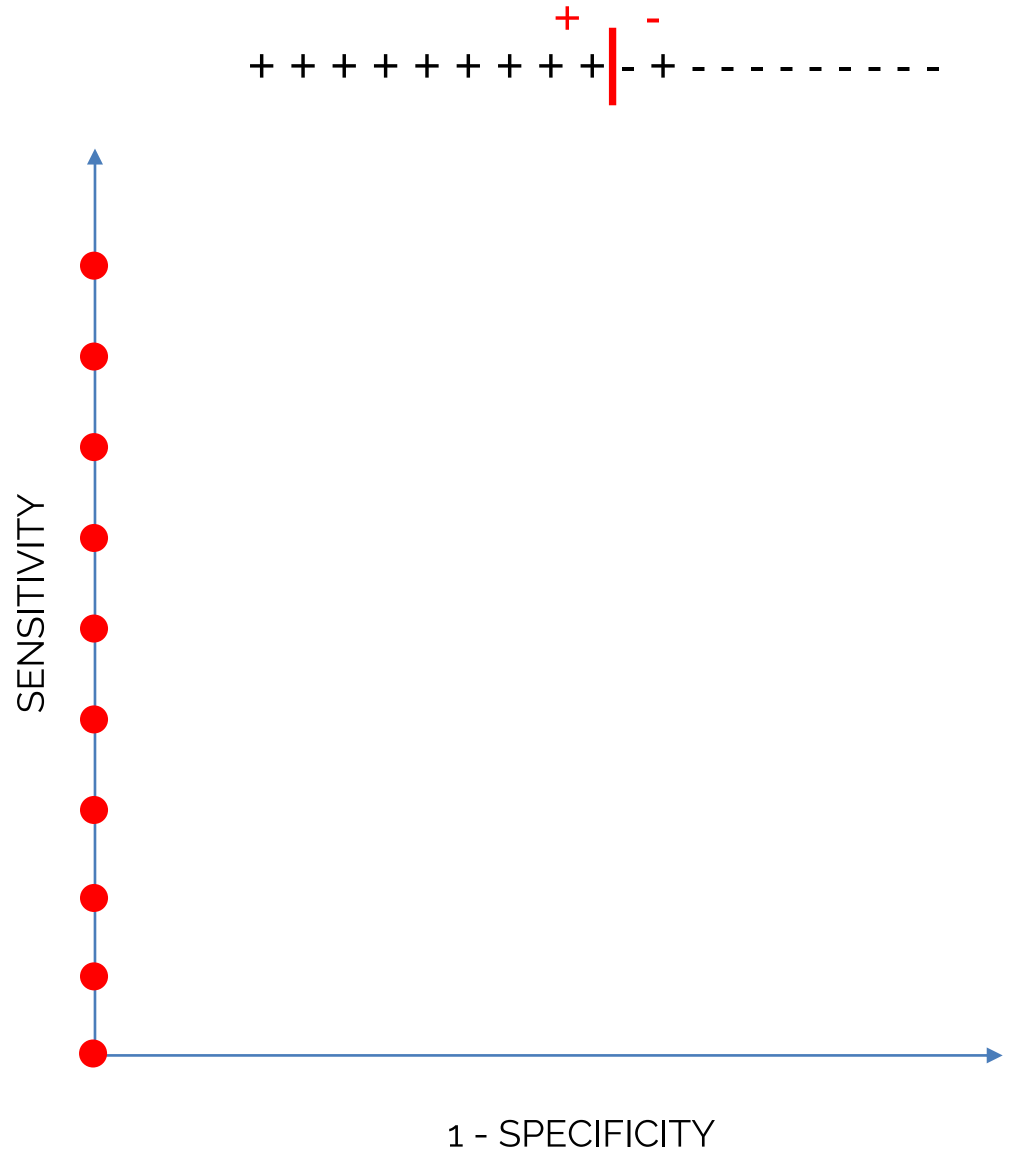


ROC analysis and Area Under the (ROC) Curve

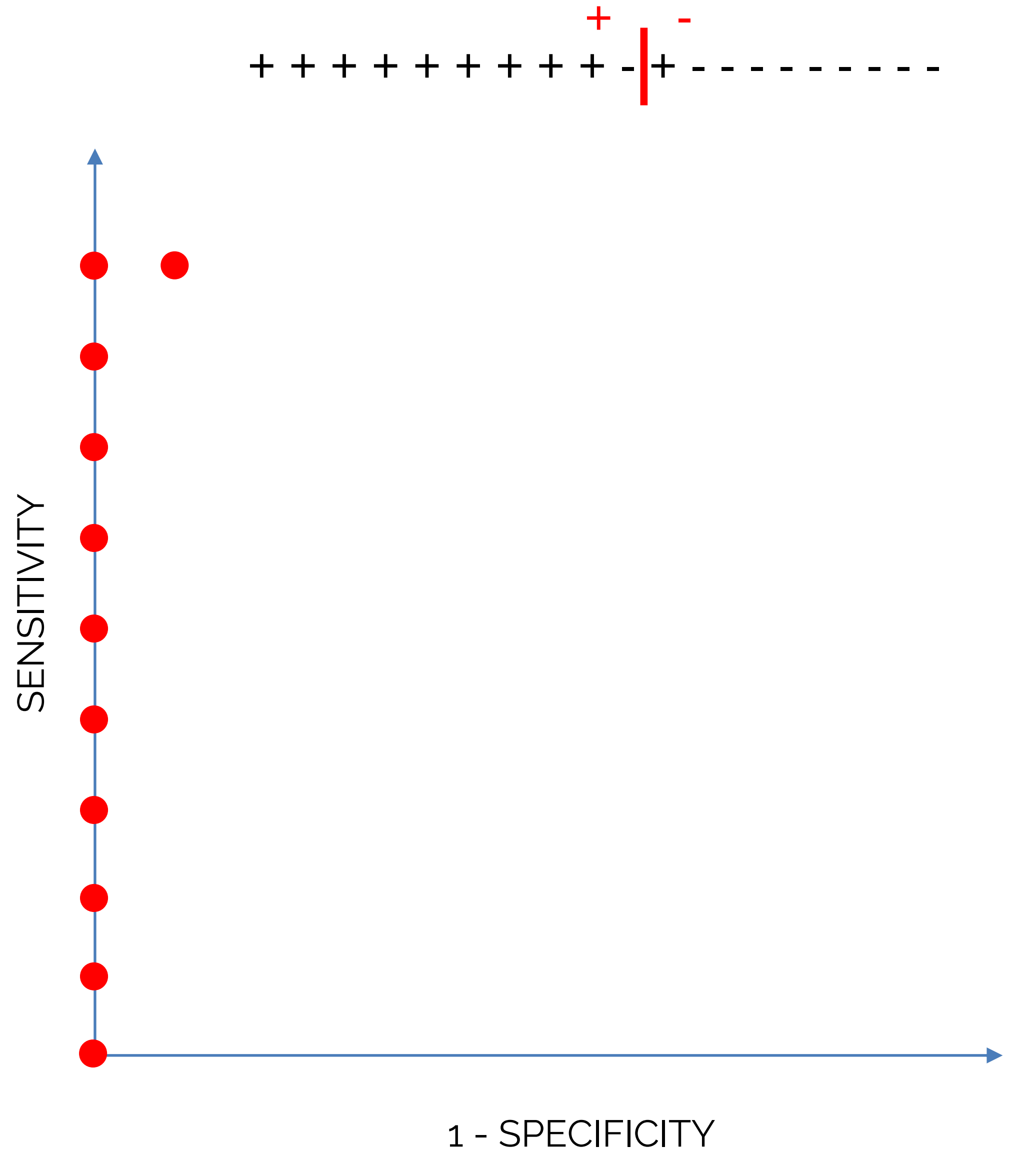
+ + + + + + + + + - + - - - - - - - - -



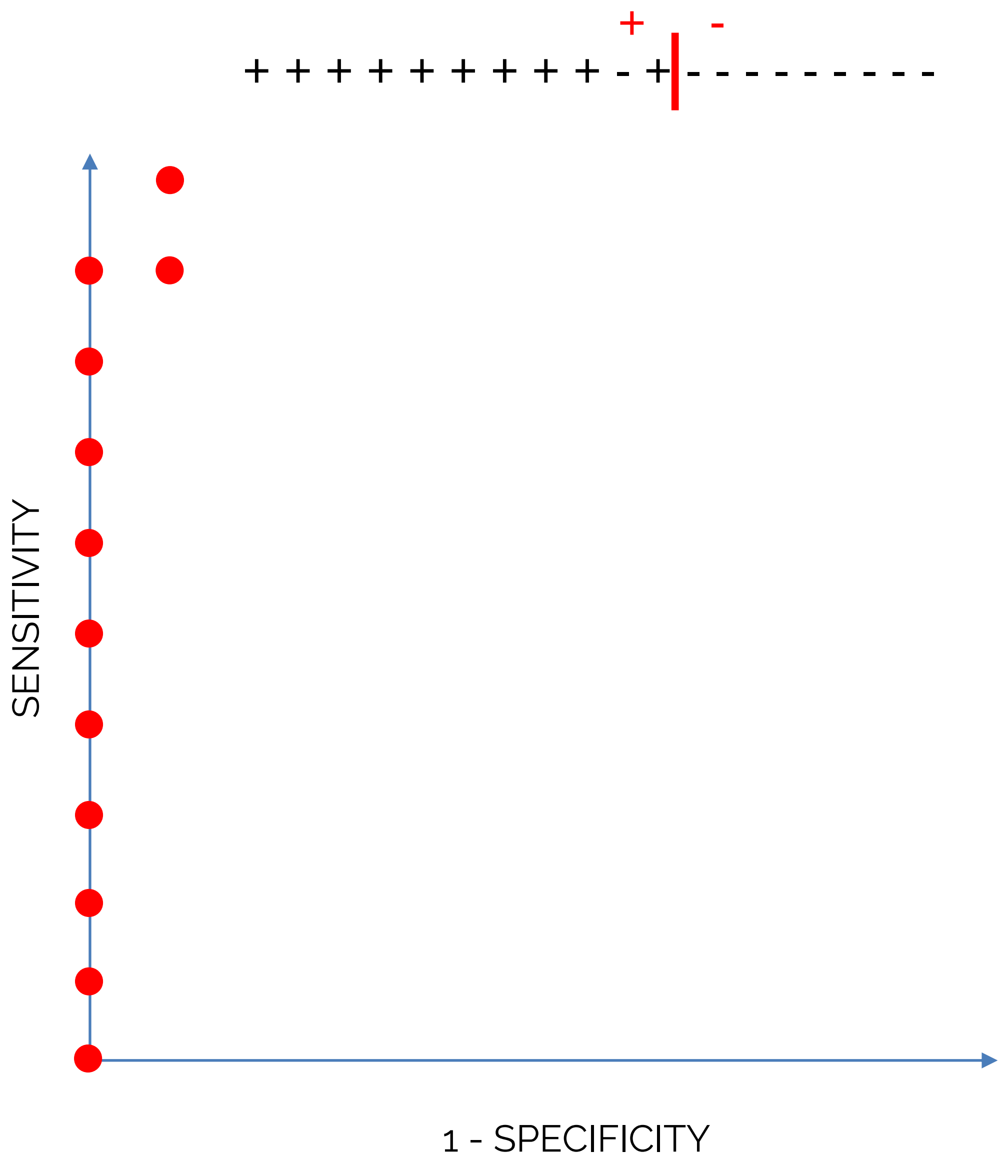
ROC analysis and Area Under the (ROC) Curve



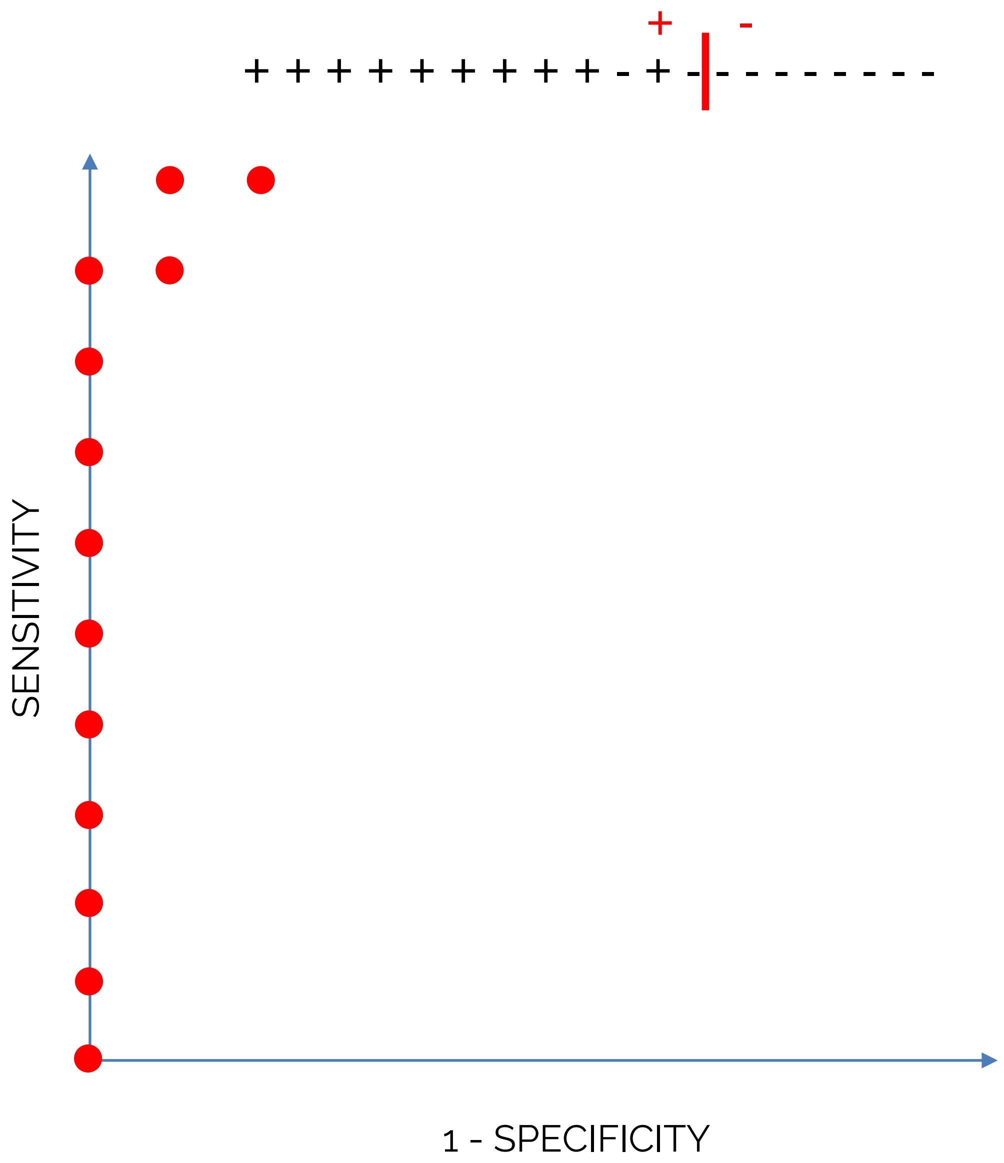
ROC analysis and Area Under the (ROC) Curve



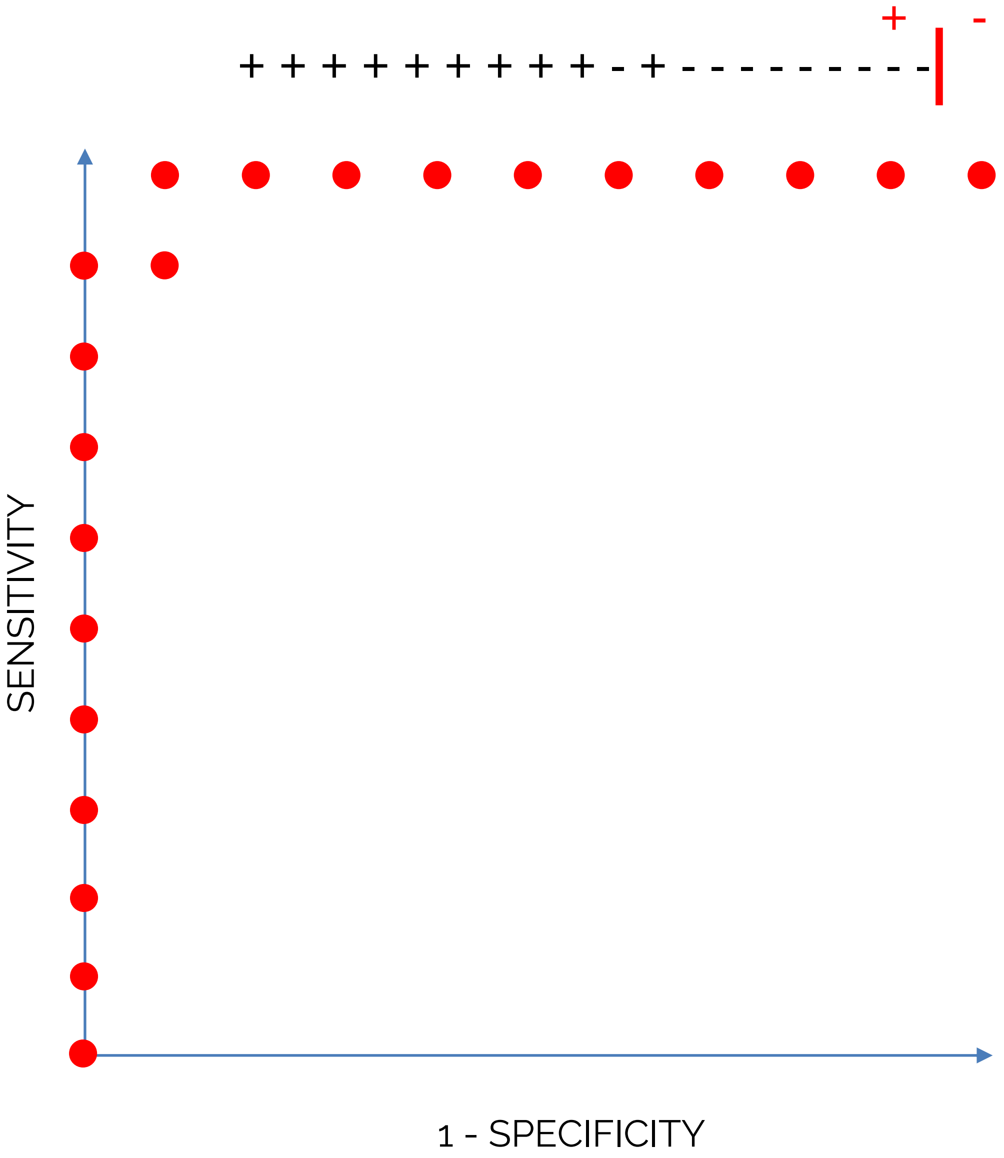
ROC analysis and Area Under the (ROC) Curve



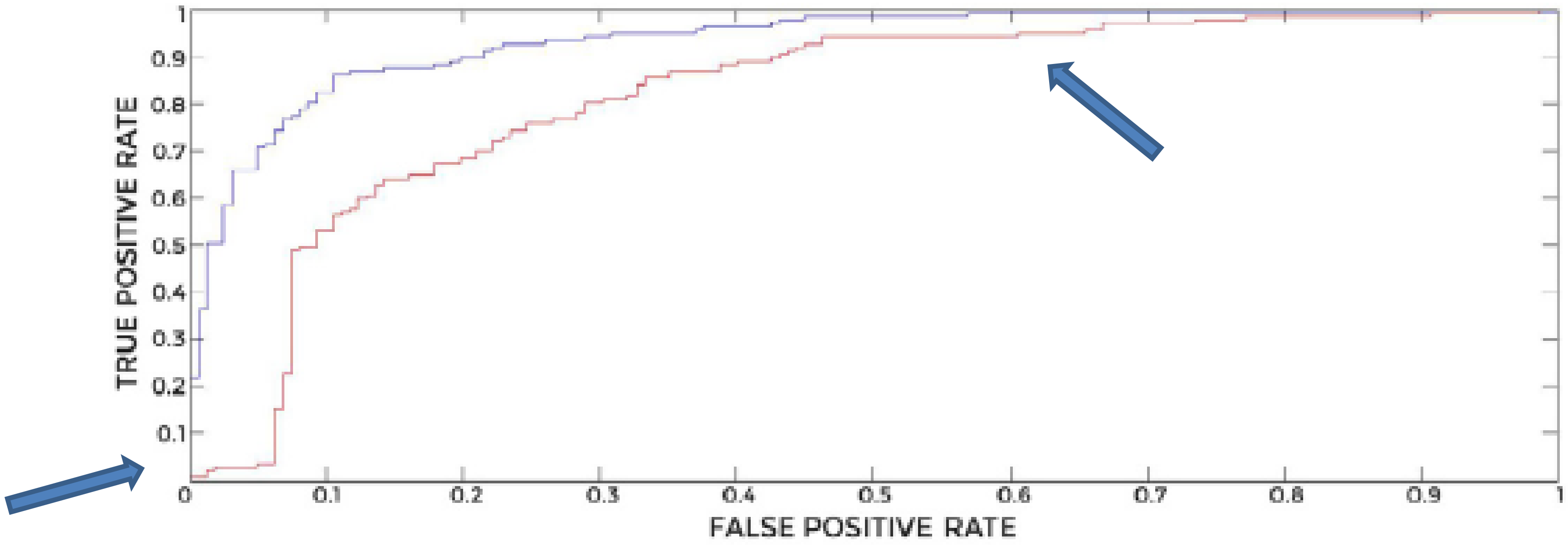
ROC analysis and Area Under the (ROC) Curve

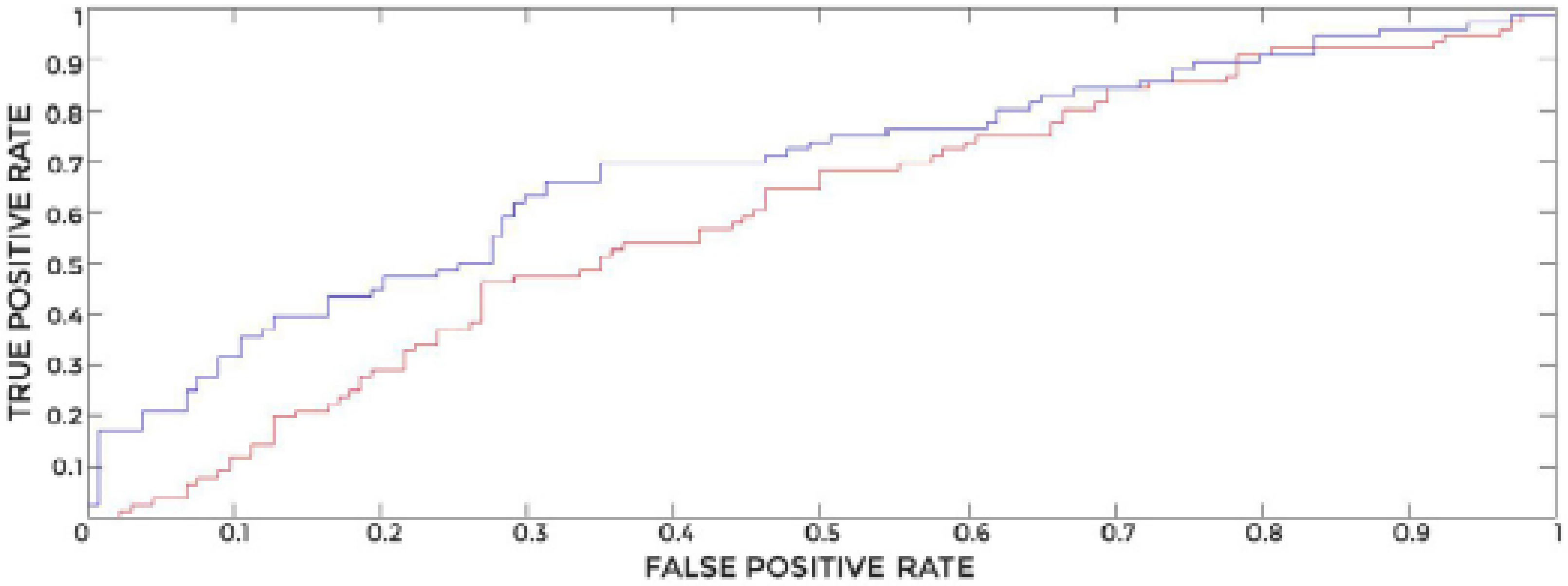


ROC analysis and Area Under the (ROC) Curve



ROC analysis and Area Under the (ROC) Curve





christian.salvatore@iusspavia.it

https://christiansalvatore.github.io/machinelearning-iuss/