Classification assessment methods: a detailed tutorial

Presentation · September 2018

CITATIONS
0

READS
5,865

1 author:

Alaa Tharwat
Frankfurt University of Applied Sciences
107 PUBLICATIONS **1,040** CITATIONS

SEE PROFILE

Classification assessment method: a detailed tutorial

Alaa Tharwat

Email: engalaatharwat@hotmail.com

- Introduction
- Classification Performance
- Classification metrics with imbalanced data
- Different Assessment methods
- Illustrative example
- Receiver Operating Characteristics (ROC)
- Area under the ROC curve (AUC)
- Precision-Recall (PR) curve
- Biometrics measures
- Conclusions

- Introduction
- Classification Performance
- Classification metrics with imbalanced data
- Different Assessment methods
- Illustrative example
- Receiver Operating Characteristics (ROC)
- Area under the ROC curve (AUC)
- Precision-Recall (PR) curve
- Biometrics measures
- Conclusions

- Classification techniques have been applied to many applications in various fields of sciences
- In classification models, the training data are used for building a classification model to predict the class label for a new sample
- The outputs of classification models can be discrete as in the decision tree classifier or continuous as the Naive Bayes classifier
- The outputs of learning algorithms need to be assessed and analyzed carefully and this analysis must be interpreted correctly, so as to evaluate different learning algorithms.
- The classification performance is represented by scalar values as in different metrics such as accuracy, sensitivity, and specificity
- Graphical assessment methods such as Receiver operating characteristics (ROC) and Precision-Recall curves give different interpretations of the classification performance
- Some of the measures which are derived from the confusion matrix for evaluating a diagnostic test

- Introduction
- Classification Performance
- Classification metrics with imbalanced data
- Different Assessment methods
- Illustrative example
- Receiver Operating Characteristics (ROC)
- Area under the ROC curve (AUC)
- Precision-Recall (PR) curve
- Biometrics measures
- Conclusions

- According to the number of classes, there are two types of classification problems, namely, **binary classification** where there are only two classes, and **multi-class classification** where the number of classes is higher than two
- Assume we have two classes, i.e., binary classification, P for *positive* class and N for *negative* class
- An unknown sample is classified to P or N
- The classification model that was trained in the training phase is used to predict the true classes of unknown samples
- This classification model produces continuous or discrete outputs. The discrete output that is generated from a classification model represents the predicted discrete class label of the unknown/test sample, while continuous output represents the estimation of the sample's class membership probability

- There are four possible outputs which represent the elements of a 2×2 *confusion matrix* or a *contingency table*
  - If the sample is positive and it is classified as positive, i.e., correctly classified positive sample, it is counted as a *true positive* (TP); if it is classified as negative, it is considered as a *false negative* (FN)
  - If the sample is negative and it is classified as negative it is considered as *true negative* (TN); if it is classified as positive, it is counted as *false positive* (FP), *false alarm*
  - The confusion matrix is used to calculate many common classification metrics
- The green diagonal represents correct predictions and the pink diagonal indicates the incorrect predictions

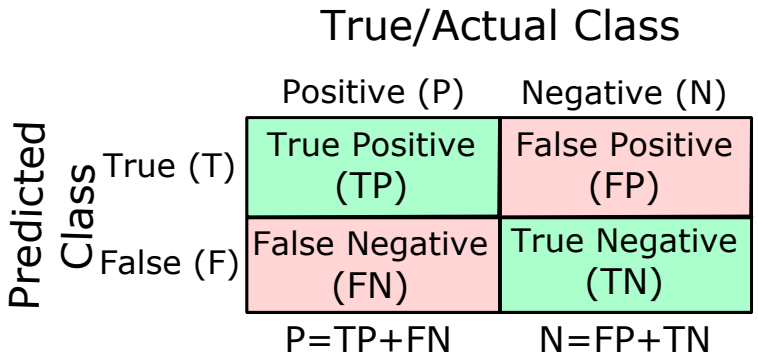
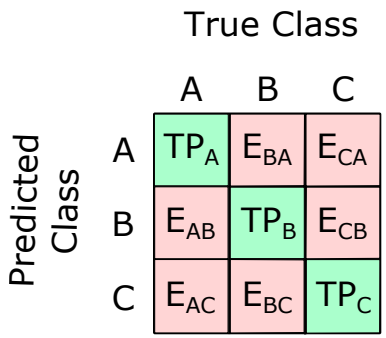


Figure: An illustrative example of the 2×2 confusion matrix. There are two true classes P and N. The output of the predicted class is true or false.

- The figure below shows the confusion matrix for a multi-class classification problem with three classes (A, B, and C)
- TP_A is the number of true positive samples in class A, i.e., the number of samples that are correctly classified from class A
- E_{AB} is the samples from class A that were incorrectly classified as class B, i.e., misclassified samples. Thus, the false negative in the A class (FN_A) is the sum of E_{AB} and E_{AC} ($FN_A = E_{AB} + E_{AC}$) which indicates the sum of all class A samples that were incorrectly classified as class B or C
- Simply, FN of any class which is located in a column can be calculated by adding the errors in that class/column
- The false positive for any predicted class which is located in a row represents the sum of all errors in that row



- Introduction
- Classification Performance
- Classification metrics with imbalanced data
- Different Assessment methods
- Illustrative example
- Receiver Operating Characteristics (ROC)
- Area under the ROC curve (AUC)
- Precision-Recall (PR) curve
- Biometrics measures
- Conclusions

- Different assessment methods are sensitive to the imbalanced data when the samples of one class in a dataset outnumber the samples of the other class(es)
- The class distribution is the ratio between the positive and negative samples ($\frac{P}{N}$) represents the relationship between the left column to the right column
- Any assessment metric that uses values from both columns will be sensitive to the imbalanced data
  - For example, some metrics such as accuracy and precision use values from both columns of the confusion matrix; thus, as data distributions change, these metrics will change as well, even if the classifier performance does not
- This fact is partially true because there are some metrics such as Geometric Mean (GM) and Youden's index (YI) use values from both columns and these metrics can be used with balanced and imbalanced data
- This can be interpreted as that the metrics which use values from one column cancel the changes in the class distribution

- However, some metrics which use values from both columns are not sensitive to the imbalanced data because the changes in the class distribution cancel each other
  - For example, the accuracy is defined as follows, $Acc = \frac{TP+TN}{TP+TN+FP+FN}$ and the GM is defined as follows, $GM = \sqrt{TPR \times TNR} = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}}$; thus, both metrics use values from both columns of the confusion matrix
  - Changing the class distribution can be obtained by increasing/decreasing the number of samples of negative/positive class
  - With the same classification performance, assume that the negative class samples are increased by α times; thus, the TN and FP values will be αTN and αFP, respectively; thus, the accuracy will be, $Acc = \frac{TP+\alpha TN}{TP+\alpha TN+\alpha FP+FN} \neq \frac{TP+TN}{TP+TN+FP+FN}$. This means that the accuracy is affected by the changes in the class distribution
  - On the other hand, the GM metric will be, $GM = \sqrt{\frac{TP}{TP+FN} \times \frac{\cancel{\alpha} TN}{\cancel{\alpha} TN+\cancel{\alpha} FP}} = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}}$ and hence the changes in the negative class cancel each other. This is the reason why the GM metric is suitable for the imbalanced data

- Introduction
- Classification Performance
- Classification metrics with imbalanced data
- Different Assessment methods
- Illustrative example
- Receiver Operating Characteristics (ROC)
- Area under the ROC curve (AUC)
- Precision-Recall (PR) curve
- Biometrics measures
- Conclusions

- **Accuracy** (Acc) is one of the most commonly used measures for the classification performance

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

where P and N indicate the number of positive and negative samples, respectively.

- The complement of the accuracy metric is the **Error rate** (ERR) or *misclassification rate*. This metric represents the number of misclassified samples from both positive and negative classes, and it is calculated as follows,
$EER = 1 - Acc = (FP + FN)/(TP + TN + FP + FN)$
- Both accuracy and error rate metrics are sensitive to the imbalanced data
- Another problem with the accuracy is that two classifiers can yield the same accuracy but perform differently with respect to the types of correct and incorrect decisions they provide

- **Sensitivity**, *True positive rate* (TPR), *hit rate*, or *recall*, of a classifier represents the positive correctly classified samples to the total number of positive samples, and it is estimated
- **specificity**, *True negative rate* (TNR), or inverse recall is expressed as the ratio of the correctly classified negative samples to the total number of negative samples

$$TPR = \frac{TP}{TP + FN} = \frac{TP}{P} \quad , \quad TNR = \frac{TN}{FP + TN} = \frac{TN}{N}$$

- Generally, we can consider sensitivity and specificity as two kinds of accuracy, where the first for actual positive samples and the second for actual negative samples
- Sensitivity depends on TP and FN which are in the same column of the confusion matrix, and similarly, the specificity metric depends on TN and FP which are in the same column; hence, both sensitivity and specificity can be used for evaluating the classification performance with imbalanced data

- The accuracy can also be defined in terms of sensitivity and specificity

$$
\begin{aligned}
Acc &= \frac{TP + TN}{TP + TN + FP + FN} \\
&= TPR \times \frac{P}{P+N} + TNR \times \frac{N}{P+N} \\
&= \frac{TP}{TP+FN}\frac{P}{P+N} + \frac{TN}{TN+FP}\frac{N}{P+N} \\
&= \frac{TP}{P+N} + \frac{TN}{P+N} = \frac{TP+TN}{TP+TN+FP+FN}
\end{aligned}
$$

- **False positive rate** (FPR) is also called *false alarm rate* (FAR), or Fallout, and it represents the ratio between the incorrectly classified negative samples to the total number of negative samples. It is the proportion of the negative samples that were incorrectly classified. Hence, it complements the specificity

$$FPR = 1 - TNR = \frac{FP}{FP + TN} = \frac{FP}{N}$$

- The **False negative rate** (FNR) or *miss rate* is the proportion of positive samples that were incorrectly classified. Thus, it complements the sensitivity measure

$$FNR = 1 - TPR = \frac{FN}{FN + TP} = \frac{FN}{P}$$

- Both FPR and FNR are not sensitive to changes in data distributions and hence both metrics can be used with imbalanced data

- Predictive values (positive and negative) reflect the performance of the prediction
- **Positive prediction value** (PPV) or *precision* represents the proportion of positive samples that were correctly classified to the total number of positive predicted samples

$$PPV = \text{Precision} = \frac{TP}{FP + TP} = 1 - FDR$$

- **Negative predictive value** (NPV), inverse precision, or true negative accuracy (TNA) measures the proportion of negative samples that were correctly classified to the total number of negative predicted samples

$$NPV = \frac{TN}{FN + TN} = 1 - FOR$$

- These two measures are sensitive to the imbalanced data
- *False discovery rate* (FDR) and *False omission rate* (FOR) measures complements the PPV and NPV, respectively

- The accuracy can be defined in terms of precision and inverse precision

$$
\begin{aligned}
Acc &= \frac{TP+FP}{P+N} \times PPV + \frac{TN+FN}{P+N} \times NPV \\
&= \frac{TP+FP}{P+N} \times \frac{TP}{TP+FP} + \frac{TN+FN}{P+N} \times \frac{TN}{TN+FN} \\
&= \frac{TP+TN}{TP+TN+FP+FN}
\end{aligned}
$$

- The **likelihood ratio** combines both sensitivity and specificity, and it is used in diagnostic tests
- In that tests, not all positive results are true positives and also the same for negative results; hence, the positive and negative results change the probability/likelihood of diseases. Likelihood ratio measures the influence of a result on the probability
- **Positive likelihood** $(LR+)$ measures how much the odds of the disease increases when a diagnostic test is positive. Similarly, **Negative likelihood** $(LR-)$ measures how much the odds of the disease decreases when a diagnostic test is negative

$$LR+ = \frac{TPR}{1-TNR} = \frac{TPR}{FPR} \ , \ LR- = \frac{1-TPR}{TNR}$$

- Both measures depend on the sensitivity and specificity measures; thus, they are suitable for balanced and imbalanced data

- Both $LR+$ and $LR-$ are combined into one measure which summarizes the performance of the test, this measure is called *Diagnostic odds ratio* (DOR)
- The DOR metric represents the ratio between the positive likelihood ratio to the negative likelihood ratio
- This measure is utilized for estimating the discriminative ability of the test and also for comparing between two diagnostic tests

$$DOR = \frac{LR+}{LR-} = \frac{TPR}{1-TNR} \times \frac{TNR}{1-TPR} = \frac{TP \times TN}{FP \times FN}$$

- **Youden's index** (YI) or *Bookmaker Informedness* (BM) metric is one of the well-known diagnostic tests
- It evaluates the discriminative power of the test
- The formula of Youden's index combines the sensitivity and specificity as in the DOR metric

$$YI = TPR + TNR - 1$$

- The YI metric is ranged from zero when the test is poor to one which represents a perfect diagnostic test. It is also suitable with imbalanced data
- One of the major disadvantages of this test is that it does not change concerning the differences between the sensitivity and specificity of the test
  - For example, given two tests, the sensitivity values for the first and second tests are 0.7 and 0.9, respectively, and the specificity values for the first and second tests are 0.8 and 0.6, respectively; the YI value for both tests is 0.5

There are many different metrics that can be calculated from the previous metrics

- **1- Matthews correlation coefficient** (MCC): this metric represents the correlation between the observed and predicted classifications, and it is calculated directly from the confusion matrix

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$
$$= \frac{\frac{TP}{N} - TPR \times PPV}{\sqrt{PPV \times TPR(1 - TPR)(1 - PPV)}}$$

  - A coefficient of $+1$ indicates a perfect prediction, -1 represents total disagreement between prediction and true values and zero means that no better than random prediction
  - This metric is sensitive to imbalanced data

- **2-Discriminant power** (DP): this measure depends on the sensitivity and specificity
- This metric evaluates how well the classification model distinguishes between positive and negative samples

$$DP = \frac{\sqrt{3}}{\pi}(log(\frac{TPR}{1-TNR}) + log(\frac{TNR}{1-TPR}))$$

  - Since this metric depends on the sensitivity and specificity metrics; it can be used with imbalanced data.

- **3-F-measure:** this is also called F_1-score, and it represents the harmonic mean of precision and recall

$$F\text{-measure} = \frac{2PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$$

- The value of F-measure is ranged from zero to one, and high values of F-measure indicate high classification performance

- F-measure has another variant which is called F_β-measure. This variant represents the weighted harmonic mean between precision and recall

$$F_\beta\text{-measure} = (1+\beta^2)\frac{PPV.TPR}{\beta^2 PPV + TPR} = \frac{(1+\beta^2)TP}{(1+\beta^2)TP + \beta^2 FN + FP}$$

- This metric is sensitive to changes in data distributions
  - Assume that the negative class samples are increased by α times; thus, the F-measure is calculated as follows, $F\text{-measure} = \frac{2TP}{2TP+\alpha FP+\alpha FN}$ and hence this metric is affected by the changes in the class distribution

- The F-measures used only three of the four elements of the confusion matrix and hence two classifiers with different TNR values may have the same F-score
- Therefore, the **Adjusted F-measure (AGF)** metric is introduced to use all elements of the confusion matrix and provide more weights to samples which are correctly classified in the minority class

$$AGF = \sqrt{F_2.InvF_{0.5}}$$

where F_2 is the F-measure where $\beta = 2$ and $InvF_{0.5}$ is calculated by building a new confusion matrix where the class label of each sample is switched (i.e. positive samples become negative and vice versa)

- **4-Markedness**(MK): this is defined based on PPV and NPV metrics

$$MK = PPV + NPV - 1$$

- This metric sensitive to data changes and hence it is not suitable for imbalanced data
- This is because the Markedness metric depends on PPV and NPV metrics and both PPV and NPV are sensitive to changes in data distributions

- **5-Balanced classification rate** or **balanced accuracy** (BCR): this metric combines the sensitivity and specificity metrics

$$BCR = \frac{1}{2}(TPR + TNR) = \frac{1}{2}(\frac{TP}{TP + FN} + \frac{TN}{TN + FP})$$

- Also, *Balance error rate* (BER) or *Half total error rate* ($HTER$) represents $1 - BCR$
- Both BCR and BER metrics can be used with imbalanced datasets

- **6-Geometric Mean** (GM): The main goal of all classifiers is to improve the sensitivity, without sacrificing the specificity
- The *Geometric Mean* (GM) metric aggregates both sensitivity and specificity measures

$$GM = \sqrt{TPR \times TNR} = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}}$$

- *Adjusted Geometric Mean* (AGM) is proposed to obtain as much information as possible about each class

$$AGM = \begin{cases} \frac{GM+TNR(FP+TN)}{1+FP+TN} & \text{if } TPR > 0 \\ 0 & \text{if } TPR = 0 \end{cases}$$

- GM metric can be used with imbalanced datasets
- However, changing the distribution of negative class has a small influence on the AGM metric and hence it is not suitable with the imbalanced data
  - For example, assume the negative class samples are increased by α times. Thus, the AGM metric is calculated as follows, $AGM = \frac{GM+TNR(\alpha FP+\alpha TN)}{1+\alpha FP+\alpha TN}$; as a consequence, the AGM metric is slightly affected by the changes in the class distribution

- **7-Optimization precision** (OP)

$$OP = Acc - \frac{|TPR - TNR|}{TPR + TNR}$$

where the second term $\frac{|TPR-TNR|}{TPR+TNR}$ computes how balanced both class accuracies are and this metric represents the difference between the global accuracy and that term

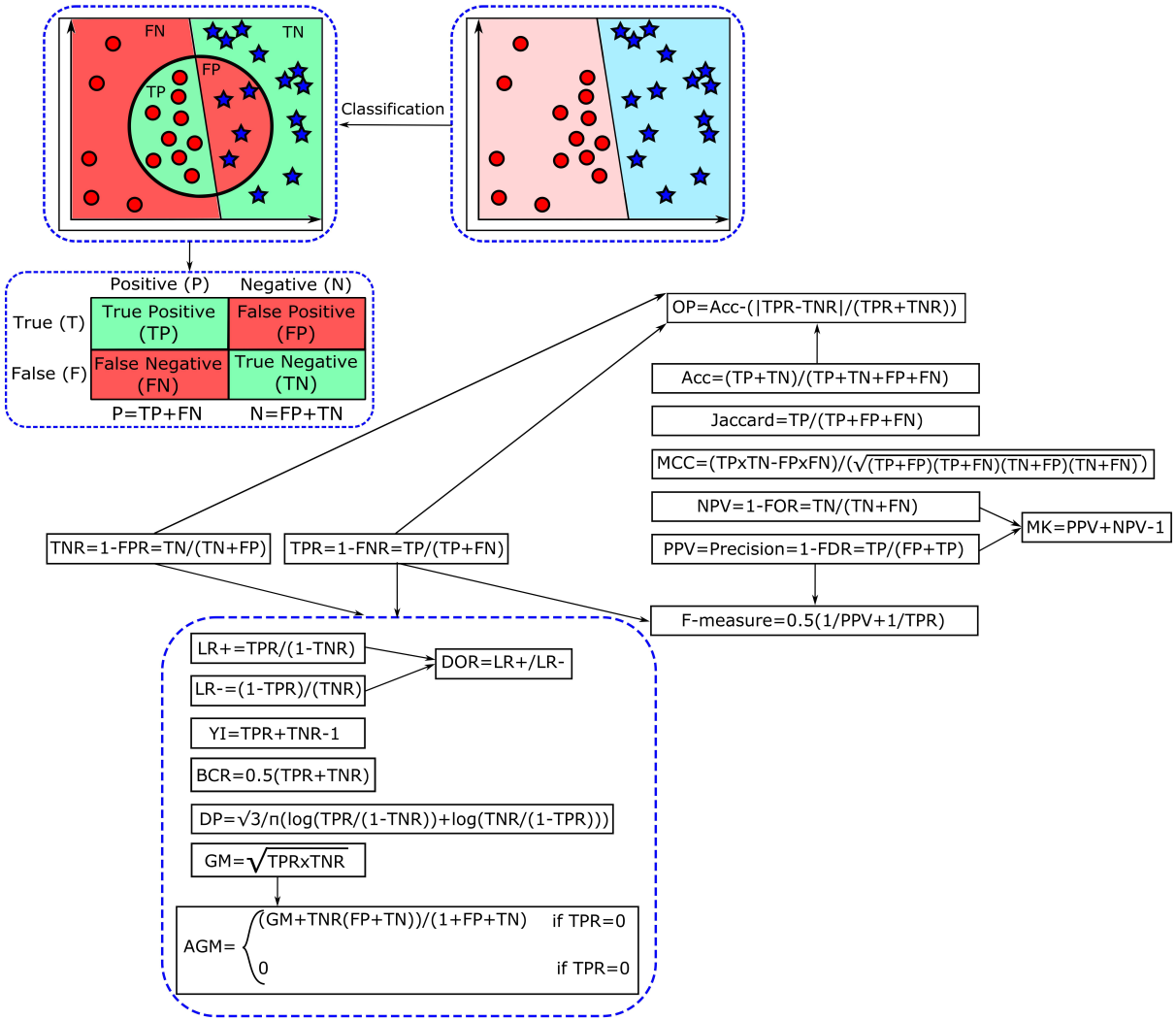
- High OP value indicates high accuracy and well-balanced class accuracies
- Since the OP metric depends on the accuracy metric, it is not suitable for imbalanced data

- **8-Jaccard**: This metric is also called Tanimoto similarity coefficient
- Jaccard metric explicitly ignores the correct classification of negative samples as follows

$$Jaccard = \frac{TP}{TP + FP + FN}$$

- Jaccard metric is sensitive to changes in data distributions

Visualization of different metrics and the relations between these metrics. Given two classes, red class and blue class. The black circle represents a classifier that classifies the sample inside the circle as red samples (belong to the red class) and the samples outside the circle as blue samples (belong to the blue class). Green regions indicate the correctly classified regions and the red regions indicate the misclassified regions.



- Introduction
- Classification Performance
- Classification metrics with imbalanced data
- Different Assessment methods
- Illustrative example
- Receiver Operating Characteristics (ROC)
- Area under the ROC curve (AUC)
- Precision-Recall (PR) curve
- Biometrics measures
- Conclusions

In this section, two examples are introduced. These examples explain how to calculate classification metrics using **two classes** or **multiple classes**

Binary classification example:

- Assume we have two classes (A and B), i.e., binary classification, and each class has 100 samples
- The A class represents the positive class while the B class represents the negative class. The number of correctly classified samples in class A and B are 70 and 80, respectively. Hence, the values of TP, TN, FP, and FN are 70, 80, 20, and 30, respectively
- The values of different classification metrics are as follows,
  $Acc = \frac{70+80}{70+80+20+30} = 0.75$, $TPR = \frac{70}{70+30} = 0.7$,
  $TNR = \frac{80}{80+20} = 0.8$, $PPV = \frac{70}{70+20} \approx 0.78$, $NPV = \frac{80}{80+30} \approx 0.73$,
  $Err = 1 - Acc = 0.25$, $BCR = \frac{1}{2}(0.7 + 0.8) = 0.75$,
  $FPR = 1 - 0.8 = 0.2$, $FNR = 1 - 0.7 = 0.3$,
  $F - measure = \frac{1}{2}(\frac{1}{0.78} + \frac{1}{0.7}) \approx 1.36$,
  $OP = Acc - \frac{|TPR-TNR|}{TPR+TNR} = 0.75 - \frac{|0.7-0.8|}{0.7+0.8} \approx 0.683$,
  $LR+ = \frac{0.7}{1-0.8} = 3.5$, $LR- = \frac{1-0.7}{0.8} = 0.375$, $DOR = \frac{3.5}{0.375} \approx 9.33$,
  $YI = 0.7 + 0.8 - 1 = 0.5$, and $Jaccard = \frac{70}{70+20+30} \approx 0.583$

Imbalanced data:

- We increased the number of samples of the B class to 1000 to show how the classification metrics are changed when using imbalanced data, and there are 800 samples from class B were correctly classified
- The values of TP, TN, FP, and FN are 70, 800, 200, and 30, respectively
- Only the values of accuracy, precision/PPV, NPV, error rate, Optimization precision, F-measure, and Jaccard are changed as follows, $Acc = \frac{70+800}{70+800+200+30} \approx 0.79$, $PPV = \frac{70}{70+200} \approx 0.26$, $NPV = \frac{800}{800+30} \approx 0.96$, $Err = 1 - Acc = 0.21$, $OP = 0.79 - \frac{|0.7-0.8|}{0.7+0.8} \approx 0.723$, $F - measure = \frac{1}{2}(\frac{1}{0.26} + \frac{1}{0.7}) \approx 2.64$, and $Jaccard = \frac{70}{70+200+30} \approx 0.233$
- Thus, the accuracy, precision, NPV, F-measure, and Jaccard metrics are sensitive to imbalanced data

Multi-classification example:

Predicted Class \ True Class	A	B	C
A	80	15	0
B	15	70	10
C	5	15	90

Figure: Results of a multi-class classification test (our example).

- In this example, there are three classes A, B, and C, the results of a classification test are shown in the figure
- The values of TP_A, TP_B, and TB_C are 80, 70, and 90, respectively, which represent the diagonal
- The values of false negative for each class (true class) are calculated as mentioned before by adding all errors in the column of that class.

- For example, $FN_A = E_{AB} + E_{AC} = 15 + 5 = 20$, and similarly $FN_B = E_{BA} + E_{BC} = 15 + 15 = 30$ and $FN_C = E_{CA} + E_{CB} = 0 + 10 = 10$. The values of false positive for each class (predicted class) are calculated as mentioned before by adding all errors in the row of that class
- For example, $FP_A = E_{BA} + E_{CA} = 15 + 0 = 15$, and similarly $FP_B = E_{AB} + E_{CB} = 15 + 10 = 25$ and $FP_C = E_{AC} + E_{BC} = 5 + 15 = 20$
- The value of true negative for the class A (TN_A) can be calculated by adding all columns and rows excluding the row and column of class A; this is similar to the TN in the 2×2 confusion matrix
- The value of TN_A is calculated as follows, $TN_A = 70 + 90 + 10 + 15 = 185$, and similarly $TN_B = 15 + 10 + 90 + 5 = 120$ and $TN_C = 15 + 15 + 70 + 5 = 105$. Using TP, TN, FP, and FN we can calculate all classification measures

- The accuracy is $\frac{80+70+90}{100+100+100} = 0.8$
- The sensitivity and specificity are calculated for each class. For example, the sensitivity of A is $\frac{TP_A}{TP_A+FN_A} = \frac{80}{80+15+5} = 0.8$, and similarly the sensitivity of B and C classes are $\frac{70}{70+15+15} = 0.7$ and $\frac{90}{90+0+10} = 0.9$, respectively, and the specificity values of A, B, and C are $\frac{185}{185+15} \approx 0.93$, $\frac{120}{120+25} \approx 0.82$, and $\frac{105}{105+20} = 0.84$, respectively

- Introduction
- Classification Performance
- Classification metrics with imbalanced data
- Different Assessment methods
- Illustrative example
- Receiver Operating Characteristics (ROC)
- Area under the ROC curve (AUC)
- Precision-Recall (PR) curve
- Biometrics measures
- Conclusions

- The *receiver operating characteristics* (ROC) curve is a two-dimensional graph in which the TPR represents the y-axis and FPR is the x-axis
- Any classifier that has discrete outputs such as decision trees is designed to produce only a class decision, i.e., a decision for each testing sample, and hence it generates only one confusion matrix which in turn corresponds to one point into the ROC space
- Continuous output classifiers such as the Naive Bayes classifier, the output is represented by a numeric value, i.e., score, which represents the degree to which a sample belongs to a specific class

- The ROC curve is generated by changing the threshold on the confidence score; hence, each threshold generates only one point in the ROC curve

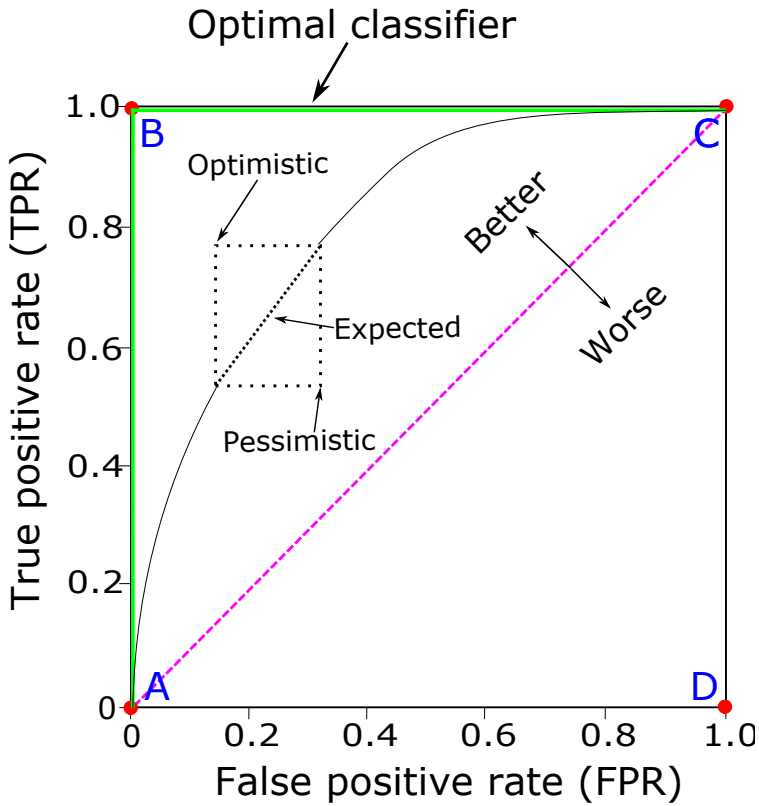
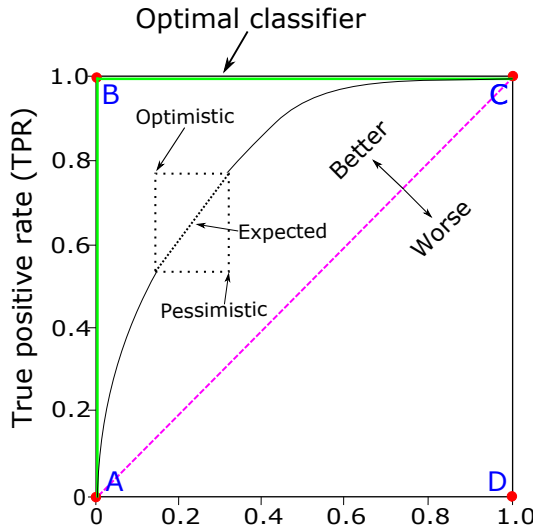


Figure: A basic ROC curve showing important points, and the optimistic, pessimistic and expected ROC segments for equally scored samples.

- There are four important points in the ROC curve
  - The point A, in the lower left corner $(0, 0)$ represents a classifier where there is no positive classification, while all negative samples are correctly classified and hence $TPR = 0$ and $FPR = 0$
  - The point C, in the top right corner (1,1), represents a classifier where all positive samples are correctly classified, while the negative samples are misclassified
  - The point D in the lower right corner $(1, 0)$ represents a classifier where all positive and negative samples are misclassified
  - The point B in the upper left corner $(0, 1)$ represents a classifier where all positive and negative samples are correctly classified; thus, this point represents the perfect classification or the *Ideal operating point*.



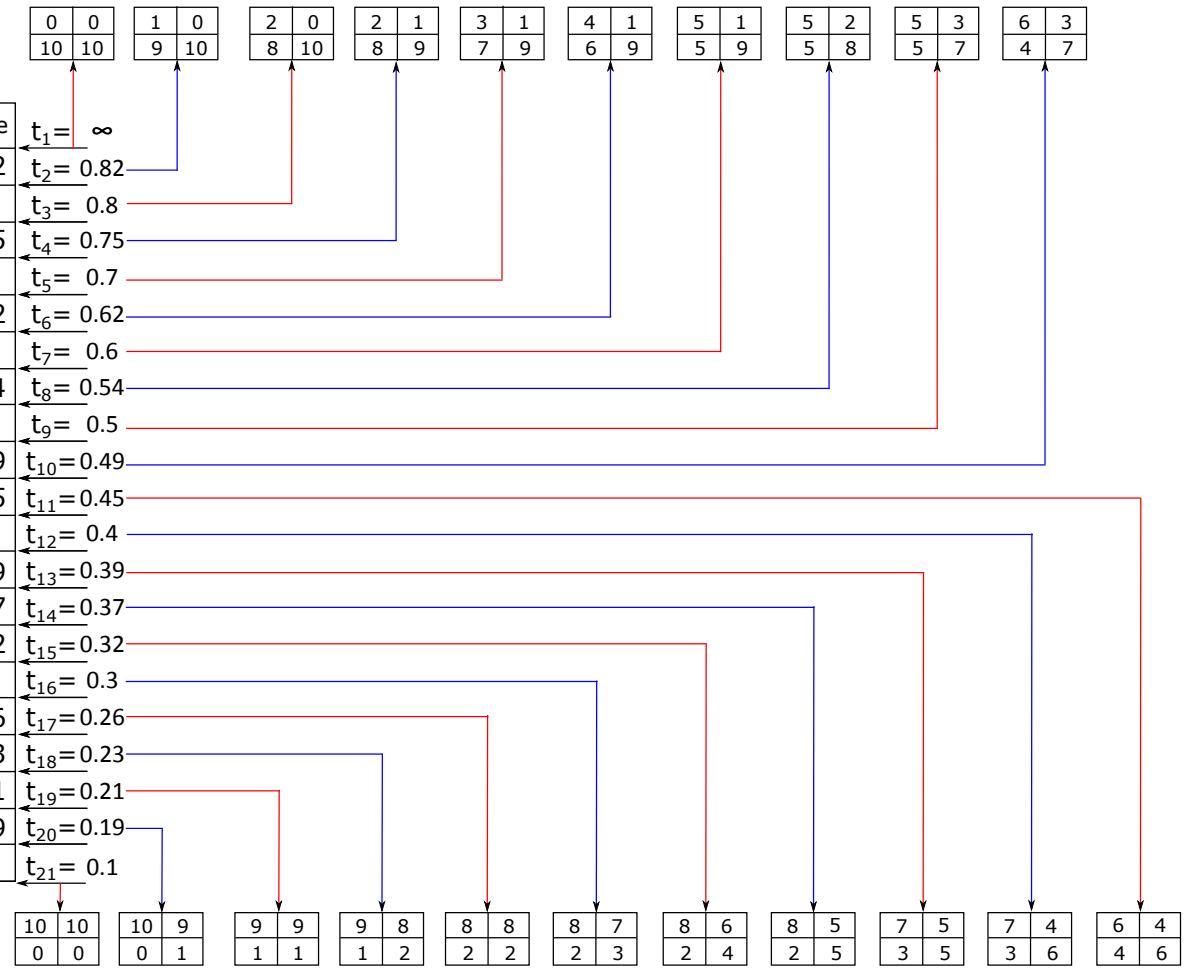
- The perfect classification performance is the green curve which rises vertically from (0,0) to (0,1) and then horizontally to (1,1). This curve reflects that the classifier perfectly ranked the positive samples relative to the negative samples
- A point in the ROC space is better than all other points that are in the southeast, i.e., the points that have lower TPR, higher FPR, or both

- In the example below, a test set consists of 20 samples from two classes; each class has ten samples, i.e., ten positive and ten negative samples

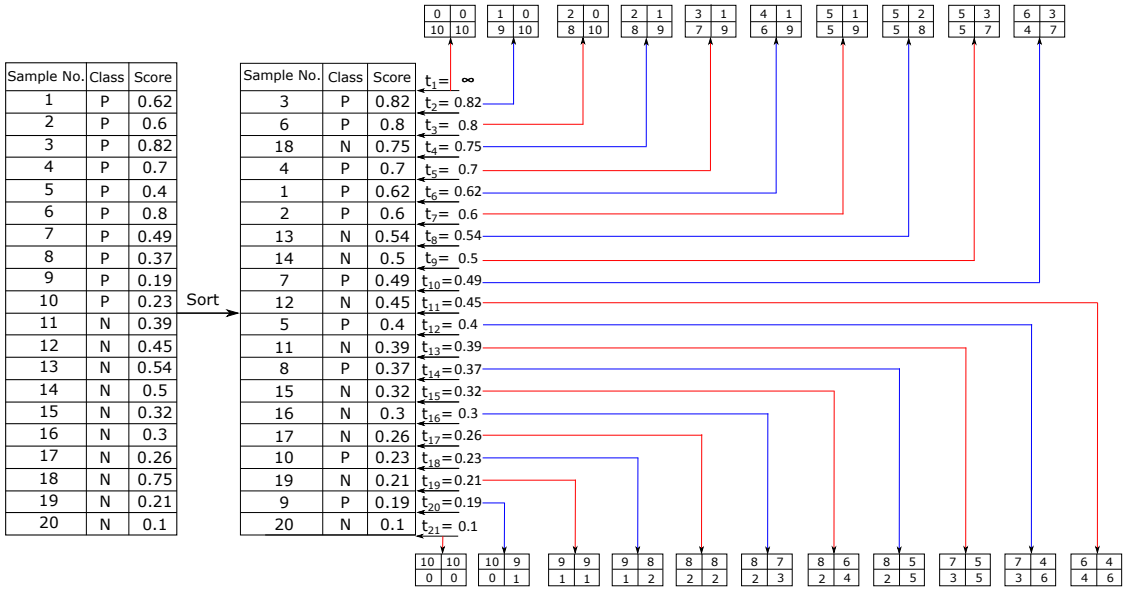
Sample No.	Class	Score
1	P	0.62
2	P	0.6
3	P	0.82
4	P	0.7
5	P	0.4
6	P	0.8
7	P	0.49
8	P	0.37
9	P	0.19
10	P	0.23
11	N	0.39
12	N	0.45
13	N	0.54
14	N	0.5
15	N	0.32
16	N	0.3
17	N	0.26
18	N	0.75
19	N	0.21
20	N	0.1

Sort

Sample No.	Class	Score	
3	P	0.82	$t_2 = 0.82$
6	P	0.8	$t_3 = 0.8$
18	N	0.75	$t_4 = 0.75$
4	P	0.7	$t_5 = 0.7$
1	P	0.62	$t_6 = 0.62$
2	P	0.6	$t_7 = 0.6$
13	N	0.54	$t_8 = 0.54$
14	N	0.5	$t_9 = 0.5$
7	P	0.49	$t_{10} = 0.49$
12	N	0.45	$t_{11} = 0.45$
5	P	0.4	$t_{12} = 0.4$
11	N	0.39	$t_{13} = 0.39$
8	P	0.37	$t_{14} = 0.37$
15	N	0.32	$t_{15} = 0.32$
16	N	0.3	$t_{16} = 0.3$
17	N	0.26	$t_{17} = 0.26$
10	P	0.23	$t_{18} = 0.23$
19	N	0.21	$t_{19} = 0.21$
9	P	0.19	$t_{20} = 0.19$
20	N	0.1	$t_{21} = 0.1$



- The initial step to plot the ROC curve is to sort the samples according to their scores
- Next, the threshold value is changed from maximum to minimum to plot the ROC curve
- To scan all samples, the threshold is ranged from ∞ to $-\infty$
- The samples are classified into the positive class if their scores are higher than or equal the threshold; otherwise, it is estimated as negative



- Changing the threshold value changes the TPR and FPR
- At the beginning, the threshold value is set at maximum ($t_1 = \infty$); hence, all samples are classified as negative samples and the values of FPR and TPR are zeros and the position of t_1 is in the lower left corner (the point (0,0))

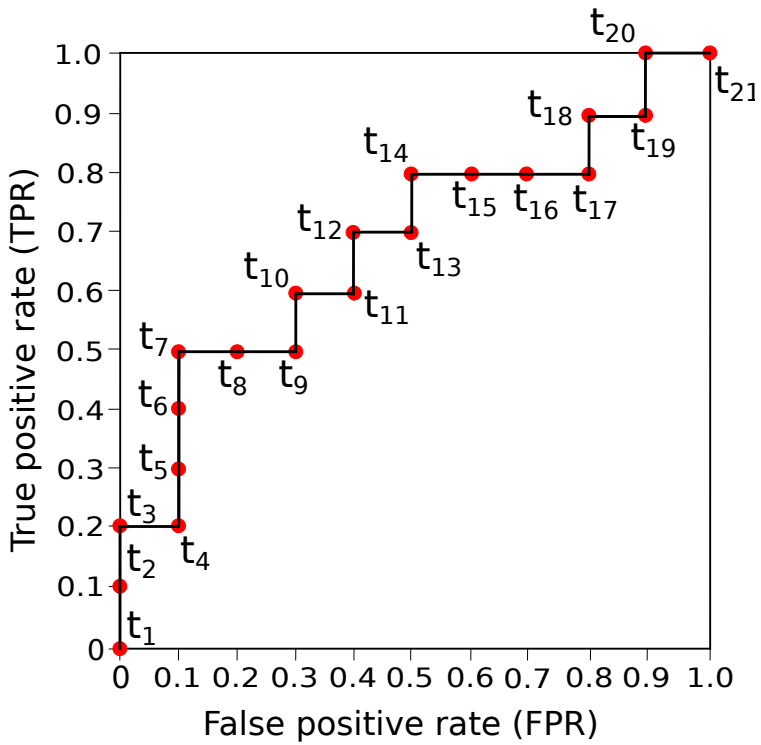
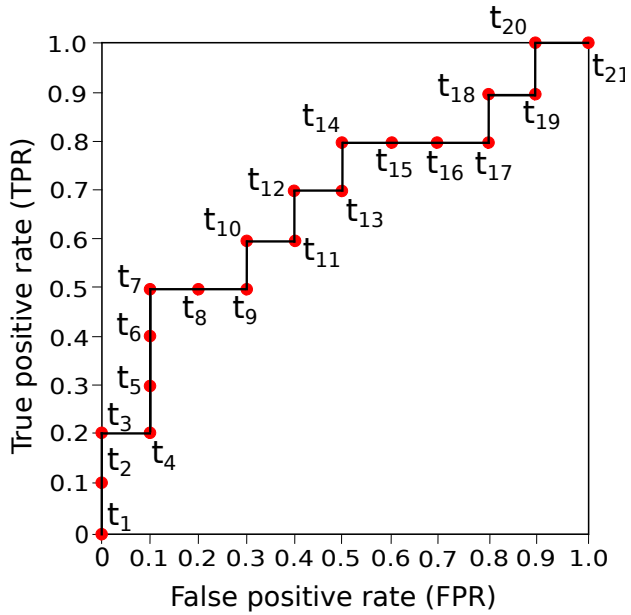


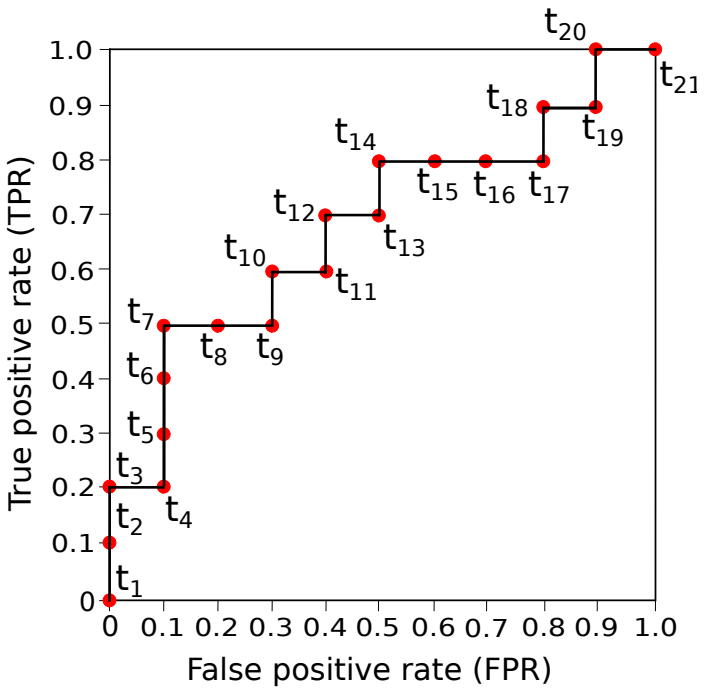
Figure: An illustrative example of the ROC curve. The values of TPR and FPR of each point/threshold are calculated in the next Table.

- Next, the threshold value is decreased to 0.82, and the first sample is classified correctly as a positive sample, and the TPR increased to 0.1, while the FPR remains zero
- Increasing the TPR moves the ROC curve up while increasing the FPR moves the ROC curve to the right as in t_4
- The ROC curve must pass through the point (0,0) where the threshold value is ∞ (in which all samples are classified as negative samples) and the point (1,1) where the threshold is $-\infty$ (in which all samples are classified as positive samples)



Threshold	TP	FN	TN	FP	TPR	FPR	FNR	PPV	Acc
$t_1 = \infty$	0	10	10	0	0	0	1	-	50
$t_2 = 0.82$	1	9	10	0	0.1	0	0.9	1.0	55
$t_3 = 0.80$	2	8	10	0	0.2	0	0.8	1.0	60
$t_4 = 0.75$	2	8	9	1	0.2	0.1	0.8	0.67	55
$t_5 = 0.70$	3	7	9	1	0.3	0.1	0.7	0.75	60
$t_6 = 0.62$	4	6	9	1	0.4	0.1	0.6	0.80	65
$t_7 = 0.60$	5	5	9	1	0.5	0.1	0.5	0.83	70
$t_8 = 0.54$	5	5	8	2	0.5	0.2	0.5	0.71	65
$t_9 = 0.50$	5	5	7	3	0.5	0.3	0.5	0.63	60
$t_{10} = 0.49$	6	4	7	3	0.6	0.3	0.4	0.67	65
$t_{11} = 0.45$	6	4	6	4	0.6	0.4	0.4	0.60	60
$t_{12} = 0.40$	7	3	6	4	0.7	0.4	0.3	0.64	65
$t_{13} = 0.39$	7	3	5	5	0.7	0.5	0.3	0.58	60
$t_{14} = 0.37$	8	2	5	5	0.8	0.5	0.2	0.62	65
$t_{15} = 0.32$	8	2	4	6	0.8	0.6	0.2	0.57	60
$t_{16} = 0.30$	8	2	3	7	0.8	0.7	0.2	0.53	55
$t_{17} = 0.26$	8	2	2	8	0.8	0.8	0.2	0.50	50
$t_{18} = 0.23$	9	1	2	8	0.9	0.8	0.1	0.53	55
$t_{19} = 0.21$	9	1	1	9	0.9	0.9	0.1	0.50	50
$t_{20} = 0.19$	10	0	1	9	1.0	0.9	0	0.53	55
$t_{21} = 0.10$	10	0	0	10	1.0	1.0	0	0.50	50

- It is clear that the ROC curve is a step function. This is because we only used 20 samples (a finite set of samples) in our example and a true curve can be obtained when the number of samples increased
- The figure also shows that the best accuracy (70%) (see the Table) is obtained at (0.1,0.5) when the threshold value was ≥ 0.6, rather than at ≥ 0.5 as we might expect with a balanced data. This means that the given learning model identifies positive samples better than negative samples.



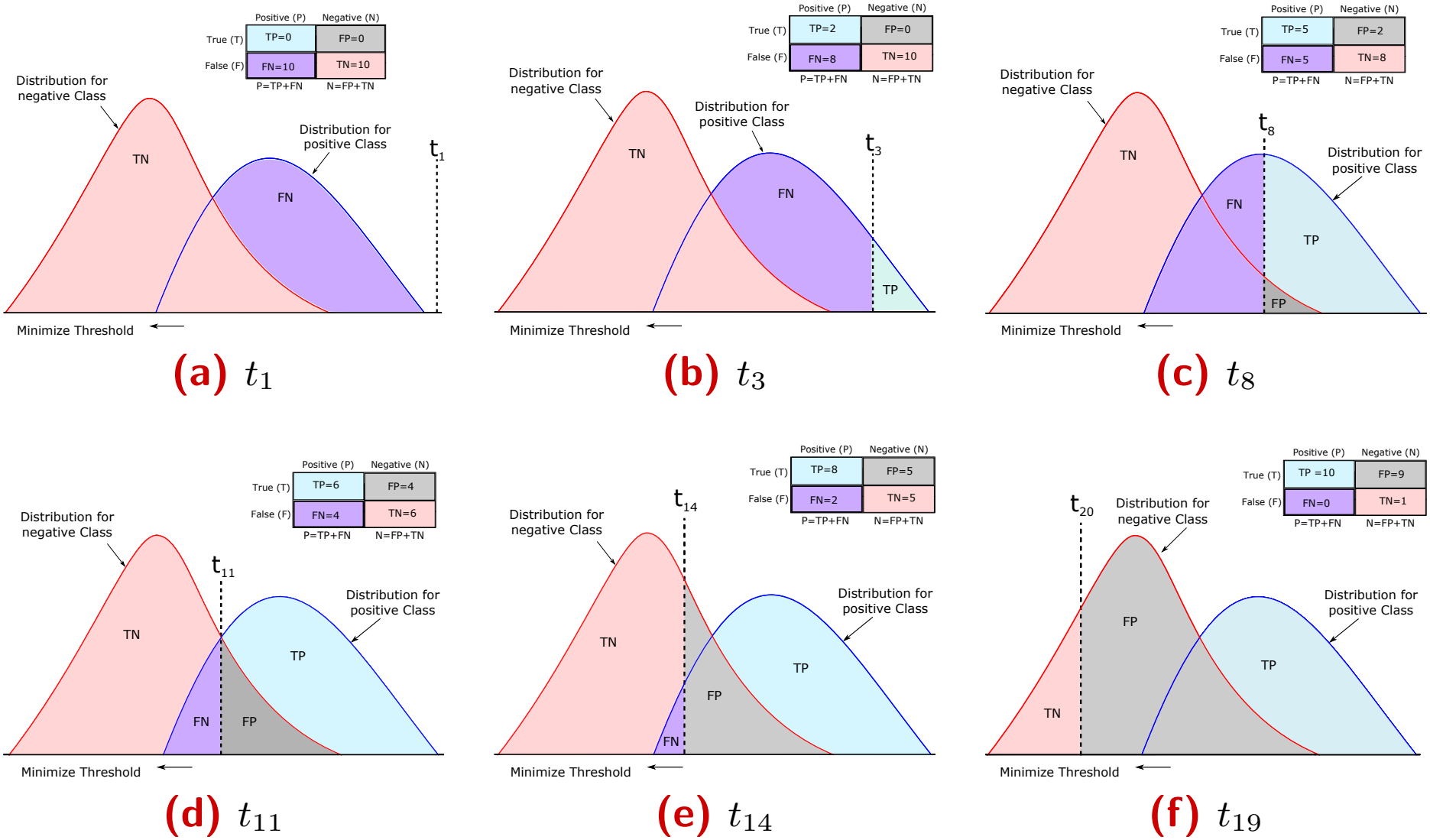
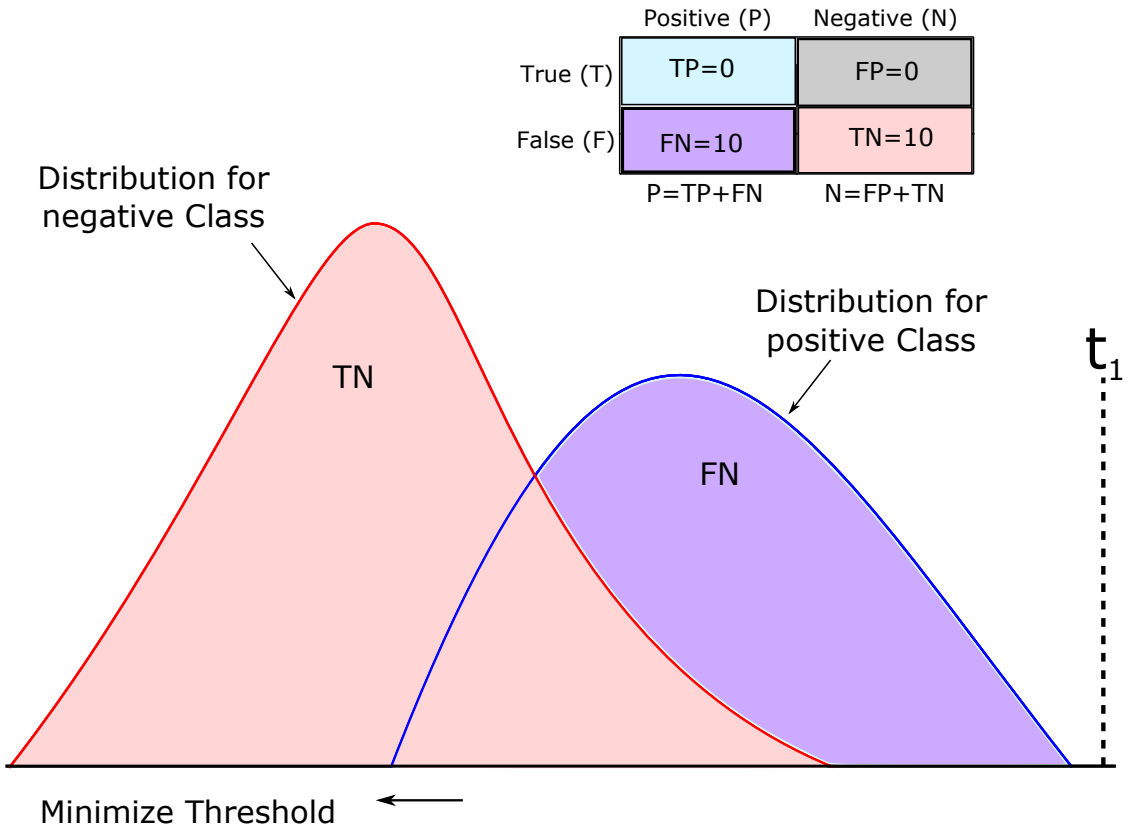
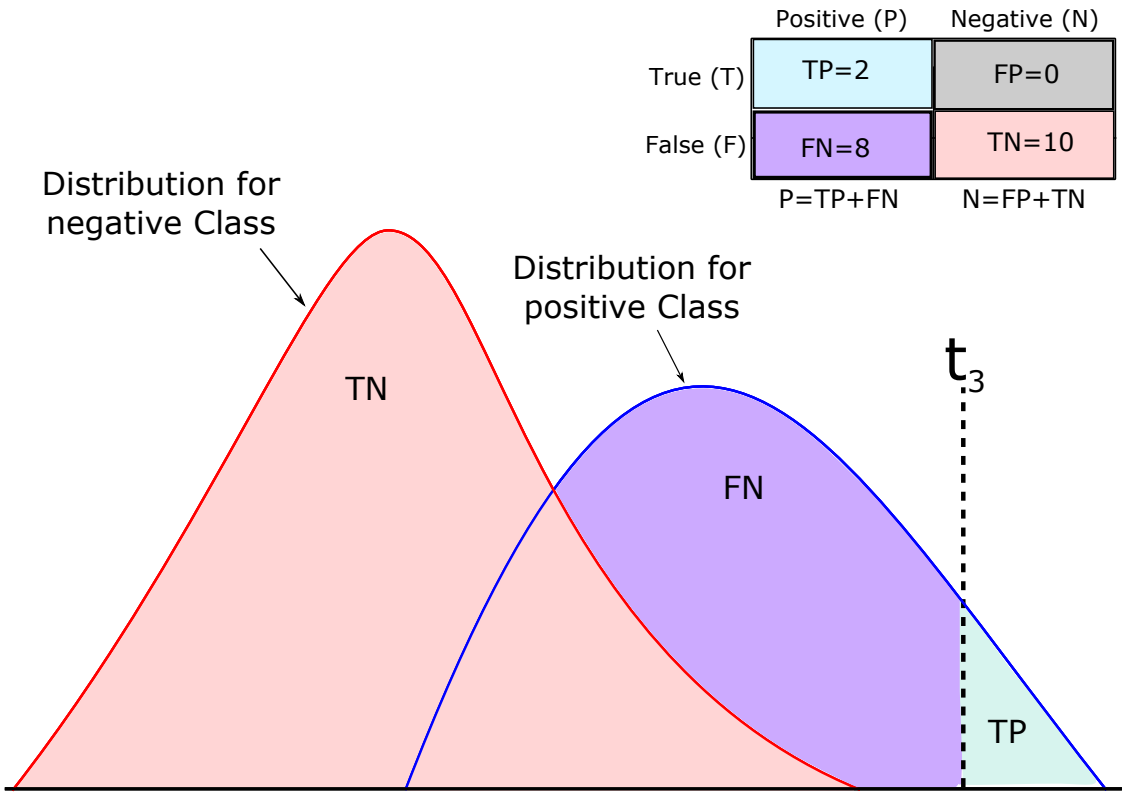


Figure: A visualization of how changing the threshold changes the TP, TN, FP, and FN values.

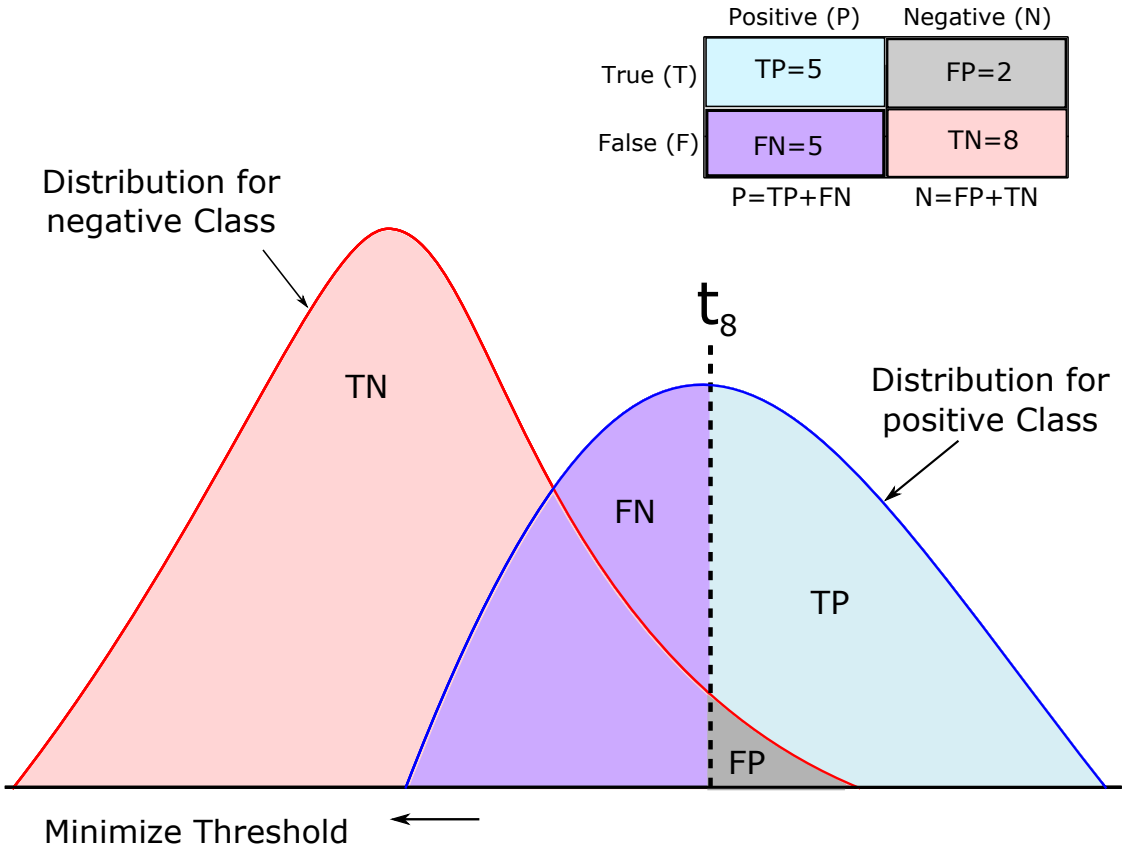
- t_1: The value of this threshold was ∞ and hence all samples are classified as negative samples. This means that (1) all positive samples are incorrectly classified; hence, the value of TP is zero, (2) all negative samples are correctly classified and hence there is no FP



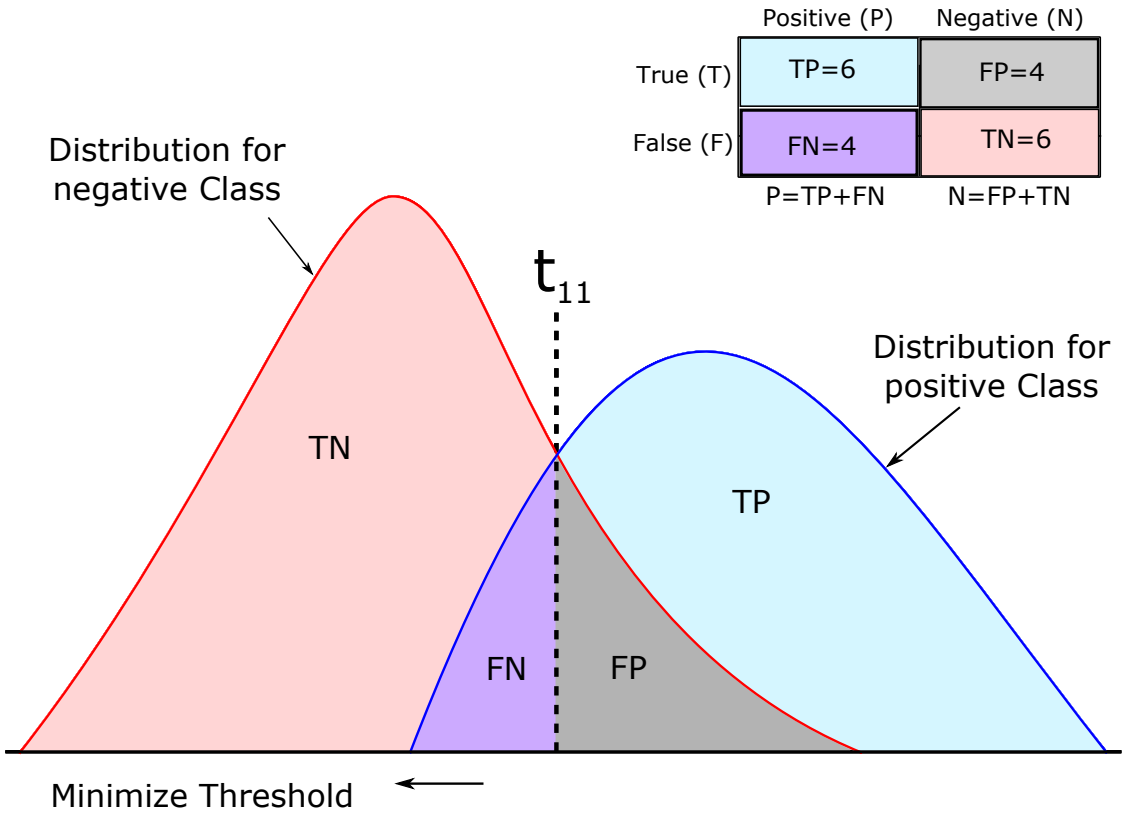
- t_3: The threshold value decreased and as shown there are two positive samples are correctly classified. Therefore, according to the positive class, only the positive samples which have scores more than or equal this threshold (t_3) will be correctly classified, i.e., TP, while the other positive samples are incorrectly classified, i.e., FN. In this threshold, also all negative samples are correctly classified; thus, the value of FP is still zero



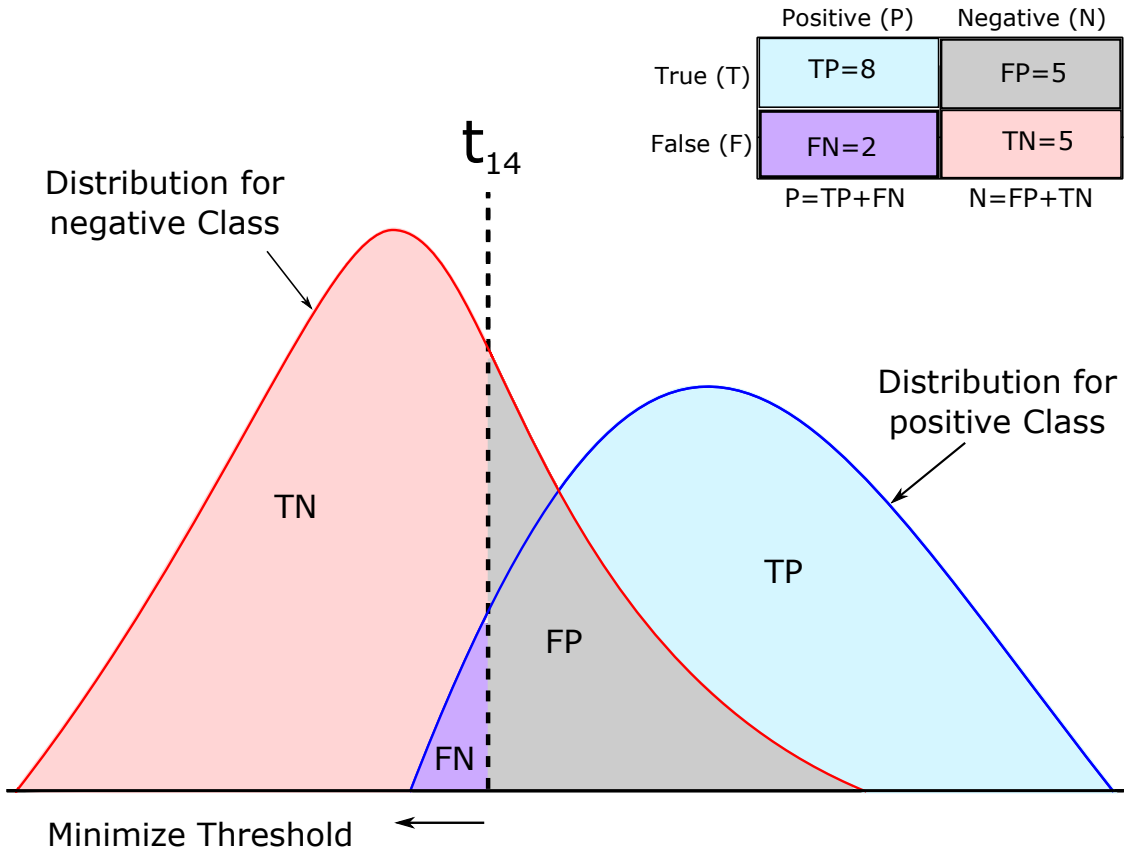
- t_8: As the threshold further decreased to be 0.54, the threshold line moves to the left. This means that more positive samples have the chance to be correctly classified; on the other hand, some negative samples are misclassified. As a consequence, the values of TP and FP are increased and the values of TN and FN decreased



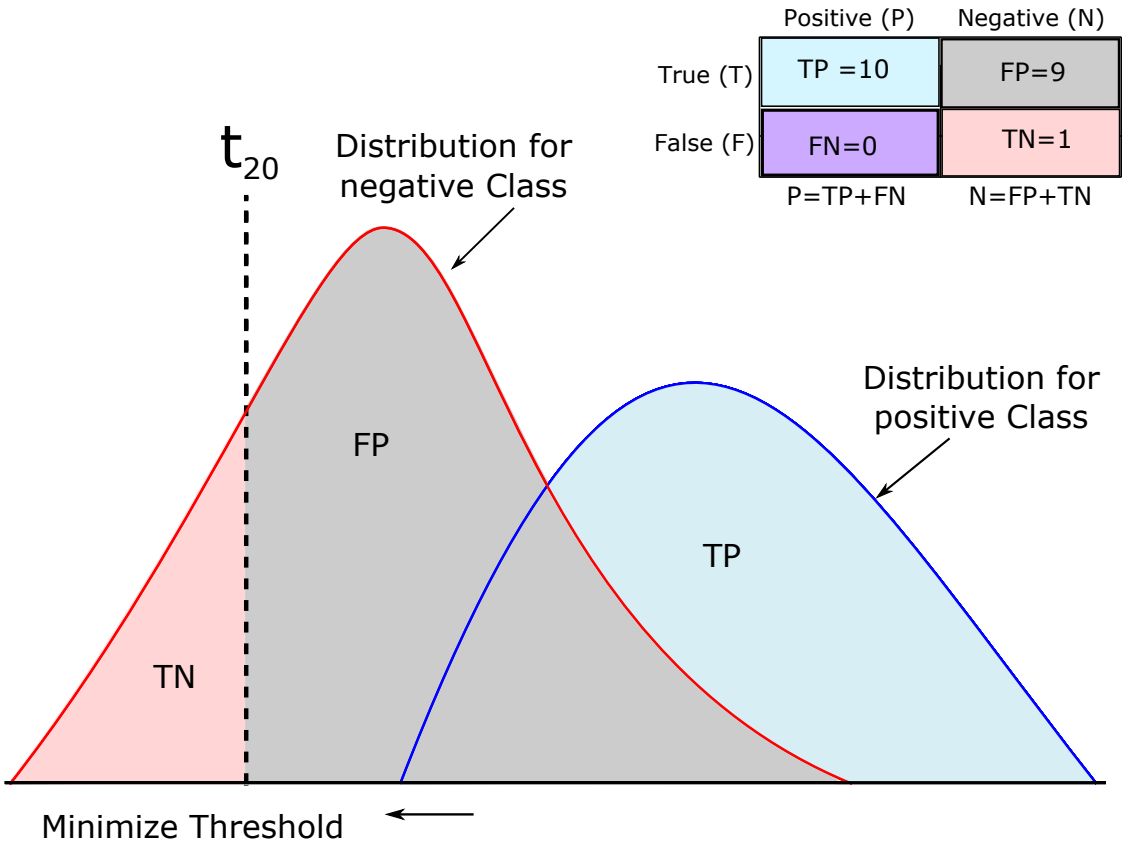
- t_{11}: This is an important threshold value where the numbers of errors from both positive and negative classes are equal (i.e., $TP = TN = 6$ and $FP = FN = 4$)



- t_{14}: Reducing the value of the threshold to 0.37 results more correctly classified positive samples and this increases TP and reduces FN. On the contrary, more negative samples are misclassified and this increases FP and reduces TN



- t_{20}: Decreasing the threshold value hides the FN area. This is because all positive samples are correctly classified. Also, from the figure, it is clear that the FP area is much larger than the area of TN. This is because 90% of the negative samples are incorrectly classified, and only 10% of negative samples are correctly classified



1: Given a set of test samples ($S_{test} = \{s_1, s_2, \ldots, s_N\}$), where N is the total number of test samples, P and N represent the total number of positive and negative samples, respectively.
2: Sort the samples corresponding to their scores, S_{sorted} is the sorted samples.
3: $FP \leftarrow 0$, $TP \leftarrow 0$, $f_{prev} \leftarrow -\infty$, and $ROC = [\,]$.
4: **for** $i = 1$ to $|S_{sorted}|$ **do**
5: **if** $f(i) \neq f_{prev}$ **then**
6: $ROC(i) \leftarrow (\frac{FP}{N}, \frac{TP}{P})$, $f_{prev} \leftarrow f(i)$
7: **end if**
8: **if** $S_{sorted}(i)$ is a positive sample **then**
9: $TP \leftarrow TP + 1$.
10: **else**
11: $FP \leftarrow FP + 1$.
12: **end if**
13: **end for**
14: $ROC(i) \leftarrow (\frac{FP}{N}, \frac{TP}{P})$.

- Introduction
- Classification Performance
- Classification metrics with imbalanced data
- Different Assessment methods
- Illustrative example
- Receiver Operating Characteristics (ROC)
- Area under the ROC curve (AUC)
- Precision-Recall (PR) curve
- Biometrics measures
- Conclusions

- Comparing different classifiers in the ROC curve is not easy. This is because there is no scalar value represents the expected performance
- Therefore, the Area under the ROC curve (AUC) metric is used to calculate the area under the ROC curve
- The AUC score is always bounded between zero and one, and there is no realistic classifier has an AUC lower than 0.5

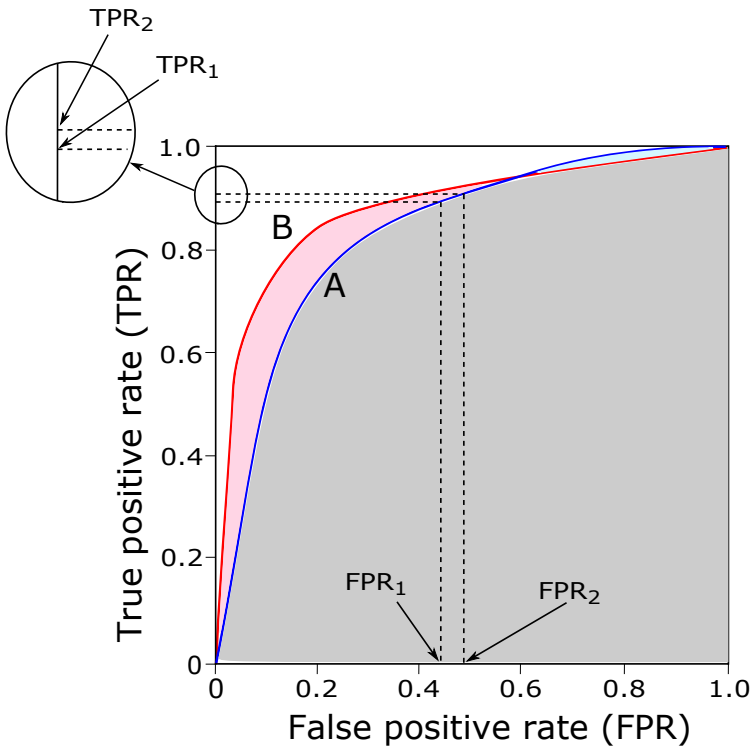
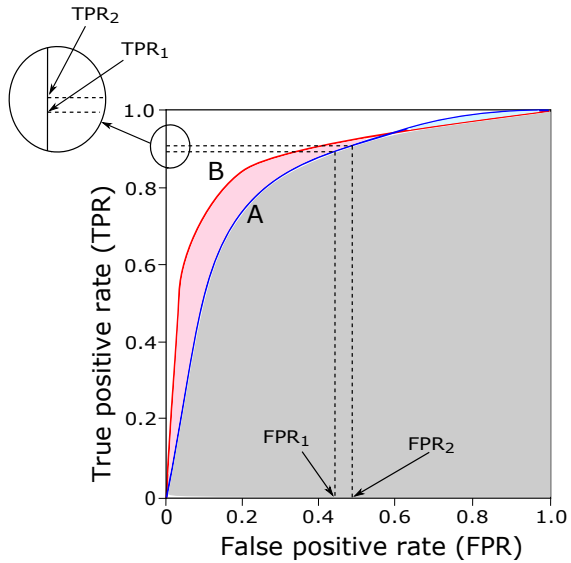


Figure: An illustrative example of the AUC metric.

- The figure below shows the AUC value of two classifiers, A and B, and the AUC of B classifier is greater than A; hence, it achieves better performance
- The gray shaded area is common in both classifiers, while the red shaded area represents the area where the B classifier outperforms the A classifier
- It is possible for a lower AUC classifier to outperform a higher AUC classifier in a specific region. For example, the classifier B outperforms A except at $FPR > 0.6$ where A has a slight difference (blue shaded area)

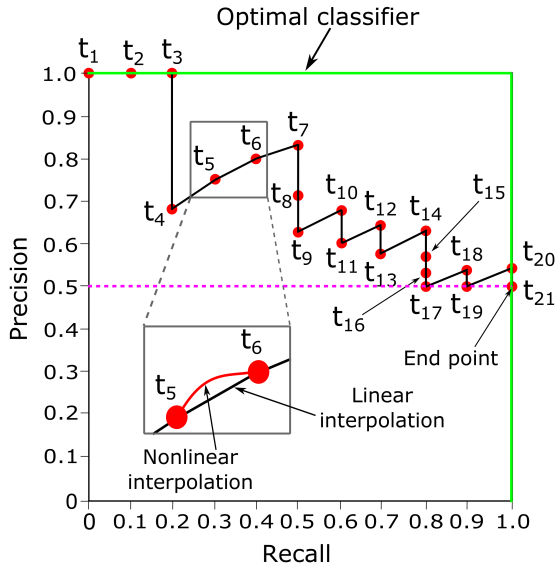


- However, two classifiers with two different ROC curves may have the same AUC score
- The steps in the AUC Algorithm represent a slight modification from the ROC Algorithm. In other words, instead of generating ROC points, while the AUC adds areas of trapezoids[1] of the ROC curve
- The AUC score can be calculated by adding the areas of trapezoids of the AUC measure
- The AUC can be also calculated under the Precision-Recall curve using the trapezoidal rule as in the ROC curve, and the AUC score of the perfect classifier in PR curves is one as in ROC curves.

[1] A trapezoid is a 4-sided shape with two parallel sides.

- Introduction
- Classification Performance
- Classification metrics with imbalanced data
- Different Assessment methods
- Illustrative example
- Receiver Operating Characteristics (ROC)
- Area under the ROC curve (AUC)
- Precision-Recall (PR) curve
- Biometrics measures
- Conclusions

- Precision and recall metrics are widely used for evaluating the classification performance. The Precision-Recall (PR) curve has the same concept of the ROC curve, and it can be generated by changing the threshold as in ROC
- However, the ROC curve shows the relation between sensitivity/recall (TPR) and 1-specificity (FPR) while the PR curve shows the relationship between recall and precision
- Thus, in the PR curve, the x-axis is the recall and the y-axis is the precision, i.e., the x-axis of ROC curve is the y-axis of PR curve. Hence, in the PR curve, there is no need for the TN value



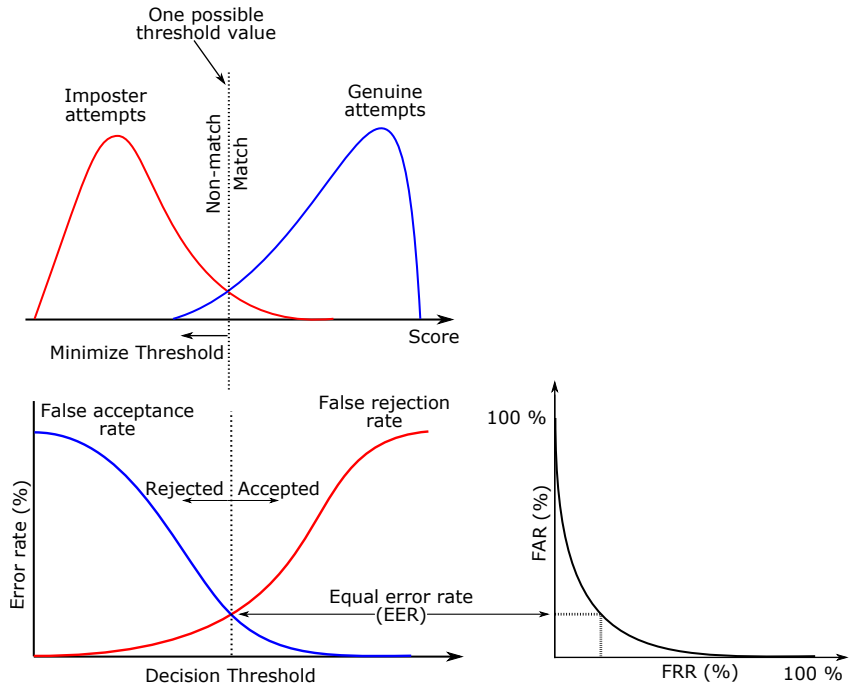
- In the PR curve, the precision value for the first point is undefined because the number of positive predictions is zero, i.e., $TP = 0$ and $FP = 0$. This problem can be solved by estimating the first point in the PR curve from the second point. There are two cases for estimating the first point depending on the value of TP of the second point:
  1. The number of true positives of the second point is zero: In this case, since the second point is (0,0), the first point is also (0,0)
  2. The number of true positives of the second point is not zero: this is similar to our example where the second point is (0.1, 1.0). The first point can be estimated by drawing a horizontal line from the second point to the y-axis. Thus, the first point is estimated as (0.0, 1.0)

- The PR curve is often zigzag curve; hence, PR curves tend to cross each other much more frequently than ROC curves
- In the PR curve, a curve above the other has a better classification performance
- The perfect classification performance in the PR curve is represented by a green curve, and this curve starts from the (0,1) horizontally to (1,1) and then vertically to (1,0), where (0,1) represents a classifier that achieves 100% precision and 0% recall, (1,1) represents a classifier that obtains 100% precision and sensitivity and this is the ideal point in the PR curve, and (1,0) indicates the classifier obtains 100% sensitivity and 0% precision
- The PR curve is to the upper right corner, the better the classification performance is. Since the PR curve depends only on the precision and recall measures, it ignores the performance of correctly handling negative examples (TN)

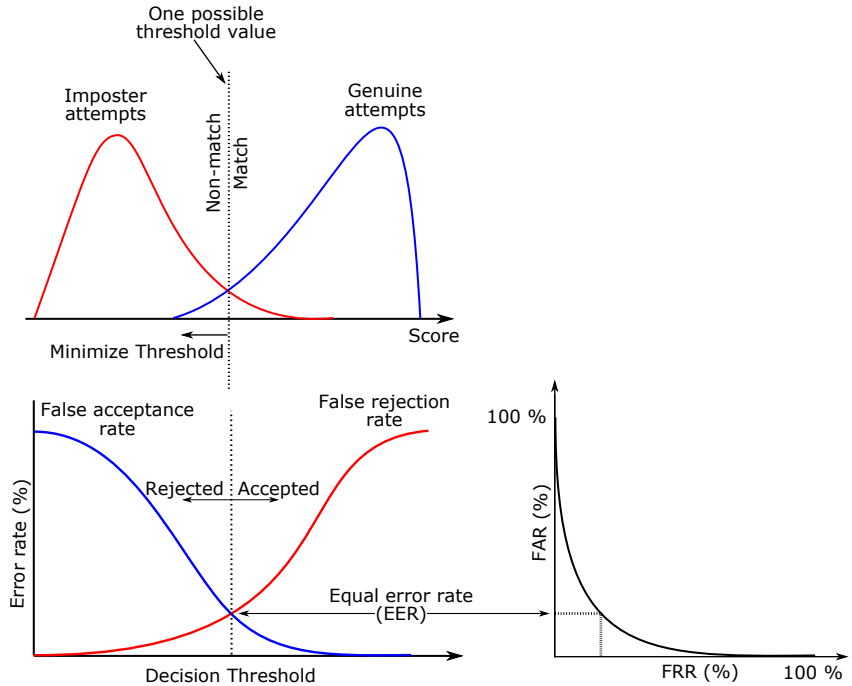
- Introduction
- Classification Performance
- Classification metrics with imbalanced data
- Different Assessment methods
- Illustrative example
- Receiver Operating Characteristics (ROC)
- Area under the ROC curve (AUC)
- Precision-Recall (PR) curve
- Biometrics measures
- Conclusions

- Biometrics matching is slightly different than the other classification problems and hence it is sometimes called two-instance problem
- In this problem, instead of classifying one sample into one of c groups or classes, biometric determines if the two samples are in the same group. This can be achieved by identifying an unknown sample by matching it with all the other known samples
- The model assigns the unknown sample to the person which has the most similar score. If this level of similarity is not reached, the sample is rejected
- Theoretically, scores of clients (persons known by the biometric system) should always be higher than the scores of imposters (persons who are not known by the system)
- In biometric systems, a single threshold separates the two groups of scores, i.e., clients and imposters
- In real applications, sometimes imposter samples generate scores higher than the scores of some client samples. Accordingly, it is a fact that however the classification threshold is perfectly chosen, some classification errors occur

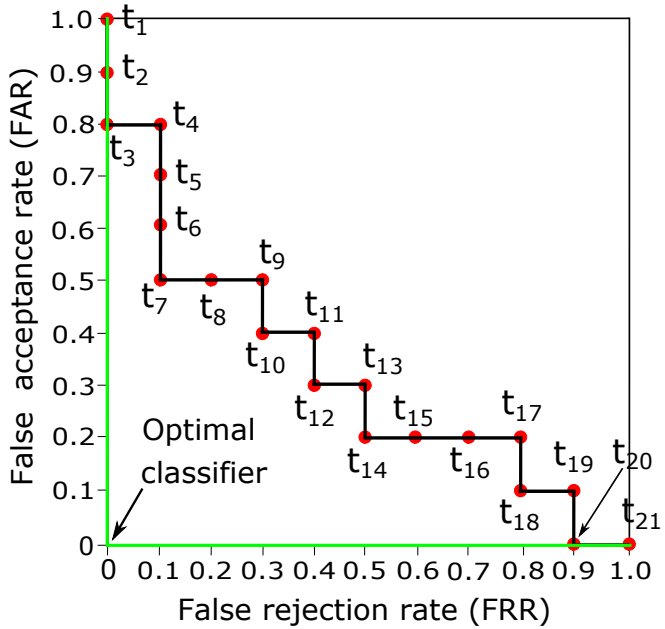
- Two of the most commonly used measures in biometrics are the *False acceptance rate* (FAR) and *False rejection/recognition rate* (FRR)
- The FAR is also called *false match rate* (FMR) and it is the ratio between the number of false acceptance to the total number of imposters attempts
- Hence, to prevent imposter samples from being easily correctly identified by the model, the similarity score has to exceed a certain level



- The FRR or *false non-match rate* ($FNMR$) measures the likelihood that the biometric model will incorrectly reject a client, and it represents the ratio between the number of false recognitions to the total number of clients' attempts
- Equal error rate (EER) measure solves the problem of selecting a threshold value partially, and it represents the failure rate when the values of FMR and $FNMR$ are equal



- Detection Error Trade-off (DET) curve is used for evaluating biometric models
- In this curve, as in the ROC and PR curves, the threshold value is changed and the values of FAR and FRR are calculated at each threshold
- This curve shows the relation between FAR and FRR
- The ideal point in this curve is the origin point where the values of both FRR and FAR are zeros and hence the perfect classification performance in the DET curve is represented by a green curve



- Introduction
- Classification Performance
- Classification metrics with imbalanced data
- Different Assessment methods
- Illustrative example
- Receiver Operating Characteristics (ROC)
- Area under the ROC curve (AUC)
- Precision-Recall (PR) curve
- Biometrics measures
- Conclusions

- For more details, download this paper "Alaa Tharwat Classification Assessment Methods, Applied Computing and Informatics, 2018".
  Link to the paper
  https://www.sciencedirect.com/science/article/pii/S2210832718301546