Confusion matrix

From Wikipedia, the free encyclopedia

In the field of machine learning and specifically the problem of statistical classification, a **confusion matrix**, also known as an error matrix,[10] is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one (in unsupervised learning it is usually called a **matching matrix**). Each row of the matrix represents the instances in an actual class while each column represents the instances in a predicted class, or vice versa – both variants are found in the literature.[11] The name stems from the fact that it makes it easy to see whether the system is confusing two classes (i.e. commonly mislabeling one as another).

It is a special kind of contingency table, with two dimensions ("actual" and "predicted"), and identical sets of "classes" in both dimensions (each combination of dimension and class is a variable in the contingency table).

Contents [hide]

Terminology and derivations from a confusion matrix

condition positive (P)
the number of real positive cases in the data

condition negative (N)
the number of real negative cases in the data

true positive (TP)
A test result that correctly indicates the presence of a condition or characteristic

true negative (TN)
A test result that correctly indicates the absence of a condition or characteristic

false positive (FP)
A test result which wrongly indicates that a particular condition or attribute is present

false negative (FN)
A test result which wrongly indicates that a particular condition or attribute is absent

sensitivity, recall, hit rate, or true positive rate (TPR)

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

specificity, selectivity or true negative rate (TNR)

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

precision or positive predictive value (PPV)

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

negative predictive value (NPV)

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

miss rate or false negative rate (FNR)

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

fall-out or false positive rate (FPR)

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

false discovery rate (FDR)

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

false omission rate (FOR)

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

Positive likelihood ratio (LR+)

$$\mathrm{LR+} = \frac{\mathrm{TPR}}{\mathrm{FPR}}$$

Negative likelihood ratio (LR-)

$$\mathrm{LR-} = \frac{\mathrm{FNR}}{\mathrm{TNR}}$$

prevalence threshold (PT)

$$\mathrm{PT} = \frac{\sqrt{\mathrm{FPR}}}{\sqrt{\mathrm{TPR}}+\sqrt{\mathrm{FPR}}}$$

threat score (TS) or critical success index (CSI)

$$\mathrm{TS} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}}$$

Prevalence

$$\frac{\mathrm{P}}{\mathrm{P}+\mathrm{N}}$$

accuracy (ACC)

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{P}+\mathrm{N}} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$$

balanced accuracy (BA)

$$\mathrm{BA} = \frac{TPR+TNR}{2}$$

F1 score

is the harmonic mean of precision and sensitivity:

$$\mathrm{F}_1 = 2\times\frac{\mathrm{PPV}\times\mathrm{TPR}}{\mathrm{PPV}+\mathrm{TPR}} = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}$$

phi coefficient (φ or r_φ) or Matthews correlation coefficient (MCC)

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

Fowlkes–Mallows index (FM)

$$\mathrm{FM} = \sqrt{\frac{TP}{TP+FP}\cdot\frac{TP}{TP+FN}} = \sqrt{PPV\times TPR}$$

informedness or bookmaker informedness (BM)

$$\mathrm{BM} = \mathrm{TPR}+\mathrm{TNR}-1$$

markedness (MK) or deltaP (Δp)

$$\mathrm{MK} = \mathrm{PPV}+\mathrm{NPV}-1$$

Diagnostic odds ratio (DOR)

$$\mathrm{DOR} = \frac{\mathrm{LR+}}{\mathrm{LR-}}$$

Sources: Fawcett (2006),[1] Piryonesi and El-Diraby (2020),[2] Powers (2011),[3] Ting (2011),[4] CAWCR,[5] D. Chicco & G. Jurman (2020, 2021),[6][7] Tharwat (2018).[8] Balayla (2020)[9]

Example [edit]

Given a sample of 12 individuals, 8 that have been diagnosed with cancer and 4 that are cancer-free, where individuals with cancer belong to class 1 (positive) and non-cancer individuals belong to class 0 (negative), we can display that data as follows:

Individual Number	1	2	3	4	5	6	7	8	9	10	11	12
Actual Classification	1	1	1	1	1	1	1	1	0	0	0	0

Assume that we have a classifier that distinguishes between individuals with and without cancer in some way, we can take the 12 individuals and run them through the classifier. The classifier then makes 9 accurate predictions and misses 3: 2 individuals with cancer wrongly predicted as being cancer-free (sample 1 and 2), and 1 person without cancer that is wrongly predicted to have cancer (sample 9).

Individual Number	1	2	3	4	5	6	7	8	9	10	11	12
Actual Classification	1	1	1	1	1	1	1	1	0	0	0	0
Predicted Classification	0	0	1	1	1	1	1	1	1	0	0	0

Notice, that if we compare the actual classification set to the predicted classification set, there are 4 different outcomes that could result in any particular column. One, if the actual classification is positive and the predicted classification is positive (1,1), this is called a true positive result because the positive sample was correctly identified by the classifier. Two, if the actual classification is positive and the predicted classification is negative (1,0), this is called a false negative result because the positive sample is incorrectly identified by the classifier as being negative. Third, if the actual classification is negative and the predicted classification is positive (0,1), this is called a false positive result because the negative sample is incorrectly identified by the classifier as being positive. Fourth, if the actual classification is negative and the predicted classification is negative (0,0), this is called a true negative result because the negative sample gets correctly identified by the classifier.

We can then perform the comparison between actual and predicted classifications and add this information to the table, making correct results appear in green so they are more easily identifiable.

Individual Number	1	2	3	4	5	6	7	8	9	10	11	12
Actual Classification	1	1	1	1	1	1	1	1	0	0	0	0
Predicted Classification	0	0	1	1	1	1	1	1	1	0	0	0
Result	FN	FN	TP	TP	TP	TP	TP	TP	FP	TN	TN	TN

The template for any binary confusion matrix uses the four kinds of results discussed above (true positives, false negatives, false positives, and true negatives) along with the positive and negative classifications. The four outcomes can be formulated in a 2×2 *confusion matrix*, as follows:

Sources: [12][13][14][15][16][17][18][19] view·talk·edit

		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)

The color convention of the three data tables above were picked to match this confusion matrix, in order to easily differentiate the data.

Now, we can simply total up each type of result, substitute into the template, and create a confusion matrix that will concisely summarize the results of testing the classifier:

		Predicted condition	
	Total 8 + 4 = 12	Cancer 7	Non-cancer 5
Actual condition	Cancer 8	6	2
	Non-cancer 4	1	3

In this confusion matrix, of the 8 samples with cancer, the system judged that 2 were cancer-free, and of the 4 samples without cancer, it predicted that 1 did have cancer. All correct predictions are located in the diagonal of the table (highlighted in green), so it is easy to visually inspect the table for prediction errors, as values outside the diagonal will represent them. By summing up the 2 rows of the confusion matrix, one can also deduce the total number of positive (P) and negative (N) samples in the original dataset, i.e. $P = TP + FN$ and $N = FP + TN$.

Table of confusion [edit]

In predictive analytics, a **table of confusion** (sometimes also called a **confusion matrix**) is a table with two rows and two columns that reports the number of *true positives*, *false negatives*, *false positives*, and *true negatives*. This allows more detailed analysis than simply observing the proportion of correct classifications (accuracy). Accuracy will yield misleading results if the data set is unbalanced; that is, when the numbers of observations in different classes vary greatly.

For example, if there were 95 cancer samples and only 5 non-cancer samples in the data, a particular classifier might classify all the observations as having cancer. The overall accuracy would be 95%, but in more detail the classifier would have a 100% recognition rate (sensitivity) for the cancer class but a 0% recognition rate for the non-cancer class. F1 score is even more unreliable in such cases, and here would yield over 97.4%, whereas informedness removes such bias and yields 0 as the probability of an informed decision for any form of guessing (here always guessing cancer).

According to Davide Chicco and Giuseppe Jurman, the most informative metric to evaluate a confusion matrix is the Matthews correlation coefficient (MCC).[20]

Other metrics can be included in a confusion matrix, each of them having their significance and use.

Sources: [21][22][23][24][25][26][27][28][29] view·talk·edit

		Predicted condition			
	Total population = P + N	Positive (PP)	Negative (PN)	Informedness, bookmaker informedness (BM) = TPR + TNR − 1	Prevalence threshold (PT) $= \frac{\sqrt{TPR\times FPR}-FPR}{TPR-FPR}$
Actual condition	Positive (P)	**True positive (TP)**, **hit**	**False negative (FN)**, **type II error**, **miss**, **underestimation**	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate $= \frac{FN}{P} = 1 - TPR$
	Negative (N)	**False positive (FP)**, **type I error**, **false alarm**, **overestimation**	**True negative (TN)**, **correct rejection**	False positive rate (FPR), probability of false alarm, fall-out $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$
	Prevalence $= \frac{P}{P+N}$	Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$
	Accuracy (ACC) $= \frac{TP+TN}{P+N}$	False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) $= \frac{TN}{PN} = 1 - FOR$	Markedness (MK), deltaP (Δp) = PPV + NPV − 1	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$
	Balanced accuracy (BA) $= \frac{TPR+TNR}{2}$	F_1 score $= \frac{2PPV\times TPR}{PPV+TPR} = \frac{2TP}{2TP+FP+FN}$	Fowlkes–Mallows index (FM) $= \sqrt{PPV\times TPR}$	Matthews correlation coefficient (MCC) $= \sqrt{TPR\times TNR\times PPV\times NPV} - \sqrt{FNR\times FPR\times FOR\times FDR}$	Threat score (TS), critical success index (CSI), Jaccard index $= \frac{TP}{TP+FN+FP}$

Confusion matrices with more than two categories [edit]

Confusion matrix is not limited to binary classification and can be used in multi-class classifiers as well.[30] The confusion matrices discussed above have only two conditions: positive and negative. For example, the table below summarizes communication of a whistled language between two speakers, zero values omitted for clarity.[31]

Perceived vowel / Vowel produced	i	e	a	o	u
i	**15**		1		
e	1		1		
a			**79**	5	
o			4	**15**	3
u				2	**2**

See also [edit]

- Positive and negative predictive values

References [edit]

1. ^ Fawcett, Tom (2006). "An Introduction to ROC Analysis" (PDF). *Pattern Recognition Letters*. **27** (8): 861–874. doi:10.1016/j.patrec.2005.10.010.
2. ^ Piryonesi S. Madeh; El-Diraby Tamer E. (2020-03-01). "Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index". *Journal of Infrastructure Systems*. **26** (1): 04019036. doi:10.1061/(ASCE)IS.1943-555X.0000512.
3. ^ Powers, David M. W. (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation". *Journal of Machine Learning Technologies*. **2** (1): 37–63.
4. ^ Ting, Kai Ming (2011). Sammut, Claude; Webb, Geoffrey I. (eds.). *Encyclopedia of machine learning*. Springer. doi:10.1007/978-0-387-30164-8. ISBN 978-0-387-30164-8.
5. ^ Brooks, Harold; Brown, Barb; Ebert, Beth; Ferro, Chris; Jolliffe, Ian; Koh, Tieh-Yong; Roebber, Paul; Stephenson, David (2015-01-26). "WWRP/WGNE Joint Working Group on Forecast Verification Research". *Collaboration for Australian Weather and Climate Research*. World Meteorological Organisation. Retrieved 2019-07-17.
6. ^ Chicco D.; Jurman G. (January 2020). "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation". *BMC Genomics*. **21** (1): 6-1–6-13. doi:10.1186/s12864-019-6413-7. PMC 6941312. PMID 31898477.
7. ^ Chicco D.; Toetsch N.; Jurman G. (February 2021). "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation". *BioData Mining*. **14** (13): 1-22. doi:10.1186/s13040-021-00244-z. PMC 7863449. PMID 33541410.
8. ^ Tharwat A. (August 2018). "Classification assessment methods". *Applied Computing and Informatics*. doi:10.1016/j.aci.2018.08.003.
9. ^ Balayla, Jacques (2020). "Prevalence threshold (ϕe) and the geometry of screening curves". *PLoS One*. **15** (10). doi:10.1371/journal.pone.0240215.
10. ^ Stehman, Stephen V. (1997). "Selecting and interpreting measures of thematic classification accuracy". *Remote Sensing of Environment*. **62** (1): 77–89. Bibcode:1997RSEnv..62...77S. doi:10.1016/S0034-4257(97)00083-7.
11. ^ Powers, David M. W. (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation". *Journal of Machine Learning Technologies*. **2** (1): 37–63. S2CID 55767944.
12. ^ Fawcett, Tom (2006). "An Introduction to ROC Analysis" (PDF). *Pattern Recognition Letters*. **27** (8): 861–874. Bibcode:2006PaReL..27..861F. doi:10.1016/j.patrec.2005.10.010. S2CID 2027090.
13. ^ Piryonesi S. Madeh; El-Diraby Tamer E. (2020-03-01). "Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index". *Journal of Infrastructure Systems*. **26** (1): 04019036. doi:10.1061/(ASCE)IS.1943-555X.0000512. S2CID 213782055.
14. ^ Powers, David M. W. (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation". *Journal of Machine Learning Technologies*. **2** (1): 37–63.
15. ^ Ting, Kai Ming (2011). Sammut, Claude; Webb, Geoffrey I. (eds.). *Encyclopedia of machine learning*. Springer. doi:10.1007/978-0-387-30164-8. ISBN 978-0-387-30164-8.
16. ^ Brooks, Harold; Brown, Barb; Ebert, Beth; Ferro, Chris; Jolliffe, Ian; Koh, Tieh-Yong; Roebber, Paul; Stephenson, David (2015-01-26). "WWRP/WGNE Joint Working Group on Forecast Verification Research". *Collaboration for Australian Weather and Climate Research*. World Meteorological Organisation. Retrieved 2019-07-17.
17. ^ Chicco D, Jurman G (January 2020). "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation". *BMC Genomics*. **21** (1): 6-1–6-13. doi:10.1186/s12864-019-6413-7. PMC 6941312. PMID 31898477.
18. ^ Chicco D, Toetsch N, Jurman G (February 2021). "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation". *BioData Mining*. **14** (13): 13. doi:10.1186/s13040-021-00244-z. PMC 7863449. PMID 33541410.
19. ^ Tharwat A. (August 2018). "Classification assessment methods". *Applied Computing and Informatics*. **17**: 168–192. doi:10.1016/j.aci.2018.08.003.
20. ^ Chicco D., Jurman G. (January 2020). "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation". *BMC Genomics*. **21** (1): 6-1–6-13. doi:10.1186/s12864-019-6413-7. PMC 6941312. PMID 31898477.
21. ^ Balayla, Jacques (2020). "Prevalence threshold (ϕe) and the geometry of screening curves". *PLoS One*. **15** (10). doi:10.1371/journal.pone.0240215.
22. ^ Fawcett, Tom (2006). "An Introduction to ROC Analysis" (PDF). *Pattern Recognition Letters*. **27** (8): 861–874. doi:10.1016/j.patrec.2005.10.010.
23. ^ Piryonesi S. Madeh; El-Diraby Tamer E. (2020-03-01). "Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index". *Journal of Infrastructure Systems*. **26** (1): 04019036. doi:10.1061/(ASCE)IS.1943-555X.0000512.
24. ^ Powers, David M. W. (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation". *Journal of Machine Learning Technologies*. **2** (1): 37–63.
25. ^ Ting, Kai Ming (2011). Sammut, Claude; Webb, Geoffrey I. (eds.). *Encyclopedia of machine learning*. Springer. doi:10.1007/978-0-387-30164-8. ISBN 978-0-387-30164-8.
26. ^ Brooks, Harold; Brown, Barb; Ebert, Beth; Ferro, Chris; Jolliffe, Ian; Koh, Tieh-Yong; Roebber, Paul; Stephenson, David (2015-01-26). "WWRP/WGNE Joint Working Group on Forecast Verification Research". *Collaboration for Australian Weather and Climate Research*. World Meteorological Organisation. Retrieved 2019-07-17.
27. ^ Chicco D, Jurman G (January 2020). "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation". *BMC Genomics*. **21** (1): 6-1–6-13. doi:10.1186/s12864-019-6413-7. PMC 6941312. PMID 31898477.
28. ^ Chicco D, Toetsch N, Jurman G (February 2021). "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation". *BioData Mining*. **14** (13): 1-22. doi:10.1186/s13040-021-00244-z. PMC 7863449. PMID 33541410.
29. ^ Tharwat A. (August 2018). "Classification assessment methods". *Applied Computing and Informatics*. doi:10.1016/j.aci.2018.08.003.
30. ^ Piryonesi S. Madeh; El-Diraby Tamer E. (2020-03-01). "Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index". *Journal of Infrastructure Systems*. **26** (1): 04019036. doi:10.1061/(ASCE)IS.1943-555X.0000512. S2CID 213782055.
31. ^ Rialland, Annie (August 2005). "Phonological and phonetic aspects of whistled languages". *Phonology*. **22** (2): 237–271. CiteSeerX 10.1.1.484.4384. doi:10.1017/S0952675705000552. S2CID 18615779.

Receiver operating characteristic

See also

- Brier score
- Coefficient of determination
- Constant false alarm rate
- Detection error tradeoff
- Detection theory
- F1 score
- False alarm
- Hypothesis tests for accuracy
- Precision and recall
- ROCCET
- Sensitivity and specificity
- Total operating characteristic

Mathematics portal

References

1. ^ Fawcett, Tom (2006). "An Introduction to ROC Analysis" (PDF). *Pattern Recognition Letters*. **27** (8): 861–874. doi:10.1016/j.patrec.2005.10.010.
2. ^ Piryonesi S. Madeh; El-Diraby Tamer E. (2020-03-01). "Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index". *Journal of Infrastructure Systems*. **26** (1): 04019036. doi:10.1061/(ASCE)IS.1943-555X.0000512.
3. ^ Powers, David M. W. (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation". *Journal of Machine Learning Technologies*. **2** (1): 37–63.
4. ^ Ting, Kai Ming (2011). Sammut, Claude; Webb, Geoffrey I. (eds.). *Encyclopedia of machine learning*. Springer. doi:10.1007/978-0-387-30164-8. ISBN 978-0-387-30164-8.
5. ^ Brooks, Harold; Brown, Barb; Ebert, Beth; Ferro, Chris; Jolliffe, Ian; Koh, Tieh-Yong; Roebber, Paul; Stephenson, David (2015-01-26). "WWRP/WGNE Joint Working Group on Forecast Verification Research". *Collaboration for Australian Weather and Climate Research*. World Meteorological Organisation. Retrieved 2019-07-17.
6. ^ Chicco D.; Jurman G. (January 2020). "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation". *BMC Genomics*. **21** (1): 6-1–6-13. doi:10.1186/s12864-019-6413-7. PMC [redacted]. PMID 31898477.
7. ^ Chicco D.; Toetsch N.; Jurman G. (February 2021). "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation". *BioData Mining*. **14** (13): 1-22. doi:10.1186/s13040-021-00244-z. PMC [redacted]. PMID 33541410.
8. ^ Tharwat A. (August 2018). "Classification assessment methods". *Applied Computing and Informatics*. doi:[redacted].
9. ^ Balayla, Jacques (2020). "Prevalence threshold (ϕe) and the geometry of screening curves". *PLoS One*. **15** (10). doi:10.1371/journal.pone.0240215.
10. ^ *a* *b* "Detector Performance Analysis Using ROC Curves - MATLAB & Simulink Example". *www.mathworks.com*. Retrieved 11 August 2016.
11. ^ Peres, D. J.; Cancelliere, A. (2014-12-08). "Derivation and evaluation of landslide-triggering thresholds by a Monte Carlo approach". *Hydrol. Earth Syst. Sci.* **18** (12): 4913–4931. Bibcode:2014HESS...18.4913P. doi:[redacted]. ISSN 1607-7938.
12. ^ Murphy, Allan H. (1996-03-01). "The Finley Affair: A Signal Event in the History of Forecast Verification". *Weather and Forecasting*. **11** (1): 3–20. Bibcode:1996WtFor..11....3M. doi:[redacted]. ISSN 0882-8156.
13. ^ Peres, D. J.; Iuppa, C.; Cavallaro, L.; Cancelliere, A.; Foti, E. (2015-10-01). "Significant wave height record extension by neural networks and reanalysis wind data". *Ocean Modelling*. **94**: 128–140. Bibcode:2015OcMod..94..128P. doi:10.1016/j.ocemod.2015.08.002.
14. ^ Swets, John A.; *Signal detection theory and ROC analysis in psychology and diagnostics : collected papers*, Lawrence Erlbaum Associates, Mahwah, NJ, 1996
15. ^ Sushkova, Olga; Morozov, Alexei; Gabova, Alexandra; Karabanov, Alexei; Illarioshkin, Sergey (2021). "A Statistical Method for Exploratory Data Analysis Based on 2D and 3D Area under Curve Diagrams: Parkinson's Disease Investigation". *Sensors*. **21** (14): 4700. doi:[redacted]. PMC [redacted]. PMID 34300440.
16. ^ Balayla, Jacques (2020). "Prevalence threshold (ϕe) and the geometry of screening curves". *PLoS One*. **15** (10). doi:10.1371/journal.pone.0240215.
17. ^ Fawcett, Tom (2006). "An Introduction to ROC Analysis" (PDF). *Pattern Recognition Letters*. **27** (8): 861–874. doi:10.1016/j.patrec.2005.10.010.
18. ^ Piryonesi S. Madeh; El-Diraby Tamer E. (2020-03-01). "Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index". *Journal of Infrastructure Systems*. **26** (1): 04019036. doi:10.1061/(ASCE)IS.1943-555X.0000512.
19. ^ Powers, David M. W. (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation". *Journal of Machine Learning Technologies*. **2** (1): 37–63.
20. ^ Ting, Kai Ming (2011). Sammut, Claude; Webb, Geoffrey I. (eds.). *Encyclopedia of machine learning*. Springer. doi:10.1007/978-0-387-30164-8. ISBN 978-0-387-30164-8.
21. ^ Brooks, Harold; Brown, Barb; Ebert, Beth; Ferro, Chris; Jolliffe, Ian; Koh, Tieh-Yong; Roebber, Paul; Stephenson, David (2015-01-26). "WWRP/WGNE Joint Working Group on Forecast Verification Research". *Collaboration for Australian Weather and Climate Research*. World Meteorological Organisation. Retrieved 2019-07-17.
22. ^ Chicco D, Jurman G (January 2020). "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation". *BMC Genomics*. **21** (1): 6-1–6-13. doi:10.1186/s12864-019-6413-7. PMC [redacted]. PMID 31898477.
23. ^ Chicco D, Toetsch N, Jurman G (February 2021). "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation". *BioData Mining*. **14** (13): 1-22. doi:10.1186/s13040-021-00244-z. PMC [redacted]. PMID 33541410.
24. ^ Tharwat A. (August 2018). "Classification assessment methods". *Applied Computing and Informatics*. doi:[redacted].
25. ^ "classification - AUC-ROC of a random classifier". *Data Science Stack Exchange*. Retrieved 2020-11-30.
26. ^ *a* *b* *c* Powers, David MW (2012). "ROC-ConCert: ROC-Based Measurement of Consistency and Certainty" (PDF). *Spring Congress on Engineering and Technology (SCET)*. Vol. 2. IEEE. pp. 238–241.[dead link]
27. ^ Fogarty, James; Baker, Ryan S.; Hudson, Scott E. (2005). "Case studies in the use of ROC curve analysis for sensor-based estimates in human computer interaction". *ACM International Conference Proceeding Series, Proceedings of Graphics Interface 2005*. Waterloo, ON: Canadian Human-Computer Communications Society.
28. ^ Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2nd ed.).
29. ^ Fawcett, Tom (2006); *An introduction to ROC analysis*, Pattern Recognition Letters, 27, 861–874.
30. ^ Hanley, James A.; McNeil, Barbara J. (1982). "The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve". *Radiology*. **143** (1): 29–36. doi:10.1148/radiology.143.1.7063747. PMID 7063747. S2CID 10511727.
31. ^ *a* *b* Mason, Simon J.; Graham, Nicholas E. (2002). "Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation" (PDF). *Quarterly Journal of the Royal Meteorological Society*. **128** (584): 2145–2166. Bibcode:2002QJRMS.128.2145M. CiteSeerX [redacted]. doi:10.1256/003590002320603584. Archived from the original (PDF) on 2008-11-20.
32. ^ Calders, Toon; Jaroszewicz, Szymon (2007). Kok, Joost N.; Koronacki, Jacek; Lopez de Mantaras, Ramon; Matwin, Stan; Mladenič, Dunja; Skowron, Andrzej (eds.). "Efficient AUC Optimization for Classification". *Knowledge Discovery in Databases: PKDD 2007*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. **4702**: 42–53. doi:[redacted]. ISBN 978-3-540-74976-9.
33. ^ Hand, David J.; and Till, Robert J. (2001); *A simple generalization of the area under the ROC curve for multiple class classification problems*, Machine Learning, 45, 171–186.
34. ^ Provost, F.; Fawcett, T. (2001). "Robust classification for imprecise environments". *Machine Learning*. **42** (3): 203–231. arXiv:[redacted]. doi:10.1023/a:1007601015854. S2CID 5415722.
35. ^ Flach, P.A.; Wu, S. (2005). "Repairing concavities in ROC curves." (PDF). *19th International Joint Conference on Artificial Intelligence (IJCAI'05)*. pp. 702–707.
36. ^ Hanley, James A.; McNeil, Barbara J. (1983-09-01). "A method of comparing the areas under receiver operating characteristic curves derived from the same cases". *Radiology*. **148** (3): 839–843. doi:[redacted]. PMID 6878708.
37. ^ Hanczar, Blaise; Hua, Jianping; Sima, Chao; Weinstein, John; Bittner, Michael; Dougherty, Edward R (2010). "Small-sample precision of ROC-related estimates". *Bioinformatics*. **26** (6): 822–830. doi:[redacted]. PMID 20130029.
38. ^ Lobo, Jorge M.; Jiménez-Valverde, Alberto; Real, Raimundo (2008). "AUC: a misleading measure of the performance of predictive distribution models". *Global Ecology and Biogeography*. **17** (2): 145–151. doi:10.1111/j.1466-8238.2007.00358.x. S2CID 15206363.
39. ^ Hand, David J (2009). "Measuring classifier performance: A coherent alternative to the area under the ROC curve". *Machine Learning*. **77**: 103–123. doi:[redacted].
40. ^ Flach, P.A.; Hernandez-Orallo, J.; Ferri, C. (2011). "A coherent interpretation of AUC as a measure of aggregated classification performance." (PDF). *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. pp. 657–664.
41. ^ Hernandez-Orallo, J.; Flach, P.A.; Ferri, C. (2012). "A unified view of performance metrics: translating threshold choice into expected classification loss" (PDF). *Journal of Machine Learning Research*. **13**: 2813–2869.
42. ^ Powers, David M.W. (2012). "The Problem of Area Under the Curve". *International Conference on Information Science and Technology*.
43. ^ Powers, David M. W. (2003). "Recall and Precision versus the Bookmaker" (PDF). *Proceedings of the International Conference on Cognitive Science (ICSC-2003), Sydney Australia, 2003, pp. 529–534*.
44. ^ Powers, David M. W. (2012). "The Problem with Kappa" (PDF). *Conference of the European Chapter of the Association for Computational Linguistics (EACL2012) Joint ROBUS-UNSUP Workshop*. Archived from the original (PDF) on 2016-05-18. Retrieved 2012-07-20.
45. ^ McClish, Donna Katzman (1989-08-01). "Analyzing a Portion of the ROC Curve". *Medical Decision Making*. **9** (3): 190–195. doi:10.1177/0272989X8900900307. PMID 2668680. S2CID 24442201.
46. ^ Dodd, Lori E.; Pepe, Margaret S. (2003). "Partial AUC Estimation and Regression". *Biometrics*. **59** (3): 614–623. doi:10.1111/1541-0420.00071. PMID 14601762.
47. ^ Karplus, Kevin (2011); *Better than Chance: the importance of null models*, University of California, Santa Cruz, in Proceedings of the First International Workshop on Pattern Recognition in Proteomics, Structural Biology and Bioinformatics (PR PS BB 2011)
48. ^ "C-Statistic: Definition, Examples, Weighting and Significance". *Statistics How To*. August 28, 2016.
49. ^ Pontius, Robert Gilmore; Parmentier, Benoit (2014). "Recommendations for using the Relative Operating Characteristic (ROC)". *Landscape Ecology*. **29** (3): 367–382. doi:10.1007/s10980-013-9984-8. S2CID 15924380.
50. ^ Pontius, Robert Gilmore; Si, Kangping (2014). "The total operating characteristic to measure diagnostic ability for multiple thresholds". *International Journal of Geographical Information Science*. **28** (3): 570–583. doi:10.1080/13658816.2013.862623. S2CID 29204880.
51. ^ Navratil, J.; Klusacek, D. (2007-04-01). *On Linear DETs. 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*. Vol. 4. pp. IV–229–IV–232. doi:10.1109/ICASSP.2007.367205. ISBN 978-1-4244-0727-9. S2CID 18173315.
52. ^ Dev P. Chakraborty (December 14, 2017). "double+probability+paper"&pg=PT214 *Observer Performance Methods for Diagnostic Imaging: Foundations, Modeling, and Applications with R-Based Examples*. CRC Press. p. 214. ISBN 9781351230711. Retrieved July 11, 2019.
53. ^ MacMillan, Neil A.; Creelman, C. Douglas (2005). *Detection Theory: A User's Guide* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates. ISBN 978-1-4106-1114-7.
54. ^ Glanzer, Murray; Kisok, Kim; Hilford, Andy; Adams, John K. (1999). "Slope of the receiver-operating characteristic in recognition memory". *Journal of Experimental Psychology: Learning, Memory, and Cognition*. **25** (2): 500–513. doi:10.1037/0278-7393.25.2.500.
55. ^ Ratcliff, Roger; McCoon, Gail; Tindall, Michael (1994). "Empirical generality of data from recognition memory ROC functions and implications for GMMs". *Journal of Experimental Psychology: Learning, Memory, and Cognition*. **20** (4): 763–785. CiteSeerX [redacted]. doi:10.1037/0278-7393.20.4.763.
56. ^ Zhang, Jun; Mueller, Shane T. (2005). "A note on ROC analysis and non-parametric estimate of sensitivity". *Psychometrika*. **70**: 203–212. CiteSeerX [redacted]. doi:10.1007/s11336-003-1119-8. S2CID 122355230.
57. ^ Yonelinas, Andrew P.; Kroll, Neal E. A.; Dobbins, Ian G.; Lazzara, Michele; Knight, Robert T. (1998). "Recollection and familiarity deficits in amnesia: Convergence of remember-know, process dissociation, and receiver operating characteristic data". *Neuropsychology*. **12** (3): 323–339. doi:10.1037/0894-4105.12.3.323. PMID 9673991.
58. ^ *a* *b* Green, David M.; Swets, John A. (1966). *Signal detection theory and psychophysics*. New York, NY: John Wiley and Sons Inc. ISBN 978-0-471-32420-1.
59. ^ "Using the Receiver Operating Characteristic (ROC) curve to analyze a classification model: A final note of historical interest" (PDF). *Department of Mathematics, University of Utah*. Department of Mathematics, University of Utah. Archived (PDF) from the original on 2020-08-22. Retrieved May 25, 2017.
60. ^ Zweig, Mark H.; Campbell, Gregory (1993). "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine" (PDF). *Clinical Chemistry*. **39** (8): 561–577. doi:[redacted]. PMID 8472349.
61. ^ Pepe, Margaret S. (2003). *The statistical evaluation of medical tests for classification and prediction*. New York, NY: Oxford. ISBN 978-0-19-856582-6.
62. ^ Obuchowski, Nancy A. (2003). "Receiver operating characteristic curves and their use in radiology". *Radiology*. **229** (1): 3–8. doi:10.1148/radiol.2291010898. PMID 14519861.
63. ^ Spackman, Kent A. (1989). "Signal detection theory: Valuable tools for evaluating inductive learning". *Proceedings of the Sixth International Workshop on Machine Learning*. San Mateo, CA: Morgan Kaufmann. pp. 160–163.
64. ^ Kharin, Viatcheslav (2003). "On the ROC score of probability forecasts". *Journal of Climate*. **16** (24): 4145–4150. Bibcode:2003JCli...16.4145K. doi:[redacted].
65. ^ Till, D.J.; Hand, R.J. (2001). "A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems". *Machine Learning*. **45** (2): 171–186. doi:[redacted].
66. ^ Mossman, D. (1999). "Three-way ROCs". *Medical Decision Making*. **19** (1): 78–89. doi:10.1177/0272989x9901900110. PMID 9917023. S2CID 24623127.
67. ^ Ferri, C.; Hernandez-Orallo, J.; Salido, M.A. (2003). "Volume under the ROC Surface for Multi-class Problems". *Machine Learning: ECML 2003*. pp. 108–120.
68. ^ Bi, J.; Bennett, K.P. (2003). "Regression error characteristic curves" (PDF). *Twentieth International Conference on Machine Learning (ICML-2003). Washington, DC*.
69. ^ Hernandez-Orallo, J. (2013). "ROC curves for regression". *Pattern Recognition*. **46** (12): 3395–3411. doi:10.1016/j.patcog.2013.06.014. hdl:[redacted].

External links

- ROC demo
- another ROC demo
- ROC video explanation
- An Introduction to the Total Operating Characteristic: Utility in Land Change Model Evaluation
- How to run the TOC Package in R
- TOC R package on Github
- Excel Workbook for generating TOC curves

Further reading

Wikimedia Commons has media related to *Receiver operating characteristic*.

- Balakrishnan, Narayanaswamy (1991); *Handbook of the Logistic Distribution*, Marcel Dekker, Inc., ISBN 978-0-8247-8587-1
- Brown, Christopher D.; Davis, Herbert T. (2006). "Receiver operating characteristic curves and related decision measures: a tutorial". *Chemometrics and Intelligent Laboratory Systems*. **80**: 24–38. doi:10.1016/j.chemolab.2005.05.004.
- Rotello, Caren M.; Heit, Evan; Dubé, Chad (2014). "When more data steer us wrong: replications with the wrong dependent measure perpetuate erroneous conclusions" (PDF). *Psychonomic Bulletin & Review*. **22** (4): 944–954. doi:[redacted]. PMID 25384892. S2CID 6046065.
- Fawcett, Tom (2004). "ROC Graphs: Notes and Practical Considerations for Researchers" (PDF). *Pattern Recognition Letters*. **27** (8): 882–891. CiteSeerX [redacted]. doi:10.1016/j.patrec.2005.10.012.
- Gonen, Mithat (2007); *Analyzing Receiver Operating Characteristic Curves Using SAS*, SAS Press, ISBN 978-1-59994-298-8
- Green, William H., (2003) *Econometric Analysis*, fifth edition, Prentice Hall, ISBN 0-13-066189-9
- Heagerty, Patrick J.; Lumley, Thomas; Pepe, Margaret S. (2000). "Time-dependent ROC Curves for Censored Survival Data and a Diagnostic Marker". *Biometrics*. **56** (2): 337–344. doi:10.1111/j.0006-341x.2000.00337.x. PMID 10877287. S2CID 8822160.
- Hosmer, David W.; and Lemeshow, Stanley (2000); *Applied Logistic Regression*, 2nd ed., New York, NY: Wiley, ISBN 0-471-35632-8
- Lasko, Thomas A.; Bhagwat, Jui G.; Zou, Kelly H.; Ohno-Machado, Lucila (2005). "The use of receiver operating characteristic curves in biomedical informatics". *Journal of Biomedical Informatics*. **38** (5): 404–415. CiteSeerX [redacted]. doi:10.1016/j.jbi.2005.02.008. PMID 16198999.
- Mas, Jean-François; Filho, Britaldo Soares; Pontius, Jr, Robert Gilmore; Gutiérrez, Michelle Farfán; Rodrigues, Hermann (2013). "A suite of tools for ROC analysis of spatial models". *ISPRS International Journal of Geo-Information*. **2** (3): 869–887. Bibcode:2013IJGI....2..869M. doi:[redacted].
- Pontius, Jr, Robert Gilmore; Parmentier, Benoit (2014). "Recommendations for using the Relative Operating Characteristic (ROC)". *Landscape Ecology*. **29** (3): 367–382. doi:10.1007/s10980-013-9984-8. S2CID 15924380.
- Pontius, Jr, Robert Gilmore; Pacheco, Pablo (2004). "Calibration and validation of a model of forest disturbance in the Western Ghats, India 1920–1990". *GeoJournal*. **61** (4): 325–334. doi:10.1007/s10708-004-5049-5. S2CID 155073463.
- Pontius, Jr, Robert Gilmore; Batchu, Kiran (2003). "Using the relative operating characteristic to quantify certainty in prediction of location of land cover change in India". *Transactions in GIS*. **7** (4): 467–484. doi:10.1111/1467-9671.00159. S2CID 14452746.
- Pontius, Jr, Robert Gilmore; Schneider, Laura (2001). "Land-use change model validation by a ROC method for the Ipswich watershed, Massachusetts, USA". *Agriculture, Ecosystems & Environment*. **85** (1–3): 239–248. doi:10.1016/S0167-8809(01)00187-6.
- Stephan, Carsten; Wesseling, Sebastian; Schink, Tania; Jung, Klaus (2003). "Comparison of Eight Computer Programs for Receiver-Operating Characteristic Analysis". *Clinical Chemistry*. **49** (3): 433–439. doi:[redacted]. PMID 12600955.
- Swets, John A.; Dawes, Robyn M.; and Monahan, John (2000); *Better Decisions through Science*, Scientific American, October, pp. 82–87
- Zou, Kelly H.; O'Malley, A. James; Mauri, Laura (2007). "Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models". *Circulation*. **115** (5): 654–7. doi:[redacted]. PMID 17283280.
- Zhou, Xiao-Hua; Obuchowski, Nancy A.; McClish, Donna K. (2002). *Statistical Methods in Diagnostic Medicine*. New York, NY: Wiley & Sons. ISBN 978-0-471-34772-9.

v·t·e **Statistics** [hide]
Outline · Index
Descriptive statistics [show]
Data collection [show]
Statistical inference [show]
Correlation · Regression analysis [show]
Categorical / Multivariate / Time-series / Survival analysis [show]
Applications [show]
Category · Mathematics portal · Commons · WikiProject

v·t·e **Machine learning evaluation metrics** [show]

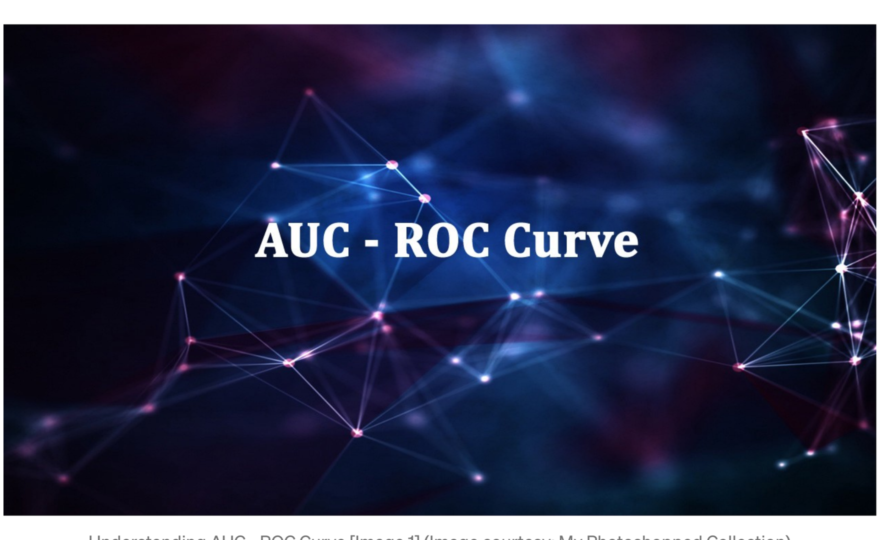
v·t·e **Public health** [show]

Authority control: National libraries Germany

Categories: Detection theory | Data mining | Biostatistics | Statistical classification | Summary statistics for contingency tables

Sarang Narkhede
Jun 26, 2018 · 5 min read · Listen

Understanding AUC - ROC Curve



Understanding AUC - ROC Curve [Image 1] (Image courtesy: My Photoshopped Collection)

In Machine Learning, performance measurement is an essential task. So when it comes to a classification problem, we can count on an AUC - ROC Curve. When we need to check or visualize the performance of the multi-class classification problem, we use the AUC (**Area Under The Curve**) ROC (**Receiver Operating Characteristics**) curve. It is one of the most important evaluation metrics for checking any classification model's performance. It is also written as AUROC (**Area Under the Receiver Operating Characteristics**)

Note: For better understanding, I suggest you read my article about Confusion Matrix.

This blog aims to answer the following questions:

1. What is the AUC - ROC Curve?

2. Defining terms used in AUC and ROC Curve.

3. How to speculate the performance of the model?

4. Relation between Sensitivity, Specificity, FPR, and Threshold.

5. How to use AUC - ROC curve for the multiclass model?

What is the AUC - ROC Curve?

AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1. By analogy, the Higher the AUC, the better the model is at distinguishing between patients with the disease and no disease.

The ROC curve is plotted with TPR against the FPR where TPR is on the y-axis and FPR is on the x-axis.

AUC - ROC Curve [Image 2] (Image courtesy: My Photoshopped Collection)

Defining terms used in AUC and ROC Curve.

TPR (True Positive Rate) / Recall /Sensitivity

Image 3

Specificity

Image 4

FPR

Image 5

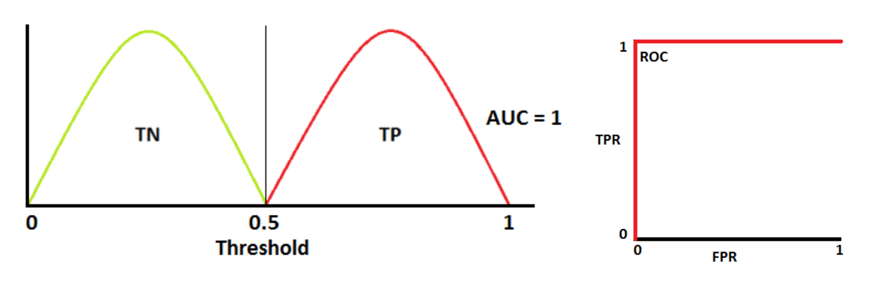
How to speculate about the performance of the model?

An excellent model has AUC near to the 1 which means it has a good measure of separability. A poor model has an AUC near 0 which means it has the worst measure of separability. In fact, it means it is reciprocating the result. It is predicting 0s as 1s and 1s as 0s. And when AUC is 0.5, it means the model has no class separation capacity whatsoever.

Let's interpret the above statements.

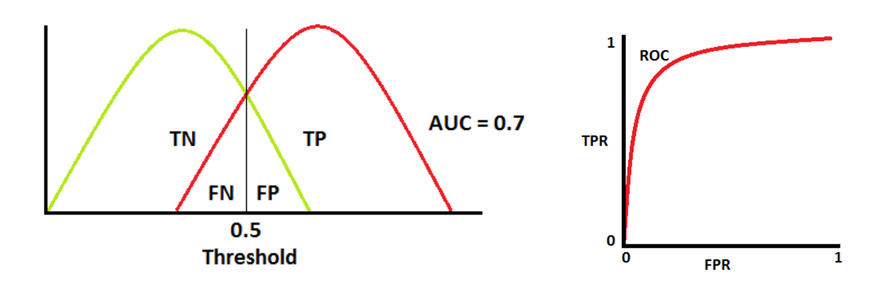
As we know, ROC is a curve of probability. So let's plot the distributions of those probabilities:

Note: Red distribution curve is of the positive class (patients with disease) and the green distribution curve is of the negative class(patients with no disease).



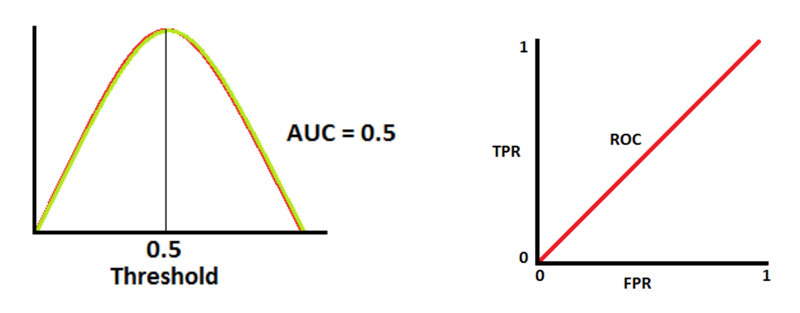
[Image 6 and 7] (Image courtesy: My Photoshopped Collection)

This is an ideal situation. When two curves don't overlap at all means model has an ideal measure of separability. It is perfectly able to distinguish between positive class and negative class.



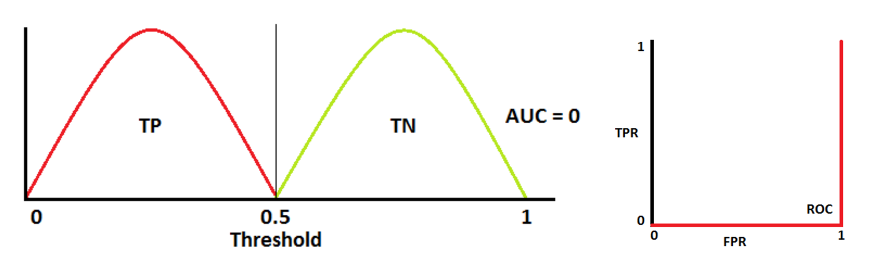
[Image 8 and 9] (Image courtesy: My Photoshopped Collection)

When two distributions overlap, we introduce type 1 and type 2 errors. Depending upon the threshold, we can minimize or maximize them. When AUC is 0.7, it means there is a 70% chance that the model will be able to distinguish between positive class and negative class.



[Image 10 and 11] (Image courtesy: My Photoshopped Collection)

This is the worst situation. When AUC is approximately 0.5, the model has no discrimination capacity to distinguish between positive class and negative class.



[Image 12 and 13] (Image courtesy: My Photoshopped Collection)

When AUC is approximately 0, the model is actually reciprocating the classes. It means the model is predicting a negative class as a positive class and vice versa.

The relation between Sensitivity, Specificity, FPR, and Threshold.

Sensitivity and Specificity are inversely proportional to each other. So when we increase Sensitivity, Specificity decreases, and vice versa.

> Sensitivity⬆️, Specificity⬇️ and Sensitivity⬇️, Specificity⬆️

When we decrease the threshold, we get more positive values thus it increases the sensitivity and decreasing the specificity.

Similarly, when we increase the threshold, we get more negative values thus we get higher specificity and lower sensitivity.

As we know FPR is 1 - specificity. So when we increase TPR, FPR also increases and vice versa.

> TPR⬆️, FPR⬆️ and TPR⬇️, FPR⬇️

How to use the AUC ROC curve for the multi-class model?

In a multi-class model, we can plot the N number of AUC ROC Curves for N number classes using the One vs ALL methodology. So for example, If you have **three** classes named **X**, **Y**, and **Z**, you will have one ROC for X classified against Y and Z, another ROC for Y classified against X and Z, and the third one of Z classified against Y and X.

· · ·

Thanks for Reading.

I hope I've given you some understanding of what exactly is the AUC - ROC Curve. If you like this post, a tad of extra motivation will be helpful by giving this post some claps 👏. I am always open to your questions and suggestions. You can share this on Facebook, Twitter, Linkedin, so someone in need might stumble upon this.

You can reach me at:

LinkedIn: https://www.linkedin.com/in/narkhedesarang/

Twitter: https://twitter.com/narkhede_sarang

Github: https://github.com/TheSarang