WIKIPEDIA

Confusion matrix

In the field of machine learning and specifically the problem of statistical classification, a **confusion matrix**, also known as an error matrix,[7] 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 a predicted class while each column represents the instances in an actual class (or vice versa).[8] The name stems from the fact that it makes it easy to see if 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

Example

Given a sample of 13 pictures, 8 of cats and 5 of dogs, where cats belong to class 1 and dogs belong to class 0,

actual = [1,1,1,1,1,1,1,1,0,0,0,0,0],

assume that a classifier that distinguishes between cats and dogs is trained, and we take the 13 pictures and run them through the classifier, and the classifier makes 8 accurate predictions and misses 5: 3 cats wrongly predicted as dogs (first 3 predictions) and 2 dogs wrongly predicted as cats (last 2 predictions).

prediction = [0,0,0,1,1,1,1,1,0,0,0,1,1]

With these two labelled sets (actual and predictions) we can create a confusion matrix that will summarize the results of testing the classifier:

		Actual class	
		Cat	**Dog**
Predicted class	**Cat**	**5**	2
	Dog	3	**3**

In this confusion matrix, of the 8 cat pictures, the system judged that 3 were dogs, and of the 5 dog pictures, it predicted that 2 were cats. All correct predictions are located in the diagonal of the table (highlighted in bold), so it is easy to visually inspect the table for prediction errors, as they will be represented by values outside the diagonal.

In abstract terms, the confusion matrix is as follows:

		Actual class	
		P	**N**
Predicted class	**P**	**TP**	FP
	N	FN	**TN**

where: P = Positive; N = Negative; TP = True Positive; FP = False Positive; TN = True Negative; FN = False Negative.

Table of confusion

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 *false positives*, *false negatives*, *true positives*, and *true negatives*. This allows more detailed analysis than mere 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 cats and only 5 dogs in the data, a particular classifier might classify all the observations as cats. The overall accuracy would be 95%, but in more detail the classifier would have a 100% recognition rate (sensitivity) for the cat class but a 0% recognition rate for the dog 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 cat).

According to Davide Chicco and Giuseppe Jurman, the most informative metric to evaluate a confusion matrix is the Matthews correlation coefficient (MCC).[9]

Assuming the confusion matrix above, its corresponding table of confusion, for the cat class, would be:

		Actual class	
		Cat	**Non-cat**
Predicted class	**Cat**	**5 True Positives**	2 False Positives
	Non-cat	3 False Negatives	**3 True Negatives**

The final table of confusion would contain the average values for all classes combined.

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)
eqv. with hit
true negative (TN)
eqv. with correct rejection
false positive (FP)
eqv. with false alarm, Type I error
false negative (FN)
eqv. with miss, Type II error

---

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)

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

Prevalence Threshold (PT)

$$PT = \frac{\sqrt{TPR(-TNR+1)}+TNR-1}{(TPR+TNR-1)}$$

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

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

---

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

Let us define an experiment from **P** positive instances and **N** negative instances for some condition. The four outcomes can be formulated in a 2×2 *confusion matrix*, as follows:

$$\mathrm{F_1} = 2 \cdot \frac{\mathrm{PPV} \cdot \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

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)

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

informedness or bookmaker informedness (BM)

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

markedness (MK) or deltaP

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

Sources: Fawcett (2006),[1] Powers (2011),[2] Ting (2011),[3] CAWCR,[4] D. Chicco & G. Jurman (2020),[5] Tharwat (2018).[6]

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

References

1. Fawcett, Tom (2006). "An Introduction to ROC Analysis" (http://people.inf.elte.hu/kiss/11dwhdm/roc.pdf) (PDF). *Pattern Recognition Letters*. **27** (8): 861–874. doi:10.1016/j.patrec.2005.10.010 (https://doi.org/10.1016%2Fj.patrec.2005.10.010).
2. Powers, David M W (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation" (https://www.researchgate.net/publication/228529307). *Journal of Machine Learning Technologies*. **2** (1): 37–63.
3. Ting, Kai Ming (2011). Sammut, Claude; Webb, Geoffrey I (eds.). *Encyclopedia of machine learning*. Springer. doi:10.1007/978-0-387-30164-8 (https://doi.org/10.1007%2F978-0-387-30164-8). ISBN 978-0-387-30164-8.
4. 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" (https://www.cawcr.gov.au/projects/verification/). *Collaboration for Australian Weather and Climate Research*. World Meteorological Organisation. Retrieved 2019-07-17.
5. Chicco D, Jurman G (January 2020). "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6941312). *BMC Genomics*. **21** (1): 6-1–6-13. doi:10.1186/s12864-019-6413-7 (https://doi.org/10.1186%2Fs12864-019-6413-7). PMC 6941312 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6941312). PMID 31898477 (https://pubmed.ncbi.nlm.nih.gov/31898477).

6. Tharwat A (August 2018). "Classification assessment methods" (https://doi.org/10.1016%2Fj.aci.2018.08.003). *Applied Computing and Informatics*. doi:10.1016/j.aci.2018.08.003 (https://doi.org/10.1016%2Fj.aci.2018.08.003).
7. Stehman, Stephen V. (1997). "Selecting and interpreting measures of thematic classification accuracy". *Remote Sensing of Environment*. **62** (1): 77–89. Bibcode:1997RSEnv..62...77S (https://ui.adsabs.harvard.edu/abs/1997RSEnv..62...77S). doi:10.1016/S0034-4257(97)00083-7 (https://doi.org/10.1016%2FS0034-4257%2897%2900083-7).
8. Powers, David M W (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation" (https://www.researchgate.net/publication/228529307). *Journal of Machine Learning Technologies*. **2** (1): 37–63. S2CID 55767944 (https://api.semanticscholar.org/CorpusID:55767944).
9. Chicco D, Jurman G (January 2020). "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6941312). *BMC Genomics*. **21** (1): 6-1–6-13. doi:10.1186/s12864-019-6413-7 (https://doi.org/10.1186%2Fs12864-019-6413-7). PMC 6941312 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6941312). PMID 31898477 (https://pubmed.ncbi.nlm.nih.gov/31898477).