March 25, 2014 · MACHINE LEARNING

Simple guide to confusion matrix terminology

A confusion matrix is a table that is often used to **describe the performance of a classification model** (or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing.

I wanted to create a **"quick reference guide" for confusion matrix terminology** because I couldn't find an existing resource that suited my requirements: compact in presentation, using numbers instead of arbitrary variables, and explained both in terms of formulas and sentences.

Let's start with an **example confusion matrix for a binary classifier** (though it can easily be extended to the case of more than two classes):

n=165	**Predicted: NO**	**Predicted: YES**
Actual: NO	50	10
Actual: YES	5	100

What can we learn from this matrix?

- There are two possible predicted classes: "yes" and "no". If we were predicting the

presence of a disease, for example, "yes" would mean they have the disease, and "no" would mean they don't have the disease.
- The classifier made a total of 165 predictions (e.g., 165 patients were being tested for the presence of that disease).
- Out of those 165 cases, the classifier predicted "yes" 110 times, and "no" 55 times.
- In reality, 105 patients in the sample have the disease, and 60 patients do not.

Let's now define the most basic terms, which are whole numbers (not rates):

- **true positives (TP):** These are cases in which we predicted yes (they have the disease), and they do have the disease.
- **true negatives (TN):** We predicted no, and they don't have the disease.
- **false positives (FP):** We predicted yes, but they don't actually have the disease. (Also known as a "Type I error.")
- **false negatives (FN):** We predicted no, but they actually do have the disease. (Also known as a "Type II error.")

I've added these terms to the confusion matrix, and also added the row and column totals:

n=165	**Predicted: NO**	**Predicted: YES**	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

This is a list of rates that are often computed from a confusion matrix for a binary classifier:

- **Accuracy:** Overall, how often is the classifier correct?
  - (TP+TN)/total = (100+50)/165 = 0.91
- **Misclassification Rate:** Overall, how often is it wrong?
  - (FP+FN)/total = (10+5)/165 = 0.09
  - equivalent to 1 minus Accuracy
  - also known as "Error Rate"
- **True Positive Rate:** When it's actually yes, how often does it predict yes?
  - TP/actual yes = 100/105 = 0.95
  - also known as "Sensitivity" or "Recall"
- **False Positive Rate:** When it's actually no, how often does it predict yes?
  - FP/actual no = 10/60 = 0.17
- **True Negative Rate:** When it's actually no, how often does it predict no?
  - TN/actual no = 50/60 = 0.83
  - equivalent to 1 minus False Positive Rate
  - also known as "Specificity"
- **Precision:** When it predicts yes, how often is it correct?
  - TP/predicted yes = 100/110 = 0.91
- **Prevalence:** How often does the yes condition actually occur in our sample?
  - actual yes/total = 105/165 = 0.64

A couple other terms are also worth mentioning:

- **Null Error Rate:** This is how often you would be wrong if you always predicted the majority class. (In our example, the null error rate would be 60/165=0.36 because if you always predicted yes, you would only be wrong for the 60 "no" cases.) This can be a useful

baseline metric to compare your classifier against. However, the best classifier for a particular application will sometimes have a higher error rate than the null error rate, as demonstrated by the **Accuracy Paradox**.

- **Cohen's Kappa:** This is essentially a measure of how well the classifier performed as compared to how well it would have performed simply by chance. In other words, a model will have a high Kappa score if there is a big difference between the accuracy and the null error rate. (**More details about Cohen's Kappa.**)
- **F Score:** This is a weighted average of the true positive rate (recall) and precision. (**More details about the F Score.**)
- **ROC Curve:** This is a commonly used graph that summarizes the performance of a classifier over all possible thresholds. It is generated by plotting the True Positive Rate (y-axis) against the False Positive Rate (x-axis) as you vary the threshold for assigning observations to a given class. (**More details about ROC Curves.**)

And finally, for those of you from the world of Bayesian statistics, here's a quick summary of these terms from **Applied Predictive Modeling**:

> In relation to Bayesian statistics, the sensitivity and specificity are the conditional probabilities, the prevalence is the prior, and the positive/negative predicted values are the posterior probabilities.

Want to learn more?

In my new 35-minute video, **Making sense of the confusion matrix**, I explain these concepts in more depth and cover more **advanced topics:**

- How to calculate precision and recall for multi-class problems
- How to analyze a 10-class confusion matrix
- How to choose the right evaluation metric for your problem
- Why accuracy is often a misleading metric

Data School Comment Policy

All comments are moderated, and will usually be approved by Kevin within a few hours. Thanks for your patience!

Engr Ali Raza • a year ago

Dear Kevin,
can you please tell me the relationship between Misclassifcations and split value or split value index??also explain that how we can calculate

the split value in single decision algorithm?

SplitValue/Set Cases Misclassifications Class Node Type
22.65 5 0.8 1.6 Decision
N/A 1 0 1.6 Terminal
8.63 4 0.6 2.7 Decision
14.4 2 0.2 2.7 Decision
34.85 2 0.2 4 Decision
N/A 1 0 2.7 Terminal
N/A 1 0 5.2 Terminal
N/A 1 0 4 Terminal
N/A 1 0 7.5 Terminal

can anyone explain this table?how it ca be generated?what formulas used behind these values

57 • Reply • Share ›

Kevin Markham Mod ➚ Engr Ali Raza
• a year ago

I'm sorry, I'm not familiar with split values or with the table you included in your comment. Good luck!

2 • Reply • Share ›

Jenica J. Wilson ➚ Engr Ali Raza
• 8 months ago

I know that this is 10 months ago but just incase you have not received an answer here it is:

When developing a measure or test you are trying to identify a unique group of people - for example, the Rosenberg Self-Esteem Test purports to identify people with low self-esteem. Good measures, particularly those used in research or assessment users would like to see them have great predictive abilities. One way to do this is to cross-validate. So splitting the data up is way to cross-validate. The small percentage of the data that is used to see how well the measure identifies people with, let say low self-esteem, is then taken or split from the data set. That subset data is set buy you. So, if you want to test

how well the measure identifies future examinee's, you can test this on 30% of your data. So your split value would be set to .70, because 70% of your data would be used as in the analysis to identify the patterns of low self-esteem. Based on the patterns that are found from the 70% of the data set, the sensitivity analysis (what you have above) would use the 30% of the data to test how accurately the measure reclassifies these participants back into their groups.

sensitivity = true positive/false positives+true positives (accuracy rates)
specificity = true negatives/ false negatives+true negatives (accuracy rates)
Prevalence = proportion of group/total N
Detection rates= measure ability to actually detect the various groups

I am not sure what the program was where you received the output above. So maybe clarity on that would help people best identify how to interpret your findings and provide you on how the information can be generated.

Reply • Share ›

ajay kumar • 4 years ago

Respected Sir,
I have two confusion matrices and I want to perform McNemar Test. It is hereby to requesting you please tell me how to generate the values of 2 by 2 matrix I means how to find the values of f11, f12, f21 and f22 from the confusion matrices.

Thank You,

54 Reply • Share ›

Kevin Markham Mod ajay kumar
• 4 years ago

I'm not familiar with McNemar's test, I'm sorry!

Reply • Share ›

Todd de Quincey • a year ago

Great summary! Thanks for putting this together. Definitely going into my notebook of tips and tricks

36 • Reply • Share ›

Kevin Markham Mod → Todd de Quincey • a year ago

You're welcome!

• Reply • Share ›

Davide • 2 years ago

Dear Kevin,
Thanks for the clear explanation.
I have a particular problem and I'm struggling to find answer. I hope your expertise can help me.
I want to compare the performance of a burned area product toward a reference one. My analysis is not pixel based so I'm not creating random points i the map and create the error matrix. This also because the proportion of total burned area vs not burned is really small. I'm therefore comparing fires detected by the product vs reference fires, and create an error matrix using the whole available data and not random samples. Also the error matrix is missing True Negatives.
Concretely:
- reference dataset has 70 fires
- tested product has 36
- 21 fires are correctly detected by the tested product (TP)
- 15 fires detected by the tested product are not true (FP)
- 49 reference fires are not detected by the tested algorithm (FN)
- I didn't add TN

Several questions rise:
- can I use the available derived metrics to quantify the results? (F1, sensitivity and precision) even if I'm not using RANDOM observations but the WHOLE AVAILABLE OBSERVATIONS in the datasets? Or should I only calculate normal rates of detection and omission?

omission ?
- should I add TN observations for the calculation of accuracy ?(how many?) they will also be not random...
- if I add an other burned area product which has, say, 48 fires, the error matrix will have a different number of observations... Are they comparable?
- Is there other statistic tools which would suits better this case?

Thank you for your attention

5 • Reply • Share ›

Kevin Markham Mod → Davide
• 2 years ago

Thanks for your detailed question! I'd be happy to help, but I'm not able to give you advice on this without having a lot more domain knowledge. For example, I don't know what a "burned area product" is, how you compare it against a "reference", what a "pixel-based" analysis might look like, and so on. This is the kind of question that I would imagine discussing with a student for 20 minutes, and only then could I give good advice! Without more information, it would be irresponisble of me to try to answer your questions. I'm sorry, and good luck!

1 • Reply • Share ›

Hritik Singh • 8 months ago

How to calculate the various metrics like precision and recall with 3 classes (eg- the iris dataset)

2 • Reply • Share ›

Kevin Markham Mod → Hritik Singh
• 8 months ago

Great question! To calculate per-class precision for the iris dataset, for example, you are answering three questions: (1) When it predicts setosa, how often is it correct? (2) When it predicts versicolor, how often is it correct? (3) When it predicts virginica

correct? (3) When it predicts virginica, how often is it correct?

To answer question 1, for example, the denominator is "number of setosa predictions" and the numerator is "how many of those were correct". You can extend that in order to answer questions 2 and 3.

Similarly, to calculate per-class recall, you would answer questions like: (1) When the true class is setosa, how often does it predict setosa? etc.

You can see a simple example of 3-class precision and recall here: http://scikit-learn.org/sta...

Hope that helps!

1 • Reply • Share ›

Jenica J. Wilson ➚ Hritik Singh
• 8 months ago

if you are using R here is the formula: This will give you the actual accuracy of how well your classifier, here it is LDA, reclassified the groups.

```
p1 <- predict(lda,dataset)$class
tab <-
table(Predict=p1,Actual=dataset$Group)#
of accurate predictions/confusion matrix
accuracy <-
sum(diag(tab))/sum(tab)#accuracy of
model
tab
accuracy
```

If you want to cross-validate in R see below.
Packages that you would need

see more

• Reply • Share ›

hana • 8 months ago

I work on credit risk project the confusion matrix is TP=187 TN=6 FP 10 AND FN=3 I calculate

some Metrix like accuracy , error sensitivity and other also I suggest the earn and lose money from confusion matrix my QU is I want some refrences for the calculation of earn and lose money only
thanks

1 • Reply • Share ›

Kevin Markham Mod ➔ hana
• 8 months ago

I'm sorry, I don't completely understand your question. Good luck!

• Reply • Share ›

Tamil Selvan • 2 years ago

can you please clarify me what is the difference between misclassification error and misclassification rate?

1 • Reply • Share ›

Kevin Markham Mod ➔ Tamil Selvan
• 2 years ago

I wouldn't recommend using the term "misclassification error". A "classification error" is a single instance in which your classification was incorrect, and a "misclassification" is the same thing, whereas "misclassification error" is a double negative.

"Misclassification rate", on the other hand, is the percentage of classifications that were incorrect.

3 • Reply • Share ›

Raj Kandala • 3 years ago

Hi, I want to calculate ROC plots using multiclass confusion matrix. Is it possible Mr. Kevin?

1 • Reply • Share ›

Kevin Markham Mod ➔ Raj Kandala
• 3 years ago

ROC curves can only be drawn for binary classification problems, meaning problems with two possible output classes. (However, you can turn a

multiclass problem into a binary problem using a "one versus all" approach.)

Also, drawing an ROC curve requires you to calculate the predicted probability of class membership for each observation, rather than just the class predictions. Therefore, a confusion matrix alone (even in the binary case) does not provide enough data in order for you to draw the ROC curve.

This post provides more information about ROC curves: http://www.dataschool.io/ro...

Hope that helps!

Reply • Share ›

Raj Kandala ➔ Kevin Markham • 3 years ago

Thank You Kevin

Reply • Share ›

Abdullah Nazzal • 3 years ago

what if we have more than two classes ... say 5 ..
how to calculate this .. ? should i convert it to binary ?
if so then how ?

1 Reply • Share ›

Kevin Markham Mod ➔ Abdullah Nazzal • 3 years ago

Many of these terms only have meaning for binary classification problems. You can convert any multi-class problem to a binary problem simply by grouping output classes together, though I wouldn't recommend doing that just so that you can use a particular metric to evaluate your model. Rather, you should only convert it from multi-class to binary if that makes sense in the context of your problem.

Reply • Share ›

Hassan Aftab Mughal
Kevin Markham • 3 years ago

I think if you have more than two classes then your transfusion matrix will be like this
n=165 Class1 Class2 Class3 Class4
Class1 10 3 3 0
Class2 5 15 1 2
Class3 0 5 20 1
Class4 0 2 1 9

2 • Reply • Share ›

Show more replies

Eshwar Bandaru • 4 years ago

Thanks a lot for the simple yet wonderful article.

I'm confused with True positive rate and Precision. Which is a better metric for evaluating the correctness of the model when the actual value is "YES"? I believe it's the True positive rate, as it evaluates the accuracy with the actual value. When exactly to use each of these and If possible, Could you share the use of these metrics with an example?

1 • Reply • Share ›

Kevin Markham Mod ➚ Eshwar Bandaru • 4 years ago

Great question, Eshwar. Neither metric is inherently "better", rather they serve different goals. If I have a spam filter (in which the positive class is "spam" and the negative class is "not spam"), I might optimize for precision or specificity because I want to minimize false positives (cases in which non-spam is sent to the spam box). If I have a metal detector (in which the positive class is "has metal"), I might optimize for sensitivity (also known as True Positive Rate) because I want to minimize false negatives (cases in which someone has metal and the detector doesn't detect it).

Hope that helps!

1 • Reply • Share ›

Sotos ➚ Kevin Markham
• 2 years ago

Thanks a lot for this article! Can you name a couple of technics to optimize a model for better precision using main algorithms, such as svm,rf,xgb.

^ | v • Reply • Share ›

Show more replies

Atif Imam • 13 days ago

One of the best tutorial . Simple easy , concise and to the point

^ | v • Reply • Share ›

Amine Musk • 4 months ago

Dear Kevin ,
Thanks for the explanation and this article. I'm having a real problem into understanding how did you fill the Table with TP and TN , FP ,FN in the example of the diseas, for me following the analogy of your description of each outcome , the confusion matrix for the example will be as such :

-TP : These are cases in which we predicted yes (they have the disease), and they do have the disease, according to the example , we predicted 110 Yes and the reality is that there is 105 person have the deseas , so the TP should be 105 , why we do have 100 ?

- The FP : We predicted yes, but they don't actually have the disease.in the example we predicted yes 110 times and we have in the reality 105 cases of the desease so the FP is 5 .

I think I'm missing something in the analogy of thinking to comprehend the variables .
Thanks for your reply