Launch a data science career!

New? Start here!

Machine Learning course

Subscribe to our newsletter

About

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

Launch a data science career!

New? Start here!

Machine Learning course

Subscribe to our newsletter

About

"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:

Launch a data science career!

New? Start here!

Machine Learning course

Subscribe to our newsletter

About

- **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

Launch a data science career!

New? Start here!

Machine Learning course

Subscribe to our newsletter

About

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:**

Launch a data science career!

New? Start here!

Machine Learning course

Subscribe to our newsletter

About

- 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

Let me know if you have any questions!

Data School Comment Policy

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

Comments Community Privacy Policy

Recommend 50 Tweet Share

Join the discussion…

New? Start here!

Machine Learning course

Subscribe to our newsletter

About

Engr Ali Raza • 3 years ago

Dear Kevin,
can you please tell me the relationship betwe
Misclassifcations and split value or split value
explain that how we can calculate the split va
decision algorithm?

SplitValue/Set Cases Misclassifications Class
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 ge
formulas used behind these values

78 • Reply • Share ›

Jenica J. Wilson → Engr Ali Raza • 2 years a

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

When developing a measure or test yo
identify a unique group of people - for
Rosenberg Self-Esteem Test purports
people with low self-esteem. Good me
particularly those used in research or
users would like to see them have gre
abilities. One way to do this is to cros
splitting the data up is way to cross-va
percentage of the data that is used to
measure identifies people with, let say
esteem, is then taken or split from the
subset data is set buy you. So, if you
well the measure identifies future exar
test this on 30% of your data. So your
would be set to .70, because 70% of y
be used as in the analysis to identify t
low self-esteem. Based on the patterr
from the 70% of the data set, the sens
(what you have above) would use the

to test how accurately the measure re
participants back into their groups.

sensitivity = true positive/false positive
(accuracy rates)
specificity = true negatives/ false nega
negatives (accuracy rates)
Prevalence = proportion of group/tota
Detection rates= measure ability to ac
various groups

I am not sure what the program was w
received the output above. So maybe
would help people best identify how t
findings and provide you on how the i
be generated.

2 ^ | v • Reply • Share ›

Kevin Markham Mod ➦ Engr Ali Raza • 3 ye

I'm sorry, I'm not familiar with split val
table you included in your comment. (

2 ^ | v • Reply • Share ›

ajay kumar • 5 years ago

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

Thank You,

56 ^ | v • Reply • Share ›

Peter ➦ ajay kumar • 10 months ago

dear friend, try `confusionMatrix` functi
package

^ | v • Reply • Share ›

Kevin Markham Mod ➦ ajay kumar • 5 year

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

^ | v • Reply • Share ›

Todd de Quincey • 2 years ago

Great summary! Thanks for putting this toget
going into my notebook of tips and tricks

38 ^ | v • Reply • Share ›

Kevin Markham Mod ➦ Todd de Quincey •

You're welcome!

You're welcome.

^ | v • Reply • Share ›

data school

Launch a data science career!

New? Start here!

Machine Learning course

Subscribe to our newsletter

About

Davide • 3 years ago

Dear Kevin,
Thanks for the clear explanation.
I have a particular problem and I'm strugglinc
hope your expertise can help me.
I want to compare the performance of a burn
toward a reference one. My analysis is not pi
not creating random points i the map and cre
matrix. This also because the proportion of tc
vs not burned is really small. I'm therefore co
detected by the product vs reference fires, ar
matrix using the whole available data and not
samples. Also the error matrix is missing True
Concretely:
- reference dataset has 70 fires
- tested product has 36
- 21 fires are correctly detected by the tested
- 15 fires detected by the tested product are
- 49 reference fires are not detected by the te
(FN)
- I didn't add TN

Several questions rise:
- can I use the available derived metrics to qu
results? (F1, sensitivity and precision) even if
RANDOM observations but the WHOLE AVAI
OBSERVATIONS in the datasets? Or should I
normal rates of detection and omission?
- should I add TN observations for the calcul
?(how many?) they will also be not random...
- if I add an other burned area product which
fires, the error matrix will have a different num
observations... Are they comparable?
- Is there other statistic tools which would su
case?

Thank you for your attention

5 ^ | v • Reply • Share ›

Kevin Markham Mod ➦ Davide • 3 years ag

Thanks for your detailed question! I'd
help, but I'm not able to give you advi
without having a lot more domain kno
example, I don't know what a "burnec
is, how you compare it against a "refe
"pixel-based" analysis might look like,
is the kind of question that I would im
with a student for 20 minutes, and onl

Launch a data science career!

New? Start here!

Machine Learning course

Subscribe to our newsletter

About

give good advice! Without more inforr
be irresponisble of me to try to answe
I'm sorry, and good luck!

1 ^ | v • Reply • Share ›

Priyansh Agarwal • 3 years ago

Hi Kevin
I am confused with Type1 and TypeII errors

false positives (FP): We predicted yes, but the
have the disease. (Also known as a "Type I er
false negatives (FN): We predicted no, but the
have the disease. (Also known as a "Type II e

Type1 error is Worst than Type 2 error and bu
Type 2 error looks worst than Type 1, becaus
when we have a disease is more bad that the

Thank

4 ^ | v • Reply • Share ›

Kevin Markham Mod → Priyansh Agarwal •

Thanks for your question. It appears tl
saying that in general, type I errors are
type II errors. That's not actually the c
is "worse" depends on the particular s
also depends on your perspective. Ho

^ | v • Reply • Share ›

manan • 8 months ago

hi there,
i m working with election predictions. i have c
model using 2018 election data-set and test
2008 election data-set. now my question is tl
the mean of all confusion matrix for three ele
single model.

2 ^ | v • Reply • Share ›

Kevin Markham Mod → manan • 8 months

I don't think I understand your goals v
provide any advice for this scenario -

^ | v • Reply • Share ›

NLR • 8 months ago

Hi Kevin, this is extremely helpful. Do you hav
number for your reference to Applied Predicti
regard to getting Bayesian calculations from
matrix?

2 ^ | v • Reply • Share ›

Launch a data science career!

New? Start here!

Machine Learning course

Subscribe to our newsletter

About

Kevin Markham Mod → NLR • 8 months ag
I don't have a page number, I'm sorry!
2 ^ | v • Reply • Share ›

Hritik Singh • 2 years ago
How to calculate the various metrics like prec
with 3 classes (eg- the iris dataset)
2 ^ | v • Reply • Share ›

Kevin Markham Mod → Hritik Singh
• 2 years ago • edited
Great question! To calculate per-class
iris dataset, for example, you are answ
questions: (1) When it predicts setosa
correct? (2) When it predicts versicolo
correct? (3) When it predicts virginica,
correct?

To answer question 1, for example, th
"number of setosa predictions" and th
"how many of those were correct". Yo
that in order to answer questions 2 an

Similarly, to calculate per-class recall,
answer questions like: (1) When the tr
setosa, how often does it predict seto

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

Hope that helps!
1 ^ | v • Reply • Share ›

Jenica J. Wilson → Hritik Singh • 2 years ag
if you are using R here is the formula:
This will give you the actual accuracy
classifier, here it is LDA, reclassified th

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

If you want to cross-validate in R see
Packages that you would need
install.packages("caret")
library("caret")
library(klaR)

Launch a data science career!

New? Start here!

Machine Learning course

Subscribe to our newsletter

About

library(klar)

You can choose your split

see more

Reply • Share ›

hana • 2 years ago

I work on credit risk project the confusion ma
TN=6 FP 10 AND FN=3 I calculate some Met
, error sensitivity and other also I suggest the
money from confusion matrix my QU is I wan
for the calculation of earn and lose money or
thanks

1 • Reply • Share ›

Kevin Markham Mod → hana • 2 years ago

I'm sorry, I don't completely understar
Good luck!

Reply • Share ›

Tamil Selvan • 3 years ago • edited

can you please clarify me what is the differen
misclassification error and misclassification r

1 • Reply • Share ›

Kevin Markham Mod → Tamil Selvan • 3 yea

I wouldn't recommend using the term
"misclassification error". A "classificat
single instance in which your classifica
incorrect, and a "misclassification" is
whereas "misclassification error" is a

"Misclassification rate", on the other h
percentage of classifications that were

3 • Reply • Share ›

Raj Kandala • 4 years ago

Dear Kevin

Sub: In multi class confusion matrix, Finding
each class.

According to accuracy defnition , Acc=
(TP+TN)/(TP+TN+FP+FN). accuracy is high e
TP=0,. How can we judge the performance ir
the same time precsion= TP/(TP+FP) is zero.
explain this situation.

1 • Reply • Share ›

Kevin Markham Mod → Raj Kandala • 4 yea

Launch a data science career!

New? Start here!

Machine Learning course

Subscribe to our newsletter

About

In a multi-class problem (meaning mo
classes), the accuracy is simply "corre
divided by "total predictions".

If you want to evaluate each class ind
option is to calculate the per-class pre
class recall. scikit-learn's classificatio
this, for example: http://scikit-learn.or

1 ^ | v • Reply • Share ›

Raj Kandala → Kevin Markham • 4 y

Thank you kevin

56 ^ | v • Reply • Share ›

Raj Kandala • 4 years ago

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

1 ^ | v • Reply • Share ›

Kevin Markham Mod → Raj Kandala • 4 yea

ROC curves can only be drawn for bir
problems, meaning problems with two
classes. (However, you can turn a mul
into a binary problem using a "one ve
approach.)

Also, drawing an ROC curve requires
the predicted probability of class mem
observation, rather than just the class
Therefore, a confusion matrix alone (e
case) does not provide enough data in
draw the ROC curve.

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

Hope that helps!

3 ^ | v • Reply • Share ›

Raj Kandala → Kevin Markham • 4 y

Thank You Kevin

1 ^ | v • Reply • Share ›

Abdullah Nazzal • 4 years ago • edited

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

1 ^ | v • Reply • Share ›

Kevin Markham Mod → Abdullah Nazzal • 4

Launch a data science career!

New? Start here!

Machine Learning course

Subscribe to our newsletter

About

Many of these terms only have meani
classification problems. You can conv
class problem to a binary problem sir
output classes together, though I wou
doing that just so that you can use a p
to evaluate your model. Rather, you sh
convert it from multi-class to binary if
sense in the context of your problem.

^ | v • Reply • Share ›

Hassan Aftab Mughal → Kevin Ma
• 4 years ago

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

2 ^ | v • Reply • Share ›

Kevin Markham Mod → H
Mughal • 4 years ago

That's correct, you can
confusion matrix for a n
problem. (My apologies
otherwise!) However, m
terminology outlined in
applies to binary classif
problems.

Here's an example of a
matrix: http://scikit-lear

As a side note, the "n='
upper left corner of the
refers to the number of
so in the case of your 4
matrix, it would be n=7

^ | v • Reply • Share