

Confusion Matrix

Confusion Matrix

Q1) How to know which method is good without data?

- logistic, random forest, knn... →
- ① dividing data into Training & testing sets
→ excellent opportunity to utilize Cross-validation
- ② Using the Training data, we train all the methods we're interested in
- ③ Test each method in testing set

{ Now, we need to identify how each method performed } creating by confusion matrix for each method.

* The rows in CM → what the ML predicts

* The columns → The Known truth

* NOTE

step 2: when we've finished comparing, and find out that there're let's say 2 confusion matrices are very similar and make it hard to choose

then, we'll use sophisticated matrices like ROC, AOC, Sensitivity, specificity..

	actual	
	Has Heart disease	Doesn't
Has	TP	FP
Doesn't	FN	TN

how many times the algorithm missed up

how many times the samples correctly

The CM have Rows & cols each row will have its own Recall any cell else is how many times the algorithm missed up

	1	2	3
1	✓		
2		✓	
3			✓

Correct prediction

(Recall)/(TPR) → TRUE POSITIVE RATE

① sensitivity: what is the Percentage of Patients with heart disease are correctly identified for the condition

$$= \frac{TP}{TP + FN}$$
 [when it's actually Yes, how often does it say Yes]

② specificity: what is the Percentage of Patient without heart disease are correctly identified.
 (TNR)

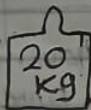
$$= \frac{TN}{TN + FP}$$
 [when it actually No, how often does it say No]

ACCURACY vs PRECISION

- accuracy: how close you become to the correct result
- your accuracy improves with tools calibrated correctly and that you're well trained on



ACCURATE

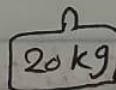


each time i try to weigh
 19.20
 21.30
 22.0

ACCURATE!



PRECISE!



19.11
 19.13
 19.20

PRECISE!

* It's important that measuring devices are accurate & precise

③ Accuracy : Overall, how often is the classifier correct??

$$\frac{TP + TN}{TP + TN + FP + FN} = 1 - \text{Miss Classification Rate}$$

④ Miss Classification Rate : overall, how often is the classifier wrong!

$$\frac{FN + FP}{FN + FP + TP + TN} \rightarrow \text{should be low}$$

⑤ FPR "False Positive Rate" : Overall, when it's actually No, how often does it Predict Yes!

$$\frac{FP}{FP + TN}$$

so, it should be low to be a good model.

⑥ Precision

when it predicts Yes, how often it's actually Yes

$$\frac{TP}{TP + FP}$$

⑦ prevalence

how often Yes occurs in our sample

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

example

Truth

Prediction

→ 7 are a man Prediction

- 4 of them are correct $TP=4$
- 3 are wrong $FP=3$

→ 3 are a Not a man Prediction

actually

- 1 is right $TN=1$
- 2 are wrong $FN=2$

	man	not a man
man	4	3
Not	2	1

* accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$

= $\frac{5}{10} = \underline{\underline{50\%}}$

* Precision = $\frac{TP}{TP+FP} = \frac{4}{7} = \underline{\underline{57\%}}$

* Recall = $\frac{TP}{TP+FN} = \frac{4}{6} = \underline{\underline{67\%}}$

* F1 = balance measure

Precision