Andrew Gard - equitable.equations@gmail.com

Beyond accuracy in classification models

Example. An algorithm for identifying potential users of a new service on a website was applied to 250 test cases with the following results.

		Predicted	
		Positive	Negative
Actual	Positive	33	22
	Negative	7	188

How good is the algorithm?

This is an example of a **confusion matrix**. The desireable results are highlighted in green while the errors are highlighted in red.

		Predicted	
		Positive	Negative
Actual	Positive	33	22
	Negative	7	188

This is an example of a **confusion matrix**. The desireable results are highlighted in green while the errors are highlighted in red.

		Predicted	
		Positive	Negative
Actual	Positive	33	22
	Negative	7	188

The upper-right and lower-left entries show false negatives and false positives, respectively.

The algorithm has an overall **accuracy** of $(33 + 188)/250 = 88.4\%$.

Actual \ Predicted	Positive	Negative
Positive	33	22
Negative	7	188

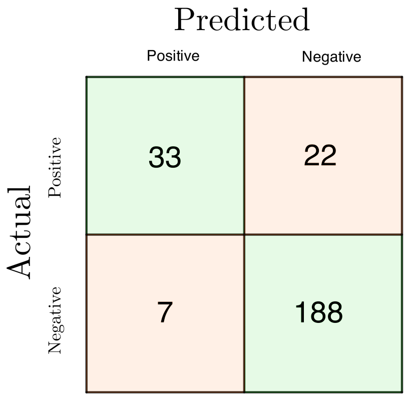
The algorithm has an overall **accuracy** of $(33 + 188)/250 = 88.4\%$.

		Predicted	
		Positive	Negative
Actual	Positive	33	22
	Negative	7	188

This might look good, but it doesn't give a full picture of the effectiveness of the algorithm.

		Predicted	
		Positive	Negative
Actual	Positive	33	22
	Negative	7	188

Of the 250 observations in this set, 195 are negatives, or 78%, so simply classifying *every* new observation as negative will result in an accuracy of 78%.



Of the 250 observations in this set, 195 are negatives, or 78%, so simply classifying *every* new observation as negative will result in an accuracy of 78%. Really, our algorithm only increases accuracy by a few points

Accuracy isn't the only way of describing the success of a classifier. In this vid, we'll cover a few important ones.

Accuracy isn't the only way of describing the success of a classifier. In this vid, we'll cover a few important ones.

- **sensitivity** and **specificify**

Accuracy isn't the only way of describing the success of a classifier. In this vid, we'll cover a few important ones.

- **sensitivity** and **specificify**
- **precision** and **recall**

Accuracy isn't the only way of describing the success of a classifier. In this vid, we'll cover a few important ones.

- **sensitivity** and **specificify**
- **precision** and **recall**

Others metrics that are commonly used include Cohen's kappa, true and positive likelihood ratios, and the F-score. Each of these answers a different question about the classifier's performance.

Sensitivity, or true positive rate, is the proportion of positives in the set that are properly identified.

Sensitivity, or true positive rate, is the proportion of positives in the set that are properly identified.

		Predicted	
		Positive	Negative
Actual	Positive	33	22
	Negative	7	188

In this example, the test's sensitivity is $33/(33+22) = 60\%$.

Sensitivity, or true positive rate, is the proportion of positives in the set that are properly identified.

		Predicted	
		Positive	Negative
Actual	Positive	33	22
	Negative	7	188

In this example, the test's sensitivity is $33/(33+22) = 60\%$. Sensitivity is complementary to the **false negative rate**, the proportion of true positives which are wrongly classified.

Specificity, or true negative rate, is the proportion of negatives in the set that are properly identified.

Specificity, or true negative rate, is the proportion of negatives in the set that are properly identified.

		Predicted	
		Positive	Negative
Actual	Positive	33	22
	Negative	7	188

In this example, the test's specificity is $188/(7+188) \approx 96.4\%$.

Specificity, or true negative rate, is the proportion of negatives in the set that are properly identified.

		Predicted	
		Positive	Negative
Actual	Positive	33	22
	Negative	7	188

In this example, the test's specificity is $188/(7 + 188) \approx 96.4\%$. Specificity is complementary to the **false positive rate**, the proportion of true negatives which are wrongly classified.

Precision and recall both consider context for the true positives in a set.

Precision and recall both consider context for the true positives in a set.

Precision represents the proportions of positive classifications which are actually positives.

		Predicted	
		Positive	Negative
Actual	Positive	33	22
	Negative	7	188

Precision and recall both consider context for the true positives in a set.

Precision represents the proportions of positive classifications which are actually positives.

		Predicted	
		Positive	Negative
Actual	Positive	33	22
	Negative	7	188

In this example, the precision is $33/(33+7) = 82.5\%$.

Precision and recall both consider context for the true positives in a set.

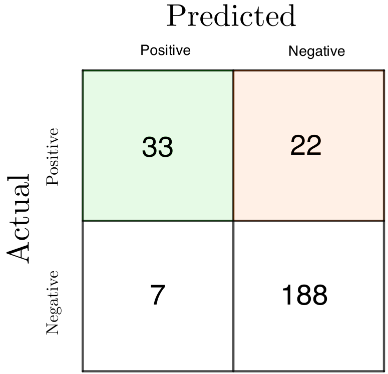
Precision represents the proportions of positive classifications which are actually positives.

		Predicted	
		Positive	Negative
Actual	Positive	33	22
	Negative	7	188

In this example, the precision is $33/(33+7) = 82.5\%$. Precision measures how reliable a positive prediction is.

Recall represents the proportion of true positives detected by the test. It is quantitatively the same as sensitivity.

Recall represents the proportion of true positives detected by the test. It is quantitatively the same as sensitivity.



In this example, the recall is $33/(33+22) = 60\%$.

There isn't a single best measure for the performance of a classification model.

Real-world considerations must be taken into account

There isn't a single best measure for the performance of a classification model. Real-world considerations must be taken into account For instance:

- How important is it to identify true positives, and what is the cost of a false positive?

There isn't a single best measure for the performance of a classification model. Real-world considerations must be taken into account For instance:

- How important is it to identify true positives, and what is the cost of a false positive?
- How important is it to identify true negatives, and what is the cost of a false negative?

There isn't a single best measure for the performance of a classification model. Real-world considerations must be taken into account For instance:

- How important is it to identify true positives, and what is the cost of a false positive?
- How important is it to identify true negatives, and what is the cost of a false negative?

Additionally, we might consider how the model compares to one that always picks the majority class or one that uses only observed proportions in some training set.