Notes 03-4: Confusion Matrix

A confusion matrix (Kohavi and Provost, 1998) contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. The following table shows the confusion matrix for a two class classifier.

The entries in the confusion matrix have the following meaning in the context of our study:

- *a* is the number of **correct** predictions that an instance is **negative**,
- *b* is the number of **incorrect** predictions that an instance is **positive**,
- *c* is the number of **incorrect** of predictions that an instance **negative**, and
- *d* is the number of **correct** predictions that an instance is **positive**.

		Predicted	
		Negative	Positive
Actual	Negative	**a**	**b**
	Positive	**c**	**d**

In formal notation, given a specific class, C_j, and a specific database tuple, t_i, a classification task may or may not assign t_i to C_j, while its actual class may or may not be C_j. With only *two* classes, there are *four* possible outcomes:

- *True positive* (`TP`): t_i is predicted to be in C_j, and is actually in C_j.
- *False positive* (`FP`): t_i is predicted to be in C_j, but is not actually in C_j.
- *True negative* (`TN`): t_i is not predicted to be in C_j, and is not actually in C_j.
- *False negative* (`FN`): t_i is not predicted to be in C_j, but is actually in C_j.

The possible outcomes can be summarized in a confusion matrix.

		Predicted Class	
		$t_i \in C_j$	$t_i \notin C_j$
Actual Class	$t_i \in C_j$	`TP`	`FN`
	$t_i \notin C_j$	`FP`	`TN`

A confusion matrix summarizes the predictive quality of the solution to a classification problem.

Quality Measures

Quality measures can be used to describe the relationships between the predicted and actual classifications.

Accuracy (`AC`): The proportion of the total number of instances that were correctly classified to the total number of instances (a.k.a. *predictive accuracy*).

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

Recall (`R`): The proportion of positive instances that were correctly classified to the total number of instances that are actually positive (a.k.a. *true positive rate*, *hit rate*, *sensitivity*).

$$TPR = \frac{TP}{TP + FN}$$

False positive rate (`FPR`): The proportion of negative instances that were incorrectly classified as positive to the total number of instances that are actually negative.

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

True negative rate (`TNR`): The proportion of negative instances that were correctly classified as negative to the total number of instances that are actually negative (a.k.a. *specificity*).

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

False negative rate (`FNR`): The proportion of positive instances that were incorrectly classified as negative to the total number of instances that are actually positive.

$$FNR = \frac{FN}{FN + TP}$$

Precision (P): The proportion of positive instances that were correctly classified as positive to the total number of instances classified as positive.

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

Error rate (E): The proportion of instances that were incorrectly classified to the total number of instances.

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

The measure that is most appropriate may vary depending on the nature of the problem domain.

ADD F1 MEASURE

Examples of Quality Measures – To Be Done

Assume a dataset of 10,000 instances, where 100 are labeled positive, and a classifier that predicts negative for every instance.

EXAMPLE = Classification.B.2.c1

Assume the classifier now predicts positive for every instance.

EXAMPLE = Classification.B.2.c2

Assume 9,900 instances are labeled as positive, and a classifier that predicts positive for every instance.

EXAMPLE = Classification.B.2.c3

Assume the classifier now predicts negative for every instance.

EXAMPLE = Classification.B.2.c4

Geometric Mean Measures

Accuracy is not an adequate measure when the number of negative instances is much greater than the number of positive instances. To account for this, some measures, such as *geometric mean*, include TP in a product.

$$GM_1 = \sqrt{TP \times P}$$
$$GM_2 = \sqrt{TP \times TN}$$

Any classifier using GM_1 or GM_2 as a measure of performance will result in $GM_1 = GM_2 = 0$ if all positive instances are incorrectly classified.