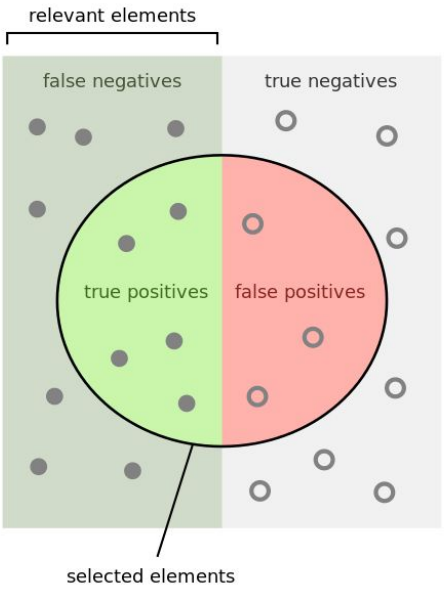
PQP Report - 4/25/2018

Griffin Bishop, Sam Pridotkas, Leo Grande, Harry Sadoyan

What is an F1-score?

When the data categories are highly unbalanced, with 90% of type A and 10% of type B, a simple accuracy can be misleading. A model could trivially achieve an accuracy of 90% by simply classifying everything as an A.

The F_1 score is a measure of accuracy that allows us to solve this problem by considering both precision *p* and recall *r*. Precision is the number of true positives, divided by the number of total positives. Recall is the number of true positives, divided by the number of all values that *should* have been positives i.e. true positives + false negatives. See Figure 1 for a visual representation.



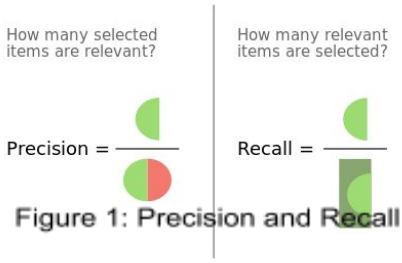


Figure 1: Precision and Recall

The general formula for an arbitrary β is:

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

An F_1 score is simply where $\beta = 1$ and looks like:

$$F_1 = \frac{2}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Commonly used variants are the F_2 score, which emphasizes recall and false negatives, and the $F_{1/2}$ score, which emphasizes precision and false positives.

F1-scores, examples of false negatives/positives for resumes vs movie reviews

$F_1 - score$: 99.89%

True Positive	997
True Negative	48
False Positive	1
False Negative	1

Three-Way Classification: Newsgroups, movie reviews, resumes

PCA with 20 components and random forest classifier (Max depth of 3):

Accuracy: 99.8326%

F1 Score: .998323

Autoencoder and random forest classifier:

Accuracy:74.55%

F1 Score: .9944

Manager versus Acting resumes

PCA with 20 components, random forest classifier (Max depth of 4):

- Accuracy: 78.87%
- F1 Score: 0.8148