Accuracy, Precision, and Recall Implementation Issues

Consider an instance where we use accuracy, precision, and recall (APR) for information retrieval. Rachel googles "restaurants near me" and in the first minute she has 17.5 million results. Let us assume that of these 17.5 million results, relevant links to Rachel's query were 4 million links. Let us also assume that there were around 5 million more relevant links that Google did not show. Rachel calculated the precision for this system by taking the ratio between the relevant links that surfaced and the total number of links that surfaced, which is 4/ 17.5. Rachel calculated the recall value for this system as the ratio between the relevant links that showed up and the sum of the relevant links that showed up and the relevant links that Google did not show, which is 4/ 5. This means the probability that Google would show all the relevant links was 0.8, which is the recall value, and the probability that the retrieved links were relevant is 0.23, which is the precision.

The following are the implementation issues the APR methods pose:

- The performance, of using APR techniques, on a training data set only informs us about what the model is supposed to learn. This is not a good indicator of performance for new/ unseen data. Rachel was able to calculate the precision and recall values based on the data she already possessed, but she cannot figure out a function to make predictions for new data.
- If a user's objectives are unclear during web document retrieval, precision and recall cannot be optimized. For example, when Rachel browse on the web, search results do not have to be very good: recall is not as important (if you get a few good hits), and

precision is not that important (if there are a few hits on the first page).

- For an unbalanced dataset/ highly-skewed data, APR measures are not informative since they are sensitive to class distribution. If Rachel was dealing with imbalanced data, she would not be able to reach a substantial result since the system using APR techniques is sensitive to the dataset.

$$Accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}$$

$$Precision = \frac{T_p}{T_p + F_p}$$

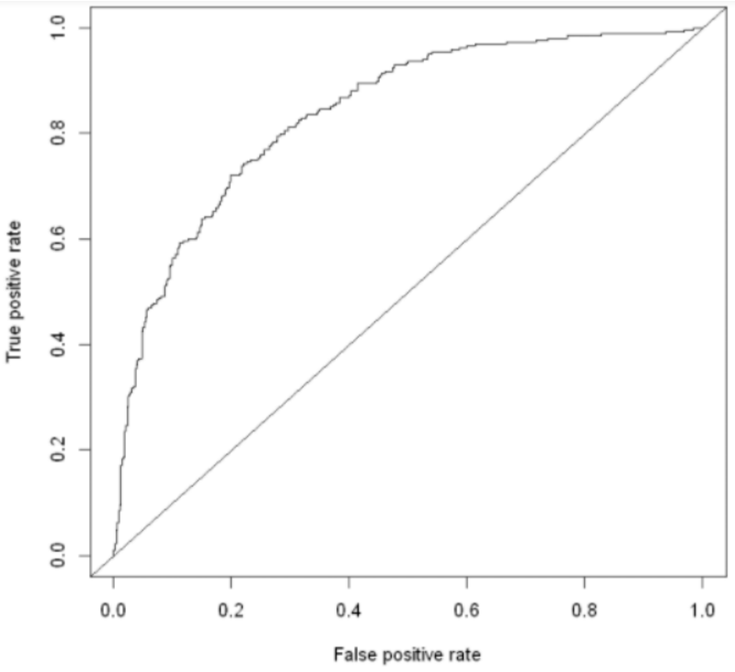
$$Recall = \frac{T_p}{T_p + T_n}$$

The figure above provides the formulas to calculate accuracy, precision, and recall. Here, ’T_p’ is true positive, ‘T_n’ is true negative, ‘F_p’ is false positive, and ‘F_n’ is false negative

- The uncertainty coefficient verifies correlation and not correctness i.e. it will prohibit an algorithmic model for being inconsistent and not for predicting the wrong class unlike accuracy tests with precision and recall, since it simply rearranges the classes (Grimmett, 2016).

- The uncertainty coefficient measures the validity of a statistical classification algorithm unlike the more basic accuracy measures with precision and recall, since it is not affected by various classes' relative fractions.

- For clustering algorithms using APR measures do not make sense. This is so because you will have to know in advance which data points belong to which cluster, and clustering is mostly an unsupervised machine learning algorithm.
- Using a receiver operating characteristic (ROC) curve over an APR curve makes sense when you want to determine the performance of a classifier when the baseline probabilities keep changing. In this scenario, ROC makes more sense since the curve does not change even if the baseline prior probability changes.



The figure above shows a ROC curve with the diagonal across it. The curve is an example model which has performed well since it is not close to the diagonal line

- The ROC curve has a unique property that the diagonal (where true positive rate or TPR is equal to false positive rate or FPR) represents chance, that the distance above the chance line (DAC) represents the probability of an informed decision, and the area under the curve (AUC) represents the probability of correct pairwise ranking. These results are not valid for the APR curve, and the AUC gets distorted for high TPR values. For a ROC curve, the AUC would help Rachel measure the test's discriminative ability i.e. how good the test is in a given situation (Hill, 2012).

Grimmett, H. Introspective Classification for Robot Perception and Decision Making. (2016, August 11). Retrieved from: http://www.robots.ox.ac.uk

Hill, S. Sparse Graphical Models for Cancer Signaling. (2012, May). Retrieved from: http://wrap.warwick.ac.uk