Notes

CSC356 Machine Learning

27 January 2021

accuracy is the fraction of all predictions that are correct.

skewed dataset has a non-uniform distribution of classes—some kinds of objects appear with greater frequency than do other kinds of objects

A confusion matrix. . .

- is a square array of integers

- $\#rows = \#columns =$ number of classes of objects in our world

- the entry in row i, column j is the number of objects that really belong to class i that our classifer said (predicted) belong to class j

- $row \rightarrow actual\ class$

- $column \rightarrow predicted\ class$

- a perfect classifier identifies every object correctly—no false positives, no false negatives

- confusion matrix for a perfect classifier contains zeros everywhere execpt on the main diagonal (the line that runs from the upper left corner to the lower right corner)—the number at row i, column i is just the number of object that belong to class i

- we can produce a confusion matrix in our Python program by calling sklearn's confusion_matrix function with an array of actual values and the corresponding array of predicted values (that our model has computed)

TP A True Positive is a correct claim that an object x is a member of class A.

TN A True Negative is a correct claim that an object x is not a member of class A.

FP A False Positive is an incorrect claim that an object x is a member of class A.

FN A False Negative is an incorrect claim that an object x is not a member of class A.

Precision tells us what fraction of claims that an object x is a member of class A are correct.

Recall tells us what fraction of all members of class A are correctly identified as members of class A.

Sensitivity is a synonym for Recall.

TPR True Positive Rate is another synonym for Recall.

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

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

$$F1 = haromonic\_mean(precision, recall)$$
$$= \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$$

All 3 statistics are numbers between 0.0 and 1.0.

$$0.0 \leq precision \leq 1.0$$
$$0.0 \leq recall \leq 1.0$$
$$0.0 \leq F1 \leq 1.0$$

ROC (Receiver Operating Characteristic)...

- TPR (the True Positive Rate is the Recall) on one axis
- FPR (False Positive Rate) on the other axis
- TNR is the True Negative Rate—it is the fraction of claims that an object x does not belong to class A that are correct
- TNR is also called specificity
- $FPR = 1 - TNR$
- $FPR = 1 - specificity$
- ROC curve shows us the relationship between *sensitivity (recall)* and $1 - sensitivity$