Evaluating Model Performance for Classification

Dr. Muhammad Wasim

Evaluating Classifier Performance

The confusion matrix or contingency table

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

- **Example:** If a hypothesis has learned SICK concept, and for new unseen examples, it predicts an example as SICK (positive) which is not correct, then we have a false positive.

Calculating Accuracy

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

- Problem with Accuracy: Accuracy does not perform well with imbalanced data. For example, if you have 95 negative examples and 5 positive examples, classifying all as negative given 0.95 or 95% accuracy score.

Precision and Recall

- **Precision:** % of predicted correct examples that are correct
- **Recall:** % of correct examples that are predicted correct

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

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

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

A combined measure – F Measure

- A combined measure that assesses the P/R tradeoff is F measure (weighted harmonic mean):

$$F = \frac{1}{\alpha \times \left(\frac{1}{P}\right) + (1 - \alpha) \times \frac{1}{R}}$$

- Normally, balanced F_1 is used (i.e. $\alpha = \frac{1}{2}$)

$$F_1 = \frac{2PR}{P + R}$$

Confusion Matrix for more than two classes

- Let us suppose, we have a ten documents belonging to three different categories: Politics, Sports, Crime

	Predicted Politics	Predicted Sports	Predicted Crime
Actual Politics	2	0	0
Actual Sports	1	3	1
Actual Crime	0	2	1

Micro vs Macro-Average

- If we have more than one class, how do we combine multiple performance measures into one quantity?
- **Macro-averaging:** Compute performance of each class, then average.
- **Micro-averaging:** Collect decisions for all classes, compute contingency table, evaluate.

Micro vs Macro Averaging

Politics	Predicted True	Predicted False
Actual True	2	0
Actual False	1	4

Sports	Predicted True	Predicted False
Actual True	3	2
Actual False	2	3

Crime	Predicted True	Predicted False
Actual True	1	2
Actual False	1	5

	Predicted Politics	Predicted Sports	Predicted Crime
Actual Politics	2	0	0
Actual Sports	1	3	1
Actual Crime	0	2	1

$$P_p = \frac{2}{2+1} = 0.67 \qquad P_s = \frac{3}{3+2} = 0.6 \qquad P_c = \frac{1}{1+1} = 0.5$$

- Macro-average precision: $\frac{0.67+0.6+0.5}{3} = 0.58$

All	Predicted True	Predicted False
Actual True	6	4
Actual False	4	12

- Micro-average precision: $\frac{6}{6+4} = 0.6$
- The micro-average will adequately capture the class imbalance