Evaluation

F-1 Score / Measure

- We can combine the precision and recall into one value called the F-1 score
- This is the harmonic mean of precision and recall

$$F\text{-}1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

- The F-1 score is a specific case of the general F score

$$F\text{-score} = \left(1 + \beta^2\right) \times \frac{(\text{precision} \times \text{recall})}{(\beta^2 \times \text{precision} + \text{recall})}$$

- β is a hyperparameter that increases the importance of recall as β increases
- $\beta = 1$ for F-1 score, which gives equal weight to the precision and recall

		Predicted	
		Planet	Not Planet
Actual	Planet	$tp = 45$	$fn = 8$
	Not Planet	$fp = 23$	$tn = 27$

- Precision $= \frac{tp}{tp+fp} = \frac{45}{45+23} = 0.662$
- Recall $= \frac{tp}{tp+fn} = \frac{45}{45+8} = 0.849$
- F-1 $= 2 \times \frac{0.662 \times 0.849}{0.662+0.849} = 0.744$
- Accuracy $= \frac{tp+tn}{tp+fp+tn+fn} = \frac{45+27}{45+23+27+8} = 0.699$
- Classification tasks can have more than two classes
  - Sentiment analysis (positive, negative, neutral)
  - We can define precision and recall for multiple classes, such as the table below where we predicts an email as urgent, normal, or spam

		Actual / Gold Labels			
		Urgent	Normal	Spam	
Predicted	Urgent	8	10	1	precision_u
	Normal	5	60	50	precision_n
	Spam	3	30	200	precision_s
		recall_u	recall_n	recall_s	

- $\text{precision}_u = \frac{8}{8+10+1}$
- $\text{precision}_n = \frac{60}{5+60+50}$
- $\text{precision}_s = \frac{200}{3+30+200}$
- $\text{recall}_u = \frac{8}{8+5+3}$

- $\text{recall}_n = \frac{60}{10+60+30}$
- $\text{recall}_s = \frac{200}{1+50+200}$

Macroaveraging

- Macroaveraging precision and recall – Mean of previously calculated per-class precision and recall

$$\begin{aligned}\text{Macroaverage Precision} &= \frac{\text{precision}_u + \text{precision}_n + \text{precision}_s}{3} \\ &= \frac{0.42 + 0.52 + 0.8}{3} \\ &= 0.60\end{aligned}$$

Microaveraging

- Create per class confusion matrices and then pool values
- All classes that are not the class of interest will be labeled as the not class

Urgent		Actual	
		Urgent	Not
Predicted	Urgent	8	11
	Not	8	340

Normal		Actual	
		Normal	Not
Predicted	Normal	60	55
	Not	40	212

Spam		Actual	
		Spam	Not
Predicted	Spam	200	33
	Not	51	83

Pooled		Actual	
		Yes	Not
Predicted	Yes	$8 + 60 + 200$	$11 + 55 + 99$
	Not	$8 + 40 + 51$	$340 + 212 + 83$

Pooled		Actual	
		Yes	Not
Predicted	Yes	268	99
	Not	99	635

$$\text{Microaverage Precision} = \frac{268}{268 + 99} = 0.73$$

Other Options for Measuring Performance

- Some other options might be considered for measuring performance

- Error-rate $= 1 -$ accuracy
- Sensitivity (True Positive Rate) $= \frac{tp}{tp+fn} =$ Recall
- Specificity (True Negative Rate) $= \frac{tn}{tn+fp}$
- Top-n Accuracy
  - If the correct value is among the top n results, it is considered correct
- Log Loss (Cross-Entropy Loss)
  - Uses probabilities given to classes
- Receiver-Operating Characteristic (ROC) Curve
  - Calculates the true positive rate and false positive rate across different thresholds

Similarity

Precursor to our first classifier

- Before we can discuss about our first classifier, we need to determine a way to measure similarity and distances between instances
- Similarity – How close two objects / instances are
- Distance – How far apart two objects / instances are
- Smaller distances between objects indicates higher similarity
- We will refer to methods to calculate distance as measurements of similarity

Similarity and Distance Metrics

- Cosine Similarity
- Euclidean Distance
- Manhattan Distance
- Jaccard Similarity
- Simple Matching Coefficient

Cosine Similarity

- Euclidean dot product with vectors X and Y

$$X \cdot Y = ||X||\ ||Y|| \cos(\sigma)$$

$$\cos(\sigma) = \frac{X \cdot Y}{||X||\ ||Y||}$$

- Cosine similarity ranges from -1 to +1
- Perfect similarity is +1
- Orthogonal vectors is 0
- Cosine similarity of vectors X and Y is represented as $S_c(X, Y)$