Methods and metrics: Classifier metrics

Christopher Potts

Stanford Linguistics

CS224u: Natural language understanding

Overview

- Different evaluation metrics **encode different values**.
- Choosing a metric is a crucial aspect to experimental work.
- You should feel free to motivate new metrics and specific uses of existing metrics, depending on what your goals are.
- For established tasks, there is usually pressure to use specific metrics, but you should feel empowered to push back.
- Areas can stagnate due to poor metrics, so we must be vigilant!

Confusion matrices

		Predicted			
		pos	neg	neutral	Support
	pos	15	10	100	125
Gold	neg	10	15	10	35
	neutral	10	100	1000	1110

A threshold was imposed for these categorical predictions.

Accuracy

The correct predictions divided by the total number of examples.

		Predicted		
		pos	neg	neutral
	pos	15	10	100
Gold	neg	10	15	10
	neutral	10	100	1000

- Bounds: $[0, 1]$, with 0 the worst and 1 the best.
- Value encoded: how often is the system correct?
- Weaknesses:
  - No per-class metrics.
  - Failure to control for class size.

Accuracy and the cross-entropy loss

Accuracy is inversely proportional to the negative log-loss (a.k.a. cross entropy loss; sklearn link):

$$-\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{i,k}\log(p_{i,k})$$

Precision

For class k: the correct predictions for k divided by the sum of all guesses for k.

		Predicted		
		pos	neg	neutral
Gold	pos	**15**	10	100
	neg	**10**	15	10
	neutral	**10**	100	1000
	Precision	0.43	0.12	0.90

Precision for pos: 15 / (15 + 10 + 10) = 0.43

- Bounds: $[0, 1]$, with 0 the worst and 1 the best. (Caveat: undefined values resulting from dividing by 0 need to be mapped to 0.)
- Value encoded: penalize incorrect guesses.
- Weakness: Achieve high precision for k simply by rarely guessing k.

Recall

For class *k*: the correct predictions for *k* divided by the sum of all true members of *k*.

		Predicted			
		pos	neg	neutral	Recall
	pos	**15**	**10**	**100**	0.12
Gold	neg	10	15	10	0.43
	neutral	10	100	1000	0.90

Recall for pos: 15 / (15 + 10 + 100) = 0.12

- Bounds: $[0, 1]$, with 0 the worst and 1 the best.
- Value encoded: penalize missed true cases.
- Weakness: Achieve high recall for *k* simply by always guessing *k*.

F scores

$$\mathrm{F}_{\beta}(k) = (\beta^2 + 1) \cdot \frac{\mathrm{Precision}(k) \cdot \mathrm{Recall}(k)}{(\beta^2 \cdot \mathrm{Precision}(k)) + \mathrm{Recall}(k)}$$

		Predicted			
		pos	neg	neutral	F_1
	pos	15	10	100	0.19
Gold	neg	10	15	10	0.19
	neutral	10	100	1000	0.90

- Bounds: $[0, 1]$, with 0 the worst and 1 the best; always between precision and recall.
- Value encoded: how much do predictions for k align with true instances of k, with β controlling the weight places on precision vs. recall
- Weaknesses:
  - No normalization for the size of the dataset.
  - Ignores the values off the row and column for k.

Averaging F scores

- Macro-averaging
- Weighted averaging
- Micro-averaging

Macro-averaged F scores

		Predicted			
		pos	neg	neutral	F_1
	pos	15	10	100	**0.19**
Gold	neg	10	15	10	**0.19**
	neutral	10	100	1000	**0.90**
					0.43

- Bounds: $[0, 1]$, with 0 the worst and 1 the best.
- Value encoded: same values as F scores plus the assumption that all classes are equal.
- Weaknesses:
  - A classifier that does well only on small classes might not do well in the real world.
  - A classifier that does well only on large classes might do poorly on small but vital smaller ones.

Weighted average F scores

		Predicted				
		pos	neg	neutral	Support	F_1
Gold	pos	15	10	100	125	**0.19**
	neg	10	15	10	35	**0.19**
	neutral	10	100	1000	1110	**0.90**
						0.43

$$\frac{0.19 \cdot 125 + 0.19 \cdot 35 + 0.90 \cdot 1110}{125 + 35 + 1110}$$

- Bounds: $[0, 1]$, with 0 the worst and 1 the best.
- Value encoded: same values as F_β plus the assumption that class size matters.
- Weaknesses: Large classes will dominate.

Micro-averaged F scores

		Predicted		
		pos	neg	neutral
Gold	pos	15	10	100
	neg	10	15	10
	neutral	10	100	1000

	yes	no
yes	15	110
no	20	1125

	yes	no
yes	15	20
no	110	1125

	yes	no
yes	1000	110
no	110	50

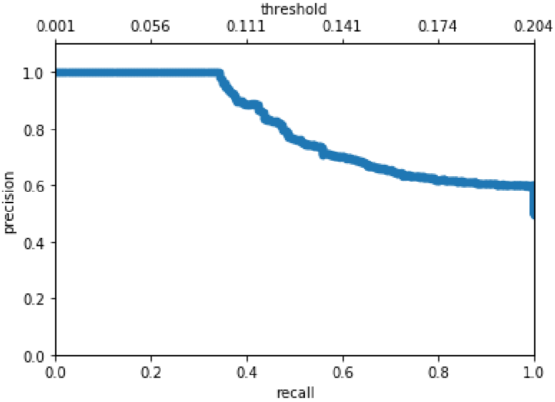
	yes	no	F_1
yes	1030	240	0.81
no	240	2300	0.91

Micro-averaged F scores

- Bounds: $[0, 1]$, with 0 the worst and 1 the best.
- Value encoded: Micro-averaged F_1 for “yes” = accuracy.
- Weaknesses:
  - Same as for weighted F scores, plus
  - a score for “yes” and “no”, hence no single summary number.

Precision–recall curves

Summarizes the relationship between precision and recall by using each predicted probability as a potential threshold:



Average precision provides a summary of the curve.

References I