Inter-annotator agreement

LING83800

1 Introduction

When our models consist of supervised machine learning trained on human annotations, agreement between *multiple human annotators* imposes a upper bound for performance, if only because it is difficult if not impossible to measure system performance exceeding that of human annotators. And, when human annotators do not show substantial agreement, this suggests the annotation task may be ill-defined, or too challenging.

One feature shared by many inter-annotator agreement metrics is the *chance correction*. While one can simply compute a probability that raters will agree, agreement metrics also account for the possibility that annotators will agree by chance.

Bibliographic note

Artstein and Poesio (2008) provide a detailed review of inter-annotator agreement metrics.

2 Cohen's κ

Cohen's κ is a chance-corrected measure of inter-annotator agreement for two raters using nominal (i.e., categorical, or binary) ratings.

Definition Cohen's κ is a function of two probabilities:

- P_o, the probability the two coders agreeing, and
- P_e, the probability of chance agreement.

Let $r_{i,j}$ be rater i's rating for item j, let N be the number of items, and let $C_i(k)$ is the number of times rater i predicted category $k \in \mathbb{K}$. Then,

$$P_o = \frac{1}{N} \sum_{j=1}^{N} \begin{cases} 1 \text{ if } r_{1,j} = r_{2,j} \\ 0 \text{ otherwise} \end{cases}$$

$$P_e = \frac{1}{N^2} \sum_{k \in \mathbb{K}} C_1(k) C_2(k)$$

$$\kappa = \frac{P_o - P_e}{1 - P_e}.$$

Interpretation When $\kappa = 1$, the two raters are in complete agreement; when $\kappa \leq 0$, the raters agree no more often than would be expected by chance. For $0 < \kappa < 1$, the interpretation is less clear, but Landis and Koch (1977:165) propose the following well-known qualitative guidelines:

0.00–0.20: slight agreement
0.21–0.40: fair agreement
0.41–0.60: moderate agreement
0.61–0.80: substantial agreement
0.81–1.00: almost perfect agreement

Problem Compute, and interpret, κ for the following 2×2 contingency table:

		$i = 2$
$i = 1$	20	5
	10	15

Solution

$$N = 20 + 5 + 10 + 15 = 50$$
$$P_o = \frac{20 + 15}{50} = .7$$
$$P_e = \frac{30 \cdot 25 + 25 \cdot 20}{50^2} = .5$$
$$\kappa = \frac{.7 - .5}{1 - .5} = .4$$

This corresponds "fair" agreement on the Landis and Koch qualitative scale.

3 Krippendorff's α

Cohen's κ is only defined for two raters. Krippendorff's α generalizes chance-corrected agreement to an arbitrary number of raters, to missing data, and arbitrary scales of measurement.

Definition Krippendorff's α is a function of:

- D_o, the amount of pairwise disagreement observed, and
- D_e, the amount of pairwise disagreement expected due to chance.

The formulae for these quantities are complex—and challenging to compute by hand—but simplify considerably when there is no missing data, and when ratings are nominal or binary.[1] Then,

$$\alpha = 1 - \frac{D_o}{D_e}$$

The NLTK module `nltk.metrics.agreement` contains a complete implementation.

[1]However, α does not reduce to κ in the two-rater, no-missing-data, nominal-scale case, because κ makes additional statistical independency assumptions.

Interpretation The interpretation is similar to Cohen's κ, though there are no established qualitative guidelines for $0 < \alpha < 1$.

References

Artstein, Ron, and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34:555–596.

Landis, J. Richard, and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33:159–174.