Reliability of Data and Reporting Performance

COMP34812: Natural Language Understanding
Week 4
Riza Batista-Navarro

Learning Outcomes

To measure data reliability based on inter-annotator agreement coefficients, e.g., Kappa

To discuss ways for summarising and comparing performance

Data Reliability

We rely on humans to label or annotate data BUT humans have different perspectives

Annotator agreement: measured to help us decide whether we can trust the labels

- Intra-annotator agreement: whether the same human consistently annotates the same item when presented at different times
- Inter-annotator agreement: whether multiple humans consistently annotate the same item even when working independently

Inter-Annotator Agreement (IAA)

The agreement between human annotators (labellers/coders)

- serves as an indication of the difficulty of the task or how well-defined it is
- serves as an upper bound on the performance of automated methods

Simplistic approach: observed agreement

- ratio of the number of items on which annotators agree, to total number of items
- does not take into account agreement by chance (random agreement)

Observed Agreement

Funny or not?

Text	A1	A2	Agree?
The problem with trouble shooting is that trouble shoots back.	Y	N	✗
A clean desk is a sign of a cluttered desk drawer.	N	Y	✗
What's Blonde and dead in a closet? The Hide and Seek Champion from 1995.	N	N	✓
Moses had the first tablet that could connect to the cloud.	Y	Y	✓
Apparently I snore so loudly that it scares everyone in the car I'm driving.	Y	Y	✓

Source: https://onelinefun.com/

3 times A1 and A2 agree out of 5 = 0.6; not very different from random agreement:

- both A1 and A2 randomly choose Y = 0.5 x 0.5 = 0.25
- both A1 and A2 randomly choose N = 0.5 x 0.5 = 0.25
- expected agreement by chance = 0.5

Cohen's Kappa coefficient

a measure of chance-corrected agreement

$$\kappa = \frac{P(a) - P(e)}{1 - P(e)}$$

P(a) is the **observed agreement**, proportion of times annotators agreed

P(e) is the **expected agreement**, proportion of times annotators expected to agree by chance

Cohen's Kappa coefficient: Example

Assume:

we have two annotators *A1* and *A2*

they are providing annotations for a binary classification task: does a sample belong to some class *c*? Yes or No

$$\kappa = \frac{P(a) - P(e)}{1 - P(e)}$$

P(a) = P(A1=Yes, A2=Yes) + P(A1=No, A2=No)

P(e) = P(A1=Yes)*P(A2=Yes) + P(A1=No)*P(A2=No)

Cohen's Kappa coefficient: Example

	Annotator 1		
	Yes	No	total
			40

Cohen's Kappa coefficient: Example

	Annotator 1		
	Yes	No	total
	33	7	40

Cohen's Kappa coefficient: Example

		Annotator 1		
		Yes	No	total
Annotator 2	Yes			
	No			
	total	33	7	40

Cohen's Kappa coefficient: Example

		Annotator 1		
		Yes	No	total
Annotator 2	Yes	31		
	No	2		
	total	33	7	40

Cohen's Kappa coefficient: Example

		Annotator 1		
		Yes	No	total
Annotator 2	Yes	31	1	
	No	2	6	
	total	33	7	40

Cohen's Kappa coefficient: Example

		Annotator 1		
		Yes	No	total
Annotator 2	Yes	31	1	32
	No	2	6	8
	total	33	7	40

Cohen's Kappa coefficient: Example

		Annotator 1		
		Yes	No	total
Annotator 2	Yes	31	1	32
	No	2	6	8
	total	33	7	40

$P(a)$ = P(A1=Yes, A2=Yes) + P(A1=No, A2=No)

= (31/40) + (6/40)

= 0.925

Cohen's Kappa coefficient: Example

		Annotator 1		
		Yes	No	total
Annotator 2	Yes	31	1	32
	No	2	6	8
	total	33	7	40

$P(e)$ = P(A1=Yes)*P(A2=Yes) + P(A1=No)*P(A2=No)

= ((33/40) * (32/40)) + ((7/40) * (8/40))

= 0.695

Cohen's Kappa coefficient: Example

		Annotator 1		
		Yes	No	total
Annotator 2	Yes	31	1	32
	No	2	6	8
	total	33	7	40

Kappa = (P(a)-P(e))/(1-P(e))

= (0.925-0.695)/(1-0.695) = 0.754

Cohen's Kappa coefficient: Interpretation

Generally: a negative value means disagreement; 0 means no agreement

Landis and Koch, 1977

slight < 0.2 < fair < 0.4 < moderate < 0.6 < substantial < 0.8 < perfect

Grove et al., 1981 (psychiatric community)

0.6 < acceptable

Krippendorff, 1980

0.67 < tentative conclusions < 0.8 < definite conclusions

Rietveld and van Hout, 1993

0.4 < moderate < 0.6 < substantial < 0.8

Green, 1997

low < 0.4 < fair/good < 0.75 < high

Other coefficients for IAA

Scott’s Pi

P(e): different chance for different categories

Fleiss’ Kappa

multi-annotator generalisation of (Cohen's) Kappa and Scott’s Pi

BUT:

These would work if we can define negative cases; for some tasks this is too difficult, e.g., **NER**

Other coefficients for IAA

NER as sequence labelling

A	member	of	the	Democratic	Party	,	Obama	served	as	a	US	president

O	O	O	O	B-ORG	I-ORG	O	B-PER	O	O	O	B-GPE	O

For such tasks, **F-score** is reported instead

- the annotations from one of the annotators is considered as gold standard (reference)
- the annotations from another annotator is considered as response, whose F-score is measured against the reference

Considerations so far...

- ✓ How data is partitioned (into fixed splits, or *k* folds)
- ✓ Whether data is representative
- ✓ Whether data is imbalanced
- ✓ Whether data is reliable
- ❏ How to summarise and compare performance

Evaluation metric

also known as evaluation measure and figure of merit

typical structure for reporting summarised results of evaluations:

	Measure 1	*Measure 2*	*Combined Measure*
Baseline 1	M_1^{B1}	M_2^{B1}	M_c^{B1}
Baseline 2	M_1^{B2}	M_2^{B2}	M_c^{B2}
Variation 1	M_1^{V1}	M_2^{V1}	M_c^{V1}
Variation 2	M_1^{V2}	M_2^{V2}	M_c^{V2}
Upper Bound	M_1^{U}	M_2^{U}	M_c^{U}

The performance we want to improve upon

Our proposed solution (and any variations)

The highest possible performance

Reporting evaluation results: Example

Performance of different humour detection methods on the same dataset

Method	Configuration	Accuracy	Precision	Recall	F1
Decision Tree		0.786	0.769	0.821	0.794
SVM	sigmoid, gamma=1.0	0.872	0.869	0.880	0.874
Multinomial NB	alpha=0.2	0.876	0.863	0.902	0.882
XGBoost		0.720	0.753	0.777	0.813
XLNet	XLNet-Large-Cased	0.916	0.872	0.973	0.920
Proposed		0.982	0.990	0.974	0.982

Annamoradnejad, Issa, and Gohar Zoghi. "ColBERT: Using BERT sentence embedding for humor detection." arXiv preprint arXiv:2004.12765 (2020).

Statistical significance

Not all differences between scores matter

Is an improvement statistically significant, i.e., unlikely to be the result of chance variation?

Accuracy	Precision	Recall	F1
0.786	0.769	0.821	0.794
0.872	0.869	0.880	0.874
0.876	0.863	0.902	0.882
0.720	0.753	0.777	0.813
0.916	0.872	0.973	0.920
0.982	0.990	0.974	0.982

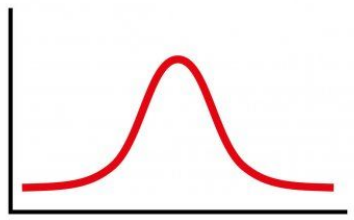
Conventionally, an improvement (or difference) is statistically significant only if the likelihood of it having occurred by chance is less than 5%, i.e., $p < 0.05$

Statistical significance

Null hypothesis statistical testing: can be performed on two samples of data, e.g., different accuracy values

null hypothesis: that there is no difference between the distribution of the two samples of data (i.e., that any variation is due to chance)

statistical test: many different types, depending on whether (1) the data distribution is normal or not, and (2) the data is paired or not



Normal: parametric test
Non-normal: non-parametric test

Drawn from the same "subjects", i.e., test instances

Statistical significance

Statistical test

possible options:

	Unpaired	**Paired**
Parametric (Normal dist.)	Independent t-test (Student's or Welch's)	Paired t-test (Student's or Welch's)
Non-parametric (Non-normal dist.)	Mann-Whitney U test	Wilcoxon signed-ranked test

results in a p-value: if below a threshold (e.g., 0.05), the null hypothesis is rejected

Considerations

- ✓ How data is partitioned (into fixed splits, or *k* folds)
- ✓ Whether data is representative
- ✓ Whether data is imbalanced
- ✓ Whether data is reliable
- ✓ How to summarise and compare performance
- ❑ Which metrics for which tasks (Up next!)