6: Uncertainty and Human Agreement

Machine Learning and Real-world Data (MLRD)

Simone Teufel

Last session: we implemented cross-validation and investigated overtraining

Over the last 5 sessions we have improved our classifier and evaluation method:

- We have created a smoothed NB classifier.
- We can now train and test our classifier using stratified cross-validation.
- We evaluate in a methodologically sound manner.

But we have artificially simplified the classification problem

- In reality there are many reviews that are neither positive nor negative.

Many movie reviews are neither positive nor negative

So far, your data sets have contained only the clearly positive or negative reviews

- Only reviews with extreme star-rating were used.
- This is a clear simplification of the real task.
- If we consider the middle range of star-ratings, things gets more uncertain.

In Session 1 you classified Review 1

What is probably the best part of this film, GRACE, is the pacing. It does not set you up for any roller-coaster ride, nor does it has a million and one flash cut edits, but rather moves towards its ending with a certain tone that is more shivering than horrific....
GRACE is well made and designed, and put together by first time director Paul Solet who also wrote the script, is a satisfying entry into the horror genre. Although there is plenty of blood in this film, it is not really a gory film, nor do I get the sense that this film is attempting at exploiting the genre in any way, which is why it came off more genuine than other horror films. I think the film could be worked out to be scarier, perhaps by building more emotional connection to the characters as they seemed a little on the two dimensional side. They had motivations for their actions, but they did not seem to be based on anything other than because the script said so.
For me, this title is a better rental than buying as I dont feel like its a movie I would return to often. I might give it one more watch to flesh out my thoughts on it, but otherwise it did not leave me with a great impression, other than that it has greater potential than what is presented.

MLRD 2021/22: NEGATIVE=42 POSITIVE=63 MLRD 2020/21: NEGATIVE=39 POSITIVE=74
MLRD 2019/20: NEGATIVE=35 POSITIVE=82
MLRD 2018/19: NEGATIVE=46 POSITIVE=87

- Let the middle range of star-ratings constitute a third class: NEUTRAL
- The ground truth for Review 1 is actually NEUTRAL

Today we will build a 3-class classifier

We will extend our classifier to cope with neutral reviews

- Your first task today will be to train and test a 3-class classifier—classifying positive, negative, neutral reviews.

Do we end up with two kinds of *neutral reviews*?

- Luke-warm reviews (reviews that contain neutral words i.e. reviews that can be characterised as saying that the movie is *ok* or *not too bad*)
- Pro-con reviews (i.e. reviews that list the good points and bad points of the movie)

Can we be certain what the true category of a review should be?

Let us return to 2 class situation to consider this problem
By assigning ground-truth based on star rating we are ignoring some issues:

- Inter-personal differences in interpretation of the rating scale
- Reader's perception vs. writer's perception

Human agreement is one possible source of truth

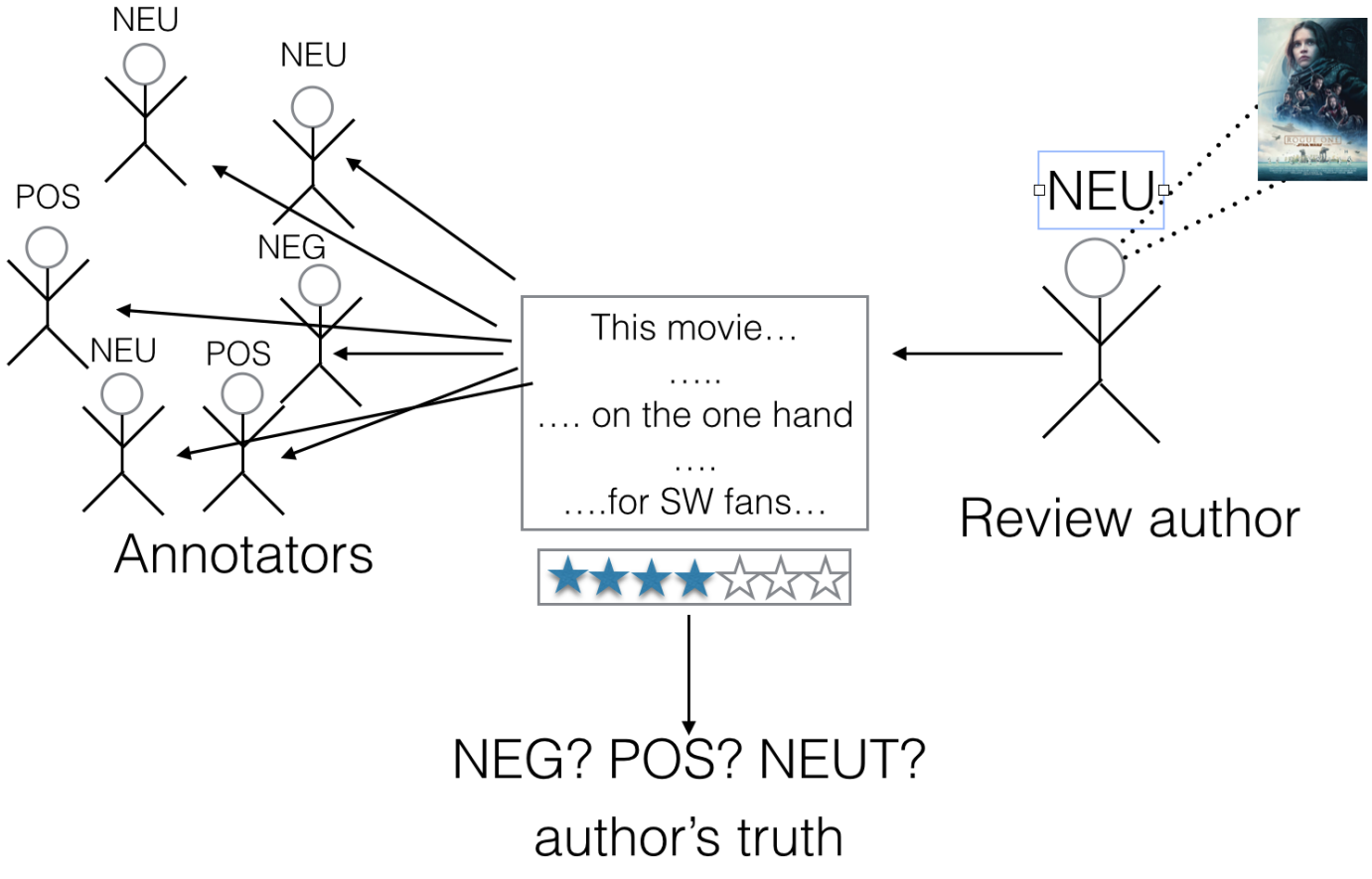
Who is to say what the true category of a review should be?

- Writer's perception is lost to us, but we can get many readers to judge sentiment afterwards.

Claim:
Human agreement is the **only** empirically available source of truth in decisions which are influenced by subjective judgement.

- Something is 'true' if several humans agree on their judgement, independently from each other
- The more they agree they more 'true' it is

Human annotation



Your classification results from Session 1

2020–21	POS	NEG
Review 1	63	42
Review 2	99	6
Review 3	103	2
Review 4	3	102

For your second task today you will quantify how much you as a class agree amongst yourselves.

Previous years' classifications

	2020–21		2019–20		2018–19		2017–18	
	POS	NEG	POS	NEG	POS	NEG	POS	NEG
Review 1	74	39	66	35	87	46	82	35
Review 2	5	108	8	93	11	122	8	109
Review 3	2	111	1	100	2	131	3	114
Review 4	106	7	96	5	130	3	112	5

We can use agreement metrics when we have multiple judges

- Accuracy required a single ground-truth
- We cannot use accuracy because it cannot be used to measure agreement between our 105 judges
- 104 agreeing with each other, 1 judge disagreeing would count as a “wrong decision”
- Instead we calculate $\bar{P}_a$, the observed agreement:

$$\bar{P}_a = \frac{1}{N} \sum_{i=0}^{N-1} \left(\frac{\text{observed rater–rater pairs in agreement on item } i}{\text{possible rater–rater pairs}} \right)$$

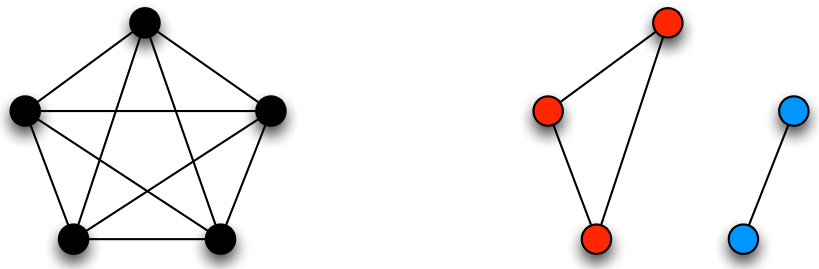
N: number of items

$\bar{P}_a$ observed agreement

- Pairwise observed agreement $\bar{P}_a$: average ratio of observed to possible rater-rater agreements
- There are $\binom{k}{2} = \frac{k!}{2!\cdot(k-2)!} = \frac{k\cdot(k-1)}{2}$ possible pairwise agreements between k judges

E.g. For one item (in our case a review) with 5 raters:

possible: $\frac{5(5-1)}{2} = 10$ observed: $\frac{3(3-1)}{2} + \frac{2(2-1)}{2} = 4$



ratio:$(\frac{3(3-1)}{2} + \frac{2(2-1)}{2})/(\frac{5(5-1)}{2}) = (\frac{3(3-1)+2(2-1)}{2}).(\frac{2}{5(5-1)}) = \frac{4}{10}$

A more informative metric incorporates chance agreement

- How much better is the agreement than what we would expect by chance?
- Need to calculate the proportion of a rater-rater pair agreement that we would expect by chance $\bar{P}_e$
- Our model of chance then is 2 independent judges choosing a class blindly—following the observed distribution of the classes
- The probability of them getting the same result is:

$$P(\text{both choose POSITIVE or both choose NEGATIVE})$$

$$= P(\text{POSITIVE})^2 + P(\text{NEGATIVE})^2$$

$\bar{P}_e$ is chance agreement

Chance agreement $\bar{P}_e$ affected by skewedness of distribution and number of categories

$p(C1)$	$p(C_2)$
0.5	0.5

$$\bar{P}_e = 0.5^2 + 0.5^2 = 0.5$$

$p(C1)$	$p(C_2)$
0.95	0.05

$$\bar{P}_e = 0.95^2 + 0.05^2 = 0.905$$

$p(C1)$	$p(C_2)$	$p(C3)$	$p(C_4)$
0.25	0.25	0.25	0.25

$$\bar{P}_e = 4 \cdot 0.25^2 = 0.25$$

Fleiss' Kappa measures reliability of agreement

- measures the reliability of agreement between a fixed number of raters when assigning categorical ratings
- calculates the degree of agreement over that which would be expected by chance

$$\kappa = \frac{\bar{P}_a - \bar{P}_e}{1 - \bar{P}_e}$$

- Observed agreement $\bar{P}_a$: average ratio of observed to possible pairwise agreements
- Chance agreement $\bar{P}_e$: sum of squares of probabilities of each category
- $\bar{P}_a - \bar{P}_e$ gives the agreement achieved above chance
- $1 - \bar{P}_e$ gives the agreement that is attainable above chance

Table of judgements

	Categories					
Item	1	...	j	...	n	
1	$n_{1,1}$	...	$n_{1,j}$	...	$n_{1,n}$	S_1
...	...	...	...	...	...	
i	$n_{i,1}$	...	$n_{i,j}$	...	$n_{i,n}$	S_i
...	...	...	...	...	...	
N	$n_{N,1}$	...	$n_{N,j}$	...	$n_{N,n}$	S_N
	C_1	...	C_j	...	C_n	

k: number of annotators; N: number of items; n: number of categories
$n_{i,j}$: the number of annotators which gave item i the judgement j
S_i: ratio of observed to possible pairwise agreements

Kappa, worked example (N=29, k=4, n=5)

Item	Category 1	2	3	4	5	S_i
1					4	$\frac{12}{12}$
2	2		2			$\frac{4}{12}$
3					4	$\frac{12}{12}$
4	2		2			$\frac{4}{12}$
5				1	3	$\frac{6}{12}$
6	1	1	2			$\frac{2}{12}$
7	3		1			$\frac{6}{12}$
8	3		1			$\frac{6}{12}$
9			2	2		$\frac{4}{12}$
10	3		1			$\frac{6}{12}$
11					4	$\frac{12}{12}$
12	4					$\frac{12}{12}$
13	4					$\frac{12}{12}$
14	4					$\frac{12}{12}$
15			3	1		$\frac{6}{12}$
16	1		2	1		$\frac{2}{12}$

Item	Category 1	2	3	4	5	S_i
17				2	2	$\frac{4}{12}$
18					4	$\frac{12}{12}$
19			3		1	$\frac{6}{12}$
20		1	3			$\frac{6}{12}$
21			1		3	$\frac{6}{12}$
22			3	1		$\frac{6}{12}$
23	4					$\frac{12}{12}$
24	4					$\frac{12}{12}$
25	2		2			$\frac{4}{12}$
26	1		3			$\frac{6}{12}$
27	2		2			$\frac{4}{12}$
28	2		2			$\frac{4}{12}$
29		1	2		1	$\frac{2}{12}$
AVG						.5804
C_j	42	3	37	8	26	
p_j	.362	.026	.319	.069	.224	

$$P(A) = .5804; \quad P(E) = .288; \quad K = \frac{.5804 - .288}{1 - .288} = .41 \ (N = 29, k = 4, n = 5)$$

κ values have no universally accepted interpretation

- if κ is 1 then raters are in complete agreement
- If κ is 0 then there is no agreement beyond what we would expect by chance
- κ can be negative ($\kappa \leq -1$)
- Beyond that there is no universally accepted interpretation
- Generally values of $\kappa = 0.8$ indicate very good agreement (Krippendorff, 1980)
- Note that size of κ is affected by the number of categories
- Note that κ may be misleading with a small sample size
- For use of κ in system evaluation see: `http://www.aclweb.org/anthology/W15-0625`

Today's Tasks: Tick 6

3-class classifier:

- Modify NB classifier so that you can run it on 3-way data (35,000 files).
- Calculate accuracy against the ground truth as before.

κ implementation:

- Download file with this year's class judgements on 4 reviews, and all MLRD years'
- Create an agreement table (i.e calculate for each item how many people said it was positive and how many said it was negative)
- Calculate $\bar{P}_a$, $\bar{P}_e$, κ
- Explore how κ changes (eg. across years, if you choose only a subset of the 4 reviews...)

Some extra reading...

- Siegel & Castellan (1988): Nonparametric Statistics for the Behavioral Sciences, McGraw-Hill; pages 284-289
- Krippendorff (1980): Content analysis. Sage Publications