6: Uncertainty and Human Agreement
Machine Learning and Real-world Data (MLRD)

Paula Buttery
(based on slides created by Simone Teufel)

Lent 2018

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 train and significance-test our classifier using stratified cross-validation

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 contained 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.

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.

Your classifications: NEGATIVE=35 POSITIVE=66

- Let the middle range of star-ratings to constitute a third class: NEUTRAL
- The ground truth for Review 1 is 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 assigned 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
- Movies with both positive and negative characteristics

Human agreement can be a 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.

Hypothesis:

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

	POSITIVE	NEGATIVE
Review 1	66	35
Review 2	8	93
Review 3	1	100
Review 4	96	5

For your second task today you will quantify how much you agree amongst yourselves

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 101 judges
- Instead we calculate $P(A)$, the observed agreement:

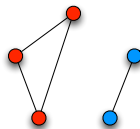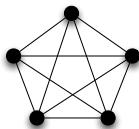$$P(A) = \text{MEAN}(\frac{\text{observed rater–rater pairs in agreement}}{\text{possible rater–rater pairs}})$$

$P(A)$ observed agreement

- Pairwise observed agreement $P(A)$: average ratio of observed to possible rater-rater agreements
- There are $\binom{n}{2} = \frac{n!}{2! \cdot (n-2)!} = \frac{n \cdot (n-1)}{2}$ possible pairwise agreements between n 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: $(\frac{3(3-1)}{2} + \frac{2(2-1)}{2})/(\frac{5(5-1)}{2}) = (\frac{3(3-1)+2(2-1)}{2}) \cdot (\frac{2}{5(5-1)}) = \frac{4}{10}$

- $P(A)$ is the mean of the proportion of prediction pairs which are in agreement for all items (i.e. sum up the ratios for all items and divide by the number of items)

A more informative metric incorporates chance agreement

- How much better is the agreement than what we would expect by chance?
- Need to calculate the probability of a rater-rater pair agreeing by chance $P(E)$
- 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:

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

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

$P(E)$ chance agreement

Chance agreement $P(E)$: sum of squares of probabilities (observed proportions) of each category

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

$$P(E) = 0.5^2 + 0.5^2 = 0.5$$

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

$$P(E) = 0.95^2 + 0.05^2 = 0.905$$

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

$$P(E) = 4 \cdot 0.25^2 = 0.25$$

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

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

- Observed agreement $P(A)$: average ratio of observed to possible pairwise agreements
- Chance agreement $P(E)$: sum of squares of probabilities of each category
- $P(A) - P(E)$ gives the agreement achieved above chance
- $1 - P(E)$ gives the agreement that is attainable above chance

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 will be negative if observed agreement is less than what would be expected by chance
- Beyond that there is no universally accepted interpretation
- Generally values of $\kappa = 0.8$ indicate very good agreement (e.g. Krippendorff)
- Note that size of κ is affected by the number of categories
- Note that κ may be misleading with a small sample size
- For information on how κ may be used (or not) 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 class' judgements on 4 reviews
- Create an agreement table
- Calculate $P(A)$, $P(E)$, κ
- Explore how κ changes with different combinations of reviews etc.

Some extra reading...

- Siegel & Castellan (1988): Nonparametric Statistics for the Behavioral Sciences, McGraw-Hill; pages 284-289
- Krippendorff (1980): Content analysis. Sage Publications
- Yannakoudakis & Cummins (2015): Evaluating the performance of Automated Text Scoring systems