6: Uncertainty and Human Agreement

Machine Learning and Real-world Data

Simone Teufel and Ann Copestake

Computer Laboratory
University of Cambridge

Lent 2017

Last session: Overtraining and cross-validation

- Your system is now quite sophisticated.
- We have a smoothed NB classifier which you can train and significance-test in a statified cross-validation setting.
- But the world we operate in is still artificially simplified.
- There are many reviews in the real world that are neither positive nor negative.

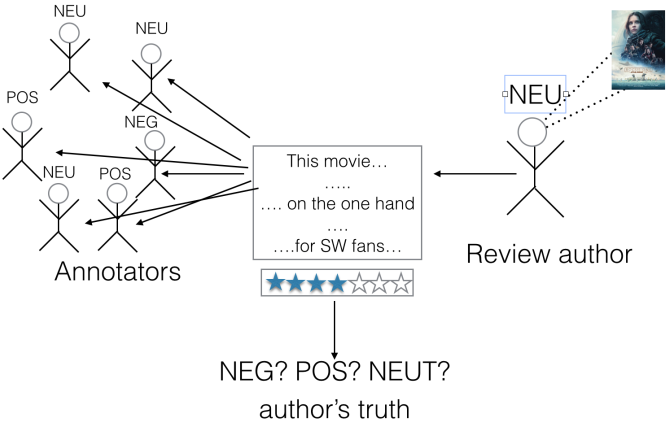
Neutral Reviews

- Let’s call both of the following “neutral reviews”:
  - Luke-warm reviews
  - Pro-con reviews
- We extend the task to include these reviews as well.
- Today you will first train and test your classifier on a trinary task – positive, negative, neutral reviews.

Uncertainty and truth

- So far, your data contained only the clearly positive or negative reviews.
- Only reviews with extreme star rating were used, a clear simplification of the real difficulty of the task.
- If we consider the middle range of star ratings, things get more uncertain:
  - Movies with both positive and negative characteristics
  - Inter-personal differences in interpretation of the rating scale
  - Reader's perception vs. writer's perception
- Who is to say what the true category of a review should be?
- Writer's perception is "lost forever", but we can get many readers to judge, at any point in time afterwards.

Human annotation



NEU

NEU

POS

NEG

NEU

POS

Annotators

This movie…

…..

…. on the one hand

….

….for SW fans…

NEU

Review author

NEG? POS? NEUT?

author’s truth

A source of truth

- 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.

Here’s how much you agreed amongst yourselves

- K = 0.70 (N=4, k=69, n=2) – overall (4 reviews)
- K = 0.89 (N=3, k=69, n=2) – Reviews 2,3,4

Here’s how much you agreed amongst yourselves

	Positive	Negative
Review 1	47	22
Review 2	2	67
Review 3	1	68
Review 4	67	2

Review 1 – NEUT

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.

It is a twisted tale of a mother’s love whose power exceeds that of anyone’s imagination. After a tragic accident leaving the unborn baby lifeless inside her, Madeline Matheson (Jordan Ladd) decides to carry out the term of her pregnancy with a mid-wife. Once born, however, her love has brought back the child to life - but with unexpected consequences. Once the grandmother gets curious to see the child, things start to get a little bit more complicated. The child seems to no longer need mothers milk, but rather blood to sustain its life. Very eerie and strange what a mother will go through to satisfy her child, and here is a film that will make you wonder to what extent you might go to save your child.

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.

Review 2 – NEUT

Upon reading the summary of this film I thought this would be a really awesome sci-fi aciton flick but I was highly mistaken.
This film is very bland and visually lacking (which I believe was done to emphasize the uncomfortable feeling of the distopian society but still). The "love story" in this film is very dry, more phyiscal than anything and it's hard to feel the chemistry.
There are lots of annoying Sci-fi noises in this film (which is common for sci-fi films that are from this time period).
This film does not need many special effects as there are no firefights/gunbattles, but I did like how the director's changes added some nice effects in certain places (although I gotta say the monkey people really threw me off).
I can see why this film was a classic but it was not up to my spoiled expectations in movies.
I would recommend this film for George Lucas fans, extreme film and Sci-Fi fans, and for fans of inherently unique feeling films.
Also the Zeotrope documentary on the bonus disc was very very interesting and neat to watch.

Review 3 – NEG

What are we going to do with Jim Carrey?
Viewers of television's "In Living Color" know this one-man cartoon from such characters as Fire Marshall Bill.
Viewers also know that "In Living Color" is a skit-show and that a little of Jim Carrey goes a long way.
Unfortunately, this fact was forgotten by the makers of the Carrey comedy ACE VENTURA: PET DETECTIVE.
Three writers, including Carrey, worked on the slapstick story, which sends a self-styled "Pet Detective" on the trail of a stolen dolphin.
The missing mammal belongs to the Miami Dolphins, who need their mascot for the upcoming Superbowl.
For plot porpoises, this story works as well as any Three Stooges short.
Carrey gets to do his "official" schtick as he snoops around greater Miami.
He leers and sneers, craning his neck to funny effect.
He even does his Captain Kirk impersonation.
Again.
All of this is pretty harmless stuff up until the point that you realize that the writers have absolutely no intention of focusing on anyone other than Carrey.
(Suggested alternate title–JIM CARREY: WILL DO ANYTHING FOR A LAUGH.)
Export it to France and you may have a hit.
As it stands, ACE VENTURA isn't even good kid's stuff.
The profanity count, alone, is too high.
Which is ironic, since children are, probably, Carrey's best audience.
The film doesn't even have the goofball charm of Chris Elliott's recent no-brainer CABIN BOY.
Sure, Carrey has his moments.
But what can you say about a film whose high-points include watching Carrey slink around to the theme from "Mission Impossible?"
ACE VENTURA has one glaring incongruity.
Amid the butt-jokes and double-takes, the script takes great pains to set-up an elaborate and rather funny "Crying Game" gag.

And, for this intended audience, that takes (ahem) cojones.

Review 4 – POS

At one point in the story townsperson Karen asks hero Tom what happened to her church.
He replies something like: “The church is flooded but at least the floodwaters put out the big fire. Well, the fire wasn’t that bad, since, while the church was burning, looters apparently thought it was safe enough to break through all the priceless stained glass windows. " In HARD RAIN a small town is nearly deserted due to flooding.
Everyone has had to evacuate because it’s raining, and now floodwaters are rising so high that buildings are being submerged and the nearby dam is about to break.
Enter a working class smart-alecky new armored car driver named Tom (Christian Slater).
Suddenly his security truck carrying over three million dollars gets stuck on the flooded street and is waylaid by armed looters.
Tom has no choice but to grab the bag full of money, hide it, and swim for his life.
This makes for an action movie full of jet skis, speedboat chases, and gun battles, as Tom tries to evade and outsmart corrupt cops and armed looters until the National Guard can answer his distress call.
He is befriended and aided by a spunky churchgoing young woman from town named Karen (Minnie Driver).
But unknown to Tom and Karen, the National Guard never heard Tom’s initial distress call.
Will Tom and Karen survive the natural and manmade disasters?
Opinion: Don’t expect thought-provoking issues or dramatics.
There’s not much more to this movie than hiding, running, swimming, shooting, and saving handcuffed heroes from drowning, but that’s what makes it escapist and fun.
Relax, take your shoes off, and break out the popcorn.

Agreement metrics

- We cannot use accuracy, because it cannot be used to measure agreement between 69 judges.
- 68 agreeing with each other, 1 judge disagreeing would count as a "wrong decision"
- We calculate P(A) instead, observed agreement:

$$\frac{\text{number of observed pairwise agreements}}{\text{number of possible pairwise agreements}}$$

- There are $\frac{69 \cdot (69-1)}{2}$ possible pairwise agreements between 69 judges
- In the above case, 68 of these pairwise agreements are not present; all others are.

Chance agreement $P(E)$

- We want an even more informative agreement metric
- We want to know only how much better the agreement is than what we would expect by chance
- Our model of chance is 2 judges blindly choosing, following the observed distribution
- The probability of them getting the same result is then:

$$\begin{aligned} P(E) &= P(\text{both draw } C_1 \text{ or both draw } C_2 \text{ or } \ldots \text{or both draw } C_n) \\ &= P(C_1)^2 + P(C_2)^2 + \ldots P(C_n)^2 \end{aligned}$$

Chance agreement is affected by:

- Number of classes:

C_1	C_2
0.5	0.5

C_1	C_2	C_3	C_4
0.25	0.25	0.25	0.25

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

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

- Distribution of classes:

C_1	C_2
0.5	0.5

C_1	C_2
0.95	0.05

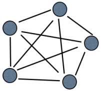
$$P(E) = 0.5^2 + 0.5^2 = 0.5$$

$$P(E) = 0.95^2 + .05^2 = 0.9050$$

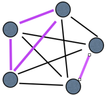
Fleiss' Kappa

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

- Pairwise observed agreement $P(A)$: Average ratio of observed to possible pair-wise agreements



possible: $\frac{5(5-1)}{2}$



observed: $\frac{3(3-1)}{2} + \frac{2(2-1)}{2}$

- Chance agreement $P(E)$: Sum of squares of probabilities of each category

Table of judgements

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

k: number of annotators; N: number of items; n: number of categories
$m_{i,j}$: the number of annotators which gave item i the judgement j

($m_{i,j} <= k$ in all cases)

Kappa values

- Kappa ranges between -1 and 1.
- If Kappa is 0, then there is no agreement beyond what we would expect by chance.
- Kappa can be negative if observed agreement is less than what would be expected by chance.
- According to a strict interpretation of values (by Krippendorff (1980)):
    - Kappa values of 0.8 indicate good agreement.
    - Kappa values of up to 0.69 indicate marginal (i.e., acceptable) agreement.

High chance agreement

	Positive	**Negative**
Review 2	2	67
Review 3	1	68

- Reviews 2 and 3: K = -0.012 (N=2, k=69, n=2)
- Why is kappa so low, despite almost all people agreeing?
- Because $P(E) > P(A)$
- In the world of R2 and R3, only negative reviews exist
- This is a small sample effect; not a problem with sufficiently high N.

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

			Class			
Item	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		$\frac{2}{12}$

			Class			
Item	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	$\frac{2}{12}$
C_j	42	3	37	8	26	
p_j	.362	.026	.319	.069	.224	

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

Task today

Task 6:

- Modify NB classifier so that you can run it on three-way data (35,000 files). Use 10-fold cross-validation as before. Observe results.
- Kappa Implementation
  - Download file with class's judgements on 4 reviews
  - First create the agreement table
  - Implement P(A); P(E); Kappa
  - Experiment with subsets of annotators; different combinations of items (reviews) etc.

Literature

- Siegel and Castellan (1988): Nonparametric Statistics for the Behavioral Sciences, McGraw-Hill; pages 284-289 (9.8; 9.8.1; not 9.8.2; 9.8.3 (except point 5); 9.8.4).
- Krippendorff (1980): Content analysis. Sage Publications, 1980