Inter-rater Reliability

Abram Hindle

2012-04-27 Tue

Outline

1. Introduction

2. Inter-rater reliability

Imagine

- We're tagging or rating items
  - commits
  - topics
  - bug reports
- We're using multiple people to tag or rate these items
- How do we ensure that we have agreement

Example

- “I made a change to the install documentation.”
  - Is this a portability relevant commit message?
- “I made a change to the install documentation regarding OSX.”
  - Is this a portability relevant commit message?
- “I made a change to the install documentation regarding OSX, Windows and Linux.”
  - Is this a portability relevant commit message?

We can disagree

- With ratings this is a problem
- Especially if we want to model or predict ratings
- How random are the ratings?
- Is there a bias?

Reliability

- So how reliability are ratings.
- Is there agreement?
- What does lack of agreement mean?

How

- We can measure correlation
  - Not really meant for it, but a good start
  - Pearson
  - Spearman
- Cohen's Kappa Statistic
  - Good for 2 raters
- Fleiss's Kappa
  - Good for more

R

- Perason cor(x,y,method="pearson")
- Spearman cor(x,y,method="spearman")
- library(irr)
- columns of a matrix
- kappa2(yourMatrix)
- kappm.fleiss(yourMatrix)

R

- Go and look at the code

What's a good Kappa or Correlation?

- 1 is good
- 0 or worse is bad
- 0.1 will get reviewers annoyed

Pay attention to

- num raters
- num subjects
- Kappa
- p-value

Threats to IRR

- Kappa of all same rank is NaN
- Lack of range in ratings 0,1 versus 1-5
- Class imbalance, if 10% is rated one way this can cause a problem