HW1- Annotation aggregation and exploratory analysis

CSCI 535 - University of Southern California

Due Date: Midnight 11:59PM, February 5, 2023

1 Goal

In this exercise, we become acquainted with some methods for gathering and analyzing annotations. We will also learn about inter-annotation agreement which permits assessing the reliability of our ratings, along with exploratory analysis of the data.

2 Fleiss' kappa (disclaimer: taken from Wikipedia with slight modifications)[1]

There are different ways of assessing inter raters agreement (e.g., Cohen's kappa and Fleiss' kappa). Cohen's Kappa is used to measure the inter-annotator agreement between two raters whereas Fleiss' kappa is a inter-annotator agreement for more than two raters.

Fleiss' kappa is calculated with the following formula:

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

Let N be the total number of items, let n be the number of ratings per item, and let k be the number of categories into which assignments are made. The items are indexed by $i = 1, ...N$ and the categories are indexed by $j = 1, ...k$. Let n_{ij} represent the number of raters who assigned the i-th item to the j-th category. First calculate p_j, the proportion of all assignments which were to the j-th category:

$$p_j = \frac{1}{Nn}\sum_{i=1}^{N} n_{ij}, \frac{1}{n}\sum_{j=1}^{k} n_{ij} = 1 \tag{2}$$

Now calculate P_i the extent to which raters agree for the i-th item (i.e., compute how many rater–rater pairs are in agreement, relative to the number of all possible rater–rater pairs):

$$P_i = \frac{1}{n(n-1)}\sum_{j=1}^{k} n_{ij}(n_{ij}-1) = \frac{1}{n(n-1)}\sum_{j=1}^{k}(n_{ij}^2 - n_{ij}) = \frac{1}{n(n-1)}[(\sum_{j=1}^{k} n_{ij}^2) - n] \tag{3}$$

Now compute $\bar{P}$, the mean of the P_i's, and $\bar{P}_e$ which go into the formula for κ:

$$\bar{P} = \frac{1}{N}\sum_{i=1}^{N} P_i = \frac{1}{Nn(n-1)}(\sum_{i=1}^{N}\sum_{j=1}^{k} n_{ij}^2 - Nn) \tag{4}$$

$$\bar{P}_e = \sum_{j=1}^{k} p_j^2 \tag{5}$$

[1] http://en.wikipedia.org/wiki/Fleiss%27_kappa

Table 1: Interpretation of Fleiss' kappa inter annotation agreement

κ	Interpretation
< 0	Poor agreement
0.01 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

Landis and Koch (1977) gave the Table 1 for interpreting values. This table is however by no means universally accepted and is subject to debate. The acceptable inter-rater reliability estimates will vary depending on the study methods and the research question.

3 Dataset

For this assignment, we will use a subset of the IPD dataset. The IPD dataset consists of nonverbal clips of people reacting to the result of a game they are playing with their partner. The subset you will be working with includes 100 videos that were given emotion labels by 100 crowd workers. Each crowd worker rated a random sample of 20 videos, resulting in each video having 20 ratings. You can read more about the dataset in Section 2 of the paper here[2], but it is not required for this assignment. The video IDs and ratings are saved in **tabulatedVotes.csv**. In this file, for each video, you can find the number of votes for each of the 7 categories of emotions (**Anger**, **Disgust**, **Fear**, **Joy**, **Neutral**, **Sadness** and **Surprise**).

4 Your tasks

You are expected to accomplish the following tasks:

1. Compute the Fleiss' kappa inter-annotator agreement on the set of videos in the provided **tabulatedVotes.csv**. Report and interpret the results using Table 1.

2. Go through a subset of 20 IPD videos in **IPD_20.csv**, see the video clips here[3]. Please pay attention to the facial expressions. Choose one facial cue which you believe to have a positive correlation with **Joy** and one facial cue with **Surprise**. For example "raised eyebrows" may have a positive correlation with the perception that the person is signaling **Surprise**. Don't limit yourself when selecting the facial cues.

3. Confirm your suspicions and subjective observations from the previous task with statistical analysis. You can do this by manually annotating the videos for each facial cue and comparing your annotations to the raters' votes. You need to annotate and calculate the p-value from the Student t-test between the group of videos with the facial cue of your choice and the group without. For simplicity, annotate the videos in **IPD_20.csv**. In **column I (cue_surprise)** and **column J (cue_joy)**, mark **Y** or **N** for the facial cue of your choice. For example, if you choose "raised eyebrows" as the cue for **Surprise**, and you see the person raising eyebrows in video 6 (first row), mark **Y** in **column I (cue_surprise)** for the first row. Once you finish the annotations, perform t-test to compare the mean number of votes from the raters for **Surprise** between the group you marked with **Y** in **column I (cue_surprise)** and the group with **N**. Repeat the process to annotate and perform t-test for **Joy**.

4. Report the details of your work for the previous tasks and describe your findings in a few paragraphs.

Note: You should implement your own function to compute the Fleiss' kappa inter annotation agreement. You can use any packages for tasks 2-4.

[2] https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10006366

[3] https://drive.google.com/drive/folders/1-at0OXhJTzUwV6T7j6V8ukLe4e-DcxVs?usp=sharing

5 Submission deadline

Please submit your report and source codes along with the csv file in a compressed folder named as **lastname_firstname** on Blackboard until **02.05.2023 at 23:59:59 Pacific time**. Late submission will incur a 10% penalty per day. If you have any questions, please email it to Su Lei (sulei@usc.edu) or use Piazza.