Task 6: Uncertainty and human agreement

Step 1: Nuanced classifier

Download the nuanced sentiment dataset which contains reviews assigned to three classes: positive, negative, neutral.

Open file 6220, which is neutral, and read it. Do you think this task has become harder or less hard by including the neutral category?

Modify your Naive Bayes classifier from Task 2 so that it can deal with a three-outcome classification. Instead of `Sentiment` use the `NuancedSentiment` provided. You can use `DataPreparation6.java` to load the nuanced dataset.

Test your system using 10-fold cross validation. What are the results?

Step 2: Human agreement

In Task 1 you were asked to manually classify four reviews as positive or negative. The file names of the reviews and the ground truth category were:

1. 6848 - neutral,
2. 9947 - neutral,
3. 937 - negative,
4. 1618 - positive.

The file `class_predictions.csv` contains the group's judgments as they were added to the database in Task 1. You can load the file using `DataPreparation6.java` provided.

Use the data to create an **agreement table** aggregating the group's judgments. For each review calculate how many people said it was positive and how many that it was negative. How do your own predictions compare with those of the entire group?

Step 3: Kappa

Human judgement can be used as a kind of truth. If no definitive decision can be taken, we cannot expect the system to agree 100% with every human but only as much as humans agree with each other. Write code to calculate **Fleiss' kappa** given by the following formula:

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

Here $\bar{P_e}$ is the mean of the squared proportions of assignments to each class:

$$\bar{P_e} = \sum_{j=1}^{k} \frac{1}{N} \sum_{i=1}^{N} (\frac{n_{ij}}{n_i})^2$$

k is the number of classes. N is the number of items (here: documents). n_i is the number of annotators that annotate item i (for this exercise, it's a constant). n_{ij} is the number of annotators who chose class j for item i.
Then $\bar{P_a}$ is the mean of the proportions of prediction pairs which are in agreement for each item:

$$\bar{P_a} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i(n_i - 1)} \sum_{j=1}^{k} n_{ij}(n_{ij} - 1)$$

You are strongly advised to make up some simple test examples to make sure your code is behaving as expected.

What κ score do you get for all four reviews taken together? What if you compare only reviews 1 and 2? What about 3 and 4 only?

Starred tick (optional)

We can investigate the behaviour of kappa by creating artificial agents which behave in different ways. For instance:

1. Random guesser: knows the overall distribution of the categories and chooses between them according to that proportion.
2. Happy random guesser: chooses positive 60% of the time, neutral 20%, negative 20%.
3. Doesn't sit on the fence: chooses positive 50% of the time, negative 50% of the time.
4. Middle of the road: chooses neutral 80% of the time.

Code some number of random agents and let them make choices for 50 examples and calculate kappa. Repeat this exercise 100 times. Can you explain the answers you get, based on your knowledge of the agents' strategies?