7 Machine Learning and Real-world Data (sht25)

In an annotation task with 4 classes (I, II, III and IV), three annotators (A, B, C) are making decisions, as in Figure 1.

	A	B	C
Item1	III	III	I
Item2	IV	I	III
Item3	II	II	I
Item4	I	IV	IV
Item5	II	IV	II
Item6	I	I	I
Item7	IV	IV	III
Item8	II	I	II

(*a*) Raw agreement amongst $k > 2$ annotators can be calculated based on pairwise agreement. Explain how this can be done, and calculate the value in the above case, showing your workings. [4 marks]

(*b*) We now want to use a chance-corrected agreement metric and choose Kappa.

(*i*) Explain why chance-corrected agreement metrics are useful. [2 marks]

(*ii*) How is chance agreement in Kappa calculated? Give the formula and calculate the value in the case above. [2 marks]

(*iii*) Give the formula for Kappa and calculate its value in our situation. [2 marks]

(*c*) New annotated data is discovered, which stems from two other annotators. Annotatator D only participated in annotation from item3 onwards, whereas Annotator E stopped annotating after item8 due to sickness. We want to use their partial annotation data, together with that from annotators A-C.

(*i*) One possible treatment is to pretend that annotators D and E were a single person, by randomly discarding one judgement for the doubly annotated items. Give at least two reasons why this is problematic. [4 marks]

(*ii*) Adapt the Kappa metric given above so that it can deal with partial annotation data. Give the motivation behind your idea as well as a formula for the final metric. [4 marks]

(*iii*) The annotation is now parcelled out into small sections (2 items each) and moved to a crowd-sourcing platform. Describe at least one potential problem with your agreement metric from (*c*)(*ii*) in this setting.

[2 marks]