9 Machine Learning and Real-world Data (pjb48)

This question concerns a sample of *English* language texts written by an author.

(*a*) When investigating a text we are concerned with its types and tokens.

(*i*) What are the types and tokens of a text? [1 mark]

(*ii*) Provide a sentence with exactly 4 types and 5 tokens. Explain any assumptions you make about the nature of tokens. [2 marks]

(*b*) Describe the expected frequency distribution of the English language types in the author's texts. Include any relevant formulas. [4 marks]

(*c*) We are interested in the written vocabulary size of the author. Could we estimate this from our sample texts? Include any relevant formulas that express the expected relationship between the size of a text and its vocabulary. [3 marks]

(*d*) Now we have another sample of English language texts written by a second author. We are interested to see if we can use a Naive Bayes classifier to automatically classify texts from the two authors.

(*i*) Define a Naive Bayes classifier for this task and describe how we use Maximum Likelihood Estimations to train the classifier. Provide equations. [4 marks]

(*ii*) Describe how the frequency distribution of types and the type/token ratio in the samples might affect the classifier. [2 marks]

(*iii*) A piece of writing has been discovered which both of our authors claim to be theirs. Could the classifier be used to settle this authorship dispute? Explain your answer. [4 marks]