8 Machine Learning and Real-world Data (sht25)

You are the program chair of a large academic conference. When papers are submitted to the conference, your job is to decide which paper should go into which topical area. For instance, a paper on computer architecture should go into the "Hardware" area. The areas are fixed before submission takes place. For each area, you have recruited an expert that will organise reviews for each paper in their area. These experts are called "area chairs". You decide to use statistical classification to route the incoming papers into the various areas. You have at your disposal several decades of papers, labelled with the area they were manually assigned to.

(*a*) Explain how you can set up a Naive Bayesian classifier for this task and derive the required parameter estimates. Give all necessary formulae. [3 marks]

(*b*) You now want to quantify how well your classifier is doing. Given that you can ask your area chairs for instant feedback, which two different evaluation methods can you realise in your setting, and how would you do this? Your answer should give details about data split and metrics. [2 marks]

(*c*) Up to now, we have assumed that areas are stable across years. You now find out that for your upcoming conference, for the first time in the history of your field, some areas have been changed. For each of the cases below, explain what would happen if you simply ran your classifier from (*a*) unchanged in the new situation, and propose the best course of action in the light of your existing classifier, giving your reasons.

(*i*) An area has become unpopular and is no longer treated in this year's conference. [2 marks]

(*ii*) An entirely new area has been proposed, treating material never before covered in your conference. [2 marks]

(*iii*) An existing area has split into two new areas. [2 marks]

(*iv*) Two existing areas have been merged into one. [2 marks]

(*d*) Rank the four situations listed in (*c*) with respect to how damaging they are to your classification strategy, giving your criteria for ranking. [2 marks]

(*e*) You want to know which areas are most similar to each other, and you want to use data to answer this question. In addition to the textual data described above, you also have access to your papers' citation network. This means you know which papers cite which existing papers, although you may not have full text available for cited papers, only identifiers. How would you determine the similarity between areas, and how would you visualise the results? [5 marks]