7 Machine Learning and Real-world Data (AAC)

Some areas of land currently covered by forest had a different previous purpose. An experiment is to be conducted to see whether areas of forest can be automatically classified according to their purpose 50 years ago. There are three categories: meadow, garden and managed woodland. The classification is to be based on the trees currently present: for instance, an area with several apple trees is relatively likely to have been a garden. There are some locations where the true category is known from historical data and the number of trees of each type observed within a fixed distance has been recorded. There is data from 25 meadows, 30 gardens and 45 woodlands. The average number of individual trees at each location is 52.

(*a*) Give formulae for two possible approaches to Naive Bayes classification for this task. [4 marks]

(*b*) How could you derive parameter estimates for use in the Naive Bayes classifiers from this type of data? [4 marks]

(*c*) How would you use the available data to train and test a Naive Bayes classifier? [5 marks]

(*d*) You are now given a large catalogue of tree species with each species manually assigned to zero or more of the categories. Describe a modification to your previous experiment which makes use of this data. [4 marks]

(*e*) You are now given data from 60 new locations. If you use this data solely for evaluation, how would you try and decide whether or not your revised method was an improvement on the original method? [3 marks]