COMS21202 Symbols, Patterns and Signals

Problem sheet: More Classification and Clustering

Q1. Suppose we have a training set consisting of 200 non-spam and 1000 spam emails. The following table indicates the numbers of emails in each class containing a particular word.

word	# non-spam containing *word*	# spam containing *word*
w_1	100	500
w_2	80	100
w_3	40	800

a) We want to use Bayesian classification to predict whether an email is spam. Estimate the likelihood ratios $P(word|spam)/P(word|\neg spam)$ and $P(\neg word|spam)/P(\neg word|\neg spam)$ from the above data.

b) How would these answers change if you applied the Laplace correction?

c) Calculating your answer using these likelihood ratios, how would an email containing all three words be classified by a maximum likelihood (ML) classifier? Would the outcome be different for a maximum a posteriori (MAP) classifier that uses the class distribution observed in the training set?

Answer the same questions for an email containing none of the three words.

d) We want to build a decision tree classifying emails as spam and non-spam, using the presence/absence of these words as boolean features. Using the numbers in the table, which feature results in the best split? Give a numerical explanation of your answer.

Q2. You are given the set of numbers $\{8, 44, 50, 58, 84\}$.

a) Give two possible clusterings you could get if you apply K-means to this data set with $K = 2$. Which one is optimal?

b) Give dendrograms using single linkage and complete linkage, and explain the differences (if any).

Q3. Imagine you are dealing with a three-class classification problem with classes A, B and C.

a) You are given a sample with 30 examples of class A, 50 examples of class B and 20 examples of class C. What is your estimate of the class priors? How would you justify this estimate?

b) Suppose you are also told that this sample is somewhat atypical and that normally classes A and B are of equal size. Using all of this knowledge, derive the class priors by maximum-likelihood estimation.

c) It is now a couple of days since you last saw the sample, and while you remember that A and B are normally of equal size, you can only remember the total size of the sample (100) and the size of class A (30). Describe how you would estimate the class distribution in this case, and give one possible answer.