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.

Answer:

word	$\frac{P(word\|spam)}{P(word\|\neg spam)}$	$\frac{P(\neg word\|spam)}{P(\neg word\|\neg spam)}$
w_1	$\frac{500/1000}{100/200} = 1$	$\frac{(1000-500)/1000}{(200-100)/200} = 1$
w_2	$\frac{100/1000}{80/200} = 1/4$	$\frac{(1000-100)/1000}{(200-80)/200} = 3/2$
w_3	$\frac{800/1000}{40/200} = 4$	$\frac{(1000-800)/1000}{(200-40)/200} = 1/4$

This means, for example, that the presence of word w_3 is four times more likely in a spam email than in a non-spam email; while its absence is four times more likely in a non-spam email than in a spam email.

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

Answer:

In this case the Laplace correction would have little effect: e.g., $\frac{P(w_2|spam)}{P(w_2|\neg spam)} = \frac{(100+1)/(1000+2)}{(80+1)/(200+2)} = 0.2514$. *The Laplace correction would make more of a difference if some counts in the table were close to zero.*

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.

Answer:

For an email containing all three words, the product of the likelihoods is $1 * 1/4 * 4 = 1$, *so for an ML classifier the email is right on the decision boundary and could be classified as either spam or non-spam. Since spam is 5 times more likely than non-spam, a MAP classifier will classify the email as spam.*

For an email containing none of the words, the product of the likelihoods is $1 * 3/2 * 1/4 = 3/8$, *so an ML classifier will classify the email as non-spam and a MAP classifier will classify it as spam.*

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.

Answer:
Full training set: 200 non-spam, 1000 spam, class ratio $1:5$. w_1 *present: 100 non-spam, 500 spam, class ratio* $1:5$. w_1 *absent: 100 non-spam, 500 spam, class ratio* $1:5$. *Both subsets have the same class ratio as the training set, so* w_2 *has zero information gain.*
w_2 *present: 80 non-spam, 100 spam, class ratio* $4:5$. w_2 *absent: 120 non-spam, 900 spam, class ratio* $2:15$.
w_3 *present: 40 non-spam, 800 spam, class ratio* $1:20$. w_3 *absent: 160 non-spam, 200 spam, class ratio* $4:5$.
The two $4:5$ *ratios cancel, but* $1:20$ *is better than* $2:15$. *So the best feature is presence/absence of* w_3. *This can be verified by calculating information gain.*

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

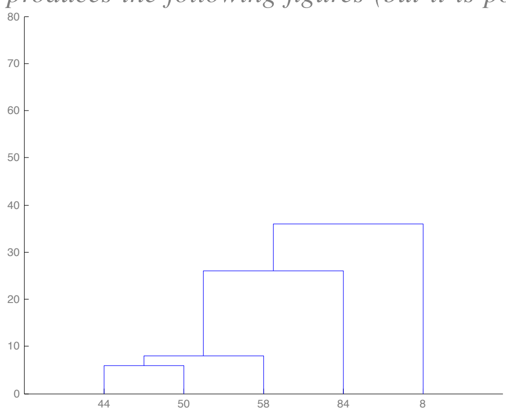
a) Give two possible clusterings you could get if you apply K-means to this data set with $K = 2$. Which one is optimal?

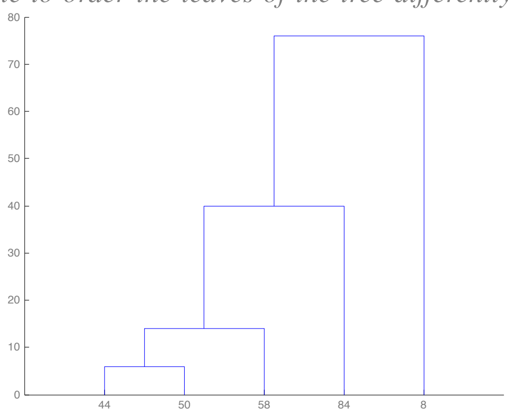
Answer:
The data has been carefully constructed such that all four possible split points yield stationary points for K-means. If you visualise the data you will see that there is a central cluster of three points with outliers on either side. 8 is further away from the central cluster of three than 84, which suggests that $\{8\}, \{44, 50, 58, 84\}$ *is the optimal clustering. This can be verified by calculating the sum of squared distances to the centroid for each cluster:*

Clusters	*Centroids*	*Total scatter*
$\{8\}, \{44, 50, 58, 84\}$	$8, 59$	$0 + 952 = 952$
$\{8, 44\}, \{50, 58, 84\}$	$26, 64$	$648 + 632 = 1280$
$\{8, 44, 50\}, \{58, 84\}$	$34, 71$	$1032 + 338 = 1370$
$\{8, 44, 50, 58\}, \{84\}$	$40, 84$	$1464 + 0 = 1464$

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

Answer:
Matlab produces the following figures (but it is possible to order the leaves of the tree differently):





Complete linkage produces larger linkages because it takes the distance between points furthest away where single linkage takes closest points: e.g., single linkage between $\{8\}$ *and* $\{44,50,58,84\}$ *is* $44-8=36$ *whereas complete linkage is* $84-8=76$.

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?

Answer:
The estimate would simply be the relative frequencies $(30/100, 50/100, 20/100) = (0.3, 0.5, 0.2)$ *(you could apply the Laplace correction but this would not make a big difference). This could be justified as the maximum-likelihood estimate of the parameters of a multinomial distribution.*

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.

Answer:
The intuitive answer is that the relative size of both A and B is estimated as the mean of $30/100$ *and* $50/100$*, i.e.,* $40/100 = 0.4$*. This can be derived as follows. The probability of a As, b Bs and c Cs is* $K\alpha^a\alpha^b(1-2\alpha)^c$ *with K some combinatorial constant and* α *the parameter to be estimated. Taking the logarithm, then the derivative wrt.* α*, setting the derivative to 0 to find the maximum and solving for* α*, we obtain* $\hat{\alpha} = \frac{a+b}{2(a+b+c)}$*. With the given numbers we obtain* $\hat{\alpha} = 80/200 = 2/5$*, and so the estimated class priors are* $(2/5, 2/5, 1/5) = (0.4, 0.4, 0.2)$.

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.

Answer:
The Expectation-Maximisation algorithm would be able to deal with this kind of missing information. The Expectation step would calculate the expected values of b and c from their sum $h = 100-30 = 70$ *and an assumed value for* α*: more precisely, the expectation of b is* $\frac{\alpha}{1-\alpha}h$ *and that of c is* $\frac{1-2\alpha}{1-\alpha}h$*. The Maximisation step would re-estimate* α *from these expectations in the same way as in the previous answer.*

EM will converge to $(0.3, 0.3, 0.4)$.