STAT 315 Optional Assignment 5

June 7, 2014

Lab: None. Due Date: 6 June, 4pm. Submit via Learn. Maximum possible marks: 9.
Note: this assignment is **optional**. Your assignment grade is made up of the best 4 of 5 assignments, so you do not need to complete this one if you have done the first four.
This assignment does not require SAS but you will need to look at the file `dolphin_plots.pdf`.

1. The religious makeup of the UK and Eire are given by the following probability distributions

	No religion	Catholic	Other Christian	Moslem	Others
UK	0.507	0.086	0.343	0.024	0.04
Eire	0.076	0.842	0.045	0.01	0.027

   A popular way to compare probability distributions is the Kullback-Leibler distance. If P and Q are probability distributions then the Kullback-Leibler distance of Q from P is

$$-\sum_x P(X = x) \log(Q(X = x)/P(X = x)).$$

   (a) Find the Kullback-Leibler distance of the UK distribution from the Eire distribution. (1)
   ```
   > UK <- c(0.507, 0.086, 0.343, 0.024, 0.04)
   > Eire <- c(0.076, 0.842, 0.045, 0.01, 0.027)
   > sum(Eire * log(UK/Eire))
   [1] -1.66597
   ```
   The distance is 1.67. Note: a minus sign was missing in an earlier version of the assignment, so negative 1.67 is also acceptable.

   (b) Find the Kullback-Leibler distance of the Eire distribution from the UK distribution. (1)
   ```
   > sum(UK * log(Eire/UK))
   [1] -1.499359
   ```
   The distance is 1.5.

   Comment: some people used logs to base 10 instead. In this case the results are 0.72 and 0.65. There is no penalty for this. It just gives you an alternative scaling for the KL distance.

   (c) Give two reasons why the Kullback-Leibler distance might be a problematic way to measure distance in an application such as clustering. (2)

   The distance from x to y is not the same as the distance from y to x. This seems wrong.

   Also, probability distributions can contain the value 0, but then the KL distance won't even be defined, because you can't take the log of zero.

   If you got a negative sign in either of the previous parts, I would also accept the answer that distance shouldn't be negative.

2. A researcher observes 94 dolphins which are seen in 10 different social groups at different times. The data are given in the file `dolphins.txt`. A 1 in row i and column j represents dolphin i being seen in group j at some time.

   A SAS dendrogram for hierarchical clustering of the data with the distance between two dolphins being the Euclidean distance and the distance between two clusters being single linkage is given in the file `dolphin_plots.pdf`.

The file also contains images of the distance matrix for the dolphins before and after clustering, with redder values denoting smaller distances. (The *distance matrix* is the 94×94 matrix whose (i, j)–entry is the distance between dolphin i and dolphin j.)

(a) Show that the Euclidean distance between dolphin i and dolphin j is the square root of the number of groups in which one appears but not the other. (2)

If dolphin i is $(x_1, \ldots, x_{10})$ and dolphin j is $(y_1, \ldots, y_{10})$ then the Euclidean distance is

$$\sqrt{\sum_{i=1}^{10}(x_i - y_i)^2}.$$

But $(x_i - y_i)^2$ is 1 exactly when one of x_i, y_i is 1 and the other is 0. In other words, exactly when one dolphin appears in the group but not the other. This proves the claim.

(b) Which dolphin or dolphins are closest to dolphin number 1? (1)

From the dendrogram, number 24.

(c) The researcher asks you to interpret these plots in terms of the social relationships between the dolphins. What is your conclusion? Explain in a down-to-earth way that a researcher who knows no statistics can understand. (2)

The dolphins seem to be broadly divided into two "clans" and groups tend to involve dolphins from one clan or the other, but not both. You can see this in the two red squares in the distance matrix. There are also possible sub-clans of dolphins which tend to form groups with each other.

Comment: these data were generated in the following way. There are actually two clans with a 90% probability of two dolphins in the same clan appearing in the same group and a 10% probability of dolphins in different clans appearing in the same group. Noise was added by randomly flipping every element of the matrix to its opposite value with probability 0.3. The important point is that any structure that you found from clustering apart from the two big groups is merely an artifact of randomness and not "real"! This shows how cautious you have to be with cluster analysis and is one reason why many people regard it with suspicion.