Statistics 6861/4866 Assignment 5

200 points

Due: December 1

The problem you solve for this assignment is the detection and correction of misclassified data. The data consist of observations that have more than one attribute measured. For this problem we restrict the attributes to have continuous measure. Each observation has been classified as belonging to class 1 or class 2. The class 1 data form a cluster and the class 2 data form a separate cluster. These clusters are well-separated.

It is possible that through measurement error, or recording error, that one or more observations have been misclassified. You can assume the misclassification rate is low.

Write a function to detect and correct misclassifications. The function prototype is:

detect.misclass <- function(class.v,p)

Here class.v is a vector containing the classification for the observations. The measured attributes for each observation are stored in p, which is a matrix. We require length(class.v) to equal nrow(p). The number of attributes measured for each observation is equal to ncol(p). If p is a mxn matrix, we have m observations and n attributes. For this problem we assume all elements of class.v are either 1 or 2, and are numeric, not character.

The output of the function is a list. The first item in the list is a logical value. The name of this logical value is “err.found”. It is set to TRUE if a misclassification was detected. Otherwise it is set to FALSE. The second item in the list is a vector containing the indices of the misclassified observations. The name of this item is “err.loc”. The last item in the list is named “new.class”. It is a vector that contains the corrected classification for an observation. So, if the observation stored in row 17 of p should have classified as a 2, and the first 16 observations were correctly classified, then err.loc[1] = 17 and new.class[1] = 2. The err.loc vector is sorted in ascending order.

When the “err.found” item is FALSE, the “err.loc” item and the “new.class” item are both set to NULL.

Suppose you call the function with these arguments.

a <- detect.misclass(old.class,p)

If you do NOT detect any misclassification. When you print a the result is:

> print(a)

$err.found

[1] FALSE

$err.loc

NULL

$new.class

NULL

Now suppose you detect that observation 5 should have been classified as a 1, and observation 569 should have been classified as a 2. Then we print a, we should get

> print(a)

$err.found

[1] TRUE

$err.loc

[1] 5 569

$new.class

[1] 1 2

Note that with this information you can correct the old.class vector. You can then form a data frame and write the corrected data to a file. If the data have only two attributes, you can plot the data before and after your function is run. You can use different colors for each classification to visually verify the correctness of your algorithm. However, this is not fool-proof since we could have over-plotting.

When you write a function such as this one, you should bear in mind that someday, someone will want to use it for data sets that have more than two classifications. That is why I have the return value specify the new classification. If you only ran this for data sets with two classifications you could use the err.loc vector to simply change the classification.

As usual, send me an email containing your function, preceded by the executable statement

name = "Your name" # note that it is name, and not Name. Note that we do not have a cat() surrounding it. Use the usual convention for naming your file.

If you have test runs in your file, comment them out.