Data Analysis and Exploration
A.A. 2010/11 Third (and last) block of exam exercises.

Every student has to solve (through a program in R that answers the questions of the exercise and include comments if useful) the practical exercise using the dataset individually assigned to each student. The R program has to be sent, through e-mail to Andrea Pugliese <pugliese@science.unitn.it>. Theoretical explanations and comments to the program, if useful, can be sent via e-mail (in some format) or handed in paper. Everything has to arrive at least 3 days before the day of the oral exam.

1. A file "remote_n.dat" where n is the number assigned to each student, (fake, but similar to real data from remote sensing) contains in the first column an index on crop type (1 = corn, 2 = soybeans, 3 = cotton, 4 = sugarbeets, 5 = clover); the remaining 4 columns contain some measurements obtained by using remote sensing on fields.

   (a) Use the discriminant analysis on these data to determine (by using equal a priori probabilities) the crop type from other measurements. Create a plot of data on canonical discriminant. variables. Find the error rate on the sample , and with the method of cross-validation.

   (b) For fields of type 1, compute the specificity (number of fields of type 1 correctly classified over the total number of fields classified as 1) and the sensitivity (number of fields of type 1 correctly classified over the total number of fields of type 1) of the classificator, both in sample and with the cross-validation.

   (c) How would a field with measures 11, 18, 26, 16 would be classified with the above method? How confident would you feel about this classification?

   (d) Repeat the analysis only on the crops of type 1 and 2 (i.e. only the first 14 rows). How does the method work? Could you use different techniques for this case?