Mixture Discriminant Analysis and Tree

Ruoyuan Li (001223313)

Data Description & Question Addressed

The data set used is called "biopsy" which can be found at **MASS** package in R (same data as I used for last assignment). It is a breast cancer data, the data assessed biopsies of breast tumors. Data "biopsy" contains 683 observations with 10 variables in total without NA and patients' ID. The 9/10 variables are predictor variables, they were scored on a scale from 1 to 10, namely: V1 is clump thinness, V2 is uniformity of cell size, V3 is uniformity of cell shape, V4 is marginal adhesion, V5 single epithelial cell size, V6 is single epithelial cell size, V7 is bland chromatin, V8 is normal nucleoli, and V9 is mitoses. The response variable is called "class" which is binary result class contain "begin" with 444 patients and "malignant" with 239 patients.

In this report, I do mixture discriminant analysis (MDA) classification and classification tree. For both method do ten different random 75/25 labelled/unlabeled splits for comparing the performance more general. Finally, a list of box plots of different numerical index, namely, AIR, RI, and diagonal rate (1-error rate) from ten random splits for both MDA and tree, and a comparison table with results of 3 numerical index values from MDA and tree which contains mean and standard deviation.

MDA & Tree Analysis

For mixture discriminant analysis, first scale predictor variables and randomly choose 25% observations which is 171 observations to be unlabelled in my case, i.e 75% of predictor variables (V1 to V9) from "biopsy" are used as training data, the 75% of response variable (class) is the giving class label for the observations in the training data ,the rest 25% observations are used as test data. By using **MclustDA** with data defines as 75% of predictor variables, and class is 75% of response variable. After getting the MDA model, then for getting test classification summary by specifying newdata to be the rest 25% predictor variables, and specifying new class to be the rest 25% unlabbled class. (Due to pages limit, I have no space to show the actual classified table). Next repeat randomly

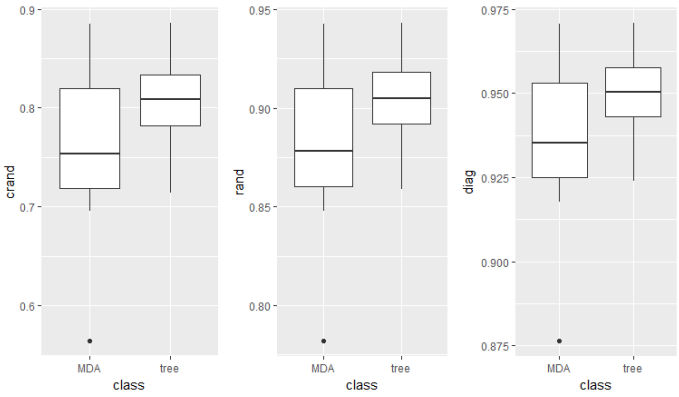

Figure 1: List of boxplot for ARI, RI, and diag for MDA and Tree

select 25% as unlablled for other 9 times, save all diagonal rate, RI, ARI for those 10 times random splits classification and the comparison comes later.

For the tree classification, also randomly choose 75% as training data, and the rest 25% as test data, using **tree** function in package **tree** to create a tree object based on training data, and predicting test data by **predict** function and specifying newdata to be testing data, then make a table of predict class with actual class. (again, no space for the classified table, so archived them). Next, do the same process for other 9 times, randomly select 25% to be unlablled and save all the diagonal rate, RI and ARI.

For comparison of MAD and tree classification method, there is a list of box plots from all ARI(crand), RI(rand), and diagonal rate (diag) from 10 times random splits for each method are shown at Figure 1. Tree obvious performs better than MDA base on those box plot. Due to limit pages, I can not show all the diagonal rate, ARI for each random split for MDA and tree, so I summarized them by mean and standard deviation as shown at Table 1. We can notice tree classification is slightly better than mixture discriminant analysis for data "biopsy" based on mean of ARI, rand index, and diagonal rate (which is 1- error rate). Moreover, tree performs better than MDA can be proved both from box plots and summarized table. In conclusion, tree is better for classification in my case for "biopsy" data.

Table 1: mean and SD of index for MDA and Tree classification

	ARI_mean	ARI_sd	RI_mean	RI_sd	Diag_mean	Diag_sd
MDA	0.76	0.093	0.88	0.045	0.936	0.027
tree	0.803	0.056	0.90	0.028	0.949	0.016