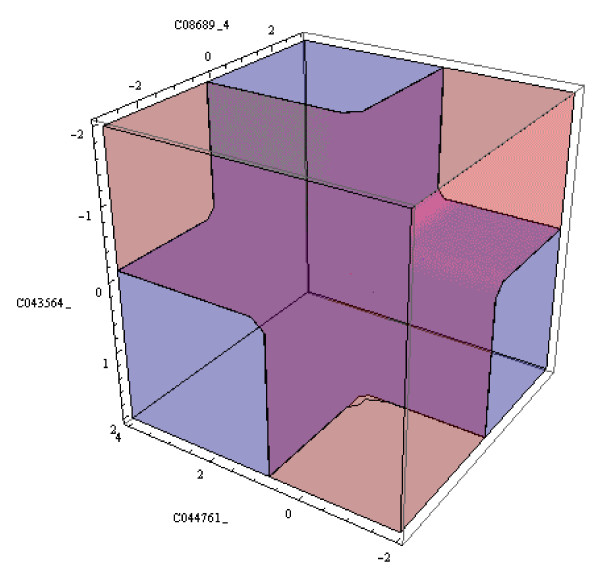
**Logical Analysis of Data**

There has recently been some interest in Operations Research in ‘the logical analysis of data’ (which we shall often abbreviate to LAD). The main emphasis so far has been on finding convenient and informative explanations of data by means of Boolean functions. The accuracy of these explanations on classifying unseen data has been estimated experimentally, and the results are impressive; certainly, the performance of the techniques is comparable with those of popular machine learning techniques. In this paper, we apply techniques from the probabilistic analysis of machine learning for example to analyse theoretically the accuracy of the LAD techniques.

The LAD model can be used for diagnosing "new" observations, i.e. observations not included in the original dataset. If the observation satisfies hof the ppositive patterns in the model, and kof the nnegative patterns, then we define the "LAD discriminate function" as the expression h/p - k/n. The observation is classified based on the sign of the discriminate function. If h/p - k/nis equal to 0, the observation remains "unclassified". The LAD discriminate function corresponding to the training data and the LAD classification model is plotted. The region colored pink corresponds to the region where the discriminate function is evaluated to have positive sign, and hence an observation in this region will be classified as positive, while the blue region corresponds to the negative sign and hence it will classify observations as negative. Observations which lie on the surface of the discriminant function will be left unclassified.

****

According to the above definition, the sensitivity of a model is defined as the proportion of correctly classified stroke patients, while the specificity of a model is defined as the proportion of correctly classified controls. The accuracy of a model is defined as the average of its so-called "corrected sensitivity" and "corrected specificity", where the corrected sensitivity of a model is the usual sensitivity plus half of the proportion of unclassified stroke patients and the corrected specificity of a model is the usual specificity plus half of the proportion of unclassified controls. Let P, Nbe the number of positive, respectively negative observations. Let p, ndenotes the number of correctly classified positive, respectively negative observations. Let up, unbe the number of unclassified positive and negative observations respectively. The formula for sensitivity, specificity, corrected sensitivity, corrected specificity, and accuracy are presented below:

Sensitivity = p/P

Specificity = n/N

Corrected sensitivity = p/P + 1/2(up/P)

Corrected specificity = n/N + 1/2(un/N)

Accuracy = 1/2(Corrected sensitivity + Corrected specificity)

The accuracy, sensitivity, specificity and hazard ratio of the diagnostic LAD system applied to the stroke data are shown in the. We present the results of testing the model on the training set, then on the validation set, and finally the average results (with 95% confidence interval) over ten random 10-folding cross-validation experiments on the training set, where in each fold a model is built on a part of the training set and tested on the remainder.