CSE 562 Final Exam

Olcay Taner YILDIZ

I. Question

For a numeric input, instead of a binary split, one can use a L-ary split with L - 1 thresholds and L branches as

$$\begin{aligned} x_j &< w_{m1} \\ w_{m1} \leq x_j &< w_{m2} \\ w_{m2} \leq x_j &< w_{m3} \\ &\dots \\ w_{m(L-2)} \leq x_j &< w_{m(L-1)} \\ x_j &\geq w_{m(L-1)} \end{aligned}$$

Propose a modification of the tree induction method to learn the L - 1 thresholds, $w_{m1}, w_{m2}, \dots, w_{m(L-1)}$ What are the advantages and the disadvantages of such a node over a binary node?

II. Question

In K-fold cross-validation, the dataset X is divided randomly into K equalsized parts, $X_i, i = 1, 2, \dots, K$. To generate each train and test set pair, we keep one of the K parts out as the test set, and combine the remaining $K - 1$ parts to form the training set. On the other hand, classes must be represented in the right proportions not to disturb the class prior probabilities (**stratification**). Propose an algorithm to do K-fold crossvalidation with stratification.

III. Question

Let us say we have L classification algorithms. How can we order these L from best to worst? (Hint: Use tests comparing two classification algorithms)

IV. Question

Given the following error rates for a four feature problem, what will be the result of the following attribute subset selection procedures?

- Stepwise forward selection
- Stepwise backward elimination

Error Rates: (F_1), 0.23; (F_2), 0.25; (F_3), 0.36; (F_4), 0.24; (F_1, F_2), 0.18; (F_1, F_3), 0.19; (F_1, F_4), 0.20; (F_2, F_3), 0.16; (F_2, F_4), 0.29; (F_3, F_4), 0.22; (F_1, F_2, F_3), 0.39; (F_1, F_2, F_4), 0.45; (F_1, F_3, F_4), 0.24; (F_2, F_3, F_4), 0.36; (F_1, F_2, F_3, F_4), 0.59.

V. Question

Solve the k-means clustering algorithm analytically in the case of k = 1.

VI. Question

A linearly separable 2-dimensional dataset with 2 classes is given. Propose a condensed 1 nearest neighbor algorithm for this dataset.