CSE 566 Final Exam

Olcay Taner Yıldız

I. Clustering

Suppose that the data mining task is to cluster the following eight points (with (x, y) representing location) into three clusters. A_1(2, 10), A_2(2, 5), A_3(8,4), A_4(5, 8), A_5(7,5), A_6(6,4), A_7(1,2), A_8(4,9). The distance function is Euclidian distance. Suppose initially we assign A_1, A_4, and A_7 as the center of each cluster, respectively. Run the k-means algorithm one epoch on this data.

II. Nonparametric Methods

- Generalize kernel smoother to multivariate data.
- All kernel functions have the requirement that $\int K(\mathbf{x})d\mathbf{x} = 1$. Prove this requirement for the weight function defined in page 156.

III. Decision Trees

Propose a tree induction algorithm with lookahead.

IV. Linear Discrimination

Run the logistic discrimination algorithm given in Figure 10.6 one epoch for the following data:

Feature 1	Feature 2	Class
1	1	1
1	0	1
0	1	1
0	0	0

V. Multilayer Perceptrons

Derive the update equations for K-class discrimination when the hidden units use tanh, instead of the sigmoid. Use the fact that $tanh' = (1 - tanh)^2$

VI. Hidden Markov Models

Let us say we have 4 states:

S_1: A, S_2: C, S_3: G, S_4: T

with initial probabilities: $\mathbf{\Pi}$ = [0.3, 0.4, 0.1, 0.2]

and the transition matrix is:

$$\mathbf{A} = \begin{vmatrix} 0.2 & 0.2 & 0.3 & 0.3 \\ 0.1 & 0.3 & 0.2 & 0.4 \\ 0.4 & 0.4 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.7 \end{vmatrix}$$

- Calculate the probability of the sequence ACCGTC.
- Given twelve sequences AAGCT, TATTA, GTAGT, TTATT, GAGCT, AAAAT, ACCTT, CTTAA, TGATT, GGTAA, TACAG, CGAAC estimate the initial probabilities for $\mathbf{\Pi}$ and $\mathbf{A}$.

VII. General

In K-fold cross-validation, the dataset X is divided randomly into K equalsized parts, $X_i, i = 1, 2, \ldots, K$. 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.