Mathematics of AI

Discretization

Given a set of N real observations $v_1, v_2, \dots, v_N$, it is desired to discretize them by partitioning into K disjoint parts $P_1, P_2, \dots, P_K$. Each part includes the values in a specific range:

$$P_1 = \{v|\ v < U_1\}, P_2 = \{v|\ U_1 \le v < U_2\}, \dots, P_k = \{v|\ U_{k-1} \le v < U_k\}, \dots, P_K = \{v|\ U_{K-1} \le v\}$$

In this project, partitioning is done differently. It is desired to determine k representative values $r_1, r_2, \dots, r_k$ each one of which represents values of corresponding part (i.e., r_k represents the values in P_k). The goal is partitioning the observations and also determining the representors in order to minimize the following objective function.

$$\|dev\|_p = \sqrt[p]{\sum_{k=1}^{K} dev_k^p}$$

where, dev_k is the deviation of values in P_k from corresponding representative value r_k as follows:

$$dev_k = \|d_k\|_q = \sqrt[q]{\sum_{v \in P_k} d_k(v)^q}$$

In this definition, $d_k(v) = |v - r_k|$ as the distance of each $v \in P_k$ to its representative value r_k. Totally, the norm of distances to the representor in each part defines the deviation of that part. Finally, norm of deviations should be minimized. These two norms may be from different degrees (i.e., $p \neq q$). This minimization problem can be solved with any heuristic optimizer (such as alternating optimization). Then, the solution should be compared with the optimum solution computed by Dynamic Programming.

The experimental results include comparison of these methods with different values of p and q in a table (values of $p = 0,1,2,\infty$ in the rows and values of $q = 0,1,2,\infty$ in the columns). Each cell includes two numbers as the objective values of two methods. This table should be filled on four sets of observations (for example four features of Iris dataset).

Good Luck!

Taheri