Mathematical programming approach in credit scoring

Mohamed Ouzineb
SI2M, National Institute of Statistics and
Applied Economics Rabat,Morocco
mohamed.ouzineb@cirrelt.net

A. *Problem description and assumptions*

The majority of lending organizations must assess the risk level of granting a credit to a new applicant. To do so, they use credit scoring models that are developed from training sets consisting of people in their records who were given loans in the past.

Given a sample of n previous borrowers or applicants, n_G of them are good and n_B are bad. Each applicant is characterized by p variables $X = (X_1, X_2, ..., X_p)$ (age, sex, salary, ...). Denote by A the set of all possible combinations of values of the variables X. The first work to do by the lending organization is to divide A into two sets : A_G represents the answers given by the good clients and A_B represents the bad ones. The response of each applicant i denoted by $(x_{i1}, x_{i2}, ..., x_{ip})$. The second work is to choose, for a given cutoff value c, weights or scores $(w_1, w_2, ..., w_p)$ so that $\sum_{j=1}^{p} w_j x_{ij} \geq c$ if the client i in the sample is a good one and $\sum_{j=1}^{p} w_j x_{ij} \leq c$ if the client i in the sample is bad.

B. *Mathematical formulation of the problem*

Generally, it is difficult to get a perfect separation of the good from the bad ones. One can allow possible errors by introducing non negative variables a_i. Therefore, $\sum_{j=1}^{p} w_j x_{ij} \geq c - a_i$ if applicant i in the sample is a good one, and, $\sum_{j=1}^{p} w_j x_{ij} \leq c + a_i$ if applicant i is a bad one. The objective is to find weights $(w_1, w_2, ..., w_p)$ that minimize the sum of the absolute values of these deviations (MSD) and the problem can be formulated as the following linear program:

$$\text{Min} \sum_{i=1}^{n} a_i \tag{1}$$

subject to :

$$\sum_{j=1}^{p} w_j x_{ij} \geq c - a_i \qquad \forall i \in G_1 \tag{2}$$

$$\sum_{j=1}^{p} w_j x_{ij} \leq c + a_i \qquad \forall i \in G_2 \tag{3}$$

$$a_i \geq 0 \quad \forall i, \quad and, \quad c \quad and \quad w_j \in R \quad \forall j. \tag{4}$$

Where:

- c: The hyperplane cutoff between the 1st and the 2nd group.
- a_i : The absolute value deviation
- w_i : The weights such as $\sum_{j=1}^{p} w_j x_{ij}$ is the equation of the hyperplane separating the two groups.
- x_{ij} : The value of feature j of client i
- n: The total number of population in the two groups
- n_G : The number of population of the 1st group (good clients)
- n_B : The number of population of the 2nd group (bad clients)
- G_1 : The set of good clients,
- G_2 : The set of bad clients.

For a problem of classification with two distinguished group, the main objective of the classifier is to determine a weighting vector $w = (w_1, w_2, ..., w_p)$ and a scalar c so that it assigns, as correctly as possible, the observations (i=1,2,…,n) from the group G_1 to other group G_2 based on the linear discriminant function which can be expressed as $Z_i = w_1 X_1 + w_2 X_2 + \ldots + w_p X_p$ using the separating value c.

In practical credit scoring, it consists to estimate the parameters w and a decision rule cutoff value c minimizing the number of misclassifications for the dataset. Generally, the parameter vector w and the value c are combined in such a discriminant function to determines the classification model or the classifier.

In practice again, the dataset is divided into two subsets: one for training used to build the model and the other one for testing used to evaluate the performance of the model.

978-1-5386-4225-2/18/$31.00 © 2018 IEEE