Miscellaneous Classification Topics

Oversampling

When classes are present in very different proportions as in 2% success (1) and 98% failure (0) and in the case that the benefits of identifying a success are much greater than the loss due to miss-specifying a failure as a success, it quite often the suggestion to “oversample” the data to attain a 50-50 divide of the successes and failures in the training data set. This 50-50 divide is likely to help the various classification methods better determine those factors (inputs) that help distinguish cases between success and failure. In contrast, if we were to simply random sample the training data set we would see very few successes relative to failures on which to train the classification models and it would be very difficult to improve on the naïve (baseline) model of always choosing the majority class, in this case failures.

XLMINER therefore supports the following oversampling procedure in its software (See SPB, p. 112):

1. First, the response (success) and nonresponse (failure) data are separated into two distinct sets, or strata.
2. Records are then randomly selected for the Training set from each stratum. Typically, one might select half the (scarce) responders for the training set, then an equal number of non-responders.
3. The remaining responders are put in the Validation set.
4. Non-responders are randomly selected for the validation set in sufficient numbers to maintain the original ratio of responders to non-responders.
5. If a Test Set is required, it can be taken randomly from the validation set.

For example, assume we have 100,000 cases with 2,000 responders (successes) and 98,000 non-responders (failures). Then the oversampling procedure would randomly select 1,000 responders from the set of responders and put them into the Training data set. Then, randomly, 1,000 non-responders would be chosen from the set of non-responders and put into the Training data set resulting in a 50-50 split of responders and non-responders in the Training data set (1,000 responders and 1,000 non-responders). Then the Validation data set is formed by putting the remaining 1,000 responders into the Validation data set and 49,000 non-responders would be randomly chosen from the remaining 97,000 non-responders. The Validation data set split of 1,000 responders to 49,000 non-responders maintains the 2% success rate of the original data set (= 1,000/(1,000 + 49,000)) and would provide a validation of competing classification methods in an environment that would be typical of an independent scoring exercise. Notice, that this oversampling gives rise to using 2,000 + 50,000 = 52,000 cases out of the 100,000 cases in the original sample. If one wishes to have a Test data set as well, it can be constructed by randomly drawing from the 50,000 cases in the initial Validation data set with the size of the Validation and Test data set sizes being equalized. Then, roughly, you will wind up with 500 responders and 24,500 in each of the Validation and Test data sets, each data set having, roughly, a 2% proportion of successes. Still, we wind up using 52,000 of the 100,000 cases when oversampling. Of course, if we started out with 1,000,000 cases the oversampled data set would be huge. You would have 10,000 responders and 10,000 non-responders in the training data set of 20,000 cases). In the Validation data set we would have 10,000 responders and 490,000 non-responders which maintain the population proportion of responders at 2%. Further random division into a Test data set would give us, roughly, 5,000 responders and 245,000 non-responders in each of the Validation and Test data sets. Then, of the original 1,000,000 observations, we wind up using an oversampling data set with 520,000 cases.

Two-Part Targeting Market Problems

Unlike the case where we assumed the **average revenue** garnered per responder is known, there may be instances when we have more information on how a person's characteristics are likely to affect the amount of merchandise that he/she orders from the company. Suppose, from previous experience with the expenditures of other customers, we are able to determine an expected expenditure equation, conditional on purchase, of the form

$$E_i = f(X_{i1}, X_{i2}, \cdots, X_{iK}) \ ,$$

where, for those individuals who purchased, E_i represents the expected expenditure of the i-th individual with the personal characteristics $X_{i1}, X_{i2}, \cdots, X_{iK}$. Then, given the probabilities of purchase of the new potential customers, we can sort over individuals, from highest to lowest, by their **unconditional expected expenditure**, say,

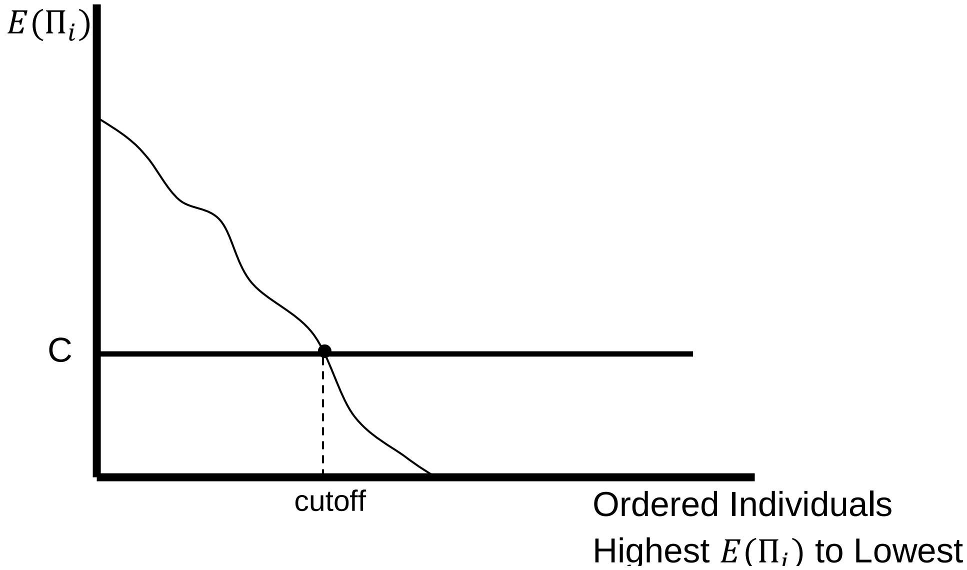
$$E(E_i) = E_i \cdot P_i \ ,$$

where P_i represents the probability of the i-th individual purchasing something. Of course, revenues do not equate to profits and therefore, if we assume a 10% profit rate on all revenue earned, we come up with an expected profit of

$$E(\Pi_i) = 0.10E(E_i)$$

Then, assuming a sufficient advertising budget, we would advertise sequentially through our sorted individuals until the expected profit of the last individual is just equal to the cost of soliciting that individual. One can represent this decision rule in the below idealized graph.

We should note that the “best” classifier to use in this case may not necessarily be the best classifier for the Average Revenue case. Recall in that case we choose the classifier that produced the maximum cumulative profit in the validation data set based on the postulated Payoff matrix. In the present case we are simply interested in choosing a classifier that does very well in the Two-Part Targeting problem. Therefore, when choosing a “best” classifier for the Two-Part Targeting problem, we should probably run several two-part experiments on the validation data set, one for each proposed classifier. Then the best classifier for the Two-Part Targeting problem would be the one that generated the largest cumulative profit in the validation data set while following the rule described above. Then, in subsequent independent scoring problems, one would use the best classifier while using the scheme described above. Of course, if the advertising budget does not allow an exhaustive solicitation to the point of the optimal cut off point (marginal profit = cost of solicitation), then one would solicit potential customers up to the point of exhaustion of the advertising budget.



Methods for Comparing Classifiers

In their textbook Introduction to Data Mining (Pearson, 2006), P. Tan, M. Steinbach, and V. Kumar have a nice section on "methods for comparing classifiers," Section 4.6, pp. 188 – 193. Their section covers (1) the calculation of a confidence interval for accuracy measures (Section 4.6.1), and (2) comparing the performance of two classifiers (Sections 4.6.2 and 4.6.3). You will find this excerpt from their book in the pdf file "Methods for Comparing Classifiers.pdf."