Solving Mixed Integer Formulation of the KS Maximization Problem – Dual Based Methods and Results from Large Practical Problems

Debashish Sarkar
Management Sciences Group
CIT, New Jersey, USA

(August 24, 2005)

Abstract

Mathematical programming-based (MP) discriminant analysis methods are non-parametric alternatives to statistical techniques, which are used to develop credit-scoring models. MP models, specifically mixed integer programming formulations, allow for the optimization of metrics such as Kolmogorov-Smirnov (KS) statistic, the area under the receiver operating curve (AUC), and misclassification costs which are typically used to measure the performance of a model. Additional constraints on model parameters due to practical considerations can also be easily incorporated into MP models. For a typical application, the KS maximization or misclassification cost minimization problem is a mixed integer program with several thousand binary variables whose difficult solution requires significant computer resources. Usually primal-based heuristics have been used to solve such problems. However, solutions from primal MP models often exhibit “flat maximum effect” (that is, different linear models are often indistinguishable in terms of their performance) and are often not as “dense” (e.g., fewer non-zero weights) as the widely popular and accepted statistical solutions (e.g. based on MLE). Non-dense solutions can lead to (i) inferior performance on the validation or test sample, or (ii) a credit scoring model without continuum risk gradation of accounts, a feature often demanded in practice. In this paper, we show that, through experiments with large practical problems, credit scoring model weights derived from the dual problem produce dense MLE-type solutions and these solutions perform equally well on the validation sample. Specifically, we propose a dual-based heuristic to solve the KS maximization problem and show that very competitive MLE-type solutions emerge when dual variables are used to infer model weights. This paper should help to increase the acceptability of the MP-based models for commercial use as a result.

1. **Introduction**

Mathematical programming-based (MP) discriminant analysis methods are non-parametric alternatives to statistical techniques, which are used to develop credit-scoring models. MP models, specifically mixed integer programming formulations, allow for the optimization of metrics such as Kolmogorov-Smirnov (KS) statistic, the area under the receiver operating curve (AUC), and misclassification costs which are typically used to measure the performance of a model. Additional constraints on model parameters due to practical considerations can also be easily incorporated into MP models. For a typical application, the KS maximization or misclassification cost minimization problem is a mixed integer program with several thousand binary variables whose difficult solution requires significant computer resources. Usually primal-based heuristics have been used to solve such problems. However, solutions from primal MP models often exhibit “flat maximum effect” (that is, different linear models are often indistinguishable in terms of their performance, see Lovie and Lovie [15]) and are often not as “dense” (e.g., fewer non-zero weights) as the widely popular and accepted statistical solutions (e.g. based on MLE). Non-dense solutions can lead to (i) inferior performance on the validation or test sample, or (ii) a credit scoring model without continuum risk gradation of accounts, a feature often demanded in practice. In this paper, we show that, through experiments with large practical problems, credit scoring model weights derived from the dual problem produce dense MLE-type solutions and these solutions perform equally well on the validation sample. Specifically, we propose a dual-based heuristic to solve the KS maximization problem and show that very competitive MLE-type solutions emerge when dual variables are used to infer model weights. This paper should help to increase the acceptability of the MP-based models for commercial use as a result.

Credit scoring is an application of general classification decisions which involve assigning observations of unknown class into specified classes by using classification models consisting of a set of variables whose values are known at decision time. Other applications of classification decisions include, among others, breast cancer diagnosis (Mangasarian, Street and Wolberg [16]), fraud detection, and pattern recognition (Warmack & Gongalez [21]). Classification models can be developed from a number of different techniques such as statistical methods (discriminant analysis, logistic and probit models), MP-based methods, neural networks, and classification trees. Model development starts with a training sample of observations whose class membership is known. The purpose of model development or discriminant analysis is to develop a function F(X), where X is the set of attributes, such that class membership is assigned depending on whether F(X) => a or F(X) < a. All methods select functions, which are "best" on training samples, hoping that this extends to the full sample. Typically, the effectiveness of a discriminant function is tested on a holdout sample.

Linear programs (LP), for example, seek to minimize some measure of the distance of each observation from the linear discriminant function F(X) = a.X or the distance that misclassified points lie on the wrong side of the function F(X) (e.g., see Erenguc & Koehler[2]; Freed & Glover[3,4]; Glover, Keene & Duca [9]; Hand [11]; Kendall [12]; Lam, Choo & Moy[13]; Mangasarian, Street & Wolberg[16]; Rubin[18]) . Mixed integer programming (MIP) models attack the underlying goal of accurate classification more directly (e.g., see Banks & Abad [1]; Gehrlein [5]; Gehrlein & Gempesaw [6]; Glen [7,8]; Loucopoulos & Pavur [14]; Rubin [17]; Stam [19]; Stam & Joachimsthaler [20]). The MIP approach has not overtaken the LP approach due to the less tractable nature of the MIP. A number of heuristics and decomposition methods have been proposed for MIP models. The MIP approach, despite its rich underlying modeling capabilities, has not gained traction among practitioners primarily for two reasons – (i) almost all heuristics have been tested on very small problems, and (ii) quality of the solution is typically far different than the MLE-type solutions (e.g., from logistic regression). The model parameters in most of the existing heuristics are derived from the "primal" version of the problem. The solution identified in the primal methods is often not dense in the number of variables and may result in inferior performance on the validation sample, or a credit scoring model without continuum risk gradation of accounts, a feature often demanded in practice. In this paper, we present a simple heuristic that is based on the dual version of the problem and show that the model weights derived from the dual version (which is a LP) lead to MLE-type solutions (both in quality and performance) for large practical problems of interest.

2. The KS Maximization Problem

2.1 Mixed Integer Programming formulation

For a two-group (good and bad) problem, the objective is to determine the weights (w_j , j=1,2,…J) for each of the J variables and a scalar c, so that the separation between the cumulative "bad" and "good" curves is largest at c. The value of each of the variables for each account is known. The jth variable value for the ith account is denoted by x_j^i . There are |G| good accounts and |B| bad accounts. Let δ_i = 1 if the ith account score is <= c; else δ_i = 0. The MIP formulation of the KS maximization problem can be written as

$$
\begin{aligned}
& Maximize\ \ \frac{1}{|B|}\sum_B \delta_i - \frac{1}{|G|}\sum_G \delta_i \\
& subject\ \ to \\
& -\sum_{j=1}^{J} w_j x_j^i \le\ -c-\varepsilon + M.\delta_i \quad ,\forall i \in G \\
& \sum_{j=1}^{J} w_j x_j^i \le\ \ c + M.(1-\delta_i\ ) \quad ,\forall i \in B \\
& c = \pm 1; \quad \delta = 0,1; \quad w_j\, unrestricted
\end{aligned}
$$

Where ε is a small number and M is a large number. By restricting c to ± 1 and ε to a small number, we eliminate the trivial solutions from the model formulation. The above formulation is referred to as the Problem (P).

2.2 Lagrangian Dual of the MIP

Let λ_i *and* θ_i be the Lagrange multipliers corresponding to the good and bad constraints in Problem (P). Then, the Langrangian dual of this problem can be written as

$$\underset{\lambda\geq 0,\theta\geq 0}{Min}\left[\underset{\delta=0,1;\ c=\pm 1;\ w_j\ unrestricted}{Max}\left\{\begin{array}{l}\frac{1}{|B|}\sum_B \delta_i - \frac{1}{|G|}\sum_G \delta_i \\ -\sum_{i\in G}\lambda_i\left(-\sum_{j=1}^{J} w_j x_j^i + c + \varepsilon - M.\delta_i\right) \\ -\sum_{i\in B}\theta_i\left(\sum_{j=1}^{J} w_j x_j^i - c - M.(1-\delta_i\ )\right)\end{array}\right\}\right]$$

Which is rewritten as

$$\underset{\lambda\geq 0,\theta\geq 0}{Min}\left[-\varepsilon\sum_G \lambda_i + M\sum_B \theta_i\ +\ \underset{\delta=0,1;\ c=\pm 1;\ w_j\ unrestricted}{Max}\left\{\begin{array}{l}\sum_G (M\lambda_i - \frac{1}{|G|})\delta_i\ +\sum_B (\frac{1}{|B|} - M\theta_i)\delta_i \\ +c(\sum_B \theta_i - \sum_G \lambda_i)\ +\sum_j (\sum_{i\in G}\lambda_i x_j^i - \sum_{i\in B}\theta_i x_j^i) w_j\end{array}\right\}\right]$$

Now, since primal feasibility implies finiteness of the dual objective, the above problem is equivalent to

$$\text{Minimize}\quad -\varepsilon\sum_G \lambda_i\ + M\sum_B \theta_i\ + \left|\sum_B \theta_i - \sum_G \lambda_i\right| + \sum_B \max(0, \frac{1}{|B|} - M\theta_i) + \sum_G \max(0, M\lambda_i - \frac{1}{|G|})$$

Subject to

$$\sum_{i\in G}\lambda_i . x_j^i\ - \sum_{i\in B}\theta_i . x_j^i\ = 0\ ,\ \ j = 1,2,\ldots J$$
$$\lambda_i\ ,\ \theta_i\ \geq 0$$

The above is a LP whose standard form representation is as follows (Problem LDLP):

$$\text{Minimize} \quad -\varepsilon\sum_G \lambda_i + M\sum_B \theta_i + Z^+ + Z^- + \sum_B b_i^+ + \sum_G g_i^+$$

subject to

$$\sum_B \theta_i - \sum_G \lambda_i - Z^+ + Z^- = 0 \qquad (2.2.1)$$

$$b_i^+ + M.\theta_i \geq \frac{1}{|B|} \quad , \; i \in B \qquad (2.2.2)$$

$$-g_i^+ + M.\lambda_i \leq \frac{1}{|G|} \quad , \; i \in G \qquad (2.2.3)$$

$$\sum_{i\in G} \lambda_i . x_j^i - \sum_{i \in B} \theta_i . x_j^i = 0 \; , \quad j = 1,2,\ldots J \qquad (2.2.4)$$

$$\lambda_i \, , \, \theta_i \, , \, g_i^+ , b_i^+ \, , \, Z^+ \, , \, Z^- \geq 0$$

Note that the dual variables corresponding to the last set of J constraints (2.2.4) have the natural interpretation of being the variable weights w_j. We exploit this observation in devising the following heuristic for the KS maximization problem. Also, we make the following observations:

(i) LDLP is the standard LP dual of the convexified Problem (P) which is obtained after replacing all the binary constraints by corresponding interval constraints, and (ii) Convexified Primal (P) is equivalent to standard LP formulations for credit scoring.

2.2 Lagrangian dual-based heuristic

Step 1 – Solve LDLP. Using the dual variables corresponding to constraints (2.2.4) as the model variable weights, identify lowest scoring q% of good accounts and r% of high scoring bad accounts. These accounts are the so-called outliers. (q and r are heuristic parameters & are <= 10%)

Step 2 – Resolve LDLP after deleting the outliers. Dual variables corresponding to (2.2.4) of the revised problem are taken to be the final model variable weights. The KS for the <u>full sample</u> is computed against these model weights.

2.3 Extensions and additional observations

The formulation of section 2.1 can easily be extended to minimize the number of misclassifications or misclassification costs. For example, $\sum_G \delta_i$ gives the total number of good accounts misclassified and ($B - \sum_B \delta_i$) denotes the total number of bad accounts misclassified. Minimizing the sum of these two quantities would correspond to the misclassification minimization problem. To minimize misclassification costs, these two quantities need to be multiplied by the corresponding costs.

At step 2 of the heuristic, one can also use other primal-based methods or introduce additional constraints or cuts to obtain better solutions. The performance of the solution from Step 2 on the full sample depends on q and r, and typically deteriorates beyond certain values of these two parameters.

3. Examples

3.1 Example 1

This example is based on a real-life credit scoring application that utilizes consumer credit data from a private credit bureau in a EU country. The development sample consisted of 4879 good and 4323 bad accounts (see Problem A in Section 4). A booked account was defined "bad" if it reached a delinquency status of 90+ days past due or worse within 2 years from the start date. A declined account was defined "bad" or "good" based on a reject inference model that utilized down-the-stream credit data.

The final logistic regression model consisted of 33 indicator variables, which were derived from 14 distinct credit bureau variables. The variable weights from the logistic regression, LP discriminant analysis, Dual (Heuristic Step 1), and Heuristic Step 2 solutions are shown in Table 1.

Table 1

Variable	Logistic regression variable weights	Primal LP discriminant model variable weights	Dual Solution variable weights (Heuristic Step 1 solution)	Heuristic Step 2 variable weights (4510 good, 4067 bad in Step 2)
W1_1	-0.5772	0	-0.9297	-1.9146
W1_2	-0.4576	0	-2.9354	-1.0216
W1_3	0.4056	0	0.1901	0.1702
W1_4	0.8303	0	0.7751	1.0216
W1_5	1.2155	0.01	1.4727	3.7560
W2_1	-0.5020	0	-1.7153	-2.1430
W2_2	-1.1510	0	-2.9092	-2.4346
W3_1	-0.3324	0	-0.1789	-0.1005
W3_2	-0.4226	0	0.1901	-1.1221
W4_1	-0.3885	0	-1.2825	-2.0432
W4_2	-0.1440	0	-0.2064	0.01
W5_1	-0.4340	-0.01	1.8755	4.0693
W5_2	-0.9836	-0.01	-1.2713	0.2608
W5_3	-1.5920	-0.01	-1.6404	-0.8514
W5_4	0.5474	0	8.5497	10.1602
W5_5	0.7234	0	6.5438	9.1386
W5_6	0.6296	0	1.9180	4.9623
W5_7	1.6390	0	5.4240	6.9458
W6_1	0.4388	0	3.7286	3.9090
W6_2	0.4901	0	1.6803	2.3457
W6_3	1.0340	0	7.0445	7.0636
W7_1	-0.2799	0	0.0112	-1.1221
W7_2	0.4595	0	1.1411	-0.1802
W8	-0.0903	0	-1.6629	-3.8882
W9_1	-0.3621	0	-3.6587	-4.0476
W9_2	-0.6188	0	-3.2896	-5.0901
W10_1	0.4339	0	5.3915	4.9995
W10_2	0.7225	0	-0.1627	2.4354
W11_1	-0.2899	0	-0.1789	-1.0216
W11_2	-0.9690	0	-3.6849	-1.9138
W12	0.3409	0	2.2222	4.5584
W13	-0.4245	0	-1.4615	-3.0160
W14	-0.3964	0	-1.1410	-0.9110
Constant	0.5687	-	-	-
KS Value (development sample)	0.544	0.4968	0.48	0.514
Validation sample KS (2073 good, 1825 bad)	0.545	0.5117	0.483	0.51

The first thing we note is the sparseness of the primal LP solution. This solution, while competitive in terms of the KS measure, does not lead to a continuum risk gradation of accounts, a feature that is often demanded from a credit-scoring model. The “continuum” nature of the risk scale is exploited for the purposes of risk-based pricing, credit limit setting, etc. The Heuristic Step 2 solution, on the other hand, is as dense as the logistic solution and is likely to be preferred to the primal LP solution.

3.2 Example 2

The second example shows that (i) denseness of LP solution can change with relatively small changes in the objective function parameters and (ii) model weights from LP solution can lead to better KS values than the logistic regression solution. Details of this example (Problem G in Section 4) are as follows:

Development sample size: 2528 good, 1831 bad

Validation sample size: 5983 good, 4279 bad

Model variables: 39 zero-one attributes from 11 distinct variables (W1-1, W1-2, W1-3; W2-1 through W2-5; W3-1 through W3-4; W4-1, W4-2, W4-3; W5-1, W5-2, W5-3; W6; W7-1, W7-2; W8-1 through W8-6; W9-1 through W9-4; W10-1 through W10-7; W11)

The primal LP discriminant model was solved for 2 cases – one with equal weights for all constraints & the second with weights proportional to 1/G or 1/B as shown in the primal formulation. The results are summarized as follows:

Model	Solution Characteristics	Development sample KS	Validation sample KS
Logistic regression	Dense in all 39 attributes	0.4782	0.48
Primal LP discriminant with equal weights in the Obj. function	4 variables in model; W2-3=W2-4=W2-5=-.01; W8-5=.01; W9-3=W9-4=-.01; W10-6=W10-7=.01	0.4059	0.4052
Primal LP discriminant with proportional weights in the Obj. function	Dense in all 39 attributes	0.4813	0.4815

We note that the KS of the proportional weight LP model (that is, the convexified problem P) is better than the KS of the logistic regression model. Also, if the variable weights in the objective function of the convexified problem P are made equal (1:1 instead of 1831:2528), the denseness of the solution drops considerably – to four variables from eleven variables for the other two cases.

4. Heuristic results on large problems

The heuristic is tested on seven real-life credit-scoring problems. The credit data for these problems came from four different bureaus (two commercial, two consumer) while the performance data was derived from either internal payment history for booked accounts or reject inferencing for declined accounts. The evaluation process consisted of the following major steps:

(1) Defining development and validation samples for each problem

(2) Developing "best" logistic regression model for each problem using the development sample

(3) Computing the MIP heuristic solution (development sample) for each problem using only the variables that were part of the final logistic model for the corresponding problem

(4) Evaluating both logistic model and MIP heuristic solutions on the corresponding validation sample

All the problems were solved using the SAS software on PC. Specifically, Proc logistic and LP were used to obtain the logistic regression and MIP solutions respectively. Table 2 provides a brief description regarding each of the problems.

Table 2

Problem	Data source for predictor variables	Development sample size	Validation sample size	Number of variables
A	Consumer credit bureau1 (Europe)	4879 good, 4323 bad	2073 good, 1825 bad	33 w_j attributes created from 14 distinct variables (all x_j^i are 0/1)
B	Consumer Credit bureau1 (Europe)	2775 good, 2457 bad	4177 good, 3691 bad	33 w_j attributes 14 distinct var (all x_j^i are 0/1)
C	Commercial credit bureau (Europe)	1628 good, 1270 bad	481 good, 393 bad	22 w_j attributes 13 distinct var (all x_j^i are 0/1)
D	Consumer Credit bureau 2 (Europe)	4808 good, 1641 bad	2032 good, 716 bad	26 w_j attributes, 9 distinct var (all x_j^i are 0/1)
E	Consumer credit bureau 2 (Europe)	6840 good, 2357 bad	2268 good, 861 bad	26 w_j attributes, 9 distinct var (all x_j^i are 0/1)
F	Com credit bureau (North America)	4245 good, 3097 bad	4266 good, 3013 bad	38 w_j attributes 11 distinct var (all x_j^i are 0/1)
G	Com credit bureau (North America)	2528 good, 1831 bad	5983 good, 4279 bad	38 w_j attributes 11 distinct var (all x_j^i are 0/1)

The MIP heuristics were run for M = 3.5 and 10.0 with ε=0.01. The final heuristic KS values were almost identical for these two different values of M. Table 3 compares the logistic and heuristic KS values for M = 3.5. The Step 2 sample sizes (q & r parameter values) are also shown in Table 3.

Table 3

Problem	Logistic model KS (Dev samp)	Heuristic Step 1 KS (Dev Samp)	Heuristic KS after Step 2 (Dev Samp)	Step2 sample size	Logistic model KS (Validation)	Heuristic KS (Validation)
A	0.5440	0.48	0.51	4510 good, 4067 bad	0.545	0.51
B	0.5429	0.51	0.54	2561 good, 2321 bad	0.5452	0.542
C	0.4306	0.41	0.4167	1606 good, 1041 bad	0.43	0.4167
D	0.569	0.535	0.558	4182 good, 1503 bad	0.569	0.5723
E	0.565	0.548	0.555	6453 good, 2180 bad	0.535	0.52
F	0.473	0.452	0.452	4245 good, 3097 bad	0.482	0.462
G	0.4782	0.441	0.443	2351 good, 1703 bad	0.48	0.448

Columns 2 and 4 in Table 3 contrast the development sample performance of the heuristic solution against the MLE solution. Similarly, the last two columns summarize the performance on the validation sample. We find that, except for problems A and G, the heuristic KS value is quite close to the MLE KS value. The heuristic solution is as dense as the MLE solution in the number of variables and it performs equally well on the validation sample. Except for problem F, we could improve the Step 1 solution in Step 2 of the heuristic.

5. Conclusions

The KS Maximization problem is solved via a 2-step heuristic where model weights are inferred from the dual variables. For a large number of practical problems, this heuristic produces robust solutions that perform equally well on the validation sample. The advantage of the dual-based solution is that it helps to identify "dense" solutions (among a host of alternatives), which not only perform well on the validation sample, but also increases the acceptability of the solution. This paper should help to increase the acceptability of the MP-based models for commercial use.

Acknowledgments: I would like to thank Dennis Wesley for his many help with the computational experiments.

6. **References**

1. W. Banks and P. Abad. An efficient optimal solution algorithm for the classification problem. Dec Sci. 22, 5 (1991) 1008-1023.
2. S. Erenguc and G. Koehler. Survey of mathematical programming models and experimental results for linear discriminant analysis. Mgmt. Dec. Econ. 11, 4 (1990) 215-225.
3. N. Freed and F. Glover. A linear programming approach to the discriminant problem. Dec. Sci. 12, 1 (1981) 68-74.
4. – and --. Simple but powerful goal programming models for discriminant problems. Eur. J. Oper. Res. 7, 1 (1981) 44-60.
5. W. V. Gehrlein. General mathematical formulations for the statistical classification problem. Oper. Rese. Letters 5 (1986) 299-304.
6. – and V. Gempesaw. A cost minimization approach to classification. In *Proceedings of the 1991 Annual Meeting of the Decision Sciences Institute,* ed. S. Melnyk (Decision Sciences Institute, Atlanta, 1991) pp. 1162-1163.
7. J. J. Glen. An iterative mixed integer programming method for classification accuracy maximizing discriminant analysis. Comp. and Oper. Res. 30 (2003) 181-198.
8. --. Categorical variable formulation: A mathematical programming approach. Presentation at Credit Scoring & Credit control VIII Conference, Edinburgh. Sept. 3-5, 2003.
9. F. Glover, S. Keene, and R. Duca. A new class of models for the discriminant problem. Dec. Sci. 19, 2 (1988) 269-280.
10. W. Gochet, A. Stam, V. Srinivasan, and S. Chen. Multigroup discriminant analysis using linear programming. Oper. Res Vol. 45, No. 2, Mar-Apr 1997.
11. D. Hand. *Discrimination and Classification* (John Wiley and Sons, 1981).
12. M. Kendall. Discrimination and Classification. In Multivariate Analysis: *Proceedings of the International Conference on Multivariate Analysis*, ed. P. Krishnaiah (Academic Press, 1966) pp. 165-185.

13. K. F. Lam, E. U. Choo, and J. W. Moy. Minimizing deviations from the group mean: a new linear programming approach for the two-group classification problem. Eur. J. Oper. Res. 88, 2 (1996) 358-367.
14. C. Loucopoulos and R. Pavur. Mixed integer programming approaches to the statistical classification problem for more than two groups. In *Proceedings of the 1990 Annual Meeting of the Decision Sciences Institute*, ed. B. Whitten and J. Gilbert (decision Sciences Institute, Atlanta, 1990) 179-181.
15. A. D. Lovie and P. Lovie. The flat maximum effect and linear scoring models for prediction. J. Forecast. 5, 159-168
16. O. L. Mangasarian, W. N. Street, and W. H. Wolberg. Breast cancer diagnosis and prognosis via linear programming. Mathematical Programming Technical Report 94-10, Dec 1994. University of Wisconsin.
17. P. Rubin. Evaluating the maximize minimum distance formulation of the linear discriminant problem. Eur. J. Oper. Res. 41, 2 (1989) 240-248.
18. P. A. Rubin. Solving mixed integer classification problems by decomposition. Notes. 1997.
19. A Stam. Extensions of mathematical programming-based classification rules: a multicriteria approach. Eur. J. Oper. Res. 48, 3 (1990) 351-361
20. – and E. Joachimsthaler. A comparison of a robust mixed-integer approach to existing methods for establishing classification rules for the discriminant problem. Eur. J. Oper. Res. 46, 1 (1990) 113-122
21. R. E. Warmack and R. C. Gonzalez. An algorithm for the optimal solution of linear inequalities and its application to pattern recognition. IEEE Trans. Comp C-22, 12 (1973) 1065-1075.