Supersparse Linear Integer Models for Predictive Scoring Systems

Berk Ustun, Stefano Tracà, Cynthia Rudin
Massachusetts Institute of Technology
77 Massachusetts Avenue, Cambridge, MA 02139, USA

Scoring systems are classification models that make predictions using a sparse linear combination of variables with integer coefficients. Such systems are frequently used in medicine because they are interpretable; that is, they only require users to add, subtract and multiply a few meaningful numbers in order to make a prediction. See, for instance, these commonly used scoring systems: (Gage et al. 2001; Le Gall et al. 1984; Le Gall, Lemeshow, and Saulnier 1993; Knaus et al. 1985). Scoring systems strike a delicate balance between accuracy and interpretability that is difficult to replicate with existing machine learning algorithms.

Current linear methods such as the lasso, elastic net and LARS are not designed to create scoring systems, since regularization is primarily used to improve accuracy as opposed to sparsity and interpretability (Tibshirani 1996; Zou and Hastie 2005; Efron et al. 2004). These methods can produce very sparse models through heavy regularization or feature selection methods (Guyon and Elisseeff 2003); however, feature selection often relies on greedy optimization and cannot guarantee an optimal balance between sparsity and accuracy. Moreover, the interpretability of scoring systems requires integer coefficients, which these methods do not produce. Existing approaches to interpretable modeling include decision trees and lists (Rüping 2006; Quinlan 1986; Rivest 1987; Letham et al. 2013).

We introduce a formal approach for creating scoring systems, called Supersparse Linear Integer Models (SLIM). SLIM produces scoring systems that are accurate and interpretable using a mixed-integer program (MIP) whose objective penalizes the training error, L_0-norm and L_1-norm of its coefficients. SLIM can create scoring systems for datasets with thousands of training examples and tens to hundreds of features - larger than the sizes of most studies in medicine, where scoring systems are often used.

Formulation

Given a dataset with N examples and P features, SLIM produces classifier $\hat{y} = \text{sign}(\mathbf{x}^T \boldsymbol{\lambda})$ where $\mathbf{x} \in \mathbb{R}^P$ is a vector of features, $\hat{y} \in \{-1, 1\}$, are predicted labels and $\boldsymbol{\lambda} \in \mathbb{Z}^P$ is a vector of coefficients. The optimization for SLIM on N

training examples is:

$$\min_{\boldsymbol{\lambda}} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[y_i \mathbf{x}_i^T \boldsymbol{\lambda} \leq 0] + C_0 \left\|\boldsymbol{\lambda}\right\|_0 + C_1 \left\|\boldsymbol{\lambda}\right\|_1$$

where C_0 and C_1 are penalties associated with the L_0-norm and L_1-norm of $\boldsymbol{\lambda}$. C_0 is the primary parameter, inducing a sparse set of coefficients, and C_1 is the secondary parameter ensuring that the coefficients are as small (and interpretable) as possible. Note C_1 does not take away from the sparsity of the solution, but promotes the smallest coefficients among equally sparse models. The objective is minimized by a MIP with $N + 3P$ variables and $2N + 6P$ constraints:

$$\min_{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}} \frac{1}{N} \sum_{i=1}^{N} \alpha_i + C_0 \sum_{j=1}^{P} \beta_j + C_1 \sum_{j=1}^{P} \gamma_j$$

$$-M\alpha_i + \epsilon \leq y_i \mathbf{x}_i^T \boldsymbol{\lambda} \leq M(1 - \alpha_i) + \epsilon \quad i = 1 \ldots N$$
$$-\Lambda \beta_j \leq \lambda_j \leq \Lambda \beta_j \qquad j = 1 \ldots P$$
$$-\gamma_j \leq \lambda_j \leq \gamma_j \qquad j = 1 \ldots P$$
$$\boldsymbol{\lambda} \in \mathcal{L}$$
$$\alpha_i \in \{0, 1\}, \beta_j \in \{0, 1\}, \gamma_j \in \mathbb{R}_+ \qquad j = 1 \ldots P$$

where $\alpha_i = \mathbb{1}[y_i \neq \hat{y}_i]$, $\beta_j = \mathbb{1}[\lambda_j \neq 0]$, and $\gamma_j = |\lambda_j|$. All feasible $\boldsymbol{\lambda}$ belong to $\{\boldsymbol{\lambda} \in \mathcal{L} : |\lambda_j| \leq \Lambda \; \forall j\}$. By default, we set $\Lambda = 100$ and $\mathcal{L} = \mathbb{Z}^P$, although we often further restrict the coefficients to have only one significant digit. Lastly, ϵ and M are scalars used in if-then constraints; we set $\epsilon = 0.1$, $M = \Lambda \cdot \max_{i,j} |x_{ij}|$. Computational factors affect the accuracy and sparsity of SLIM scoring systems: current MIP solvers can train sparse and accurate scoring systems for datasets with $N \approx 10000$ and $P \approx 100$.

Theoretical Bound

We can bound the true risk of a SLIM scoring system, $R^{\text{true}}(f) = \mathbb{E}_{X,Y} \mathbb{1}[f(X) \neq Y]$, by its empirical risk, $R^{\text{emp}}(f) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[f(\mathbf{x}_i) \neq y_i]$, as follows: For every $\delta > 0$, and every $f \in \{f_1, \cdots, f_K\}$,

$$R^{\text{true}}(f) \leq R^{\text{emp}}(f) + \sqrt{\frac{\log(K) - \log(\delta)}{2N}},$$

with probability at least $1 - \delta$. Here, $f(\mathbf{x}_i) = \text{sign}(\mathbf{x}_i^T \boldsymbol{\lambda})$ and $\boldsymbol{\lambda} \in \{\boldsymbol{\lambda} \in \mathcal{L} : |\lambda_j| \leq \Lambda\}$. Since $\mathcal{L} \subseteq \mathbb{Z}^P$, it can be shown that $\log(K) \leq P \log(2\Lambda + 1)$.

Experimental Results

We compare the accuracy and sparsity of SLIM scoring systems to classification models produced by C5.0, CART, Logistic Regression (LR), Elastic Net (EN), Random Forests (RF) and Support Vector Machines (SVM) in Table 1. Our comparison includes the breastcancer, haberman, internetad, mammo, spambase and tictactoe datasets from the UCI Machine Learning Repository, as well as the violent-crime dataset, which is derived from a study of crime among young people raised in out-of-home care, made available by the US Department of Justice Statistics (Cusick et al. 2010).

We report the mean 5-fold cross-validation (CV) test error as a measure of accuracy (top), and the median 5-fold CV model size as a measure of sparsity (bottom). Model size reflects the number of coefficients for LR, EN and SLIM, and the number of leaves for C5.0 and CART; we omit this statistic for RF and SVM as sparsity does not affect the interpretability of these methods. SLIM models were trained for 1 hour using the CPLEX 12.5 API for MATLAB; all other models were trained for default times using packages in R 2.15. We set the free parameters for most methods to values that minimized the 5-fold CV test error; for EN, we set the L_1 penalty to the value that produced the sparsest model on the L_1 regularization path, within 1-SE of the L_1 penalty that minimized the 5-fold CV test error.

Dataset	N	P	C5.0	CART	LR	EN	RF	SVM	SLIM
breastcancer	683	10	5.3% 8	5.9% 4	3.7% 9	3.5% 4	2.7% -	2.9% -	3.7% 3
haberman	306	4	27.8% 3	26.8% 6	26.5% 3	26.5% 4	28.1% -	26.2% -	23.2% 3
internetad	2359	1431	3.9% 10	4.5% 7	8.5% 616	3.1% 473	2.5% -	3.7% -	3.6% 14
mammo	961	6	18.4% 5	17.5% 4	29.4% 5	32.4% 4	18.0% -	17.2% -	17.2% 4
spambase	4601	58	7.7% 63	10.6% 7	7.3% 57	7.6% 52	4.8% -	6.5% -	7.4% 18
tictactoe	958	28	7.5% 39	11.7% 21	2.7% 18	1.7% 19	1.6% -	0.7% -	3.3% 18
violentcrime	558	108	25.1% 27	24.9% 10	22.2% 57	19.5% 108	21.3% -	19.5% -	20.1% 9

Table 1: Accuracy and sparsity of all methods.

Our results suggest that these methods produce classification models that are comparable in terms of accuracy but vary *dramatically* in terms of sparsity. In particular, SLIM consistently produces scoring systems that are both accurate and sparse. Furthermore, as shown in the following two demonstrations, SLIM also produces scoring systems that are highly interpretable.

A Scoring System to Detect Breast Cancer

When applied to the breastcancer dataset (Mangasarian and Wolberg 1990), SLIM produces a scoring system to predict whether a tumor is malignant (Class = +1) using only 3 cell-related features. Here, coefficients are restricted to $\mathcal{L} = \{0, \pm 1, \pm 5, \pm 10, \pm 50, \pm 100, \pm 500\}^{10}$:

$$\text{Score} = \text{ClumpThickness} + \text{UniformityOfCellSize}$$
$$+ \text{BareNuclei} - 10$$

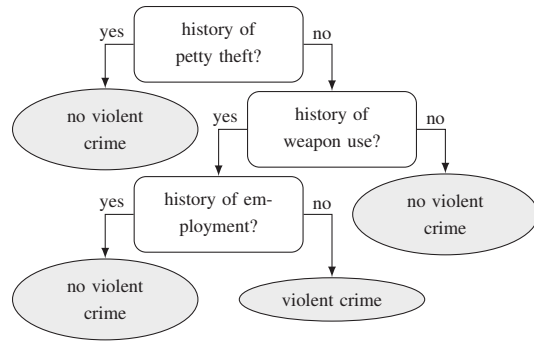Predicted Class $= \text{sign}(\text{Score})$.

Figure 1: Decision tree induced by a SLIM model.

This score would be easy for doctors to compute when analyzing patient scans.

A Scoring System to Predict Violent Crime

When applied to the violentcrime dataset, SLIM produces the following scoring system to predict whether a young person raised in out-of-home care will commit a violent crime (Class = +1) using 3 features related to their background and criminal record:

$$\text{Score} = -10\,\text{PettyTheft} + 9\,\text{WeaponUse}$$
$$- 9\,\text{Employment} - 1$$

Predicted Class $= \text{sign}(\text{Score})$.

The three features are indicators for past history of petty theft, indicator of past weapon use, and whether the person has ever been employed. In this case, we had restricted coefficients to the set $\mathcal{L} = \{0, \pm 1, \pm 9, \pm 10\}^{108}$. This model can also be visualized as a decision tree since it uses only discrete features, as shown in Figure 1.

Conclusion

We introduced SLIM as a tool to create data-driven scoring systems for binary classification. SLIM's models tend to be accurate since they are optimized, but also highly interpretable, as they are built from a small number of non-zero terms with integer coefficients.

References

Cusick, G. R.; Courtney, M. E.; Havlicek, J.; and Hess, N. 2010. *Crime during the Transition to Adulthood: How Youth Fare as They Leave Out-of-Home Care*. National Institute of Justice, Office of Justice Programs, US Department of Justice.

Efron, B.; Hastie, T.; Johnstone, I.; and Tibshirani, R. 2004. Least angle regression. *The Annals of Statistics* 32(2):407–499.

Gage, B. F.; Waterman, A. D.; Shannon, W.; Boechler, M.; Rich, M. W.; and Radford, M. J. 2001. Validation of clinical classification schemes for predicting stroke. *Journal of the American Medical Association* 285(22):2864–2870.

Guyon, I., and Elisseeff, A. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research* 3:1157–1182.

Knaus, W. A.; Draper, E. A.; Wagner, D. P.; and Zimmerman, J. E. 1985. Apache II: a severity of disease classification system. *Critical Care Medicine* 13(10):818–829.

Le Gall, J.-R.; Loirat, P.; Alperovitch, A.; Glaser, P.; Granthil, C.; Mathieu, D.; Mercier, P.; Thomas, R.; and Villers, D. 1984. A simplified acute physiology score for ICU patients. *Critical Care Medicine* 12(11):975–977.

Le Gall, J.-R.; Lemeshow, S.; and Saulnier, F. 1993. A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. *Journal of the American Medical Association* 270(24):2957–2963.

Letham, B.; Rudin, C.; McCormick, T. H.; and Madigan, D. 2013. An interpretable stroke prediction model using rules and Bayesian analysis. In *Proc. AAAI Late-Breaking Track*.

Mangasarian, O. L., and Wolberg, W. H. 1990. Cancer diagnosis via linear programming. *SIAM News* 23(5):1,18.

Quinlan, J. R. 1986. Induction of decision trees. *Machine Learning* 1(1):81–106.

Rivest, R. L. 1987. Learning decision lists. *Machine Learning* 2(3):229–246.

Rüping, S. 2006. *Learning interpretable models*. Ph.D. Dissertation, Universitäat Dortmund.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.

Zou, H., and Hastie, T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2):301–320.