Feature Subset Selection for Logistic Regression via Mixed Integer Optimization

Yuichi TAKANO (Senshu University, Japan)
Toshiki SATO (University of Tsukuba)
Ryuhei MIYASHIRO (Tokyo University of Agriculture and Technology)
Akiko YOSHISE (University of Tsukuba)

Workshop on Advances in Optimization (WAO2016)
TKP Shinagawa Conference Center, Tokyo, JAPAN
August 12-13, 2016

Outline

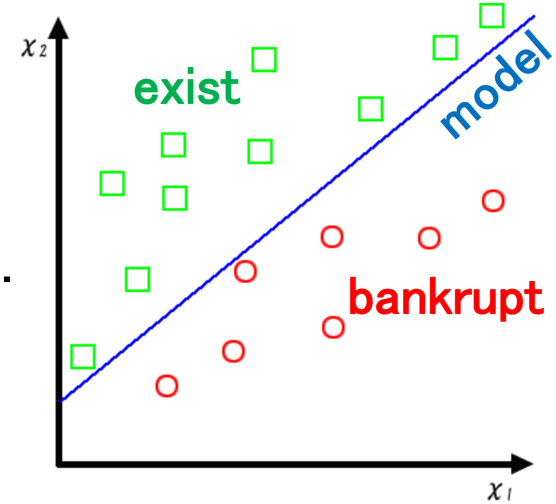
- Introduction
- Mixed Integer Optimization Formulation
- Computational Results
- Conclusions

Outline

- Introduction
- Mixed Integer Optimization Formulation
- Computational Results
- Conclusions

Binary Classification

- **Binary classification** aims at developing a model for separating two classes of samples that are characterized by numerical features.
  - Example: corporate bankruptcy prediction
  - Methods: classic discriminant analysis, logistic regression, SVM and so on



- We focus on the **feature subset selection problem for logistic regression**.

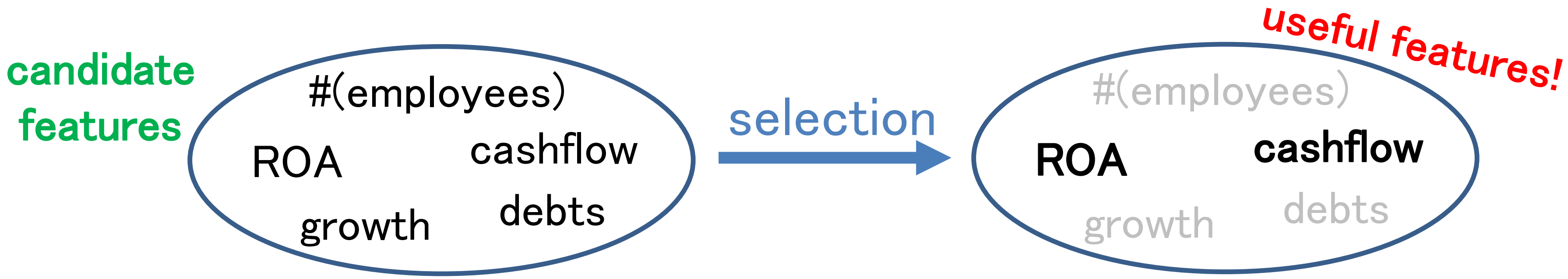
$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top$ —input→ model —predict→ $y_i \in \{-1, 1\}$

financial indicators of company i

company i will exist or go bankrupt

Feature Subset Selection

- **Feature subset selection** is the method of choosing a set of significant features for model construction.



- Potential benefits of feature subset selection are:
  - Improving predictive performance by preventing overfitting
  - Identifying a model that captures the essence of a system
  - Providing a computationally efficient set of features
- It is essential importance in statistics.
- It has recently received considerable attention in data mining and machine learning as a result of the increased size of the datasets.

Methods for Feature Subset Selection

- **Stepwise Method** (i.e., local search algorithm)
- **Metaheuristics** (e.g., tabu search, simulated annealing)
- **L_1-regularized Regression** (a.k.a. LASSO)
  - These algorithms do not necessarily provide a best subset of features under a goodness-of-fit measure (e.g., AIC, BIC, C_p)
- **Branch-and-Bound Algorithm** (e.g., Narendra & Fukunaga (1977))
  - This algorithm assumes the monotonicity of a GOF measure in its pruning process; this assumption is not satisfied by commonly used goodness-of-fit measures.
- **Our Approach: Mixed Integer Optimization (MIO)**
  - We formulate the problem as an MILO problem by making a piecewise linear approximation of the logistic loss function.

Outline

- Introduction
- Mixed Integer Optimization Formulation
- Computational Results
- Conclusions

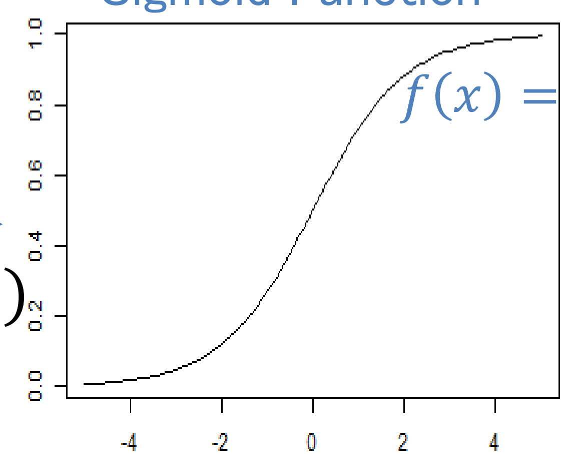
Logistic Regression Model

Logistic regression model

$$\Pr(y \mid \boldsymbol{x}) = \frac{1}{1 + \exp(-y(\boldsymbol{w}^\top \boldsymbol{x} + b))}$$

y : binary class label ($y \in \{-1,1\}$)
$\boldsymbol{x}$: p–dimensional feature vector
$\boldsymbol{w}$: p–dimensional coefficient vector (to be estimated)
b : intercept (to be estimated)

Sigmoid Function



$$f(x) = \frac{1}{1 + \exp(-x)}$$

$\boldsymbol{x} \longrightarrow \boldsymbol{w}^\top \boldsymbol{x} + b \longrightarrow$ Sigmoid Function $\longrightarrow \Pr(y \mid \boldsymbol{x})$

feature vector

$(-\infty, +\infty)$

$(0,1)$

occurrence probability

Information Criterion

- The **log likelihood function** is defined as follows:

$$\ell(b, \boldsymbol{w}) = \log\left(\prod_{i=1}^{n} \underbrace{\Pr(y_i \mid \boldsymbol{x}_i)}_{\text{occurrence probability}}\right) \cdots = -\sum_{i=1}^{n} f(y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b))$$

Logistic loss function

$$f(v) = \log(1 + \exp(-v))$$

- We will select a subset $S \subseteq \{1,2,\ldots,p\}$ of features so that the **information criterion** is minimized:

$$\mathrm{IC}(S) = -2 \underbrace{\max\{\ell(b, \boldsymbol{w}) \mid w_j = 0\ (j \notin S)\}}_{\text{maximum log likelihood}} + F\,(|S| + 1)$$

penalty parameter (F)

#features ($|S|$)

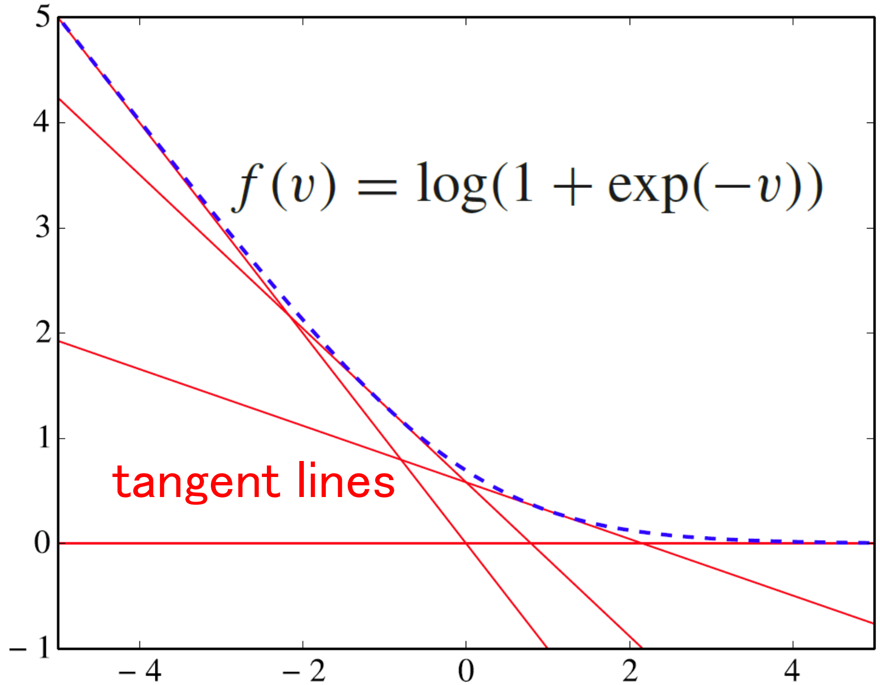
$F = 2$ → **Akaike information criterion**

$F = \log n$ → **Bayesian information criterion**

Piecewise Linear Approximation

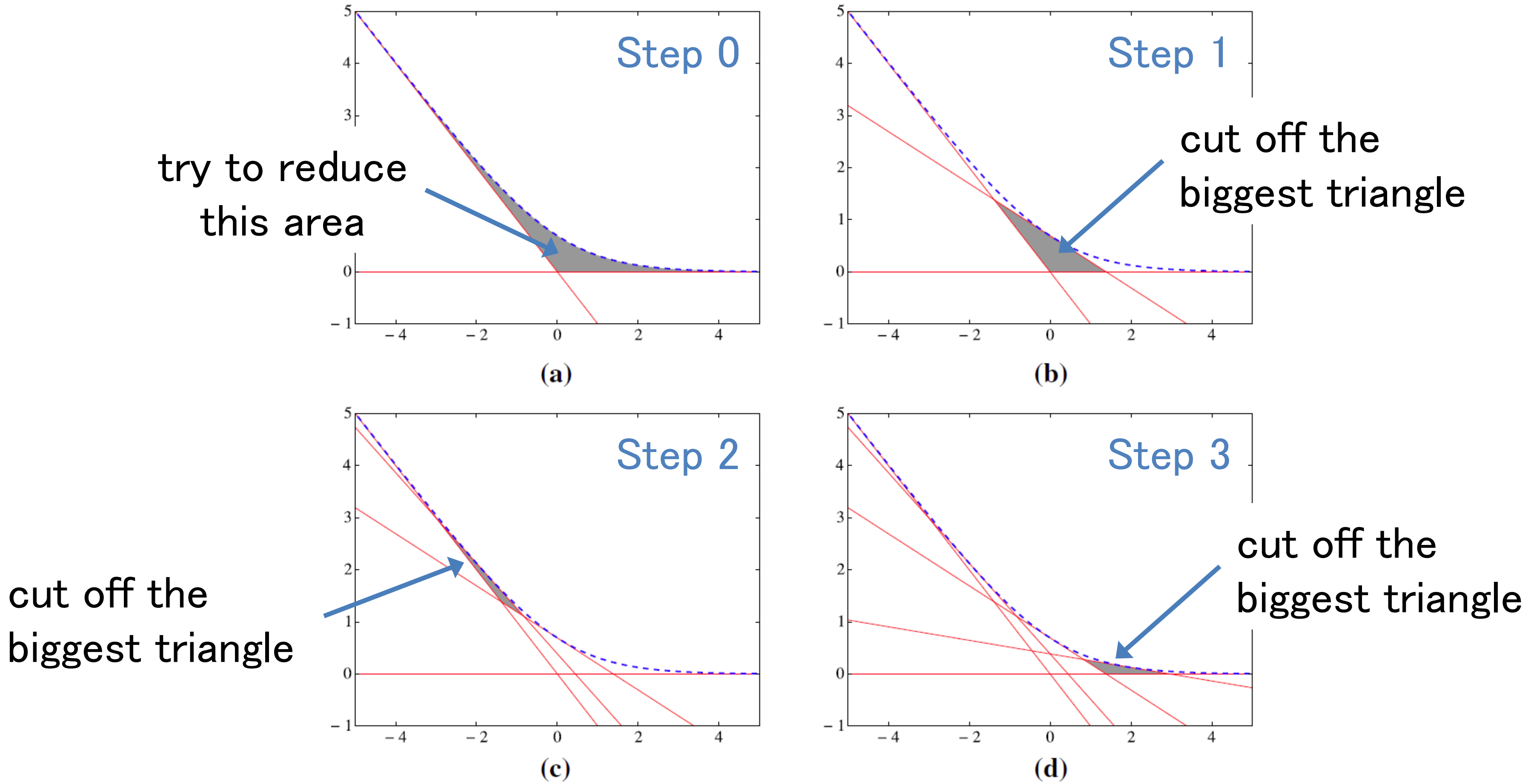
- Most MIO software cannot handle such a nonlinear function.
- The logistic loss function can be approximated by the pointwise maximum of a family of tangent lines as follows:

$$
\begin{aligned}
f(v) &\approx \max\left\{ f'(v_k)(v - v_k) + f(v_k) \mid k = 1, 2, \ldots, m \right\} \\
&= \min\left\{ t \mid t \geq f'(v_k)(v - v_k) + f(v_k) \quad (k = 1, 2, \ldots, m) \right\}
\end{aligned}
$$



Greedy Algorithm for Selecting Tangent Lines

- It is crucial to select a limited number of “good” tangent lines.
- Our greedy algorithm adds tangent lines one by one so that the biggest triangle (= approximation error) will be cut off.



Mixed Integer Linear Optimization Formulation

- Feature subset selection for logistic regression is formulated as a mixed integer linear optimization (MILO) problem as follows:

$$\underset{b,\boldsymbol{t},\boldsymbol{w},\boldsymbol{z}}{\text{minimize}} \quad 2\sum_{i=1}^{n} t_i + F\left(\sum_{j=1}^{p} z_j + 1\right)$$ min. information criterion

$$\text{subject to} \quad t_i \geq f'(v_k)\left(y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - v_k\right) + f(v_k)$$ tangent lines

$$(i = 1, 2, \ldots, n;\, k = 1, 2, \ldots, m),$$

$$z_j = 0 \;\Rightarrow\; w_j = 0 \quad (j = 1, 2, \ldots, p),$$ unselected feature is eliminated

$$z_j \in \{0, 1\} \quad (j = 1, 2, \ldots, p).$$ the j–th feature is selected or not

big–M method: $-Mz_j \leq w_j \leq Mz_j$

SOS type1: GRB.SOS_TYPE1: $\{1 - z_j, w_j\}$

Optimality Guarantee

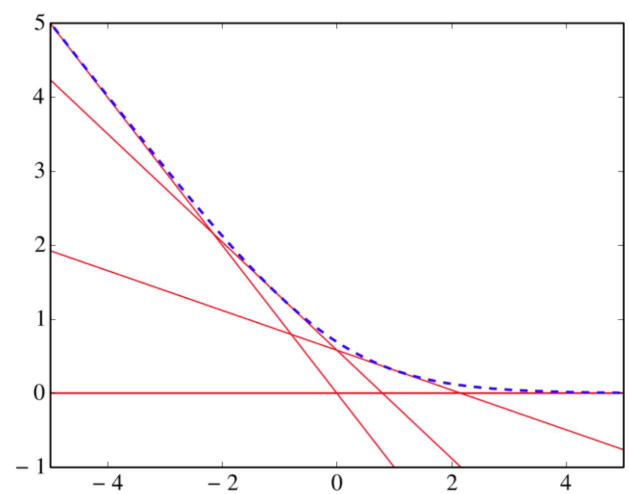
- By solving the MILO problem, we have

  $\mathrm{obj}^*_{\mathrm{MILO}}$: optimal objective value,
  S^* : subset of features.

- Let $\mathrm{IC}_{\mathrm{opt}}$ be the minimum value of the information criterion.
  - The objective function of the MILO problem is an underestimator to the information criterion.
  - S^* is not necessarily a minimizer of the information criterion.

- Thus we can give an optimality guarantee to an obtained subset S^* as follows:

$$\underbrace{\mathrm{obj}^*_{\mathrm{MILO}}}_{\text{lower bound}} \leq \mathrm{IC}_{\mathrm{opt}} \leq \underbrace{\mathrm{IC}(S^*)}_{\text{upper bound}}$$



Outline

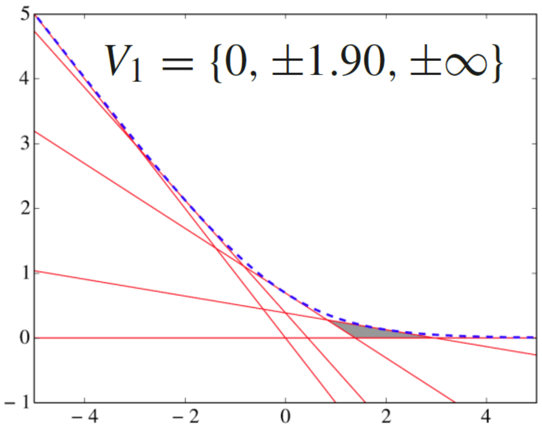
- Introduction
- Mixed Integer Optimization Formulation
- Computational Results
- Conclusions

Experimental Design for AIC minimization

- The datasets for classification were downloaded from the UCI machine learning repository.

- We compare the following methods for minimizing AIC:
  - $\mathrm{SW}_{\mathrm{const}}$: stepwise method starting with $S = \emptyset$ (`step` in R)
  - $\mathrm{SW}_{\mathrm{all}}$: stepwise method starting with $S = \{1,2,\ldots,p\}$ (`step` in R)
  - L_1−reg: L_1−regularized logistic regression (`glmnet` in R)
  - MILO(V): our MILO formulation (Gurobi Optimizer)

- We employed three sets of tangent lines computed by our greedy algorithm:



$$V_1 = \{0, \pm 1.90, \pm\infty\} \quad (|V_1| = 5, \text{ see also Fig. 2d}),$$
$$V_2 = \{0, \pm 0.89, \pm 1.90, \pm 3.55, \pm\infty\} \quad (|V_2| = 9),$$
$$V_3 = \{0, \pm 0.44, \pm 0.89, \pm 1.37, \pm 1.90, \pm 2.63, \pm 3.55, \pm 5.16, \pm\infty\} \quad (|V_3| = 17).$$

Breast Cancer Dataset ($n = 194$, $p = 33$)

Method	$\mathrm{AIC}(S)$	LB	Relgap	$\lvert S \rvert$	Time (sec.)
$\mathrm{SW}_{\mathrm{const}}$	162.94	---	---	12	1.06
$\mathrm{SW}_{\mathrm{all}}$	152.13	---	---	24	1.49
L_1–reg	157.57	---	---	24	4.67
MILO(V_1), $\lvert V_1 \rvert = 5$	147.04	137.96	6.58%	18	22.88
MILO(V_2), $\lvert V_2 \rvert = 9$	147.04	144.56	1.72%	18	57.72
MILO(V_3), $\lvert V_3 \rvert = 17$	147.04	146.41	0.43%	18	240.39

- $\mathrm{AIC}(S)$: AIC of selected subset S of features.
- LB: obj.val. of MILO problem (i.e., lower bound on min. AIC)
- Relgap: relative optimality gap, i.e., $100 \times \frac{\mathrm{AIC}(S) - \mathrm{LB}}{\mathrm{LB}}$
- $\lvert S \rvert$: the number of selected features
- Time (sec.): computation time in seconds

Breast Cancer Dataset ($n = 194$, $p = 33$)

Method	AIC(S)	LB	Relgap	$\lvert S \rvert$	Time (sec.)
SW_{const}	162.94	---	---	12	1.06
SW_{all}	152.13	---	---	24	1.49
L_1–reg	157.57	---	---	24	4.67
MILO(V_1), $\lvert V_1 \rvert = 5$	147.04	137.96	6.58%	18	22.88
MILO(V_2), $\lvert V_2 \rvert = 9$	147.04	144.56	1.72%	18	57.72
MILO(V_3), $\lvert V_3 \rvert = 17$	147.04	146.41	0.43%	18	240.39

- Stepwise methods and L_1–reg finished their computations within five seconds, but they provided low–quality solutions.
- MILO formulations successfully attained the smallest AIC value among these methods.
- As the number of tangent lines increased, Relgap got small, but its computation time also increased.

Libras Movement Dataset ($n = 360, p = 90$)

Method	AIC(S)	LB	Relgap	$\|S\|$	Time (sec.)
$\mathrm{SW}_{\mathrm{const}}$	22.00	---	---	10	6.53
$\mathrm{SW}_{\mathrm{all}}$	18.00	---	---	8	323.22
L_1−reg	28.00	---	---	13	4.75
MILO(V_1), $\|V_1\| = 5$	14.00	8.00	75.00%	6	10000.00
MILO(V_2), $\|V_2\| = 9$	16.00	8.00	100.00%	7	10000.00
MILO(V_3), $\|V_3\| = 17$	16.00	6.00	166.67%	7	10000.00

- MILO computation took very long time, so we quit its computation in 10000 seconds.
- Stepwise methods and L_1−reg finished their computations within six minutes but they still provided low−quality solutions.
- MILO formulations attained smaller AIC values than the other methods; however, Relgap was very large⋯.

Outline

- Introduction
- Mixed Integer Optimization Formulation
- Computational Results
- Conclusions

Conclusions

- This talk considered the feature subset selection problem for logistic regression.
- We formulated its approximation as a mixed integer linear optimization (MILO) problem by applying a piecewise linearization technique to the logistic loss function.
- We also developed a greedy algorithm to select good tangent lines used for piecewise linear approximation.
- Our approach has the advantage of selecting a subset of features with an optimality guarantee.
- Our method often outperformed the stepwise methods and L_1-regularized logistic regression in terms of solution quality.
- Future directions of study: specialized B&B algorithm, discrete choice model (e.g., multinomial logit model), and exact algorithm for selecting a "best" set of tangent lines.

References

T. Sato, Y. Takano, R. Miyashiro and A. Yoshise
Feature Subset Selection for Logistic Regression via Mixed Integer Optimization
Computational Optimization and Applications
Vol.64, No.3, pp.865-880 (2016).

Thank you for your attention!