ERATO感謝祭 Season IV

特徴選択のためのLasso解列挙（AAAI'17）

原 聡[1,2]、前原 貴憲[3]

1) 国立情報学研究所
2) JST, ERATO, 河原林巨大グラフプロジェクト
3) 理研AIP

研究背景

研究背景：特徴選択は完璧か？

- 『特徴選択を使うと、タスクに関連する特徴量と、タスクに関連しない特徴量とを識別することができる』と言われている。
  - Lassoを使うとモデルのスパースな表現が得られる。
  - Lassoによって選ばれた特徴量が重要な特徴量だと言われている。

研究背景：特徴選択は完璧か？

- 『特徴選択を使うと、タスクに関連する特徴量と、タスクに関連しない特徴量とを識別することができる』と言われている。
  - Lassoを使うとモデルのスパースな表現が得られる。
  - Lassoによって選ばれた特徴量が重要な特徴量だと言われている。

- しかし、機械学習に完璧はありえない。
  - 有限のデータから学習する以上、ある程度のエラーは起こりうる。
  - データ由来・学習手法由来のバイアスがのることがある。

研究背景：特徴選択は完璧か？

- 『特徴選択を使うと、タスクに関連する特徴量と、タスクに関連しない特徴量とを識別することができる』と言われている。
  - Lassoを使うとモデルのスパースな表現が得られる。
  - Lassoによって選ばれた特徴量が重要な特徴量だと言われている。

- しかし、機械学習に完璧はありえない。
  - 有限のデータから学習する以上、ある程度のエラーは起こりうる。
  - データ由来・学習手法由来のバイアスがのることがある。

機械学習は時として間違える。
機械学習がミスすると。。。

研究背景：機械学習がミスすると。。。

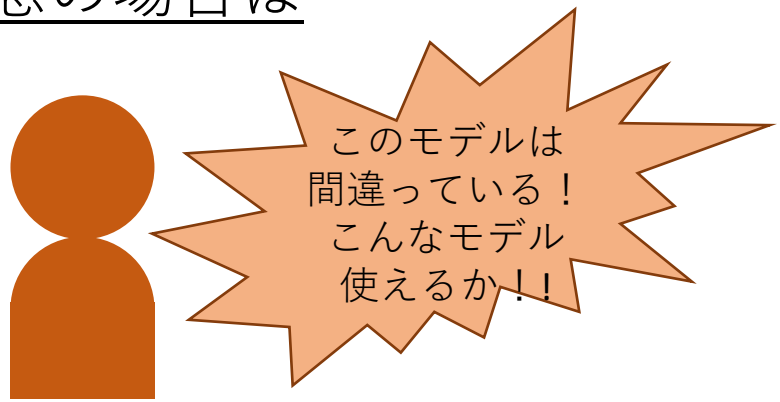
研究背景：機械学習がミスすると。。。

研究背景：機械学習がミスすると。。。

- 理想的には

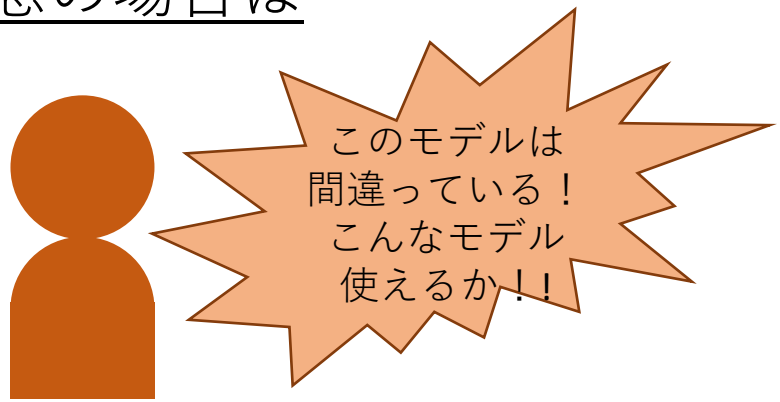
研究背景：機械学習がミスすると。。。

- 最悪の場合は



研究背景：機械学習がミスすると。。。

- 最悪の場合は



悲劇
たとえ精度の高いモデルでも、ユーザの信頼が得られないと使われない。

研究背景：ユーザに信頼される特徴選択をしたい。

- しかし、“間違えない特徴選択”は難しい。
  - Lassoは選ばれた特徴量が“真に重要な特徴量”であることが保証されない。
    - Adaptive Lassoをはじめ、様々な改善法が考案されている。
    - しかし、有限のデータから学習している以上、エラーは避けられない。

研究背景：ユーザに信頼される特徴選択をしたい。

- しかし、“間違えない特徴選択”は難しい。
  - Lassoは選ばれた特徴量が“真に重要な特徴量”であることが保証されない。
    - Adaptive Lassoをはじめ、様々な改善法が考案されている。
    - しかし、有限のデータから学習している以上、エラーは避けられない。

- 本研究のアイディア
  - そもそも“重要な特徴量の組”を一つ探そうとしているから難しい。

研究背景：ユーザに信頼される特徴選択をしたい。

- しかし、“間違えない特徴選択”は難しい。
  - Lassoは選ばれた特徴量が“真に重要な特徴量”であることが保証されない。
    - Adaptive Lassoをはじめ、様々な改善法が考案されている。
    - しかし、有限のデータから学習している以上、エラーは避けられない。

- 本研究のアイディア
  - そもそも“重要な特徴量の組”を一つ探そうとしているから難しい。
  - “重要な特徴量の組”をたくさん見つけて、それをユーザに提示したらどうか？
    → **「Lasso解を複数列挙する問題」**を考える。

研究背景：ユーザに信頼される特徴選択をしたい。

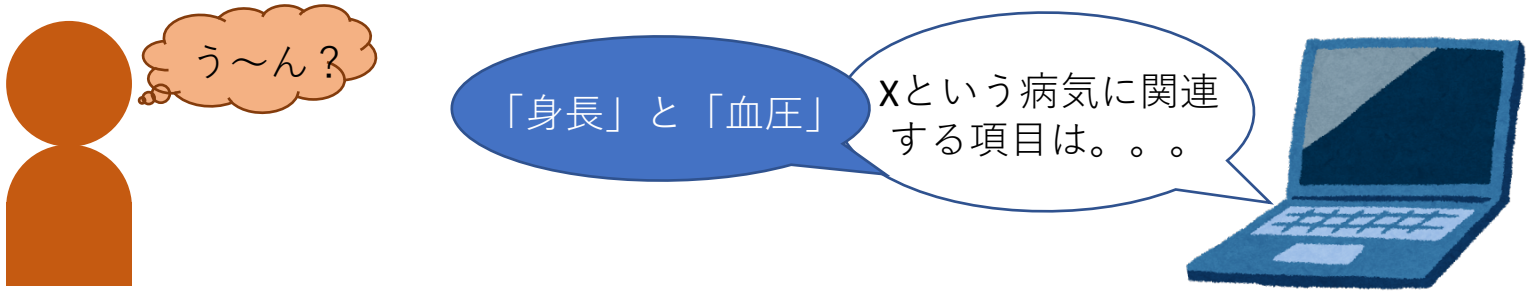
- しかし、“間違えない特徴選択”は難しい。
  - Lassoは選ばれた特徴量が“真に重要な特徴量”であることが保証されない。
    - Adaptive Lassoをはじめ、様々な改善法が考案されている。
    - しかし、有限のデータから学習している以上、エラーは避けられない。

- 本研究のアイディア
  - そもそも“重要な特徴量の組”を一つ探そうとしているから難しい。
  - “重要な特徴量の組”をたくさん見つけて、それをユーザに提示したらどうか？
    → **「Lasso解を複数列挙する問題」**を考える。

研究背景：ユーザに信頼される特徴選択をしたい。

- しかし、“間違えない特徴選択”は難しい。
  - Lassoは選ばれた特徴量が“真に重要な特徴量”であることが保証されない。
    - Adaptive Lassoをはじめ、様々な改善法が考案されている。
    - しかし、有限のデータから学習している以上、エラーは避けられない。

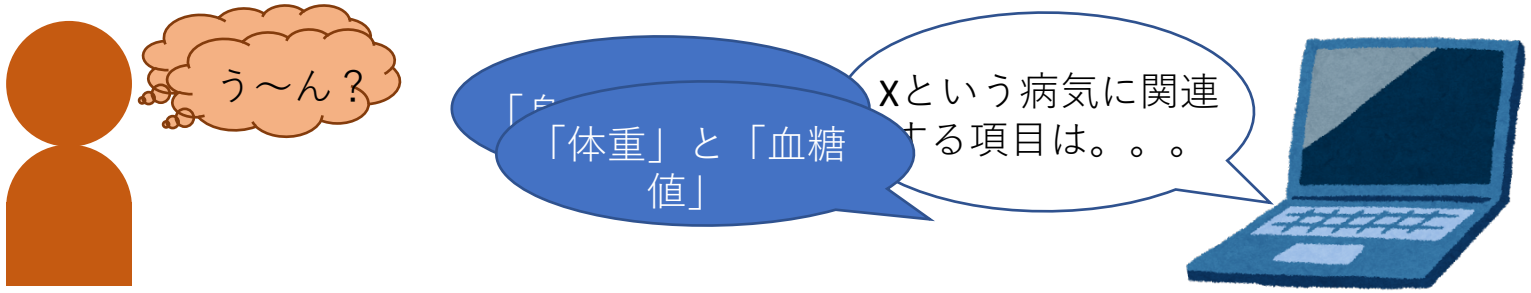
- 本研究のアイディア
  - そもそも“重要な特徴量の組”を一つ探そうとしているから難しい。
  - “重要な特徴量の組”をたくさん見つけて、それをユーザに提示したらどうか？
    → **「Lasso解を複数列挙する問題」**を考える。



研究背景：ユーザに信頼される特徴選択をしたい。

- しかし、“間違えない特徴選択”は難しい。
  - Lassoは選ばれた特徴量が“真に重要な特徴量”であることが保証されない。
    - Adaptive Lassoをはじめ、様々な改善法が考案されている。
    - しかし、有限のデータから学習している以上、エラーは避けられない。

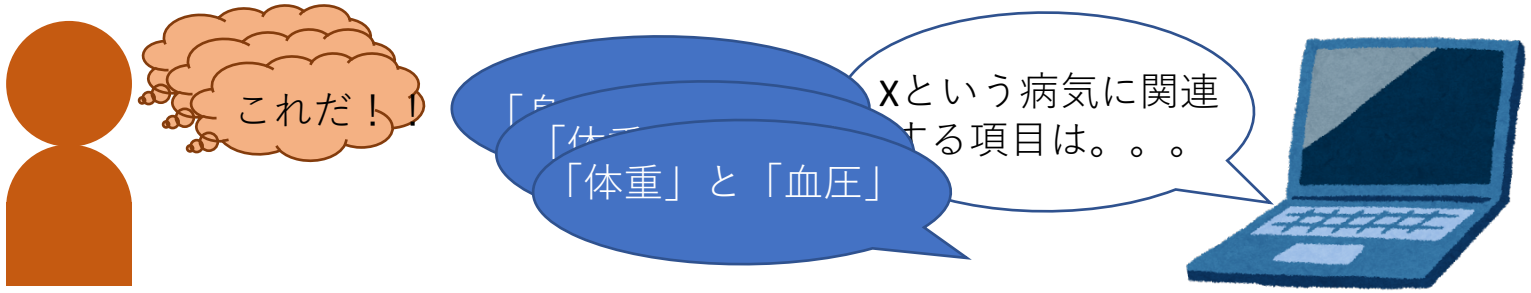
- 本研究のアイディア
  - そもそも“重要な特徴量の組”を一つ探そうとしているから難しい。
  - “重要な特徴量の組”をたくさん見つけて、それをユーザに提示したらどうか？
    → **「Lasso解を複数列挙する問題」**を考える。



研究背景：ユーザに信頼される特徴選択をしたい。

- しかし、“間違えない特徴選択”は難しい。
  - Lassoは選ばれた特徴量が“真に重要な特徴量”であることが保証されない。
    - Adaptive Lassoをはじめ、様々な改善法が考案されている。
    - しかし、有限のデータから学習している以上、エラーは避けられない。

- 本研究のアイディア
  - そもそも“重要な特徴量の組”を一つ探そうとしているから難しい。
  - “重要な特徴量の組”をたくさん見つけて、それをユーザに提示したらどうか？
    → **「Lasso解を複数列挙する問題」**を考える。



Lasso

Lassoによる特徴選択

スパース線形回帰問題

Given: 入出力のペア$(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}\ (i = 1, 2, \ldots, N)$

Find: 回帰係数$\beta \in \mathbb{R}^d$ s.t. $x_i^\top \beta \approx y\ (i = 1, 2, \ldots, N)$

ただし、βは非ゼロ要素が少ない（スパース）

- スパース性
  - 物理的要請
    - 大量の特徴量のうち、効く特徴量は少ないはずという直感。
  - 解釈性向上
    - 解から意味のある特徴量を見出したい。変数の絞込み。

解法：Lasso回帰（ℓ_1正則化）

$$\beta^* = \underset{\beta}{\operatorname{argmin}} \frac{1}{2} \|X\beta - y\|^2 + \rho\|\beta\|_1$$

- Lasso解β^*はスパース。$\operatorname{supp}(\beta^*) = \{i : \beta_i^* \neq 0\}$が重要な特徴量。

Lassoの理論的正当性：復元定理

Lasso解：$\beta^* = \underset{\beta}{\operatorname{argmin}} \frac{1}{2} \|X\beta - y\|^2 + \rho\|\beta\|_1$

適当な設定の元で、Lasso解β^*は真のパラメータβ^oのサポート（非零要素）を高い確率で復元する。

- 仮定
  - 真のモデルが$y = X\beta^o + w$; β^oはスパース, $w \sim N(0, \sigma^2)$
  - 正則化パラメータρは十分小さい。
  - Xの各行は十分独立（高相関な特徴量はない）。

- 上記仮定の元で、高い確率で真のパラメータβ^oのサポートが復元できる。

$$\operatorname{supp}(\beta^*) = \operatorname{supp}(\beta^o)$$

Lassoの限界：重要な特徴量を見落とす。

- 高次元データでは、類似した特徴量が存在することが多い。
  - 類似特徴量のうち、どれを使っても同等の予測精度のモデルが作れる。
  - 「Xの各行は十分独立（高相関な特徴量はない）」という仮定が成立しない場合に相当。

→このような場合、Lassoは類似特徴量の一部だけを使い残りを無視する。

- Lassoの限界：類似特徴量の中の重要な特徴量を見落とす。
  - 例えば、「身長」と「体重」のような相関の高い特徴量のうち、片方（例えば「身長」）だけを使ったモデルを出力する。

  →「体重」を見落とす。「体重」を使ったモデルを期待するユーザとの間に齟齬が起きる。

- 本研究：「身長」、「体重」それぞれを使ったモデルを両方出力する。

本研究の成果：Lasso解の列挙

成果1：アルゴリズム

Lassoの解を目的関数値の昇順にサポートを列挙するアルゴリズム。
列挙した解からユーザに気に入った解を選んでもらう。

成果2：列挙版の復元定理

正則化パラメータρが十分小さければ、適当な個数だけ解を列挙すれば真のパラメータβ^oのサポートを復元するものが見つかる。

- どれがサポート復元するかは特定不能（なんらかの別基準が必要）。
- 何個列挙すればいいかは問題依存（傾向は理論的にわかる）。

副次的な成果

Lassoで得られた特徴量を安易に信頼するのは危険。
実データで、実際に同等な解が無数に存在することを確認。

問題の定式化と提案法

問題の定式化：Lasso解の列挙

- 解のサポートを$S \subseteq \{x_1, x_2, \ldots, x_d\}$に制限したLasso：

$$\mathrm{Lasso}(S) = \min_{\beta} \frac{1}{2} \|X\beta - y\|^2 + \rho\|\beta\|_1 \ \text{ s.t. } \mathrm{supp}(\beta) \subseteq S$$

最適解での目的関数値を$\mathrm{Lasso}(S)$とする。

問題：Lasso解の列挙

$\mathrm{Lasso}(S)$の小さい順に極小のSをk個列挙する。
（極小：$\mathrm{supp}(\beta) = S$となるもの。それ以外は冗長。）

【注意】正則化パスに基づいた解の列挙ではない。

- 正則化パスでは疎な解から密な解へとρを変化させた時の解を列挙する。
- 本問題ではρ固定の元で、目的関数値が昇順になるように解のサポートを列挙する。
  - $\{x_1, x_2, x_4\}, \{x_1, x_2, x_3\}, \{x_1, x_3, x_5\}, \{x_1, x_2\}, \ldots$

アルゴリズム：『Lawlerのk-best列挙』

アルゴリズム概略

1. サポートSを入力して、特徴量の集合Tが出力されたとする。
2. 全ての$t \in T$について
   Sからをtを取り除いた$S' = S \setminus \{t\}$を作る 。
   $\mathrm{Lasso}(S')$の解T'を得る。
   (T', S')を解の候補としてヒープに保持する。
3. 保持している解の候補のうち、目的関数値が最小のものを出力する。
4. 以上、繰り返し。

アルゴリズム：『Lawlerのk-best列挙』

アルゴリズム概略

1. サポートSを入力して、特徴量の集合Tが出力されたとする。

2. 全ての$t \in T$について

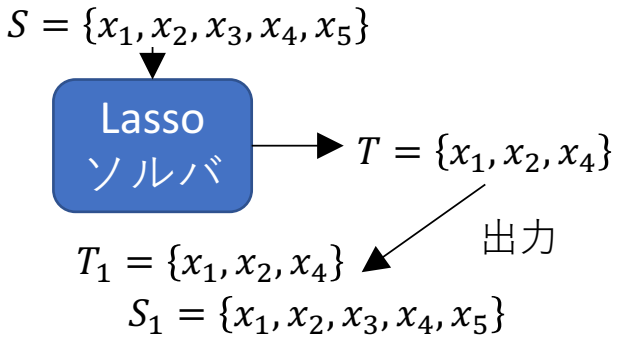
   Sからをtを取り除いた$S' = S \setminus \{t\}$を作る 。

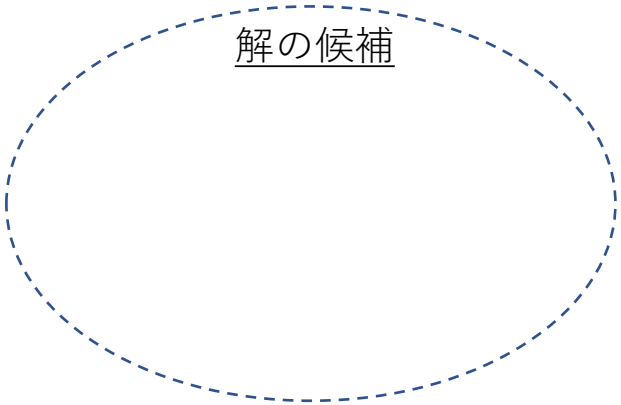
   $\mathrm{Lasso}(S')$の解T'を得る。

   (T', S')を解の候補としてヒープに保持する。

3. 保持している解の候補のうち、目的関数値が最小のものを出力する。

4. 以上、繰り返し。





アルゴリズム：『Lawlerのk-best列挙』

アルゴリズム概略

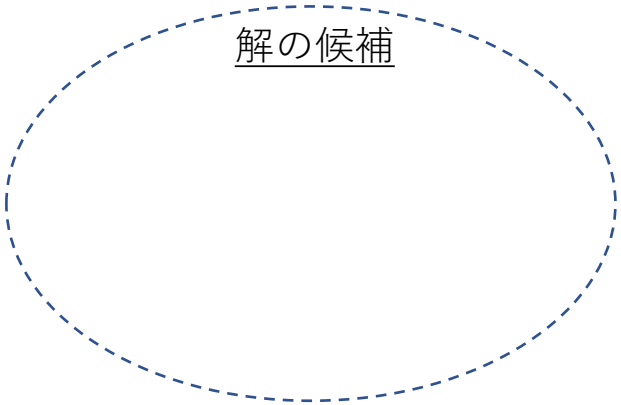
1. サポートSを入力して、特徴量の集合Tが出力されたとする。
2. **全ての$t \in T$について**
   Sからをtを取り除いた$S' = S \setminus \{t\}$を作る 。
   Lasso(S')の解T'を得る。
   (T', S')を解の候補としてヒープに保持する。
3. 保持している解の候補のうち、目的関数値が最小のものを出力する。
4. 以上、繰り返し。

Lasso
ソルバ

$T_1 = \{x_1, x_2, x_4\}$
$S_1 = \{x_1, x_2, x_3, x_4, x_5\}$

$S_1 = \{x_1, x_2, x_3, x_4, x_5\}$

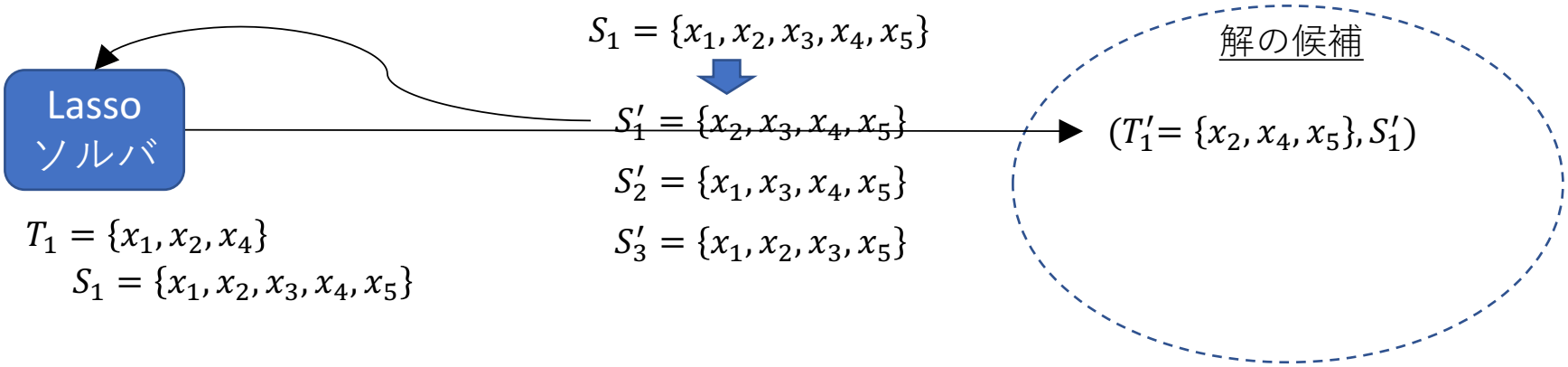
$S_1' = \{x_2, x_3, x_4, x_5\}$
$S_2' = \{x_1, x_3, x_4, x_5\}$
$S_3' = \{x_1, x_2, x_3, x_5\}$



アルゴリズム：『Lawlerのk-best列挙』

アルゴリズム概略

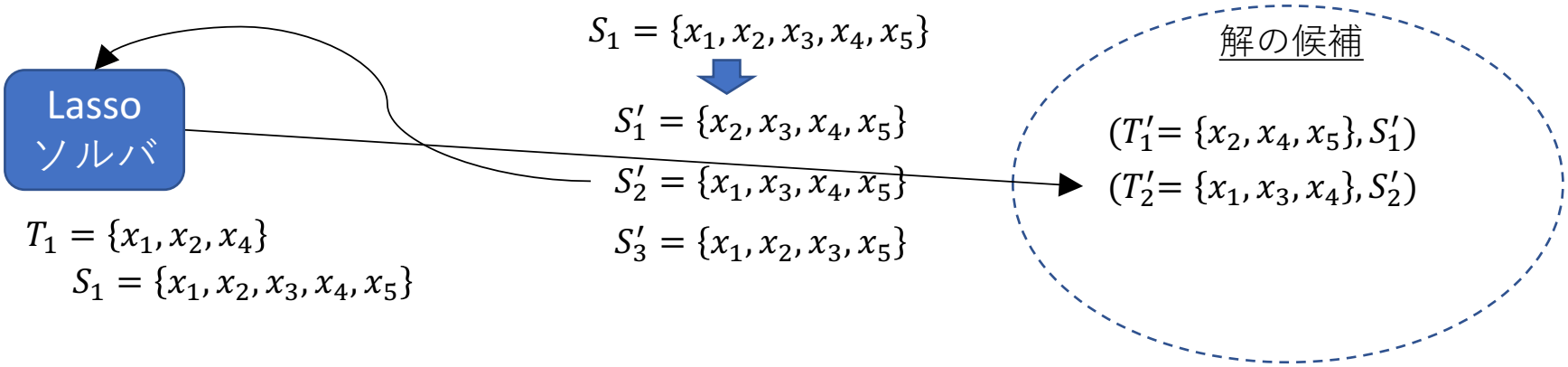
1. サポートSを入力して、特徴量の集合Tが出力されたとする。
2. **全ての$t \in T$について**
   Sからをtを取り除いた$S' = S \setminus \{t\}$を作る 。
   Lasso(S')の解T'を得る。
   (T', S')を解の候補としてヒープに保持する。
3. 保持している解の候補のうち、目的関数値が最小のものを出力する。
4. 以上、繰り返し。



アルゴリズム：『Lawlerのk-best列挙』

アルゴリズム概略

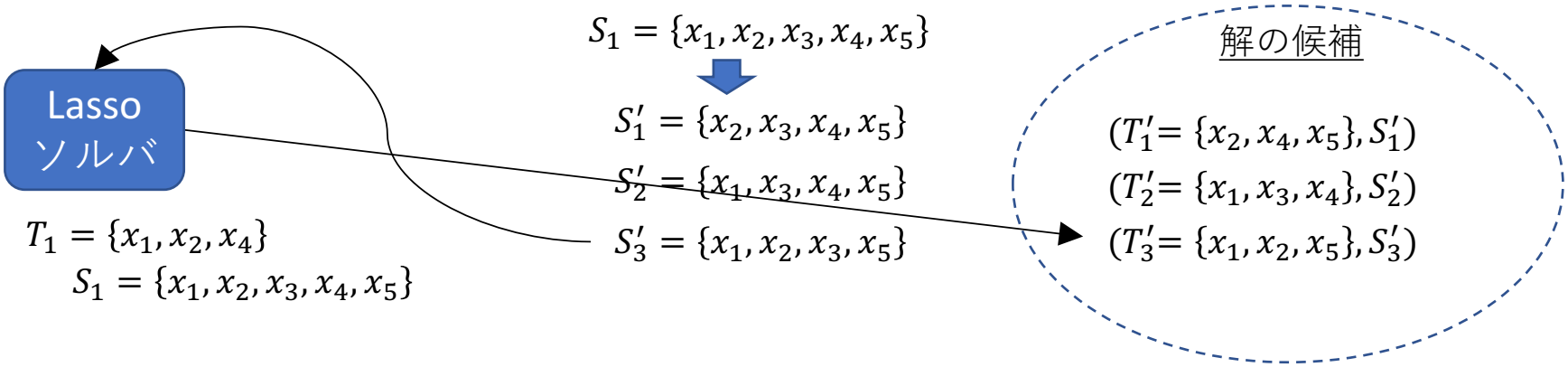
1. サポートSを入力して、特徴量の集合Tが出力されたとする。
2. **全ての$t \in T$について**
   Sからをtを取り除いた$S' = S \setminus \{t\}$を作る 。
   Lasso(S')の解T'を得る。
   (T', S')を解の候補としてヒープに保持する。
3. 保持している解の候補のうち、目的関数値が最小のものを出力する。
4. 以上、繰り返し。



アルゴリズム：『Lawlerのk-best列挙』

アルゴリズム概略

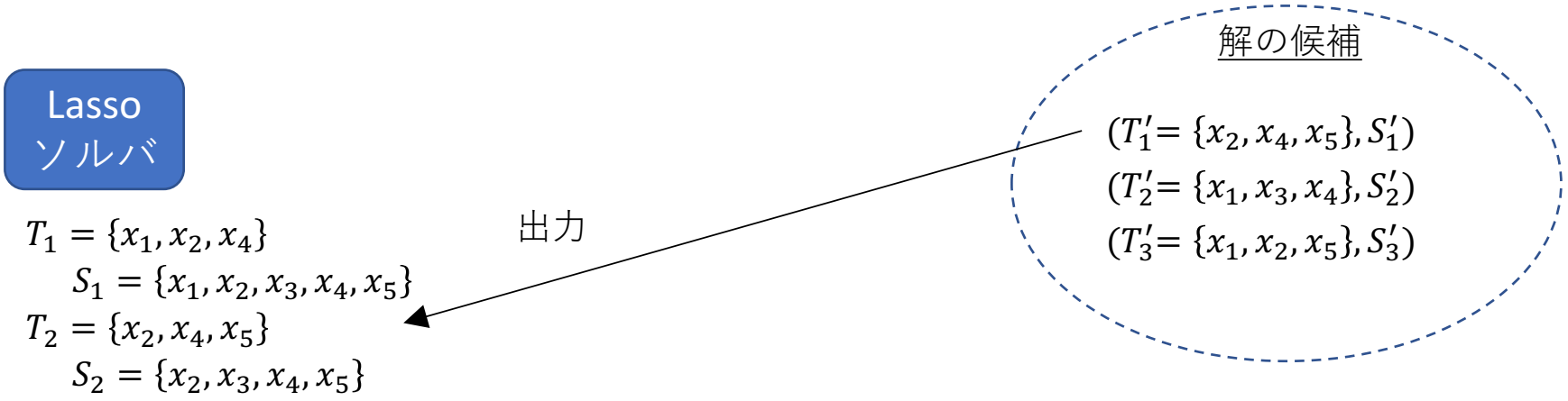
1. サポートSを入力して、特徴量の集合Tが出力されたとする。
2. **全ての$t \in T$について**
   Sからをtを取り除いた$S' = S \setminus \{t\}$を作る 。
   Lasso(S')の解T'を得る。
   (T', S')を解の候補としてヒープに保持する。
3. 保持している解の候補のうち、目的関数値が最小のものを出力する。
4. 以上、繰り返し。



アルゴリズム：『Lawlerのk-best列挙』

アルゴリズム概略

1. サポートSを入力して、特徴量の集合Tが出力されたとする。
2. **全ての$t \in T$について**
   Sからをtを取り除いた$S' = S \setminus \{t\}$を作る 。
   Lasso(S')の解T'を得る。
   (T', S')を解の候補としてヒープに保持する。
3. 保持している解の候補のうち、目的関数値が最小のものを出力する。
4. 以上、繰り返し。



アルゴリズム：『Lawlerのk-best列挙』

アルゴリズム概略

1. サポートSを入力して、特徴量の集合Tが出力されたとする。
2. **全ての$t \in T$について**
   Sからをtを取り除いた$S' = S \setminus \{t\}$を作る 。
   Lasso(S')の解T'を得る。
   (T', S')を解の候補としてヒープに保持する。
3. 保持している解の候補のうち、目的関数値が最小のものを出力する。
4. 以上、繰り返し。

Lasso
ソルバ

$T_1 = \{x_1, x_2, x_4\}$
$S_1 = \{x_1, x_2, x_3, x_4, x_5\}$
$T_2 = \{x_2, x_4, x_5\}$
$S_2 = \{x_2, x_3, x_4, x_5\}$

$S_2 = \{x_2, x_3, x_4, x_5\}$

$S_4' = \{x_3, x_4, x_5\}$
$S_5' = \{x_2, x_3, x_5\}$
$S_6' = \{x_2, x_3, x_4\}$

解の候補

$(T_2' = \{x_1, x_3, x_4\}, S_2')$
$(T_3' = \{x_1, x_2, x_5\}, S_3')$

アルゴリズムの妥当性

定理

提案法により$\mathrm{Lasso}(S)$の小さい順に極小のSを列挙できる。

- 不要な探索をスキップすることで、提案法を効率化できる。
  - 既に探索したSを重複して探索しないようにする。取り除く変数の履歴を保持する。
  - 探索したことのないSについても、Lassoの最適性条件から解が既に探索済みのものと一致することが判定できることがある。

列挙版の復元定理

定理概略

適当な仮定の元で、十分たくさん列挙すれば、高い確率で列挙した解の中に$\mathrm{supp}(\beta^o)$が含まれる。

- 適当な仮定・高い確率：
  - 正則化パラメータρは十分小さい(β^oの非ゼロ成分を下からバウンド)。
  - ノイズが小さいほど確率は高い。

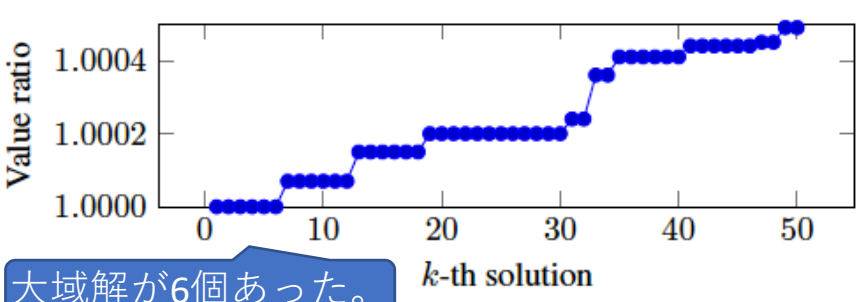
- 列挙個数について言えること：
  - 正則化パラメータρが小さいほど列挙すべき要素が増える。
  - Xの独立性が低い（高相関の特徴量が多い）ほど列挙すべき要素が増える。

実験結果

実験1. シロイズナの開花

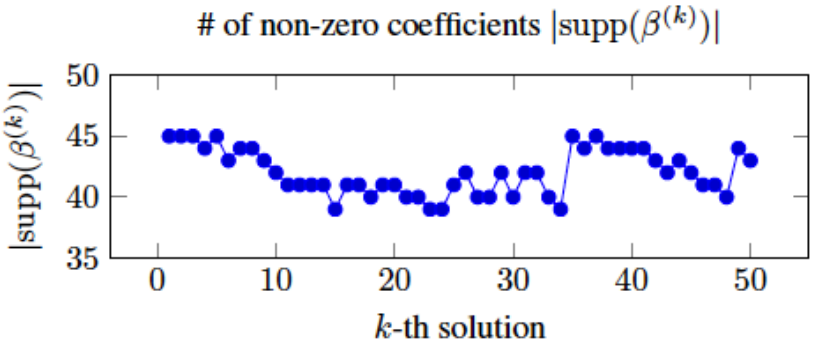
- Thaliana gene expression data (Atwell et al. ’10):
  どの遺伝子が開花に効くかを知りたい。
  - $x \in \mathbb{R}^{216130}$：遺伝子各パターンが生起しているかどうか（2 値）
  - $y \in \mathbb{R}$：発現量
  - データ数（個体数）：134

50個列挙しても、目的関数値は
0.05%しか増加しなかった。



大域解が6個あった。

解のサポートのサイズは
大体40~45くらい。



大域解が複数ある = 単純にLassoを適用すると、6個のうちの1つが見つかるだけ。他の特徴量は見落とす。

実験2. ニュース記事の分類

- 20 Newsgroups Data (Lang'95); ibm vs mac

  ニュース記事を二つのカテゴリに分類するのに特徴的な単語を知りたい。

  - $x \in \mathbb{R}^{11648}$：単語の発現（実数値、tf-idf）
  - $y \in \{\text{ibm}, \text{mac}\}$：記事のカテゴリ（2値）
  - データ数（投稿数）：1168

  → 分類問題なので、ロジスティック回帰に提案法を適用。

大域解にあった語　　列挙解で置き換わった語

Original words		Replaced	Subject
bios	→	drive	ibm
ide	→	drive	ibm
dos	→	os, drive	ibm
controller	→	drive	ibm
quadra, centris	→	040, clock	mac
windows, bios, controller	→	disk, drive	ibm
bios, help, controller	→	disk, drive	ibm
centris, pc	→	610	mac

drive, os, diskのようなibmマシン（Windows機）に特有の単語が見落とされていたのが見つかった。

040, 610のようなmacマシン（型番）に特有の単語が見落とされていたのが見つかった。

まとめ

- 問題意識：ユーザに信頼される特徴選択をしたい。
  - 単一の特徴選択結果を出力するのでなく、複数の結果を列挙して出力する。
- 「Lasso解のサポート列挙」として問題を定式化した。
- Lawlerのk-best フレームワークを適用した効率的なアルゴリズムを設計。
- 列挙版のスパース復元定理を証明した。
  - どれだけ列挙するかは問題パラメタ依存。
- 実験より，実際の特徴選択問題には「同じくらいの品質の解が大量に存在する」ことを確認。
  - Lasso で得られた特徴量を安易に信じるのは危険。

補足資料

列挙版の復元定理（完全版）

- 仮定：$y = X\beta^o + w$; β^oスパース, $w \sim N(0, \sigma\text{^}2)$
  - $\|X\beta^o - y\|_2 \leq \delta\|X\beta^* - y\|_2$ for some $\delta \geq 0$
  - $\|X\beta^* - y\|_2 \leq \epsilon$ for some $\epsilon \geq 0$
  - $\forall u \neq 0$ with $\|Xu\|_2 \leq (1+\delta)\epsilon$, $\|u_{S^o}\|_1 \leq \gamma\|u_{S^{oc}}\|_1$ for some $\gamma \geq \max\{1, \delta^2\}$

定理1：No False Inclusion

By enumerating solutions up to $L\left(\beta^{(\ell)}\right) \geq \gamma L(\beta^*)$, we can find $\left(\beta^{(k)}, S^{(k)}\right)$, $(1 \leq k \leq \ell)$ such that $\mathrm{supp}\left(\beta^{(k)}\right) \subseteq \mathrm{supp}(\beta^o) \subseteq S^{(k)}$ and $L\left(\beta^{(k)}\right) \leq L(\beta^o)$.

定理2：No False Exclusion

Let $\left(\beta^{(k)}, S^{(k)}\right)$ be an enumerated solution where $\mathrm{supp}\left(\beta^{(k)}\right) \subseteq \mathrm{supp}(\beta^o) \subseteq S^{(k)}$. If $X_{S^o}^\top X_{S^o}$ is invertible, then we have

$$\mathrm{supp}\left(\beta^{(k)}\right) \supseteq \left\{i : |\beta_i^o| > 2\rho \left\|\left(X_{S^o}^\top X_{S^o}\right)^{-1}\right\|_\infty\right\}$$

with probability $1 - |S^o|\exp\left(-\rho^2/2\sigma\sqrt{\lambda_{\max}\left(X_{S^o}^\top X_{S^o}\right)}\right)$.