大規模機械学習のための事例と特徴のセーフスクリーニング

竹内一郎

名古屋工業大学

本日の講演内容の一部は以下の方々との共同研究です

小川晃平，鈴木良規，中川和也，津田宏治，烏山昌幸，奥村翔太，柴垣篤志（敬称略）

スパースモデリング

- 線形モデル

$$f(\boldsymbol{x}) = w_1 x_1 + w_2 x_2 + \ldots + w_d x_d$$

- カーネルモデル

$$f(\boldsymbol{x}) = \alpha_1 K(\boldsymbol{x}, \boldsymbol{x}_1) + \alpha_2 K(\boldsymbol{x}, \boldsymbol{x}_2) + \ldots + \alpha_n K(\boldsymbol{x}, \boldsymbol{x}_n)$$

- スパースモデリング
  - $\{w_j\}_{j=1}^d$ の一部が零（一部の特徴のみで f を表現）
  - $\{\alpha_i\}_{i=1}^n$ の一部が零（一部の事例のみで f を表現）

スパースモデリングの例

- LASSO（L_1 正則化による特徴のスパース化）

$$\boldsymbol{w}^* := \arg\min_{\boldsymbol{w}\in\mathbb{R}^d} \lambda \underbrace{\|\boldsymbol{w}\|_1}_{L_1\ \text{正則化}} + \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i))^2$$

（$\lambda > 0$ は正則化パラメータ）

- SVM（ヒンジ損失関数による事例のスパース化）

$$\boldsymbol{\alpha}^* := \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \frac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{K}\boldsymbol{\alpha} + C\sum_{i=1}^{n} \underbrace{\max\{0, 1 - y_i f(\boldsymbol{x}_i)\}}_{\text{ヒンジ損失関数}}$$

（$C > 0$ は正則化パラメータ，$\boldsymbol{K}$ はカーネル行列）

スパース学習の計算コストとスクリーニング

- スパース学習の計算コスト
  - 学習後の最適解において，どの w_j やどの α_i が零となるかわからない
  - スパース学習の計算コストはオリジナルの次元数 d や事例数 n に依存する
- スクリーニングによる高速化
  - 学習前に $w_j = 0$ となる特徴や $\alpha_i = 0$ となる事例を推定/同定することを**スクリーニング**と呼ぶ
  - これらの特徴や事例を訓練集合から除去することにより効率的な学習が可能となる

ヒューリスティクスと安全スクリーニング

- ヒューリスティクス

  $w_j = 0$ となる特徴や $\alpha_i = 0$ となる事例を推測して取り除き，学習後に最適性を確認

  - Sure independence screening *(Fan et al., 2007)*
  - Shrinking option in `libsvm` *(Fan et al., 2005)*

- 安全スクリーニング（safe screening）

  $w_j = 0$ や $\alpha_i = 0$ となることが保障された特徴や事例を取り除く（最適性の確認は不要）

  - 安全特徴スクリーニング *(El Ghaoui et al., 2012)*
  - 安全事例スクリーニング *(Ogawa et al., 2013)*

本日の講演内容

- Part 1：SVM の安全事例スクリーニング

  Ogawa, Suzuki, and Takeuchi. Safe screening of non-support vectors in pathwise SVM computation. ICML2013.

- Part 2：高次交互作用モデルの安全特徴スクリーニング

  Nakagawa, Suzumura, Karasuyama, Tsuda, and Takeuchi. Safe feature pruning for sparse high-order interaction models. arXiv:1506.08002.

- Part 3：安全スクリーニング技術の応用
  - 感度分析

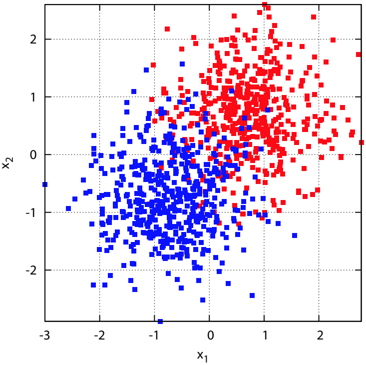
    Okumura, Suzuki, and Takeuchi. Quick sensitivity analysis for incremental data modification. KDD2015.
  - モデル選択

    Shibagaki, Suzuki, Karasuyama, and Takeuchi. Regularization Path of Cross-Validation Error Lower Bounds. NIPS2015.

Part 1

SVMの安全事例スクリーニング

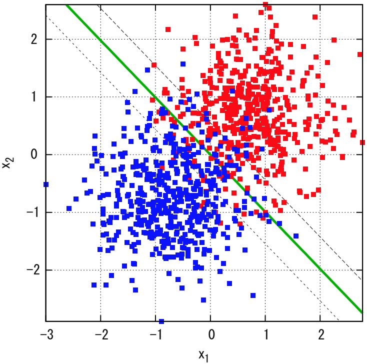
SVMの安全事例スクリーニングの例



Before safe screening

人工データ ($n = 1000$ and $d = 2$)

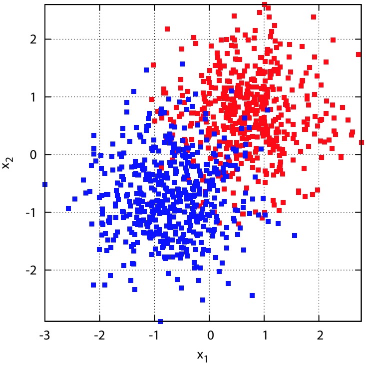
SVMの安全事例スクリーニングの例



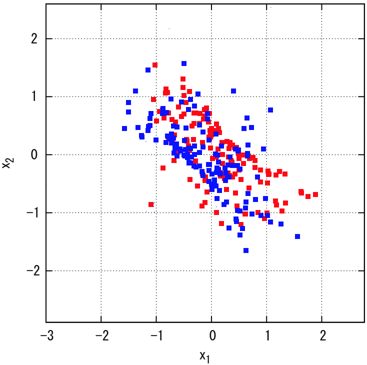
Before safe screening

人工データ ($n = 1000$ and $d = 2$)

SVMの安全事例スクリーニングの例



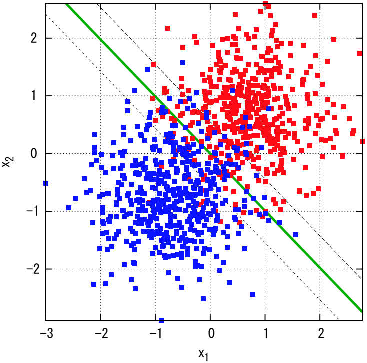
Before safe screening



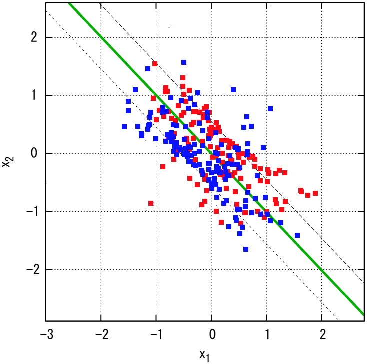
After safe screening

人工データ ($n = 1000$ and $d = 2$)

SVMの安全事例スクリーニングの例



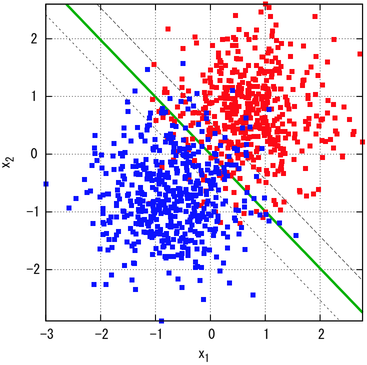
Before safe screening



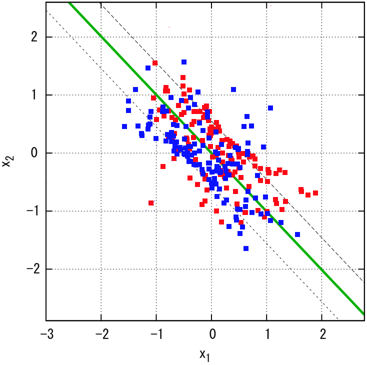
After safe screening

人工データ ($n = 1000$ and $d = 2$)

SVMの安全事例スクリーニングの例



Before safe screening



After safe screening

人工データ ($n = 1000$ and $d = 2$)

事例を削除しても解の最適性が保障される

サポートベクトルと非サポートベクトル

- SVM の分類規則

$$\hat{y} = \begin{cases} -1 & \text{if } f(\boldsymbol{x}) < 0, \\ +1 & \text{if } f(\boldsymbol{x}) \geq 0, \end{cases} \quad f(\boldsymbol{x}) = \underbrace{\boldsymbol{w}^{*\top}\boldsymbol{x}}_{\text{主形式}} = \underbrace{\sum_{i=1}^{n} \alpha_i^* y_i K(\boldsymbol{x}, \boldsymbol{x}_i)}_{\text{双対形式}}$$

ただし，$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ は訓練事例集合

- SVM は事例に関してスパース

$$\alpha_i^* = 0 \quad \Rightarrow \quad \text{分類関数 } f \text{ に影響を与えない} \quad \Rightarrow \quad \text{非 SV}$$

マージンと非サポートベクトル

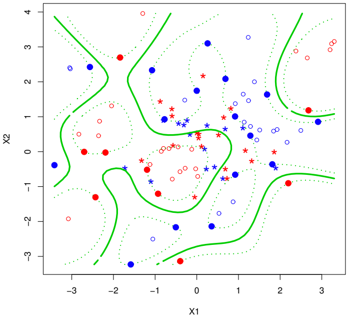
- サポートベクトル (SVs)

  $\star$ $\star$: $y_i f(\boldsymbol{x}_i) < 1 \;\Rightarrow\; \alpha_i = C$

  $\bullet$ $\bullet$: $y_i f(\boldsymbol{x}_i) = 1 \;\Rightarrow\; \alpha_i \in [0, C]$

- 非サポートベクトル (non-SVs)

  $\circ$ $\circ$: $\underbrace{y_i f(\boldsymbol{x}_i) > 1 \;\Rightarrow\; \alpha_i = 0}_{\text{事例スパース}}$



事前に最適解 $\boldsymbol{w}^*$ において，

$$y_i f(\boldsymbol{x}_i) = (y_i \boldsymbol{x}_i)^\top w^* > 1$$

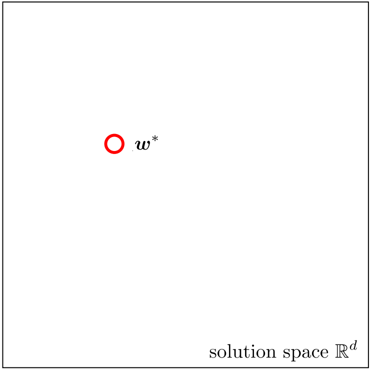
が成り立つことがわかれば，事例 $(\boldsymbol{x}_i, y_i)$ を捨ててもよい

安全事例スクリーニングの基本アイデア 1

- 主問題の解空間 $\mathbb{R}^d$ において，最適解 $\boldsymbol{w}^*$ は未知だが，最適解を含む球 $\mathcal{B}$

$$\boldsymbol{w}^* \in \mathcal{B} := \left\{ \boldsymbol{w} \;\middle|\; \|\boldsymbol{w} - \boldsymbol{m}\| \le r \right\},$$

が既知である状況を考える（$\boldsymbol{m}$：中心，r：半径）

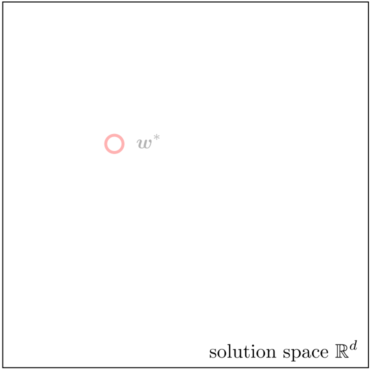


安全事例スクリーニングの基本アイデア 1

- 主問題の解空間 $\mathbb{R}^d$ において，最適解 $\boldsymbol{w}^*$ は未知だが，**最適解を含む球 $\mathcal{B}$**

$$\boldsymbol{w}^* \in \mathcal{B} := \left\{ \boldsymbol{w} \;\middle|\; \|\boldsymbol{w} - \boldsymbol{m}\| \le r \right\},$$

が既知である状況を考える（$\boldsymbol{m}$：中心，r：半径）

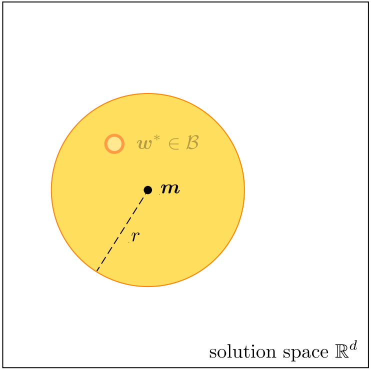


安全事例スクリーニングの基本アイデア 1

- 主問題の解空間 $\mathbb{R}^d$ において，最適解 $\boldsymbol{w}^*$ は未知だが，最適解を含む球 $\mathcal{B}$

$$\boldsymbol{w}^* \in \mathcal{B} := \left\{\boldsymbol{w} \;\middle|\; \|\boldsymbol{w} - \boldsymbol{m}\| \leq r\right\},$$

が既知である状況を考える（$\boldsymbol{m}$：中心，r：半径）

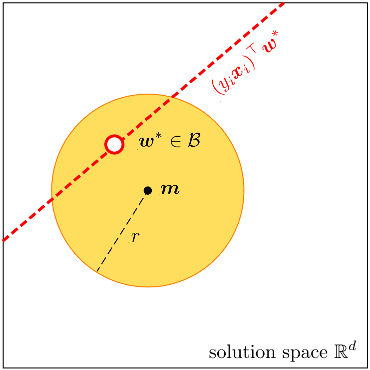


安全事例スクリーニングの基本アイデア２

- マージン $y_i f(\boldsymbol{x}_i) = (y_i \boldsymbol{x}_i)^\top \boldsymbol{w}^*$ の下限と上限：

下限： $(y_i \boldsymbol{x}_i)^\top \boldsymbol{w}^* \geq \min_{\boldsymbol{w} \in \mathcal{B}} (y_i \boldsymbol{x}_i)^\top \boldsymbol{w} = (y_i \boldsymbol{x}_i)^\top \boldsymbol{m} - \|y_i \boldsymbol{x}_i\| r,$

上限： $(y_i \boldsymbol{x}_i)^\top \boldsymbol{w}^* \leq \max_{\boldsymbol{w} \in \mathcal{B}} (y_i \boldsymbol{x}_i)^\top \boldsymbol{w} = (y_i \boldsymbol{x}_i)^\top \boldsymbol{m} + \|y_i \boldsymbol{x}_i\| r$

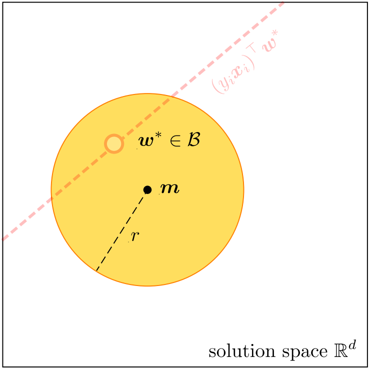


安全事例スクリーニングの基本アイデア２

- マージン $y_i f(\boldsymbol{x}_i) = (y_i \boldsymbol{x}_i)^\top \boldsymbol{w}^*$ の下限と上限：

下限： $(y_i \boldsymbol{x}_i)^\top \boldsymbol{w}^* \geq \min_{\boldsymbol{w} \in \mathcal{B}} (y_i \boldsymbol{x}_i)^\top \boldsymbol{w} = (y_i \boldsymbol{x}_i)^\top \boldsymbol{m} - \|\boldsymbol{y_i x_i}\| r,$

上限： $(y_i \boldsymbol{x}_i)^\top \boldsymbol{w}^* \leq \max_{\boldsymbol{w} \in \mathcal{B}} (y_i \boldsymbol{x}_i)^\top \boldsymbol{w} = (y_i \boldsymbol{x}_i)^\top \boldsymbol{m} + \|y_i \boldsymbol{x}_i\| r$

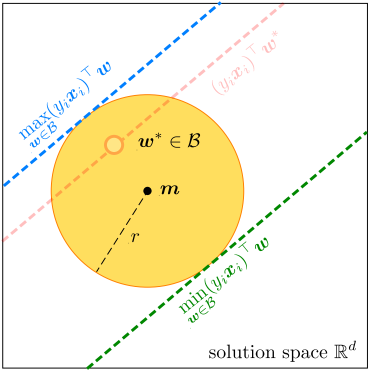


安全事例スクリーニングの基本アイデア２

- マージン $y_i f(\boldsymbol{x}_i) = (y_i \boldsymbol{x}_i)^\top \boldsymbol{w}^*$ の下限と上限：

下限： $(y_i \boldsymbol{x}_i)^\top \boldsymbol{w}^* \geq \min_{\boldsymbol{w} \in \mathcal{B}} (y_i \boldsymbol{x}_i)^\top \boldsymbol{w} = (y_i \boldsymbol{x}_i)^\top \boldsymbol{m} - \|\boldsymbol{y}_i \boldsymbol{x}_i\| r,$

上限： $(y_i \boldsymbol{x}_i)^\top \boldsymbol{w}^* \leq \max_{\boldsymbol{w} \in \mathcal{B}} (y_i \boldsymbol{x}_i)^\top \boldsymbol{w} = (y_i \boldsymbol{x}_i)^\top \boldsymbol{m} + \|y_i \boldsymbol{x}_i\| r$



安全事例スクリーニングの基本アイデア３

最適解 $\boldsymbol{w}^*$ が球 $\mathcal{B}$ に含まれている，すなわち，$\boldsymbol{w}^* \in \mathcal{B}$ ならば，

$$\underbrace{\min_{\boldsymbol{w}\in\mathcal{B}} (y_i \boldsymbol{x}_i)^\top \boldsymbol{w} > 1}_{\text{マージンの下限}>1} \;\Rightarrow\; \underbrace{(y_i \boldsymbol{x}_i)^\top \boldsymbol{w}^* > 1}_{\text{マージン}>1} \;\Rightarrow\; \underbrace{\alpha_i^* = 0}_{\text{非 SV}}$$

最適解 $\boldsymbol{w}^*$ が未知であっても，最適解を含む
球 $\mathcal{B}$ が既知であれば，非 SV を同定できる！

最適解を含む球（定理）

- 以下のクラスの最適化問題（SVM を含む）を考える：

$$\boldsymbol{w}^* := \arg\min_{\boldsymbol{w} \in \mathbb{R}^d} \; J(\boldsymbol{w}) := \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n} \ell_i(\boldsymbol{w}).$$

ただし，$\ell_i(\boldsymbol{w}) := \ell(y_i, \boldsymbol{x}_i^\top \boldsymbol{w})$ は凸な損失関数とする．

- 任意の $\in \mathbb{R}^d$ に対し，最適解 $\boldsymbol{w}^*$ は以下の球に含まれる：

$$\boldsymbol{w}^* \in \mathcal{B} := \left\{ \boldsymbol{w} \;\middle|\; \|\boldsymbol{w} - \boldsymbol{m}\| \le r \right\}.$$

ただし，球の中心と半径は，$\nabla\ell_i() \in \mathbb{R}^d$ を ℓ_i の における任意の劣勾配すると，以下のように与えられる：

$$\boldsymbol{m} := \frac{1}{2}\left(- C\sum_{i=1}^{n} \nabla\ell_i() \right), r := \frac{1}{2}\left\| + C\sum_{i=1}^{n} \nabla\ell_i() \right\|.$$

最適解を含む球（定理）

- 以下のクラスの最適化問題（SVM を含む）を考える：

$$\boldsymbol{w}^* := \arg\min_{\boldsymbol{w} \in \mathbb{R}^d} \; J(\boldsymbol{w}) := \frac{1}{2}\|\boldsymbol{w}\|^2 + C \sum_{i=1}^{n} \ell_i(\boldsymbol{w}).$$

ただし，$\ell_i(\boldsymbol{w}) := \ell(y_i, \boldsymbol{x}_i^\top \boldsymbol{w})$ は凸な損失関数とする．

- 任意の $\tilde{\boldsymbol{w}} \in \mathbb{R}^d$ に対し，最適解 $\boldsymbol{w}^*$ は以下の球に含まれる：

$$\boldsymbol{w}^* \in \mathcal{B} := \left\{ \boldsymbol{w} \;\middle|\; \|\boldsymbol{w} - \boldsymbol{m}\| \le r \right\}.$$

ただし，球の中心と半径は，$\boldsymbol{\nabla}\ell_i(\tilde{\boldsymbol{w}}) \in \mathbb{R}^d$ を ℓ_i の $\tilde{\boldsymbol{w}}$ における任意の劣勾配すると，以下のように与えられる：

$$\boldsymbol{m} := \frac{1}{2}\left(\tilde{\boldsymbol{w}} - C \sum_{i=1}^{n} \boldsymbol{\nabla}\ell_i(\tilde{\boldsymbol{w}}) \right), r := \frac{1}{2}\left\| \tilde{\boldsymbol{w}} + C \sum_{i=1}^{n} \boldsymbol{\nabla}\ell_i(\tilde{\boldsymbol{w}}) \right\|.$$

最適解を含む球（定理）

- 以下のクラスの最適化問題（SVM を含む）を考える：

$$w^* := \arg\min_{w \in \mathbb{R}^d} J(w) := \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \ell_i(w).$$

ただし，$\ell_i(w) := \ell(y_i, x_i^\top w)$ は凸な損失関数とする．

- 任意の $\tilde{w} \in \mathbb{R}^d$ に対し，最適解 w^* は以下の球に含まれる：

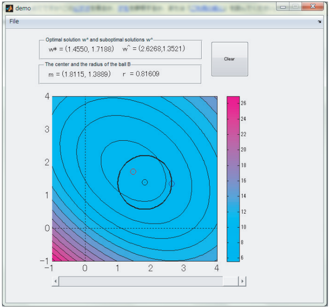
$$w^* \in \mathcal{B} := \left\{ w \mid \|w - m\| \le r \right\}.$$

ただし，球の中心と半径は，$\nabla \ell_i(\tilde{w}) \in \mathbb{R}^d$ を ℓ_i の $\tilde{w}$ における任意の劣勾配すると，以下のように与えられる：

$$m := \frac{1}{2}\left(\tilde{w} - C\sum_{i=1}^{n} \nabla \ell_i(\tilde{w})\right), r := \frac{1}{2}\left\|\tilde{w} + C\sum_{i=1}^{n} \nabla \ell_i(\tilde{w})\right\|.$$

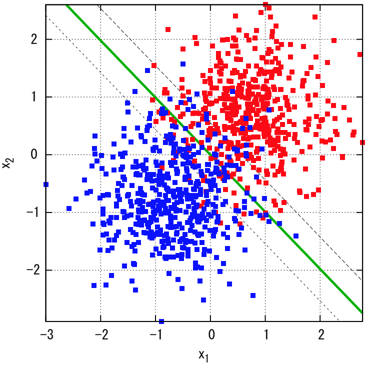
直感的には

- 最適解 w^* を含む球 $\mathcal{B}$ を構成するには任意の近似解 $\tilde{w} \in \mathbb{R}^d$ を利用する
- 近似解 $\tilde{w}$ が最適解 w^* に近いと半径 r が小さくなり，スクリーニングできる非 SV が増える

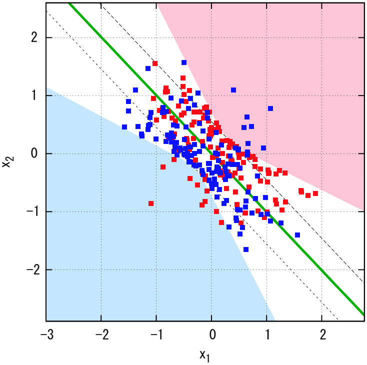


MATLAB demo

安全事例スクリーニングの例



Before safe screening



After safe screening

近似解 $\tilde{w}$ の求め方

- 正則化パラメータ C が小さいときの自明な解を近似解とする
- 異なる正則化パラメータ C における解を近似解とする
- サブサンプリングなどにより近似解を得る（Part 3 参照）

実験結果 1

Data	Sample Size n	LIBLINEAR	Sc.Rule	Sc.SVM	Sc.Total
acoustic	78,832	0.16	0.03	0.01	**0.038**
covtype	581,012	0.54	0.09	0.11	**0.199**
yahoo	1,036,492	5.55	1.13	1.39	**2.518**
url	2,396,130	10.39	1.71	6.92	**8.631**
kdd-a	8,407,752	18.20	2.37	3.37	**5.740**
kdd-b	19,264,097	37.54	4.53	2.26	**6.788**

最小の正則化パラメータ $C_0 = (||\boldsymbol{y}\boldsymbol{y}^\top \odot K||_\infty)^{-1}$ における自明な最適解を近似解として $C = C_0/0.8$ における線形 SVM の最適解を求める計算コスト（秒）

実験結果２（$C_1 < C_2 < \ldots$ の解を順次求める）

Data Set	Kernel (γ)	LIBSVM/LIBLINEAR	Sc.Rule	Sc.SVM	Sc.Total
dna	Linear	27.54	0.184	26.14	**26.32**
$n = 2,000$	RBF $(0.1/d)$	138.8	0.6979	104.4	**105.1**
$d = 180$	RBF $(1/d)$	6.13	0.6239	3.4	**4.024**
	RBF $(10/d)$	2.95	0.4729	2.43	**2.903**
DIGIT1	Linear	235.5	0.4799	231.6	**232.1**
$n = 1,500$	RBF $(0.1/d)$	801	1.302	731	**732.3**
$d = 241$	RBF $(1/d)$	72.84	1.281	63.82	**65.1**
	RBF $(10/d)$	5.57	1.096	3.08	**4.176**
satimage	Linear	115.2	0.3319	114	**114.3**
$n = 4,435$	RBF $(0.1/d)$	21345	6.465	20604	**20610**
$d = 36$	RBF $(1/d)$	448	7.966	322.3	**330.3**
	RBF $(10/d)$	32.06	8.181	5.74	**13.92**
gisette	Linear	994.1	32.34	937.4	**969.7**
$n = 6,000$	RBF $(0.1/d)$	418.9	15.19	389.5	**404.7**
$d = 5,000$	RBF $(1/d)$	55.74	14.26	37.93	**52.19**
	RBF $(10/d)$	**56.68**	10.32	54.44	64.76
mushrooms	Linear	5.22	0.5139	4.25	**4.765**
$n = 8,124$	RBF $(0.1/d)$	757.2	28.23	603.7	**631.9**
$d = 112$	RBF $(1/d)$	143.6	24.24	95.79	**120**
	RBF $(10/d)$	100.2	16.32	81.21	**97.53**
news20	Linear	**4403**	26.59	4504	4531
$n = 19,996$	RBF $(0.1/d)$	**22.15**	16.05	21.52	37.57
$d = 1,355,191$	RBF $(1/d)$	4664	83.55	3760	**3844**
	RBF $(10/d)$	59656	166.5	51310	**51477**
shuttle	Linear	81.53	1.607	73.31	**74.92**
$n = 43,500$	RBF $(0.1/d)$	58866	638.1	56118	**56756**
$d = 9$	RBF $(1/d)$	55717	692.7	49999	**50692**
	RBF $(10/d)$	51568	738.5	43053	**43792**

Part 2

スパース高次交互作用モデルの

安全特徴スクリーニング

L_1 正則化によるスパース線形モデル（*LASSO*）

- LASSO の主問題

$$\boldsymbol{w}^* := \arg\min_{\boldsymbol{w}\in\mathbb{R}^d} \ \lambda\|\boldsymbol{w}\|_1 + \sum_{i=1}^{n}(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2$$

- LASSO の双対問題

$$\boldsymbol{\gamma}^* := \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^n} \ \frac{1}{2}\left(\gamma_i - \frac{1}{\lambda}y_i\right)^2 \ \texttt{s.t.} \ \left|\sum_{i=1}^{n} x_{ij}\gamma_i\right| \leq 1, \forall j$$

- スパース性の条件

$$\left|\sum_{i=1}^{n} x_{ij}\gamma_i^*\right| < 1 \quad \Rightarrow \quad w_j^* = 0,$$

LASSOの安全特徴スクリーニング

- 双対最適解 γ^* を含む領域 $\mathcal{R}$
  - (El Ghaoui et al., 2012)
  - (Liu et al., 2014)
  - (Fercoq et al., 2015)
- 特徴スクリーニング

  $\gamma^* \in \mathcal{R}$ ならば,

$$\underbrace{\max_{\gamma \in \mathcal{R}} \left| \sum_{i=1}^{n} x_{ij} \gamma_i \right| < 1}_{\text{スコアの上限が 1 未満}} \Rightarrow \underbrace{\left| \sum_{i=1}^{n} x_{ij} \gamma_i^* \right| < 1}_{\text{スコアが 1 未満}} \Rightarrow \underbrace{w_j^* = 0}_{\text{スパース}}$$

双対最適解 γ^* が未知であっても，最適解を含む領域 $\mathcal{R}$ が既知であれば，係数が零の特徴を同定できる！

高次交互作用モデル

- オリジナルの特徴（特徴数 d 個）

$$\{(\boldsymbol{z}_i, y_i)\}_{i=1}^n, \boldsymbol{z}_i \in [0,1], y_i \in \mathbb{R}$$

- 高次交互作用モデル（特徴数 $D = \sum_{\rho=1}^r \binom{d}{\rho}$ 個）

$$\begin{aligned} f(\boldsymbol{z}_i) = {} & w_1 z_1 + w_2 z_2 + \ldots + w_d z_d \\ & + w_{1,2} z_1 z_2 + w_{1,3} z_1 z_3 + \ldots + w_{d-1,d} z_{d-1} z_d \\ & + w_{1,2,3} z_1 z_2 z_3 + w_{1,2,4} z_1 z_2 z_4 + \ldots + w_{d-2,d-1,d} z_{d-2} z_{d-1} z_d \end{aligned}$$

- 拡張入力行列 $\boldsymbol{X} \in \mathbb{R}^{n \times D}$ を考え，LASSO とスクリーニング

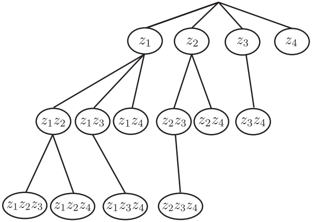
$$\underset{n \times D}{\boldsymbol{X}} := \begin{array}{c} \begin{array}{cccc} \text{(main effect)} & \text{(2}^{\text{nd}}\text{ order interactions)} & \cdots & (r^{\text{th}}\text{ order interactions)} \end{array} \\ \left[\begin{array}{ccc|ccc|c|ccc} z_1^1 & \ldots & z_d^1 & z_1^1 z_2^1 & \ldots & z_{d-1}^1 z_d^1 & \ldots & z_1^1 z_2^1 \cdots z_r^1 & \ldots & z_{d-r+1}^1 z_{d-r+2}^1 \cdots z_d^1 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ z_1^n & \ldots & z_d^n & z_1^n z_2^n & \ldots & z_{d-1}^n z_d^n & \ldots & z_1^n z_2^n \cdots z_r^n & \ldots & z_{d-r+1}^n z_{d-r+2}^n \cdots z_d^n \end{array}\right] \end{array}$$

計算コストとコスト削減のトリック

- 例えば，$d = 5000, r = 5$ の場合，$D > 10^{16}$
- すべての高次交互作用特徴のスコアの上限を計算できない

$$\max_{\gamma \in \mathcal{R}} \left| \sum_{i=1}^{n} \gamma_i x_{ij} \right|, \quad j = 1, \ldots, D.$$

- 高次交互作用項の木構造を利用

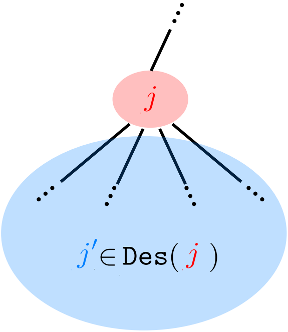


安全枝刈りルール（*Safe Pruning Rule*）

ノード（特徴）j のための安全枝刈りルール：$\mathtt{spr}(j)$

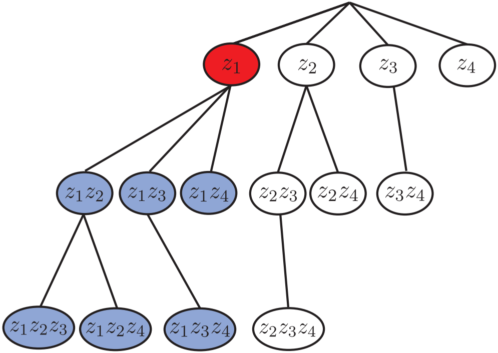
$$\mathtt{spr}(j) \text{ is true} \;\Rightarrow\; \left|\sum_{i=1}^{n} x_{ij'}\gamma_i^*\right| < 1 \;\Rightarrow\; w_{j'}^* = 0 \text{ for all } j' \in \mathtt{Des}(j),$$

ただし，$\mathtt{Des}(j)$ はノード j の子孫ノードの集合



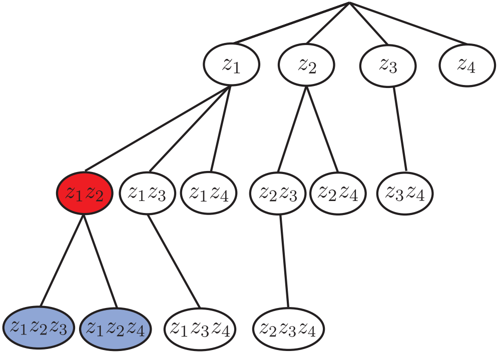
安全枝刈りルールの動作

$$\mathtt{spr}(z_1) = \mathsf{false}, \quad \mathcal{A} = \{z_1\}$$



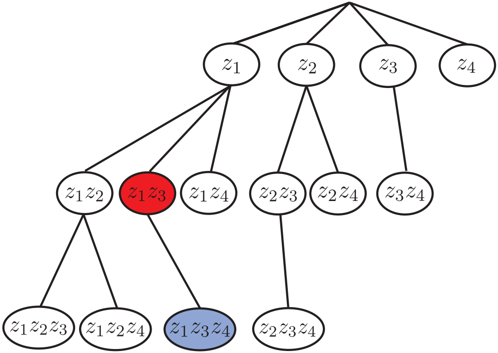
安全枝刈りルールの動作

$$\texttt{spr}(z_1 z_2) = \text{true}, \quad \mathcal{A} = \{z_1\}$$



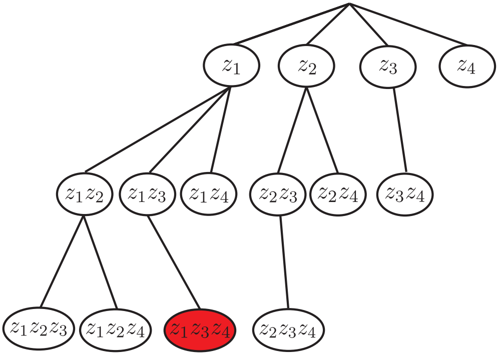
安全枝刈りルールの動作

$$\mathtt{spr}(z_1 z_3) = \mathsf{false}, \quad \mathcal{A} = \{z_1, z_1 z_3\}$$



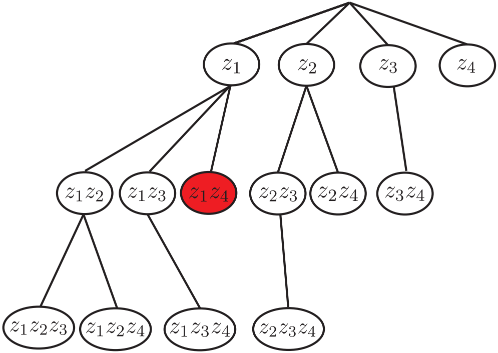
安全枝刈りルールの動作

$$\texttt{spr}(z_1 z_3 z_4) = \text{false}, \quad \mathcal{A} = \{z_1, z_1 z_3, z_1 z_3 z_4\}$$



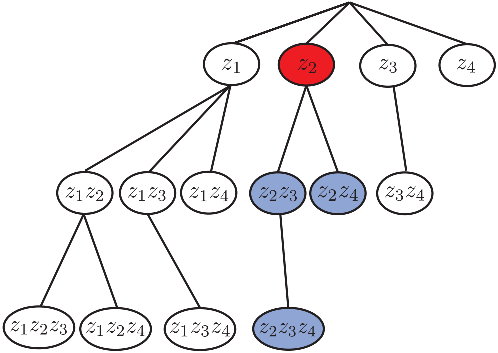
安全枝刈りルールの動作

$$\mathtt{spr}(z_1 z_4) = \mathtt{true}, \quad \mathcal{A} = \{z_1, z_1 z_3, z_1 z_3 z_4\}$$



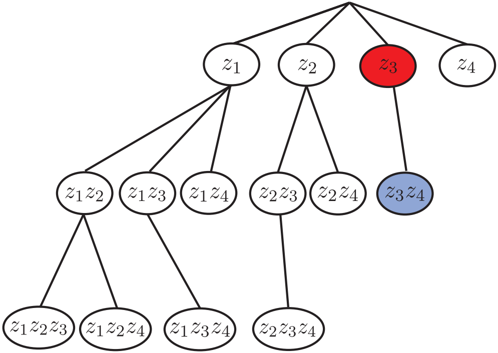
安全枝刈りルールの動作

$$\texttt{spr}(z_2) = \texttt{true}, \quad \mathcal{A} = \{z_1, z_1 z_3, z_1 z_3 z_4\}$$



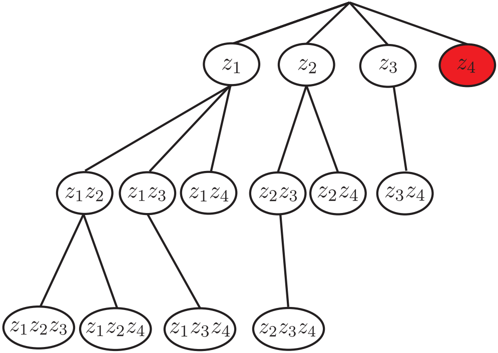
安全枝刈りルールの動作

$$\texttt{spr}({\color{red}z_3}) = \text{true}, \quad \mathcal{A} = \{z_1, z_1z_3, z_1z_3z_4\}$$



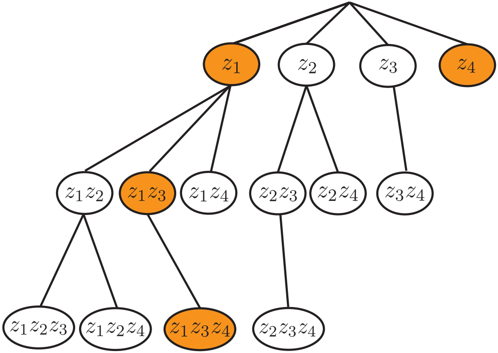
安全枝刈りルールの動作

$$\mathtt{spr}(z_4) = \mathsf{false}, \quad \mathcal{A} = \{z_1, z_1 z_3, z_1 z_3 z_4, z_4\}$$

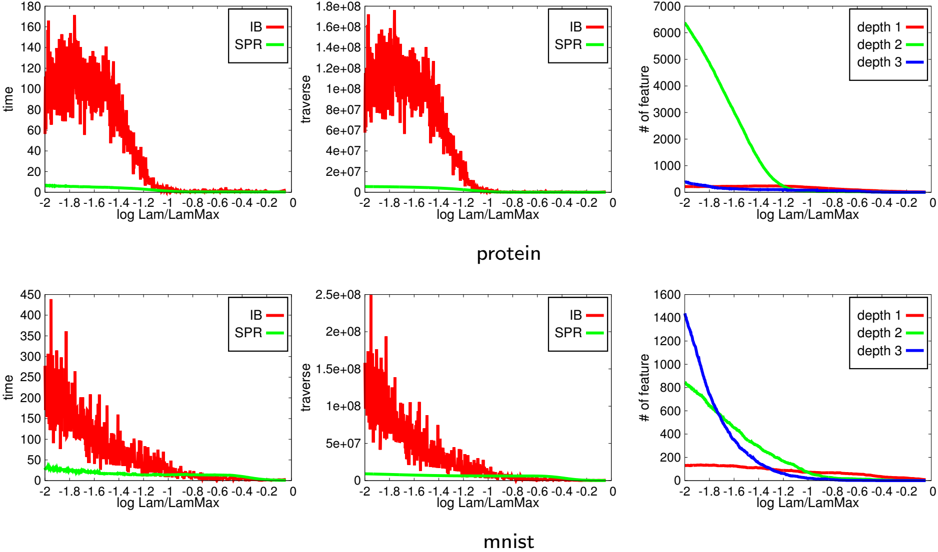


安全枝刈りルールの動作

$$\mathcal{A} = \{z_1, z_1 z_3, z_1 z_3 z_4, z_4\}$$



実験結果 （$\lambda_1 > \lambda_2 > \ldots$ の最適解の計算）



Itemset Boosting *(Saigo et al., 2006)* と比較

Part 3

安全スクリーニング技術の応用

（概要）

安全スクリーニング技術の応用

- 安全スクリーニング技術のキモ
  - よい近似解 $\tilde{\boldsymbol{w}}$ が得られているとき，
  - 最適解 $\boldsymbol{w}^*$ を含む領域 $\mathcal{B}$ がわかり，
  - 最適解の線形スコア $\boldsymbol{\theta}^\top \boldsymbol{w}^*$ の下限と上限を計算可能
- 線形スコアの例（２クラス分類問題）

$$\hat{y} = \mathrm{sign}(f(\boldsymbol{x})) = \mathrm{sign}(\boldsymbol{x}^\top \boldsymbol{w}^*) = \begin{cases} +1 & \text{if } \boldsymbol{x}^\top \boldsymbol{w}^* > 0, \\ -1 & \text{if } \boldsymbol{x}^\top \boldsymbol{w}^* < 0. \end{cases}$$

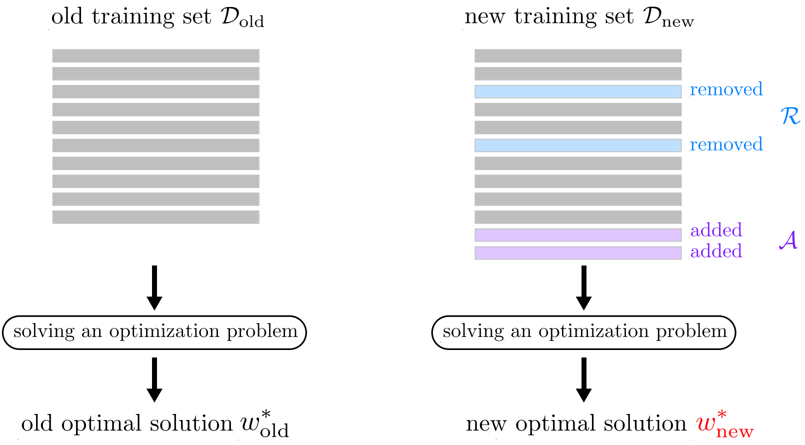
すなわち，

$$\min_{\boldsymbol{w} \in \mathcal{B}} \boldsymbol{x}^\top \boldsymbol{w} > 0 \;\Rightarrow\; \hat{y} = +1,$$

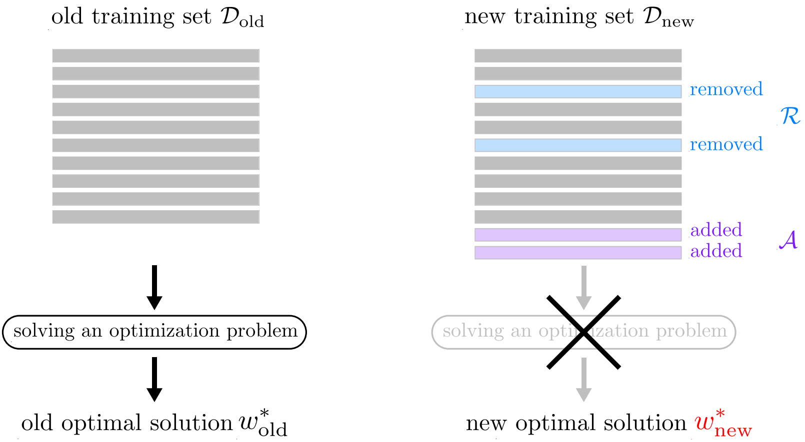
$$\max_{\boldsymbol{w} \in \mathcal{B}} \boldsymbol{x}^\top \boldsymbol{w} < 0 \;\Rightarrow\; \hat{y} = -1.$$

最適解を知らなくても２クラス分類ができる！

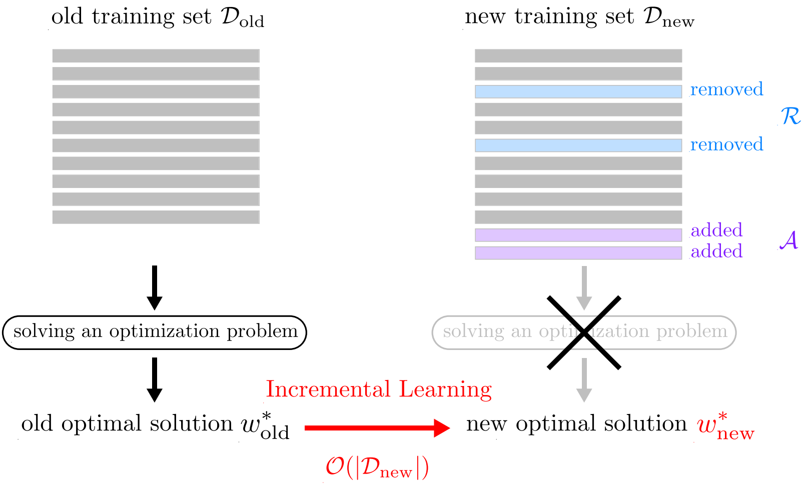
逐次学習のための感度分析 *(KDD2015)*



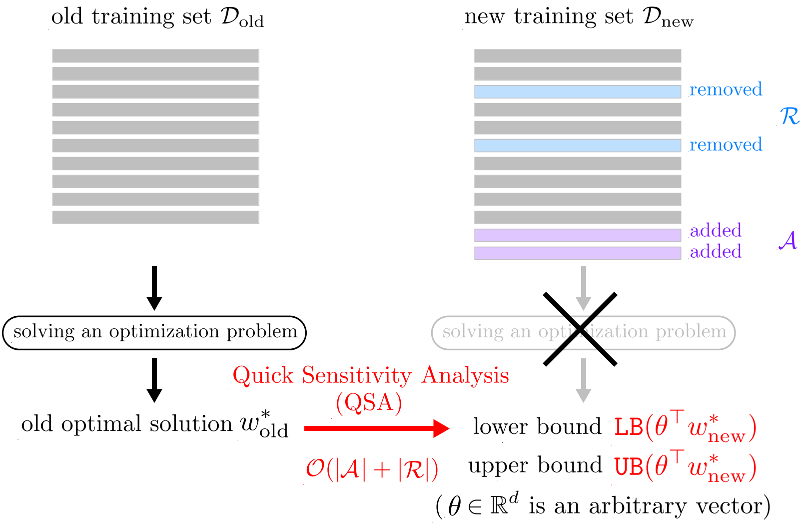
逐次学習のための感度分析 (KDD2015)



逐次学習のための感度分析 *(KDD2015)*



逐次学習のための感度分析 *(KDD2015)*



The computational cost depends only on $|\mathcal{A}| + |\mathcal{R}|$.

感度分析

- 方針

  更新前の最適解 $\boldsymbol{w}^*_{\rm old}$ を近似解 $\tilde{\boldsymbol{w}}$ とし，更新後の最適解 $\boldsymbol{w}^*_{\rm new}$ に関する感度分析（２クラス分類など）を行う

- 実験設定
  - kdd2010 ベンチマークデータ

    $$n_{\rm train} = |\mathcal{D}_{\rm old}| > 8 \text{ million} \quad \text{and} \quad n_{\rm test} > 0.5 \text{ million}$$

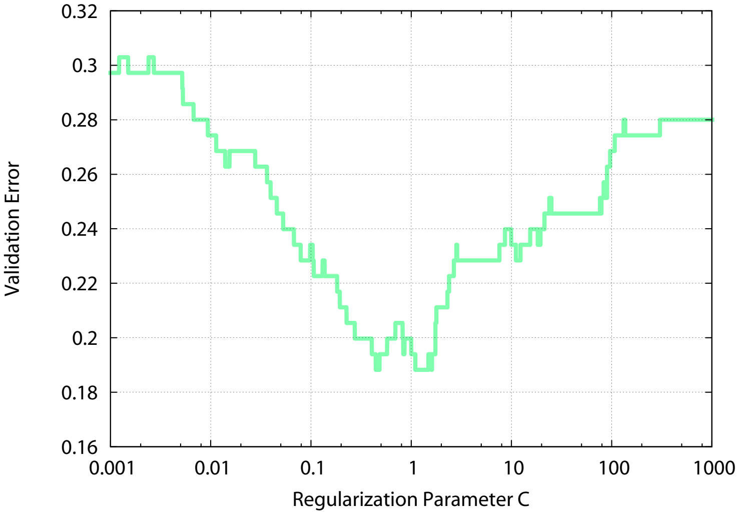
  - 訓練事例の 0.01, 0.1, 1%を更新

- 実験結果

% of updated instances	0.01%	0.1%	1%
% of label identification	99.987%	99.958%	99.867%
% of speed-ups	99.888%	99.887%	99.791%

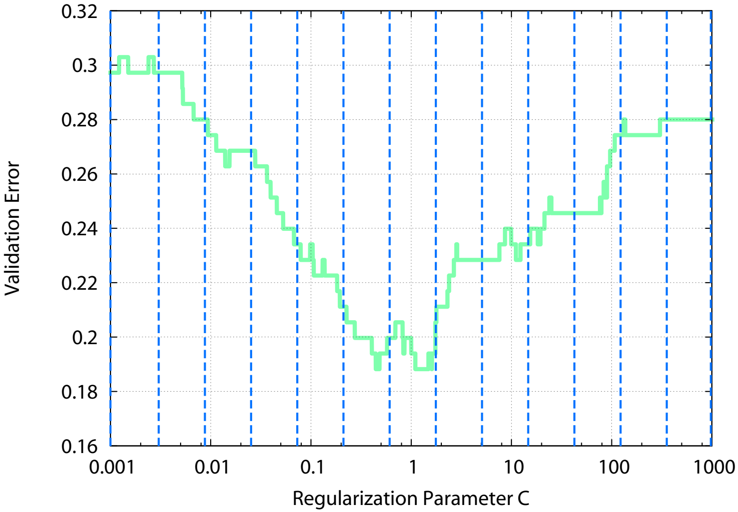
99%以上のテストラベルを 1/1000 の計算コストで確定

精度保障付クロスバリデーション *(NIPS2015)*



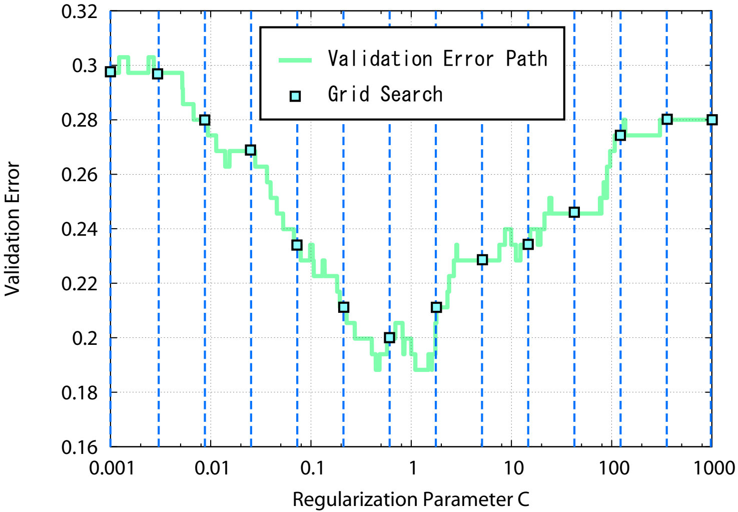
Model selection can be done with approximation guarantee.

精度保障付クロスバリデーション *(NIPS2015)*



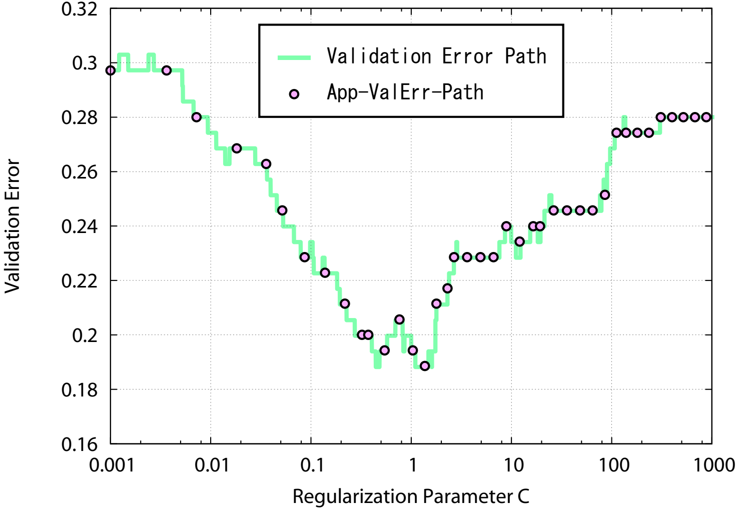
Model selection can be done with approximation guarantee.

精度保障付クロスバリデーション *(NIPS2015)*



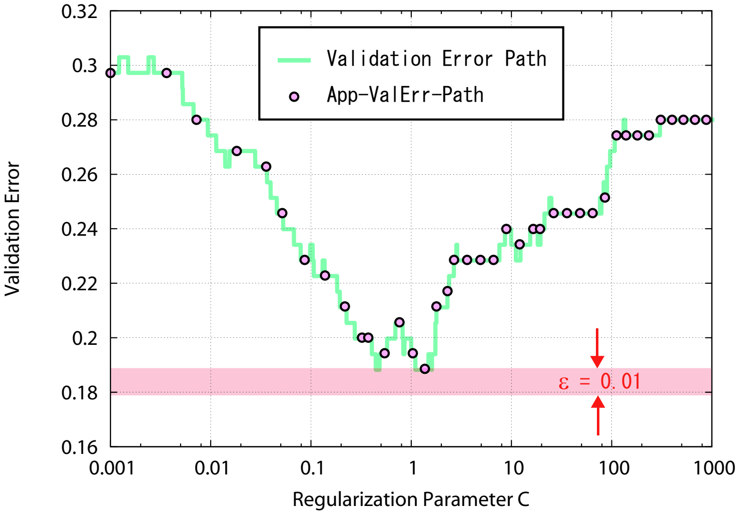
Model selection can be done with approximation guarantee.

精度保障付クロスバリデーション *(NIPS2015)*



Model selection can be done with approximation guarantee.

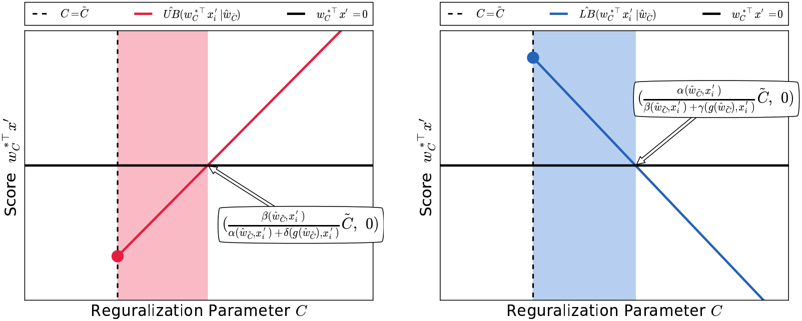
精度保障付クロスバリデーション *(NIPS2015)*



Model selection can be done with approximation guarantee.

なぜ保障できるか（その１）

- 安全スクリーニング技術を使うと，評価スコア $\boldsymbol{x}^\top \boldsymbol{w}_C^*$ の下限・上限を正則化パラメータ C の関数として表現可能
- 正則化パラメータが一定範囲内にあれば，$\boldsymbol{x}^\top \boldsymbol{w}_C^*$ の符号が不変であることを利用



なぜ保障できるか（その２）

A lower and an upper bounds of $w_C^{*\top} x'_i$ is written as

$$\begin{aligned}
\mathtt{LB}(w_C^{*\top} x'_i) &= \quad \alpha(\hat{w}_{\tilde{C}}, x'_i) - \frac{C}{\tilde{C}}(\beta(\hat{w}_{\tilde{C}}, x'_i) + \gamma(g(\hat{w}_{\tilde{C}}), x'_i)), \\
\mathtt{UB}(w_C^{*\top} x'_i) &= -\beta(\hat{w}_{\tilde{C}}, x'_i) + \frac{C}{\tilde{C}}(\alpha(\hat{w}_{\tilde{C}}, x'_i) + \delta(g(\hat{w}_{\tilde{C}}), x'_i)),
\end{aligned}$$

where, for $C \geq \tilde{C}$,

$$\begin{aligned}
\alpha(\hat{w}_{\tilde{C}}, x'_i) &:= \frac{1}{2}(\|\hat{w}_{\tilde{C}}\|\|x'_i\| + w_{\tilde{C}}^{*\top} x'_i) \geq 0, \\
\beta(\hat{w}_{\tilde{C}}, x'_i) &:= \frac{1}{2}(\|\hat{w}_{\tilde{C}}\|\|x'_i\| - w_{\tilde{C}}^{*\top} x'_i) \geq 0,
\end{aligned}$$

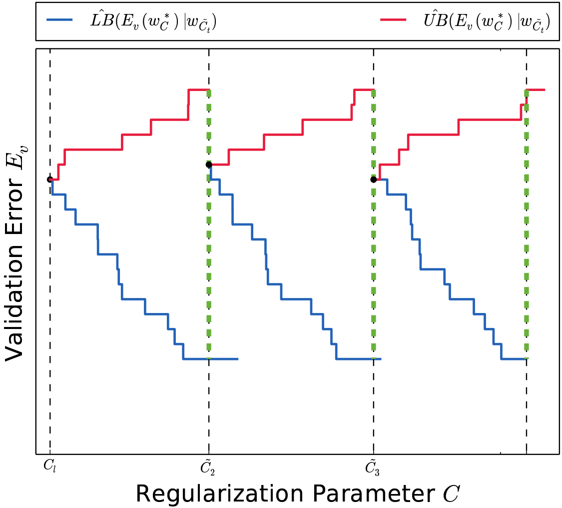
and

$$\begin{aligned}
\gamma(g(\hat{w}_{\tilde{C}}), x'_i) &:= \frac{1}{2}(\|g(\hat{w}_{\tilde{C}}), x'_i)\|\|x'_i\| + g(\hat{w}_{\tilde{C}}^{\top})x'_i) \geq 0, \\
\delta(g(\hat{w}_{\tilde{C}}), x'_i) &:= \frac{1}{2}(\|g(\hat{w}_{\tilde{C}}), x'_i)\|\|x'_i\| - g(\hat{w}_{\tilde{C}}^{\top})x'_i) \geq 0,
\end{aligned}$$

where $g(\hat{w}_{\tilde{C}}), x'_i)$ is the gradient vector of the objective function at $w = \hat{w}_{\tilde{C}}$.

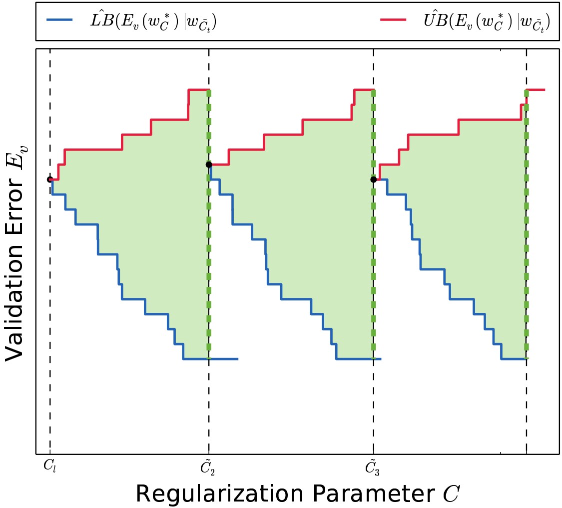
なぜ保障できるか（その３）

- 評価誤差の変化の下限・上限を正則化パラメータ C の関数として計算可能



なぜ保障できるか（その３）

- 評価誤差の変化の下限・上限を正則化パラメータ C の関数として計算可能



まとめと今後の課題

- 大規模データのスパース学習に安全スクリーニングは有効
  - 安全事例スクリーニング
  - 高次交互作用モデルの安全特徴スクリーニング
  - 特徴と事例の同時スクリーニング
- 十分によい近似解 $\Rightarrow$ 最適解の線形スコアの下限・上限
  - 感度分析
  - モデル選択
  - 差分プライバシ

参考文献

- L. El Ghaoui, V. Viallon and T. Rabbani. Safe feature elimination in sparse supervised learning. Pacific Journal of Optimization, 2012.
- J. Liu, Z. Zhao, J. Wang and J. Ye. Safe Screening with Variational Inequalities and Its Application to Lasso. ICML2014.
- J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. Journal of The Royal Statistical Society B, 70:849911, 2008.
- R. Fan, K. Chang, C. Hsieh, X. Wang and C. Lin. LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research, vol. 9, pp. 18711874, 2008.
- H. Saigo, T. Uno and K. Tsuda. Mining complex genotypic features for predicting hiv-1 drug resistance. Bioinformatics, 24:24552462, 2006.
- K. Ogawa, Y. Suzuki and I. Takeuchi. Safe screening of non-support vectors in pathwise SVM computation. ICML2013.
- K. Nakagawa, S. Suzumura, M. Karasuyama, K. Tsuda and I. Takeuchi. Safe feature pruning for sparse high-order interaction models. arXiv:1506.08002, 2015.
- S. Okumura, Y. Suzuki and I. Takeuchi. Quick sensitivity analysis for incremental data modification and its application to leave-one-out CV in linear classification problems. KDD2015.
- A. Shibagaki, Y. Suzuki, M. Karasuyama and I. Takeuchi. Regularization path of cross-Validation error lower bounds. NIPS2015.

ご静聴ありがとうございました