ブースティングに関する最近の進展について

大阪市立大学数学研究所
「情報幾何学研究集会 2009」

川喜田 雅則

九州大学
システム情報科学研究院
kawakita@csce.kyushu-u.ac.jp

Table of Contents

Table of Contents

1.1 統計的判別問題

Setting

- 特徴空間を $\mathcal{X} \in R^M$, ラベル集合を $\mathcal{Y} = \{1, -1\}$ と定義.
- 特徴量 X とラベル Y の結合密度を $p(x, y) = p(x)p(y \mid x)$ と書く.
- 今 n 個のデータ $D = \{X_i, Y_i\}_{i=1}^n$ が得られた状況を考える.

統計的判別問題

- 目的は特徴量が与えられたとき正確にラベルを推測する判別機 $g(x) : \mathcal{X} \to \mathcal{Y}$ を構成すること.
- 判別機 g の性能は**リスク** $L(g) = P(g(X) \neq Y)$ で評価.
- ベイズ判別機 $g^*(x) := \mathbf{argmax}_{y \in \mathcal{Y}}\, p(y \mid x)$ は $L(g)$ を最小化する. $L^* := L(g^*)$ をベイズリスクと呼ぶ.
- 判別法は L^* に近いリスクを持つ g の構成を目指す

1.2 ブースティングアルゴリズム

ブースティング　ブースティングとは多数の弱学習機を組み合わせて一つの高精度の弱学習機を構築する (ブーストする)

- 任意の弱学習機の集合を $C = \{f_j : \mathcal{X} \to \mathcal{Y} \mid j = 1, 2, \cdots J\}$ とする. 理論的には「任意のデータセットに対して C の中に $50(\%)$ より少しでも小さい誤判別率を持つものが見つかる集合」であれば良い.
- ブースティングは逐次的に弱学習機を線形に結合していく

$$F_T(x) = \sum_{j=1}^{J} \alpha_j f_j(x),$$

ただし $f_j(x) \in C$. F_T を **判別関数** と呼ぶ

- 最終的な判別機は $g(x) := \mathbf{sign}(F(x))$ と構成できる.

1.2 Adaboost

アルゴリズム　弱学習機 C とデータ $D = \{(X_i, Y_i)\}_{i=1}^n$ を想定

1. 時刻 0 の重みを $w_i(1) \equiv 1/n$ $(i = 1, 2, \cdots, n)$ と初期化.
2. 各時刻 $t = 1, 2, \cdots T$ において以下の処理を実行
   (a) D から確率 $w_i(t)$ によるリサンプリングで $D(t)$ を構成
   (b) C の中でデータセット $D(t)$ 上でのリスク, すなわち

$$\epsilon_t(f_j) := \sum_{i=1}^{n} w_i(t) I(f_j(X_i) \neq Y_i)$$

   を最小にする f_j を選択　　$f_{j(t)} = \underset{f \in C}{\mathbf{argmin}}\, \epsilon_t(f).$

   (c) 弱学習機の信頼度 α_t を計算:　　$\alpha_t = \frac{1}{2} \ln((1 - \epsilon_t)/\epsilon_t)$

   (d) 判別関数 F_t を更新　　$F_{t+1}(x) = F_t(x) + \alpha_t f_{j(t)}(x)$

   (e) 重み $\{w_i(t)\}$ を更新:

$$w_i(t+1) = \frac{w_i(t) \exp(-\alpha_t y_i f_{j(t)}(\boldsymbol{x}_i))}{Z_t} \qquad Z_t := \sum_{i=1}^{N} w_i(t) \exp(-\alpha_t y_i f_{j(t)}(\boldsymbol{x}_i)).$$

(f) もしアルゴリズムの停止条件を考慮している場合は条件をチェックする.

3. 最終的に判別機 $g(x)$ を

$$g(x) := \mathbf{sign}(F(x))$$

と得る.

1.2 ブースティングの実行例

決定スタンプ　深さ 1 の決定木

$$f(x; m, b, s) := s \cdot \mathbf{sign}(x_m - b).$$

- 一次元なら決定スタンプの線形和は関数空間でちゅう密
- 過学習しにくい
- GAM と同様に結果の解釈がしやすい
- 計算量が少ない

決定スタンプの用意の仕方

- 全ての共変量について用意 $(j = 1, 2, \cdots d)$
- b はデータの全ての中間点をとるように選ぶ

では デモンストレーションを御覧下さい.

1.3 AdaBoostの性質

Adaboostは以下のような性質を持つ

最悪リスク $\epsilon_{t+1}(f_{j(t)}) = \frac{1}{2}$ **が常に成り立つ.**

過学習しにくい 停止条件や弱学習機の選択が適切なら

解のスパース性 いらない多くの弱学習機の係数は 0 **のまま.**
(L_1 **ペナルティとの関係)**

解釈性 単一特徴量弱学習機を用いればGAMと対応がつく

お手軽 誤解を恐れずに言えばカーネルマシンは玄人用, ブースティングは素人用

学習の単調性 訓練誤差の上界はステップ毎に単調減少する.

1.4 歴史と文献紹介 (1/3)

動機 “弱学習可能性は強学習可能性と同値か?” questioned by Kearns and Valiant (1988).

AdaBoost(Adaptive Boosting) Freund and Schapire (1997) により提案されて以来, 盛んに研究された.

統計的解釈 Breiman (1998), Mason et al. (2000), Friedman et al. (2000), Collins et al. (2002)

マージン最大化による解釈 Schapire et al. (1998),Mason et al. (2000)

ベイズリスク一致性関連釈 Schapire et al. (1998), Mason et al. (2000), Schapire et al. (1998) はマージン最大化により精度を評価したが Breiman (1999) はその欠点を指摘. Koltchinskii and Panchenko (2002), Lugosi and Vayatis (2004) はブースティングの修正版について証明 Bartlett and Traskin (2007) が初めて完全な形で証明.

1.4 歴史と文献紹介 (2/3)

幾何学的解釈 Lebanon and Lafferty (2002) はブースティングが拡張 KL-divergence による弱学習機から連想される指数型分布への逐次射影アルゴリズムであることを示した. Collins et al. (2000) は Bregman-divergence への拡張を提案し Murata et al. (2004) は U-divergence に基づく U-Boost を提案した.

停止条件 crossvaildation or Zhang and Yu (2005)

高次元データ Bühlmann (2006) は L_2 ロスを用いたブースティングが高次元データでうまく働く事を示した.

弱学習機の選択 Zhang (2004) は決定スタンプのように一共変量に基づく弱学習機の使用が必ずしも悪くない根拠を示した.

1.4 歴史と文献紹介 (3/3)

回帰や密度推定への拡張 Freund and Schapire (1997) Rosset and Segal (2002) しかし Zhang (2004) の結果を考慮すると判別以外の目的に使用する利点があるかは注意が必要

1 クラス問題への応用 Rätsch et al. (2002)

正則化 Rätsch et al. (1999) など多数

局所化 Kotsiantis et al. (2006), Kawakita and Eguchi (2008)

カーネルマシンとの関連 Rätsch et al. (2002)Kawakita et al. (2006)

Table of Contents

2.1 統計的解釈 (Friedman et al., 2000)

指数ロス $L(F) = \frac{1}{n}\sum_{i=1}^{n} \exp(-y_i F(\boldsymbol{x}_i))$

アルゴリズム **AdaBoost は指数ロスにより簡潔に記述できる**

1. $F_0(\boldsymbol{x}) \equiv 0$
2. **時刻** t **において結合された判別関数を** F_{t-1} **とする**

$$
\begin{aligned}
f_{j(t)} &= \underset{f \in C}{\mathbf{argmin}}\, L(F_{t-1} + \alpha f) \\
\alpha_t &= \underset{\alpha > 0}{\mathbf{argmin}}\, L(F_{t-1} + \alpha f_{j(t)})
\end{aligned}
$$

3. **学習機を以下のように更新し、**2 **へ戻る**

$$F_t = F_{t-1} + \alpha_t h_t$$

準備

ロジスティック回帰 ラベルを $Y' := (Y+1)/2$ と再定義

ロジスティックモデル $$p(y \mid x;\alpha) = \frac{\exp(yF(x;\alpha))}{\sum_{y' \in \{0,1\}} \exp(y'F(x;\alpha))}$$

log-オッズ $$\ln \frac{P(+1 \mid x;\alpha)}{P(-1 \mid x;\alpha)} = F(x;\alpha)$$

判別関数 $$F(x;\alpha) := \sum_{m=1}^{M} \alpha_m x_m$$

スコア関数 特徴量毎の予測への寄与の様子を観察

仮定: 各弱学習機 f_j が特徴量 x_{m_j} にのみ依存 $f_j(x) = f_j(x_{m_j})$.

$$F_T(x) = \sum_{t=1}^{T} \alpha_t f_{j(t)}(x) = \sum_{m=1}^{M} S_m(x) \qquad S_m(x) := \sum_{t=1}^{T} \alpha_t f_{j(t)}(x) I(m_j = m)$$

準備

一般化加法モデル ラベルを $Y' := (Y+1)/2$ と再定義

ロジスティックモデル $$p(y \mid x;\alpha) = \frac{\exp(yF(x;\alpha))}{\sum_{y' \in \{0,1\}} \exp(y'F(x;\alpha))}$$

log-オッズ $$\ln \frac{P(+1 \mid x;\alpha)}{P(-1 \mid x;\alpha)} = F(x;\alpha)$$

判別関数 $$F(x;\alpha) := \sum_{m=1}^{M} s_m(x_m).$$

スコア関数 特徴量毎の予測への寄与の様子を観察

仮定: 各弱学習機 f_j が特徴量 x_{m_j} にのみ依存 $f_j(x) = f_j(x_{m_j})$.

$$F_T(x) = \sum_{t=1}^{T} \alpha_t f_{j(t)}(x) = \sum_{m=1}^{M} S_m(x) \qquad S_m(x) := \sum_{t=1}^{T} \alpha_t f_{j(t)}(x) I(m_j = m)$$

2.1 GAM との対応づけ

ロジスティック回帰

$$F_L(\boldsymbol{x}) = \log \frac{P(+1|x;\alpha)}{P(-1|x;\alpha)} = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_m x_m$$

AdaBoost

2.1 GAM との対応づけ

ロジスティック GAM

$$F_L(\boldsymbol{x}) = \log \frac{P(+1|x;\alpha)}{P(-1|x;\alpha)} = s_1(x_1) + s_2(x_2) + \cdots + s_m(x_m)$$

AdaBoost

2.1 GAM との対応づけ

ロジスティック GAM

$$F_L(\boldsymbol{x}) = \log \frac{P(+1|x;\alpha)}{P(-1|x;\alpha)} = s_1(x_1) + s_2(x_2) + \cdots + s_m(x_m)$$

AdaBoost

$$F(\boldsymbol{x}) \quad = \quad \underset{F^* \in \mathcal{F}}{\mathbf{argmin}}\, E[e^{-YF^*(\boldsymbol{x})}]$$

2.1 GAM との対応づけ

ロジスティック GAM

$$F_L(\boldsymbol{x}) = \log \frac{P(+1|x;\alpha)}{P(-1|x;\alpha)} = s_1(x_1) + s_2(x_2) + \cdots + s_m(x_m)$$

AdaBoost

$$F(\boldsymbol{x}) \quad = \quad \underset{F^* \in \mathcal{F}}{\mathbf{argmin}}\, E[e^{-YF^*(\boldsymbol{x})}] = \frac{1}{2} \log \frac{P(+1|x;\alpha)}{P(-1|x;\alpha)}$$

2.1 GAM との対応づけ

ロジスティック GAM

$$F_L(\boldsymbol{x}) = \log \frac{P(+1|x;\alpha)}{P(-1|x;\alpha)} = s_1(x_1) + s_2(x_2) + \cdots + s_m(x_m)$$

AdaBoost

$$F(\boldsymbol{x}) = \underset{F^* \in \mathcal{F}}{\mathbf{argmin}}\, E[e^{-YF^*(\boldsymbol{x})}] = \frac{1}{2} \log \frac{P(+1|x;\alpha)}{P(-1|x;\alpha)}$$

$$= \sum_{j=1}^{T} \alpha_j f_{j(t)}(x) = S_1(x_1) + S_2(x_2) + \cdots + S_m(x_m)$$

2.1 GAMとの対応づけ

ロジスティックGAM

$$F_L(\boldsymbol{x}) = \log \frac{P(+1|x;\alpha)}{P(-1|x;\alpha)} = \underbrace{s_1(x_1) + s_2(x_2) + \cdots + s_m(x_m)}_{\text{各特徴量のスコアの和}}$$

AdaBoost

$$F(\boldsymbol{x}) = \underset{F^* \in \mathcal{F}}{\mathbf{argmin}}\, E[e^{-YF^*(\boldsymbol{x})}] = \frac{1}{2} \log \frac{P(+1|x;\alpha)}{P(-1|x;\alpha)}$$

$$= \sum_{j=1}^{T} \alpha_j f_{j(t)}(x) = \underbrace{S_1(x_1) + S_2(x_2) + \cdots + S_m(x_m)}_{\text{各特徴量のスコアの和}}$$

2.2 AdaBoost の幾何学的解釈

準備 **(Lebanon and Lafferty, 2002) による解釈**

- **拡張 KL ダイバージェンス**

$$D(\mu, \nu) = \int_{\mathcal{X}} q(x) \sum_{y \in \mathcal{Y}} \left\{ \nu(y \mid x) - \mu(y \mid x) - \mu(y \mid x) \left(\log \frac{\nu(y \mid x)}{\mu(y \mid x)} \right) \right\} dx.$$

- **C から連想される指数型分布族**

$$\mathcal{M}(C) = \left\{ \mu(y|x;\theta) = \exp\left(-\frac{y}{2} \langle \theta, \Phi(x) \rangle - g(x;\theta) \right) \right\}$$

ここで $\Phi(x) = (f_1(x), f_2(x), \cdots)^T$ **とし, それと同じ次元のパラメータベクトルを** $\theta = (\theta_1, \theta_2, \cdots)^T$ **とする. また** $g(x;\theta) = -\left\langle \theta, \sum_{y' \in \mathcal{Y}} \frac{y'}{2} p(y' \mid x) \Phi(x) \right\rangle$ **とする.**

2.2 AdaBoostの幾何学的解釈

AdaBoostの幾何学的解釈 (Lebanon and Lafferty, 2002)

- 経験分布とモデル $\mathcal{M}(C)$ 上の点とのKLダイバージェンスは定数項を除いて $\widehat{A}(F(\cdot;\theta)) = \frac{1}{n}\sum_{i=1}^{n}\exp(-Y_i F(X_i;\theta))$ と一致
- AdaBoostはKLダイバージェンスの意味で経験分布 $\hat{p}$ からモデル $\mathcal{M}(C)$ 上の再近傍点 $\hat{\theta}$ を逐次的に探索する.

1. 弱判別機とその係数を以下のように選ぶ

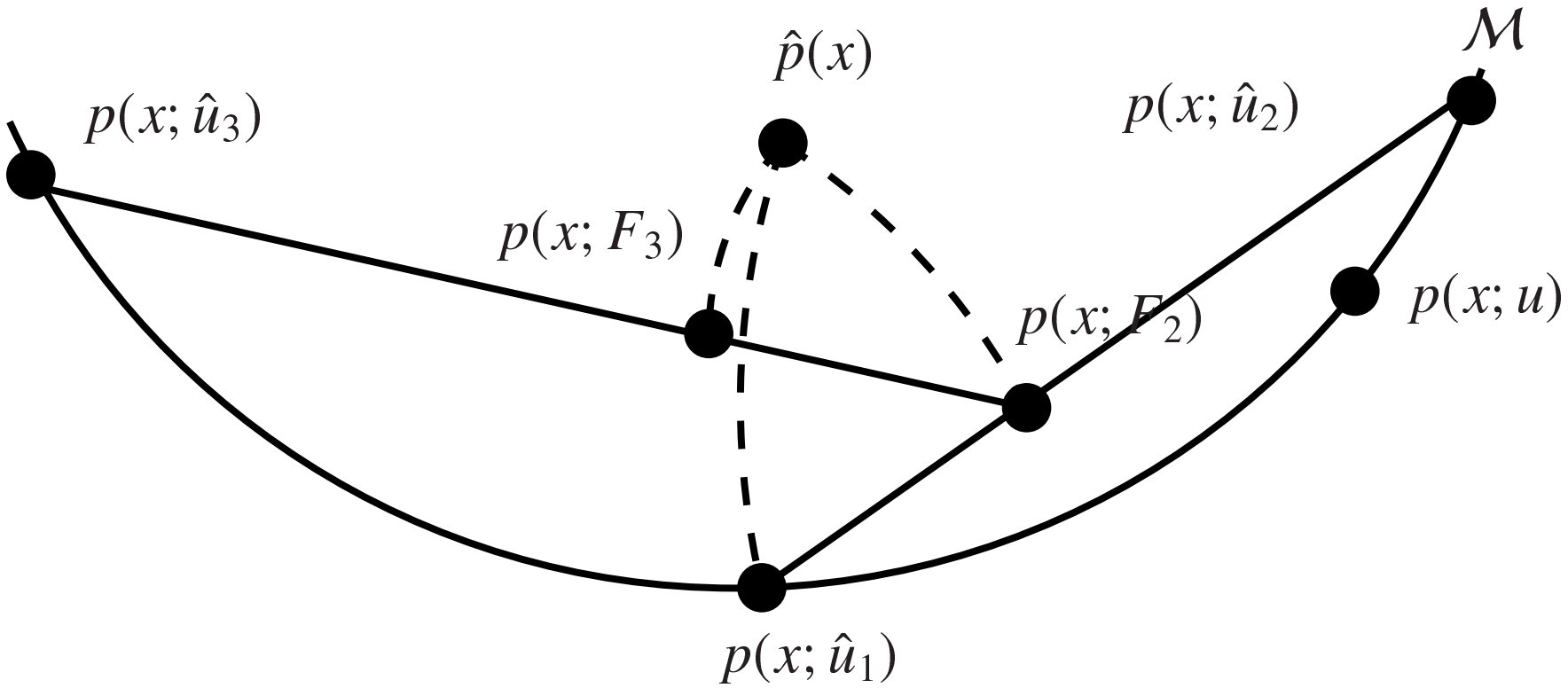
$$f_t \approx \operatorname*{argmin}_{f\in C} \widehat{A}(F_{t-1}+\alpha f) \quad \text{for any positive } \alpha,$$

$$\alpha_t = \operatorname*{argmin}_{\alpha\in R} \widehat{A}(F_{t-1}+\alpha f_t),$$

2. 判別関数を $F_t(x) = F_{t-1}(x) + \alpha_t f_t(x)$ と更新する.

2.2 ブースティングの幾何学的解釈

弱学習機モデルからの飛び出し



2.3 カーネルマシンとの関係

背景

- ブースティングとカーネルマシンは互いに近い関係にある (Rätsch et al., 2002). ここではカーネル指数型分布族 (Canu and Smola, 2006) を用いて両者を統一的視点で見てみよう.

両者の統合に向けて

- 弱学習機から連想される弱カーネルの導出
- 一応新しい正則化ブースティングアルゴリズムの導出と多クラスへの拡張
- 提案した方法はブースティングとカーネルマシン双方の特徴を併せ持っている. その性能を UCI や DELVE などのベンチマークデータセットで評価する.

Remark この研究は統数研の池田先生, 江口先生との共同研究

カーネルマシンとしてのブースティング

WL カーネル

$$K_C(x, x') = \sum_{f_j \in C} \pi_j f_j(x) f_j(x'), \tag{1}$$

ここで $0 \leq \pi_j \leq 1, \sum_j \pi_j = 1$ **とする.** π_j **が一様ならば Rätsch et al. (2000) と一致.**

WL カーネルとは カーネル関数 K_C **は以下のように解釈できる.**

- **分布** $\mathbf{Prob}(\mathbb{F} = f_j) = \pi_j$ **に従う確率変数** $\mathbb{F} \in C$ **を考える.**
- **任意の** $x, x' \in \mathcal{X}$ **が与えられたとき,**

$$\begin{aligned} K_C(x, x') &= E_{\mathbb{F}}[\mathbb{F}(x)\mathbb{F}(x')] = E_{\mathbb{F}}[2I(\mathbb{F}(x) = \mathbb{F}(x')) - 1] \\ &= 2\mathbf{Prob}(\mathbb{F}(x) = \mathbb{F}(x')) - 1 \end{aligned}$$

カーネルマシンとしてのブースティング

カーネル K_C の RKHS と Boosting の判別関数の全体が一致

$$\mathcal{H}_{K_C} = \left\{ F(x;\theta) = \sum_j \theta_j f_j(x) \,|\, \forall j\,,\ \theta_j \in R\,,\ f_j \in C \right\}$$

再生性よりブースティングの指数型分布族は

$$\mathcal{M}(C) = \left\{ \mu(y|x;\theta) = \exp\left(-\frac{y}{2} \langle F(\cdot;\theta), K_C(\cdot,x) \rangle_{\mathcal{H}_{K_C}} - g(x;\theta) \right) \middle|\ F \in \mathcal{H}_{K_C} \right\}$$

カーネル指数型分布族 Canu and Smola (2006) として書ける. ここで $g(x;\theta) = -\left\langle F(\cdot;\theta), \sum_{y' \in \mathcal{Y}} \frac{y'}{2} p(y' \,|\, x) K_C(\cdot,x) \right\rangle$.

ブースティングとカーネルマシン

- **ブースティングは本質的にはカーネルマシンと同じモデル**
- **両者の差はロス関数と最適化法の違いと捉えられる.**

無限個の弱学習機からなる WL カーネル

動機

- 複雑な問題に対しては多様な弱学習機が必要
- しかし弱学習機の増加に伴い WL カーネルの計算量が増大

いくつかの弱学習機について無限個用意しても計算量が低いシンプルな L_1 カーネルを導くことを示せる.

連続版 WL カーネル

- C が連続統計モデル $C = \{f(x;\xi) \mid \xi \in \Xi\}$ とする
- Ξ 上の任意の確率密度関数 $\pi(\xi)$ に対する K_C の定義

$$K_C(x, x') = \int_{\xi \in \Xi} \pi(\xi) f(x;\xi) f(x';\xi) d\xi$$

決定スタンプカーネル

- **各特徴量 X_m が有限の範囲 $[\ell_m, r_m]$ に値をとると仮定**
- **再定義** $C_{\mathrm{ds}} = \{f^{\mathrm{s}}(x; m, b) \,|\, m = 1, 2, \cdots, M, \forall m, b \in [\ell_m, r_m]\}$

今 f^{s} のパラメータの分布を一様分布とする:

$$\pi(m, b) = \pi(m) \cdot \pi(b \,|\, m) = \frac{1}{M} \cdot \frac{1}{r_m - \ell_m}.$$

決定スタンプカーネル

$$\begin{aligned} K_{\mathrm{ds}}(x, x') &= \sum_{m=1}^{M} \int_{\ell_m}^{r_m} \pi(m, b) f^{\mathrm{s}}(x; m, b) f^{\mathrm{s}}(x'; m, b) db \\ &= \sum_{m=1}^{M} \int_{\ell_m}^{r_m} \frac{\mathrm{sign}(x_m - b)\mathrm{sign}(x'_m - b)}{M(r_m - \ell_m)} db = 1 - \frac{2}{M} \sum_{m=1}^{M} \frac{|x_m - x'_m|}{r_m - \ell_m}. \end{aligned}$$

グラム行列による比較

カーネル関数の判別能力の視覚的比較

- **線形分離可能でかつミスラベルを** 10% **含むデータセット** $\{(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)\}$ $(n = 50)$ **を準備**
- **カーネル関数の視覚的性能比較のため, このデータセット上の** **グラム行列** $G = [G_{ij} := k(X_i, X_j)]$ **を観察する. ただし** $Y_i = -1$ $(i = 1, 2, \cdots, 28)$ **かつ** $Y_i = 1$ $(i = 29, \cdots, 50)$ **とした.**

$$\begin{pmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_{50} \end{pmatrix} = \mathbf{sign}\left(\begin{pmatrix} k(X_1, X_1) & k(X_1, X_2) & \cdots & k(X_1, X_{50}) \\ k(X_2, X_1) & k(X_2, X_2) & \cdots & k(X_2, X_{50}) \\ \vdots & \vdots & \vdots & \vdots \\ k(X_{50}, X_1) & k(X_{50}, X_2) & \cdots & k(X_{50}, X_{50}) \end{pmatrix} \cdot \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \\ \vdots \\ \hat{\theta}_{50} \end{pmatrix} \right)$$

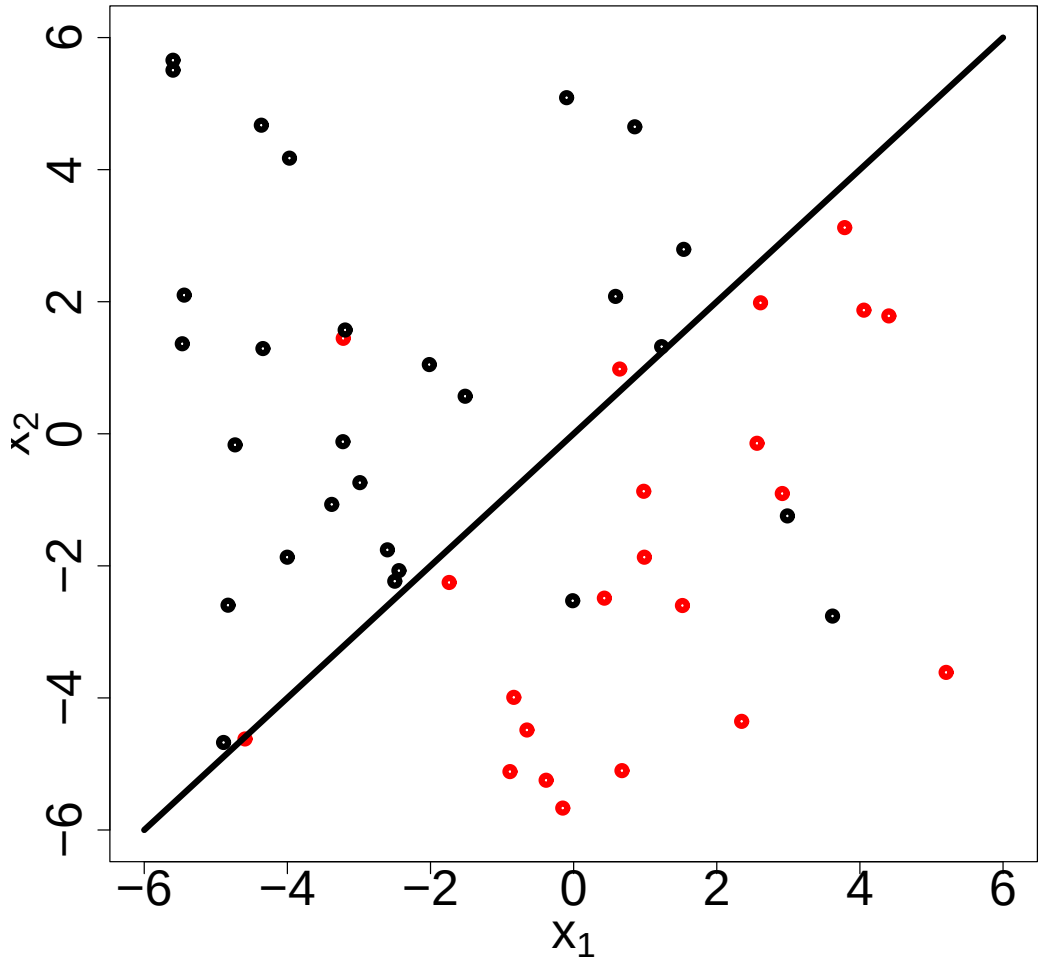
グラム行列による比較

カーネル関数の判別能力の視覚的比較

- **線形分離可能でかつミスラベルを** 10% **含むデータセット** $\{(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)\}$ $(n = 50)$ **を準備**
- **カーネル関数の視覚的性能比較のため, このデータセット上の** **グラム行列** $G = [G_{ij} := k(X_i, X_j)]$ **を観察する. ただし** $Y_i = -1$ $(i = 1, 2, \cdots, 28)$ **かつ** $Y_i = 1$ $(i = 29, \cdots, 50)$ **とした.**

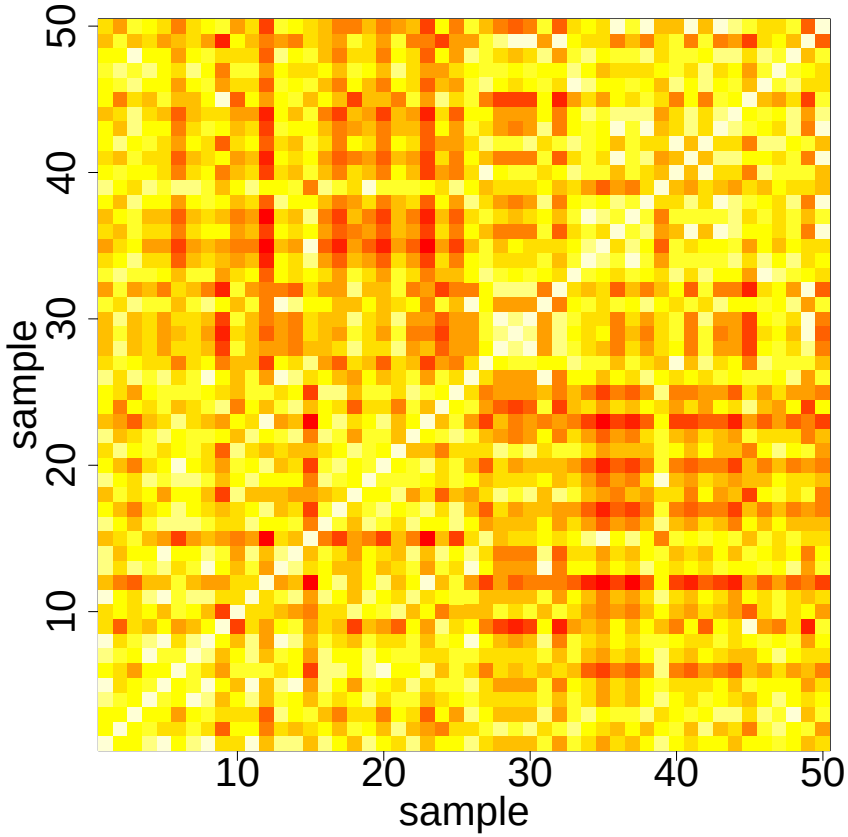
$$\begin{pmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_{50} \end{pmatrix} = \mathbf{sign}\left(\begin{pmatrix} k(X_1, X_1) & k(X_1, X_2) & \cdots & k(X_1, X_{50}) \\ k(X_2, X_1) & k(X_2, X_2) & \cdots & k(X_2, X_{50}) \\ \vdots & \vdots & \vdots & \vdots \\ k(X_{50}, X_1) & k(X_{50}, X_2) & \cdots & k(X_{50}, X_{50}) \end{pmatrix} \cdot \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \\ \vdots \\ \hat{\theta}_{50} \end{pmatrix} \right)$$

用いたデータ

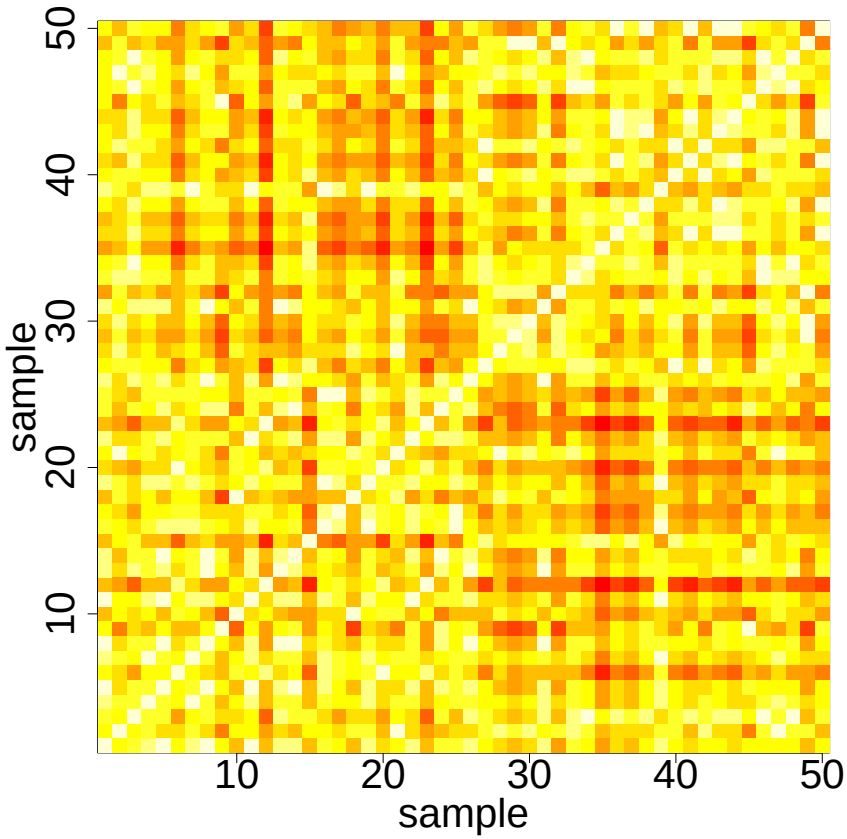


決定スタンプカーネル 離散版と連続版

$K_{C_{\mathrm{ds}}}$ と K_{ds}



(a) $K_{C_{\mathrm{ds}}}$



(b) K_{ds}

線形判別機カーネル

- 線形判別機のクラスを $\mathcal{C} = \{f(x; w) = w^T x \,|\, w \in \Re^M\}$ と定義する.
- さらに $\pi(w)$ をパラメータ w の平均 0, 分散共分散行列 V を持つ任意の分布とする.
- 線形判別機カーネル K_{lin}

$$\begin{aligned} K_{\mathrm{lin}}(x, x') &= \int_{\Re^M} \pi(w) f(x; w) f(x'; w) dw \\ &= x^T \left\{ \int_{\Re^M} \pi(w) w w^T dw \right\} x' = x^T V x'. \end{aligned}$$

これは線形代数にしばしば現れる行列 V により一般化されたユークリッド内積 (L_2 カーネル) である.

決定木のクラス

深さ d 以下の全ての決定木

$$C_{\mathrm{dt}}(d) = \{f^{\mathrm{dt}}(x;\sigma,m,b) = \sum_{\ell=1}^{L} \sigma_\ell \Pi_{j'\in\tilde{\Gamma}(\ell)} h(x;m_{j'},b_{j'},\tilde{s}(j',\ell)) \,|\, \sigma \in \mathcal{Y}^L,$$
$$\forall j = 1,2,\cdots,N,\ m_j \in \{1,2,\cdots,M\},\ b_j \in [\ell_{m_j}, r_{m_j}]\}.$$

深さ d 以下の決定木のパラメータ

$$\begin{aligned}
\pi(\sigma,m,b) &= \pi(\sigma)\pi(m,b),\\
\pi(m,b) &= \pi(b\,|\,m)\pi(m),\\
\pi(\sigma) &= 1/L, \quad \pi(m) = M^{-N},\\
\pi(b\,|\,m) &= \Pi_{j=1}^{N}\pi(b_j\,|\,m),\\
\pi(b_j\,|\,m) &= \frac{I(\ell_{m_j} \le b_j \le r_{m_j})}{r_{m_j} - \ell_{m_j}}
\end{aligned}$$

決定木$C_{\mathrm{dt}}(d)$の記法

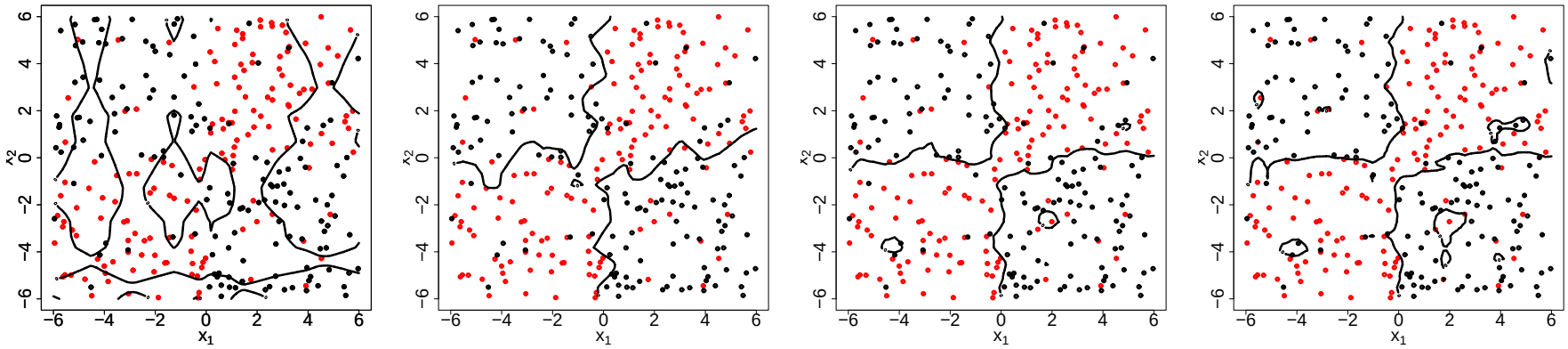
L	**the number of leaves, which is 2^d in a full tree.**
N	**the number of nodes, which is $2^d - 1$ in a full tree.**
$\sigma = (\sigma_1, \sigma_2, \cdots, \sigma_L)^T$	**Each $\sigma_\ell \in \mathcal{Y}$ is the label of the leaf ℓ.**
$m = (m_1, m_2, \cdots, m_N)^T$	**Each $m_j \in \{1, 2, \cdots, M\}$ is the feature on which the node j bases.**
$b = (b_1, b_2, \cdots, b_N)^T$	**Each $b_j \in [\ell_m, r_m]$ is the threshold value of the node j.**
$\widetilde{\Gamma}(\ell)$	$\widetilde{\Gamma}(\ell) =$ {**all ancestors of the leaf ℓ**}
RD(j)	**RD**$(j) =$ {**all right descendants of the node j**}.
LD(j)	**LD**$(j) =$ {**all left descendants of the node j**}.
$\tilde{s}(j', \ell)$	**For any** $j' \in \widetilde{\Gamma}(\ell)$, $\tilde{s}(j', \ell) = \begin{cases} 1 & (\ell \in \mathbf{RD}(j')) \\ -1 & (\ell \in \mathbf{LD}(j')) \end{cases}$
$h(x; m, b, s)$	$h(x; m, b, s) = I(f^{\mathrm{s}}(x; m, b, s) = 1)$**.**

決定木カーネル

- **決定木カーネル** K_d^{dt}

$$
\begin{aligned}
K_d^{\mathrm{dt}}(x, x') &= E_{\sigma,m,b}[f^{\mathrm{dt}}(x;\sigma,m,b) f^{\mathrm{dt}}(x';\sigma,m,b)] \\
&= \left(1 - \frac{1}{M}\sum_{m=1}^{M} \frac{|x_m - x'_m|}{r_m - \ell_m}\right)^d .
\end{aligned}
$$

- **このパラメータ分布** $\pi(\sigma, m, b)$ **は完全木より刈り取られた木に大きな確率を割り当てることに注意**



正則化ブースティング RegAdaBoost

Proposition 1. 訓練データ $\{X_i, Y_i\}_{i=1}^n$ が与えられたとする. このとき $\eta \in R^n$ が存在して

$$\hat{\theta} = \underset{\theta}{argmin}\, \widehat{A}(F(\cdot;\theta)) + \lambda \|F(\cdot;\theta)\|^2_{\mathcal{H}_{K_C}}$$

を満たす $\hat{\theta}$ が $\hat{\theta}_j = \sum_{\ell=1}^n \pi_j f_j(X_\ell)\eta_\ell$ と書ける.

ゆえに

$$\overline{F}(x;\eta) = \sum_j \overbrace{\left\{\sum_{\ell=1}^n \pi_j f_j(X_\ell)\eta_\ell\right\}}^{\hat{\theta}_j} f_j(x) = \sum_{\ell=1}^n \eta_\ell K_C(x, X_\ell).$$

$$\widehat{\mathbb{A}}(\overline{F}(\cdot;\eta)) = \frac{1}{n}\sum_{i=1}^n \exp(-Y_i \overline{F}(X_i;\eta)) + \lambda \eta^T \boldsymbol{K}_C \eta,$$

RegAdaBoost アルゴリズム

1. $\lambda > 0$ と $\eta^0 = 0$ を固定する
2. 各ステップ $t = 1, 2, \cdots, \tau$ において

   (a) 重み $\{D_t(i)\}$ を下のようにセットする.

$$D_t(i) = \frac{1}{Z}\{\exp(-Y_i \overline{F}_{t-1}(X_i; \eta^{t-1}))(-Y_i) + 2n\lambda\eta_i^{t-1}\}$$

   (b) 新しい最良弱学習機とその係数を選ぶ:

$$\begin{aligned} h_t &= \mathbf{argmin}_{h \in C'} \sum_{i=1}^{n} D_t(i) h(X_i), \\ \alpha_t &= \mathbf{argmin}\, \widehat{\mathbb{A}}(\overline{F}_{t-1}(\cdot; \eta^{t-1}) + \alpha h_t). \end{aligned}$$

   (c) 係数 $\{\eta_\ell^{t-1}\}$ を $\eta_\ell^t = \eta_\ell^{t-1} \overset{\alpha}{+} \alpha_t(I(h_t = K(\cdot, X_\ell)) - I(h_t = -K(\cdot, X_\ell)))$ と更新する.

3. 最終的に判別機 $g(x) = \mathbf{sign}(\overline{F}(x; \eta^\tau))$ を得る.

実験: 設定

- UCI と DELVE ベンチマークデータセット
- K_{ds} の範囲 $\{\ell_m, r_m\}$ の推定値として訓練データにおける各特徴量の範囲を用いた.
- AdaBoost のステップ数 τ と RegAdaBoost の λ は 10-fold crossvalidation を用いて決めた.
- RegAdaBoost のステップ数 τ は K_g の場合を除いて十分に長くとった.
- 訓練データのサイズは次表に載せた. テストデータは 1000 サンプル以内のデータはその全てを用い, 1000 サンプル以上のデータは 1000 までで切り捨てた.

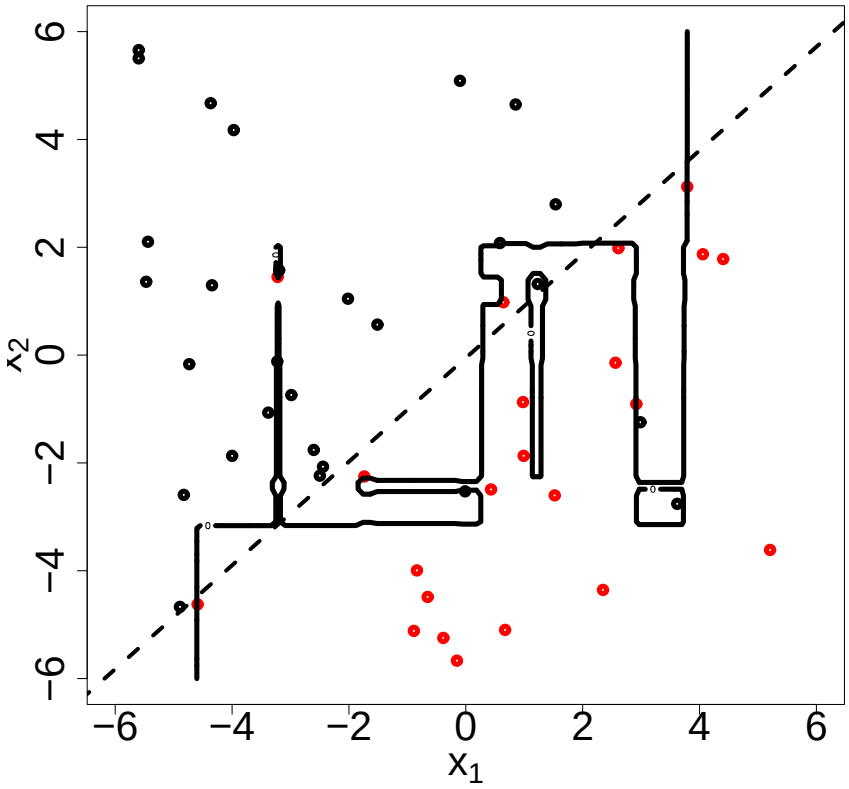
実験: 結果

データ	特徴量	標本数	テスト誤差			
			AdaBoost	RegAdaBoost		
				$K_{C_{\mathrm{ds}}}$	K_{ds}	K_g
slope	2	50	19.5	14.7	14.8	18.1
bcw	9	200	7.88	3.32	3.73	2.9
splice	60	200	13.4	12.4	12.2	17.3
thyroid	5	100	5.26	0.87	0.87	3.51
titanic	3	600	20.5	21.7	20.8	21
wine	13	100	5.19	0	0	31.16
waveform	21	200	11.1	9.8	9.3	8.7

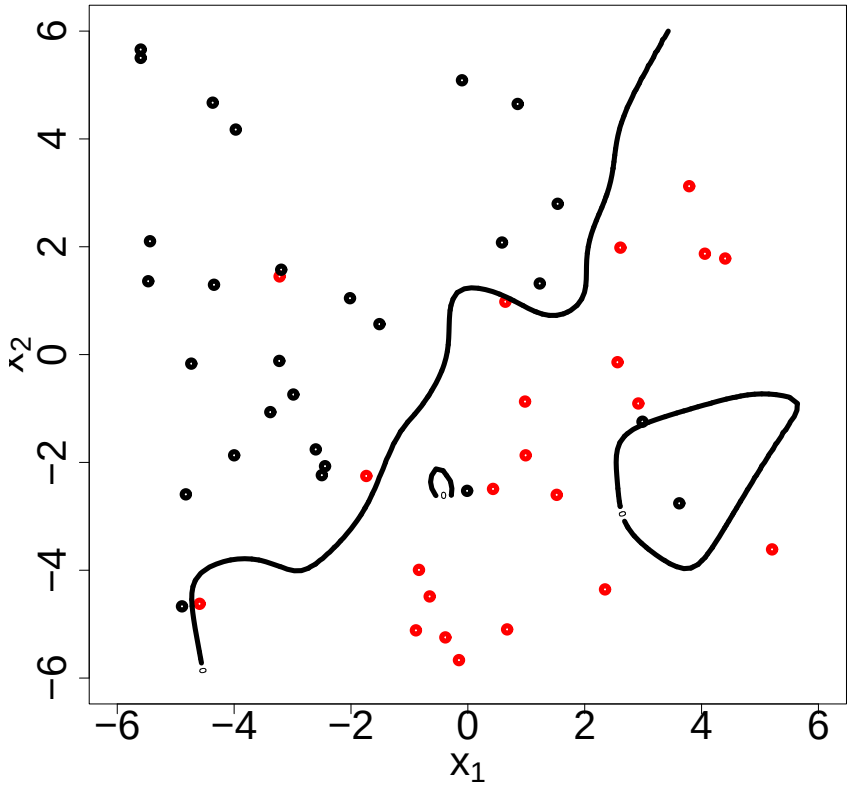
実験: 結果

データ	特徴量	標本数	テスト誤差			
			AdaBoost	RegAdaBoost		
				$K_{C_{\mathrm{ds}}}$	K_{ds}	K_g
slope	2	50	19.5	14.7	14.8	18.1
bcw	9	200	7.88	3.32	3.73	2.9
splice	60	200	13.4	12.4	12.2	17.3
thyroid	5	100	5.26	0.87	0.87	3.51
titanic	3	600	20.5	21.7	20.8	21
wine	13	100	5.19	0	0	31.16
waveform	21	200	11.1	9.8	9.3	8.7

Slope data の判別結果

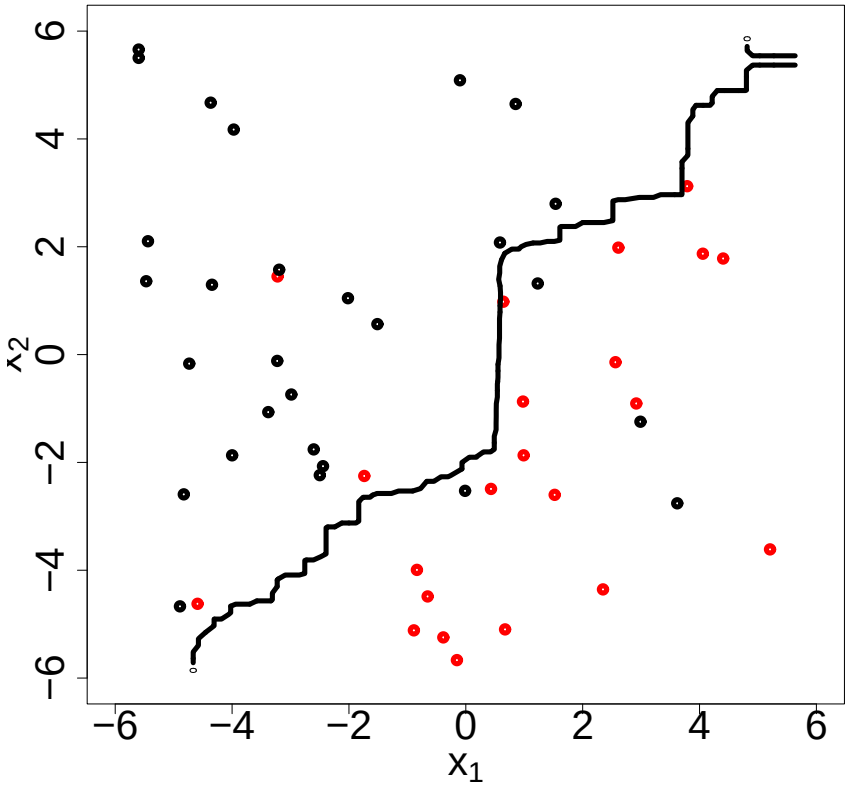


(c) AdaBoost

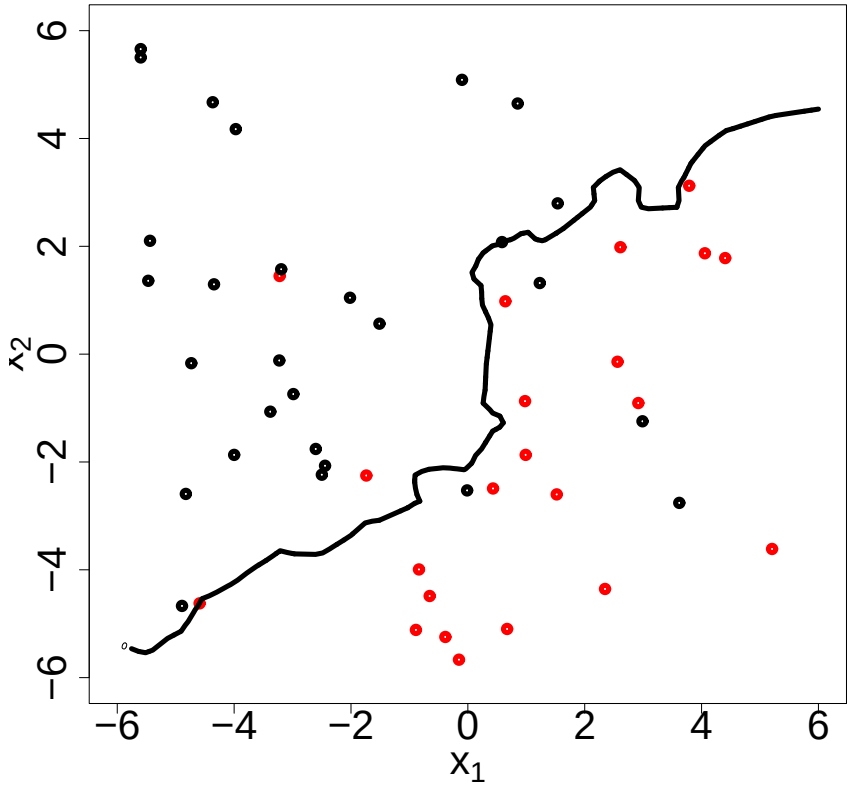


(d) K_g

Slope data の判別結果



(e) $K_{C_{\mathrm{ds}}}$



(f) K_{ds}

実験: 結果

データ	特徴量	標本数	テスト誤差			
			AdaBoost	RegAdaBoost		
				$K_{C_{\mathrm{ds}}}$	K_{ds}	K_g
slope	2	50	19.5	**14.7**	14.8	18.1
bcw	9	200	7.88	3.32	3.73	**2.9**
splice	60	200	13.4	12.4	**12.2**	17.3
thyroid	5	100	5.26	**0.87**	**0.87**	3.51
titanic	3	600	**20.5**	21.7	20.8	21
wine	13	100	5.19	**0**	**0**	31.16
waveform	21	200	11.1	9.8	9.3	**8.7**

Extension to multiclass classification (1/2)

Multinomial logistic regression **(Krishnapuram et al., 2005)**

$$p(Y = y \mid x; \theta) = \exp\left(\theta_y^T \Phi(x) - \log \sum_{y' \in \mathcal{Y}} \exp(\theta_{y'}^T \Phi(x))\right),$$

where $\theta = (\theta_1^T, \theta_2^T, \cdots, \theta_{G-1}^T)^T$ **and** $\theta_G = 0$**.**

AdaBoost.M2 loss

$$\begin{aligned}
h_j(x, \bar{y}) &= \begin{cases} \bar{y} & (f_j(x) >= 0) \\ \mathcal{Y} \backslash \bar{y} & \textbf{otherwise} \end{cases} \\
C^{\mathrm{m}} &= \{f_j(x, y, \bar{y}) = I(y \in h_j(x, \bar{y}))\} \\
\Phi(x, y; j, \bar{y}) &= (f_1(x, y, 1) \cdots f_J(x, y, 1), f_1(x, y, 2), \cdots, f_J(x, y, G))^T \\
p(Y = y \mid x; \theta) &= \exp\left(\theta^T \Phi(x, y) - \sum_{y' \in \mathcal{Y}} p(y' \mid x) \theta \Phi(x, y)\right),
\end{aligned}$$

Extension to multiclass classification (2/2)

Construction of algorithm

1. **Obtain a multiclass loss function $\widehat{A}$ by calculating $D(\hat{p}, p(Y = y \mid x; \theta))$ and removing the constant term.**
2. **Adding an RKHS regularization to $\widehat{A}$, the representer theorem gives us the form of solution and the loss $\widetilde{A}$.**
3. **Minimize $\widetilde{A}$ w.r.t. θ by any optimization method...**
   - **Functional Gradient Descent**
   - **Bound optimization (EM-type)**
4. **A resultant classifier is defined as $g(x) = \mathbf{argmax}_{y \in \mathcal{Y}} p(Y = y \mid x; \hat{\theta})$, where $\hat{\theta}$ is an obtained estimator in 2.**

カーネル関数の学習

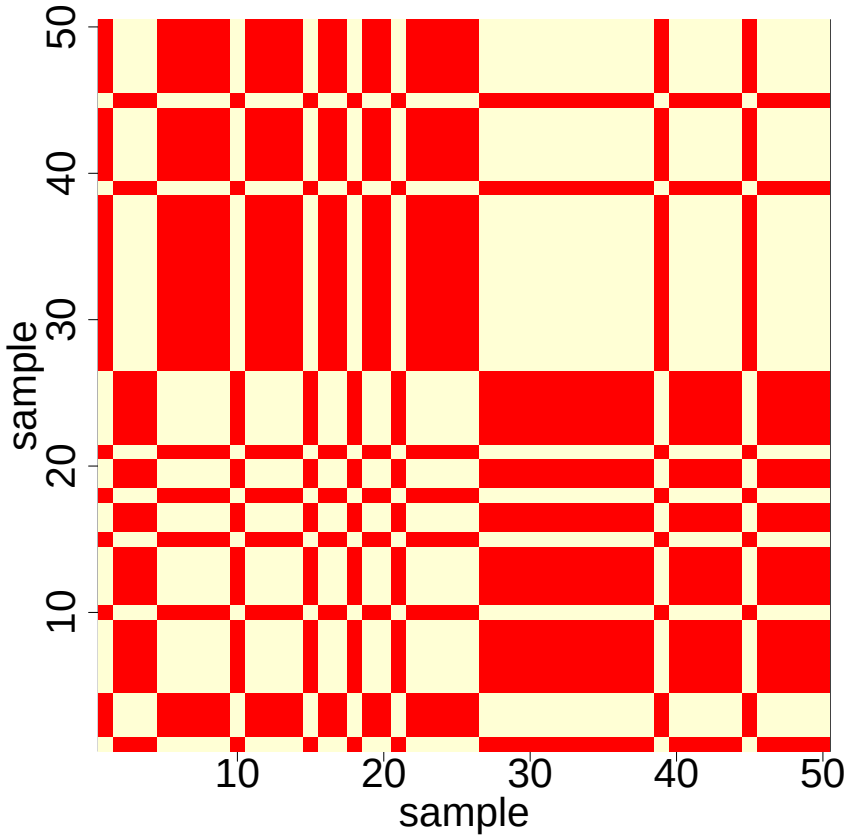
さらに弱いモデル 前述の正則化ブースティングは弱学習機の集合全体に基づく推定を行なう. しかしブースティングのように弱学習機を一つずつ加えていく過程をカーネル法に持ち込むのはどうだろうか?

カーネル関数の判別能力の変化

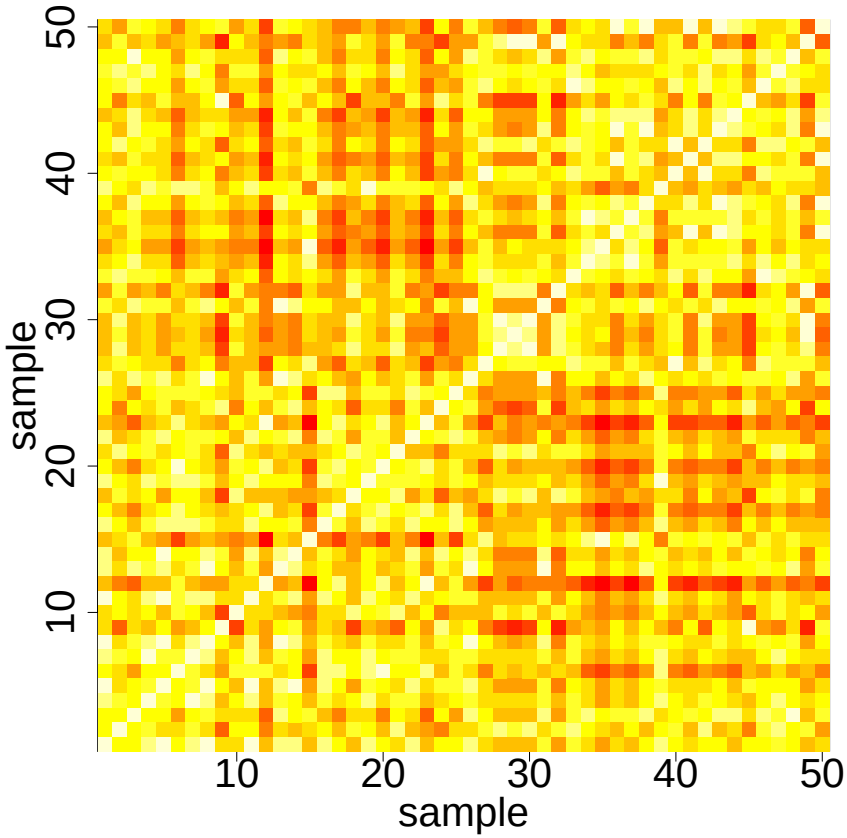
- 線形分離可能でかつミスラベルを 10% 含むデータセット $\{(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)\}$ $(n = 50)$ を考える
- 最初から全ての弱学習機を用いた WL カーネルと, ブースティングにより各ステップまでに選ばれた弱学習機のみから構成される WL カーネルをグラム行列で比較する.

カーネル関数のブースティング

Observation of boosting w.r.t. Gram matrix



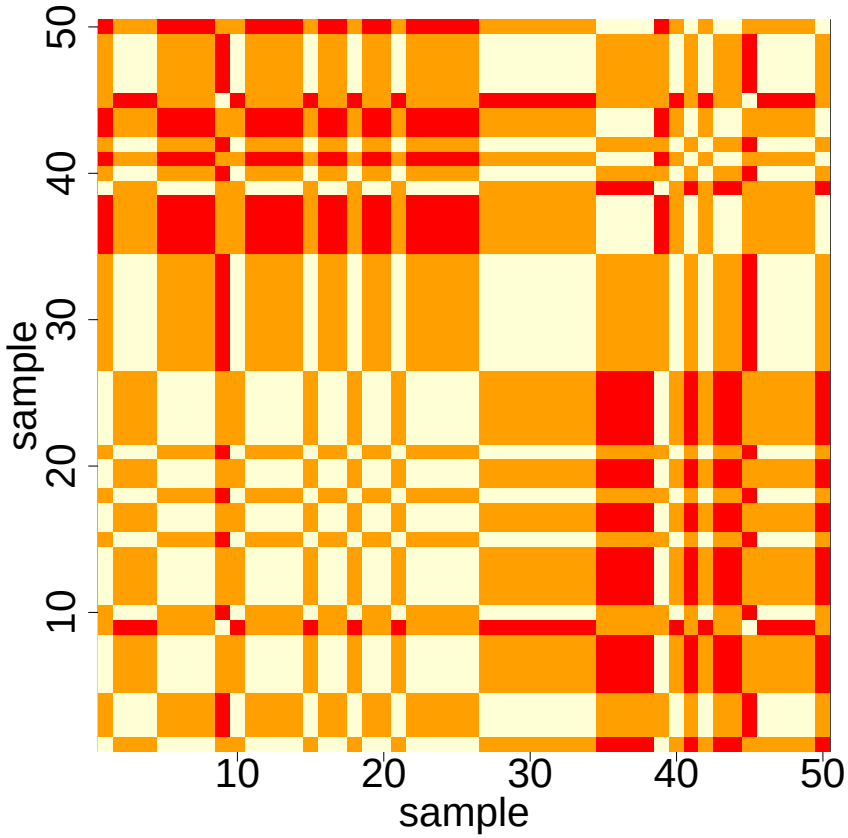
(a) $K_{C_{\mathrm{ds}}}$ **(T=1)**



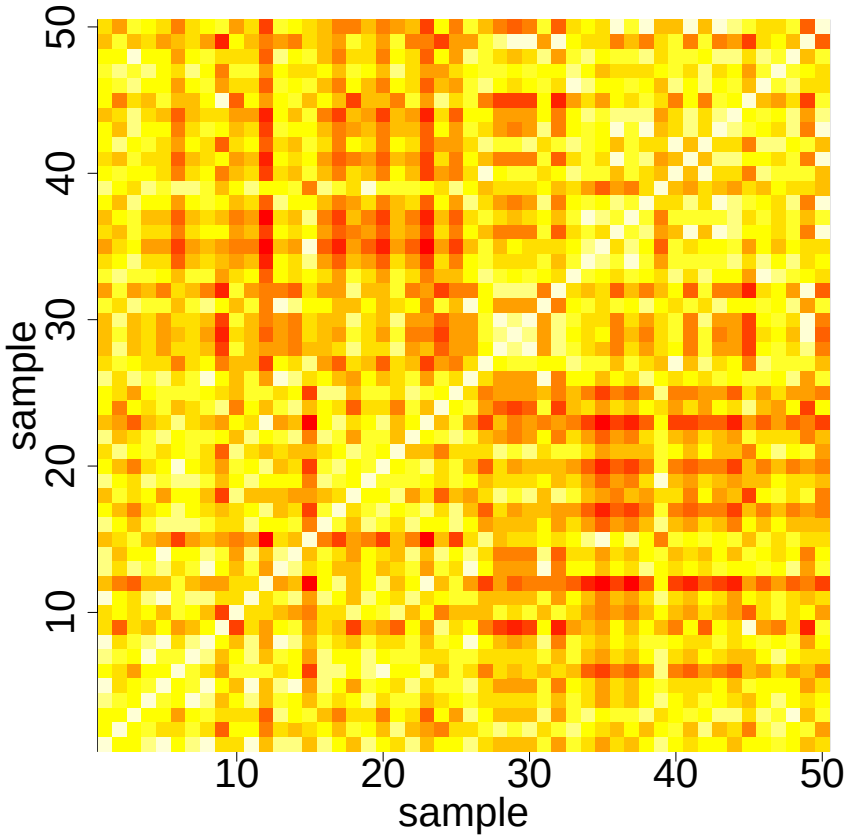
(b) $K_{C_{\mathrm{ds}}}$

カーネル関数のブースティング

Observation of boosting w.r.t. Gram matrix



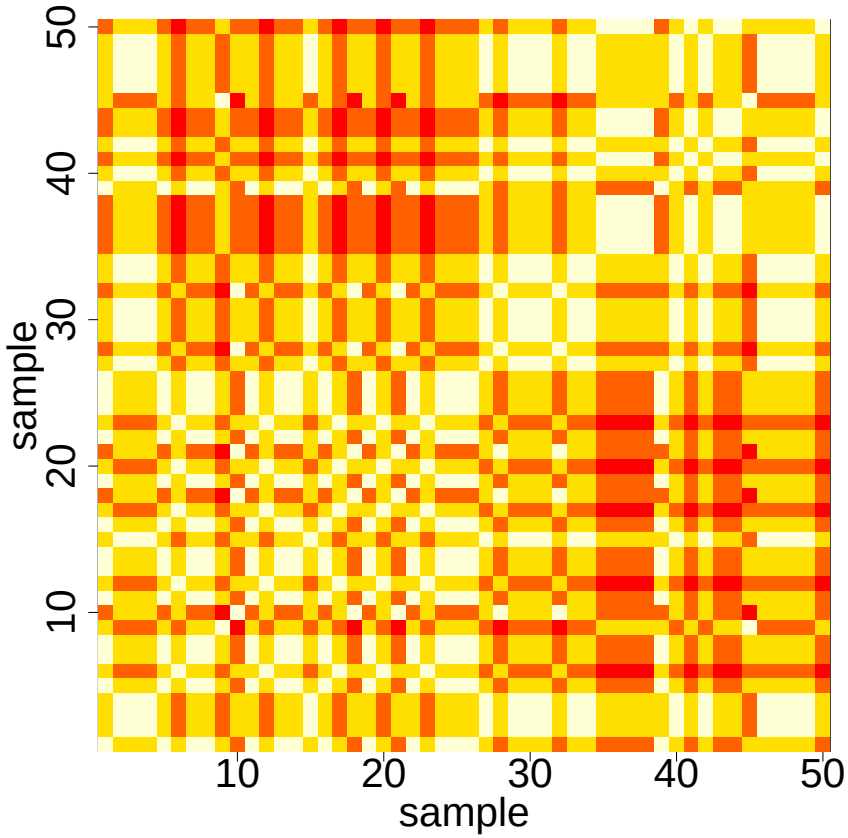
(a) $K_{C_{\mathrm{ds}}}$ **(T=2)**



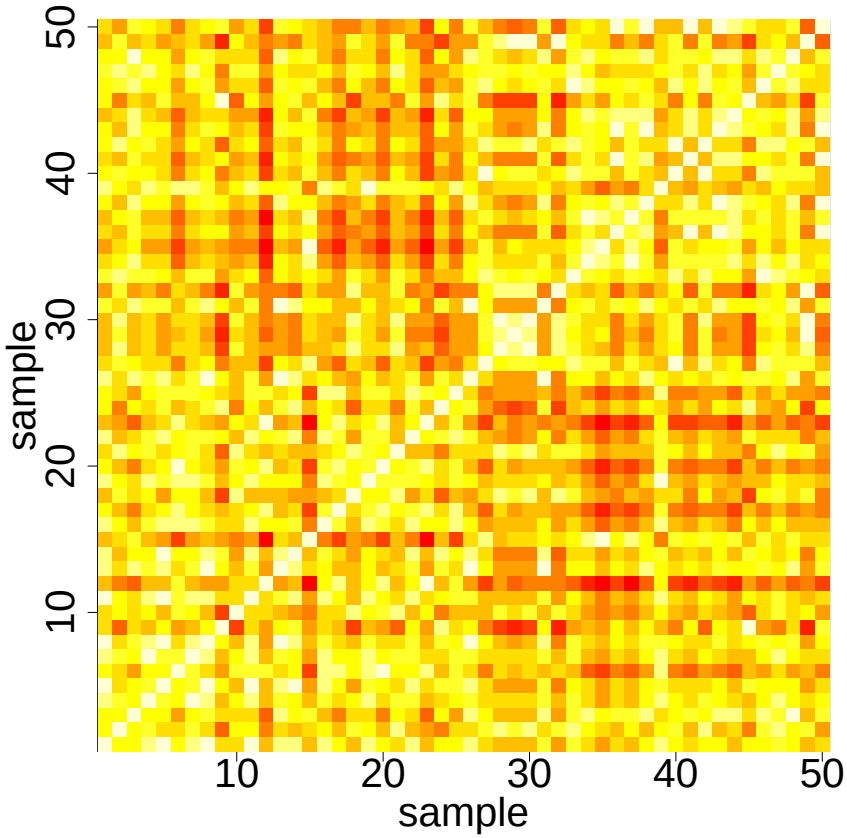
(b) $K_{C_{\mathrm{ds}}}$

カーネル関数のブースティング

Observation of boosting w.r.t. Gram matrix



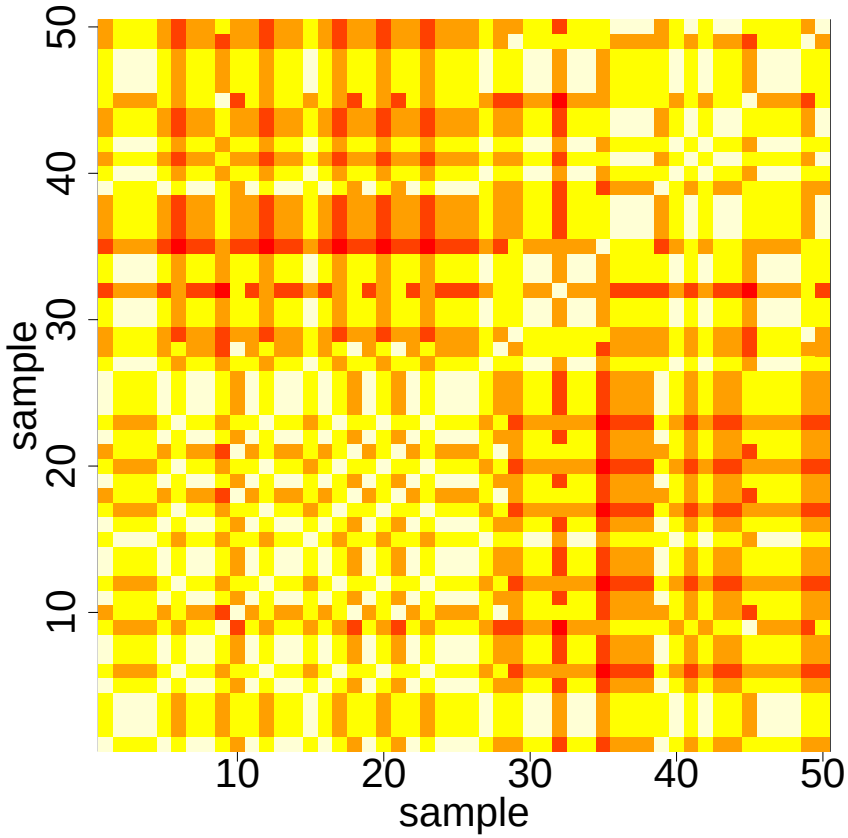
(a) $K_{C_{\mathrm{ds}}}$ **(T=3)**



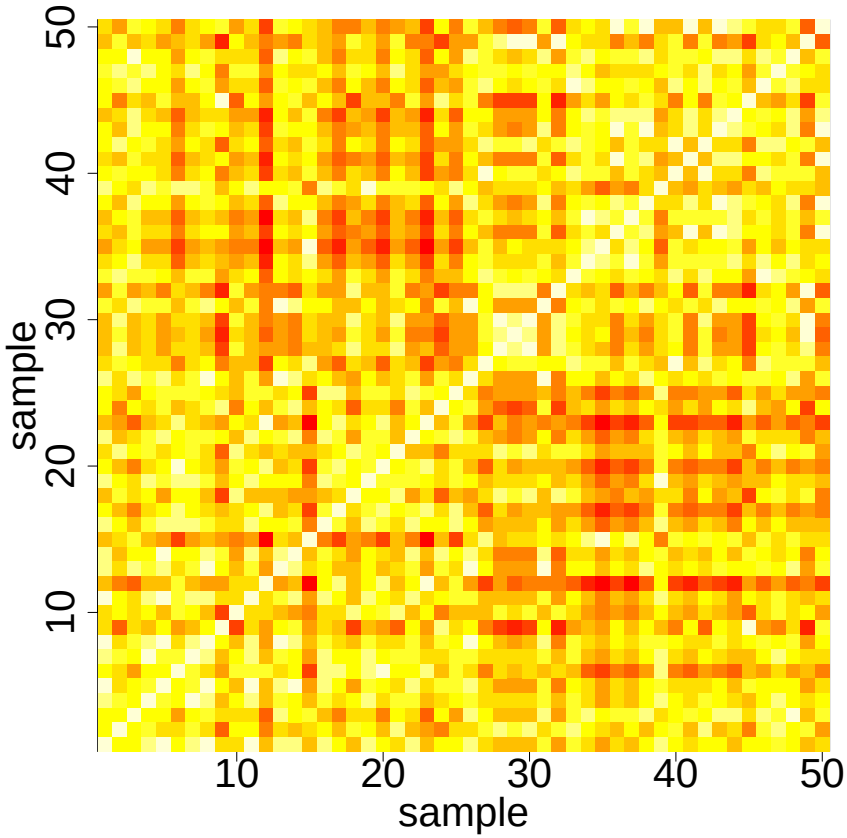
(b) $K_{C_{\mathrm{ds}}}$

カーネル関数のブースティング

Observation of boosting w.r.t. Gram matrix



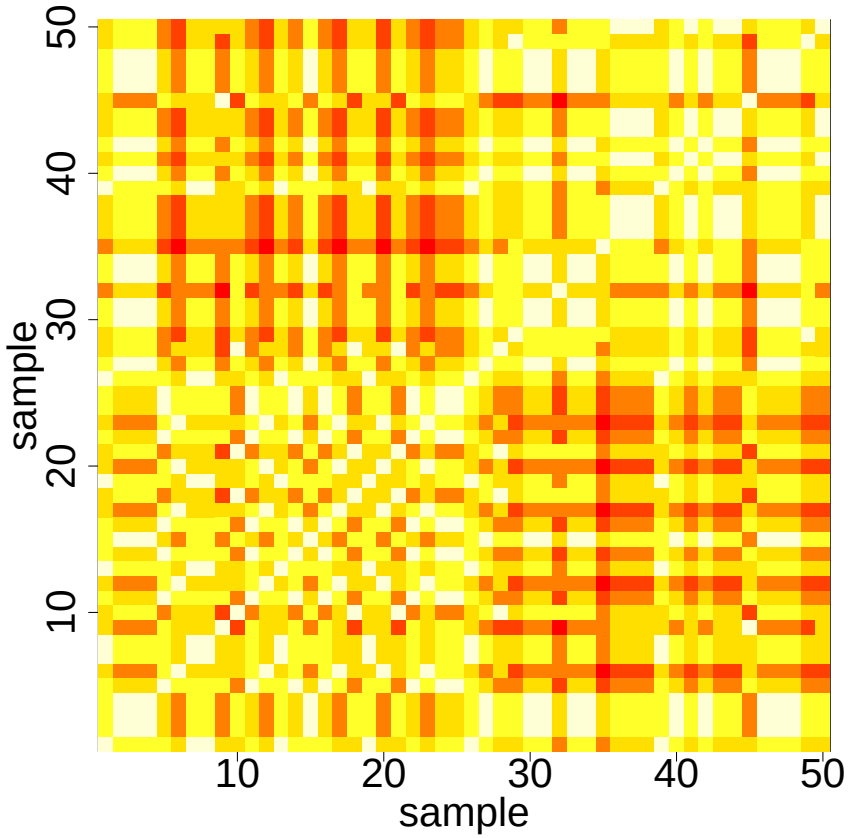
(a) $K_{C_{\mathrm{ds}}}$ **(T=4)**



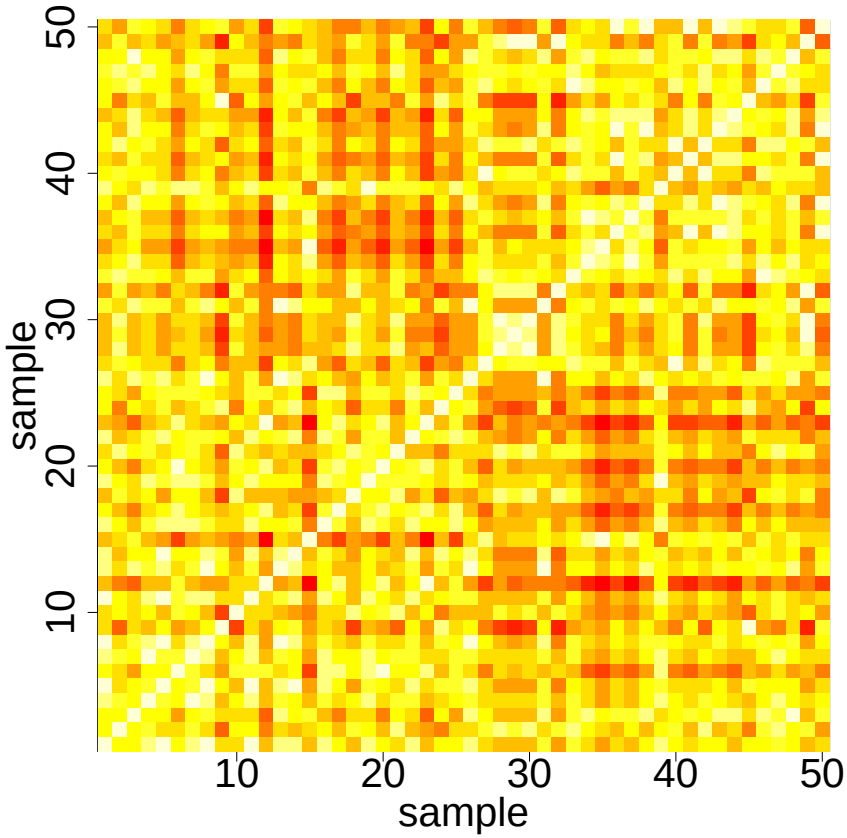
(b) $K_{C_{\mathrm{ds}}}$

カーネル関数のブースティング

Observation of boosting w.r.t. Gram matrix



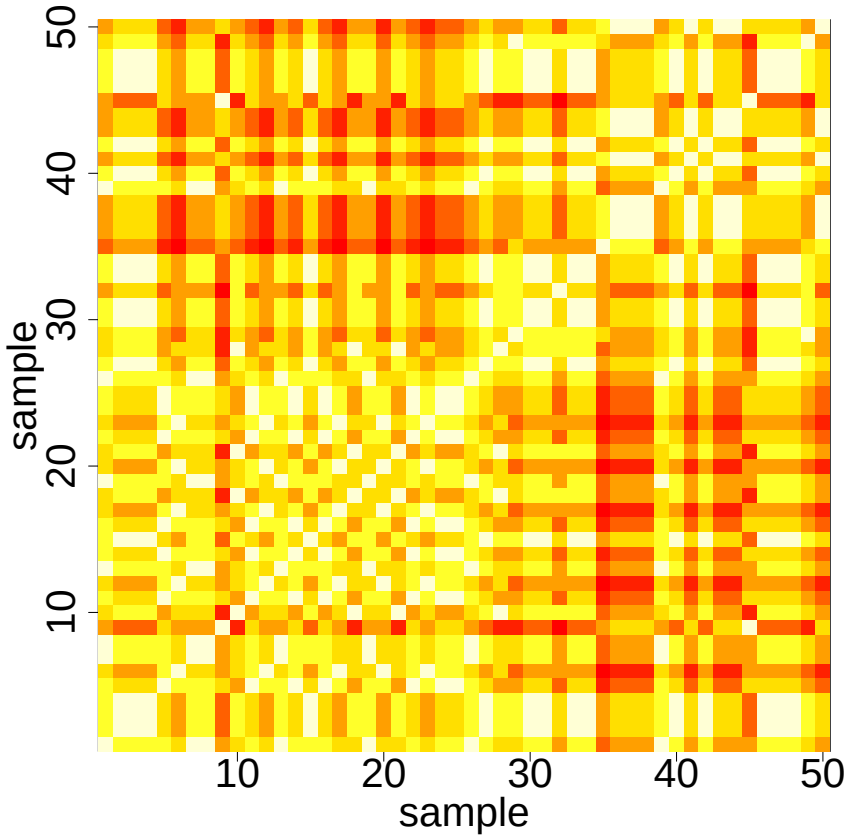
(a) $K_{C_{\mathrm{ds}}}$ **(T=5)**



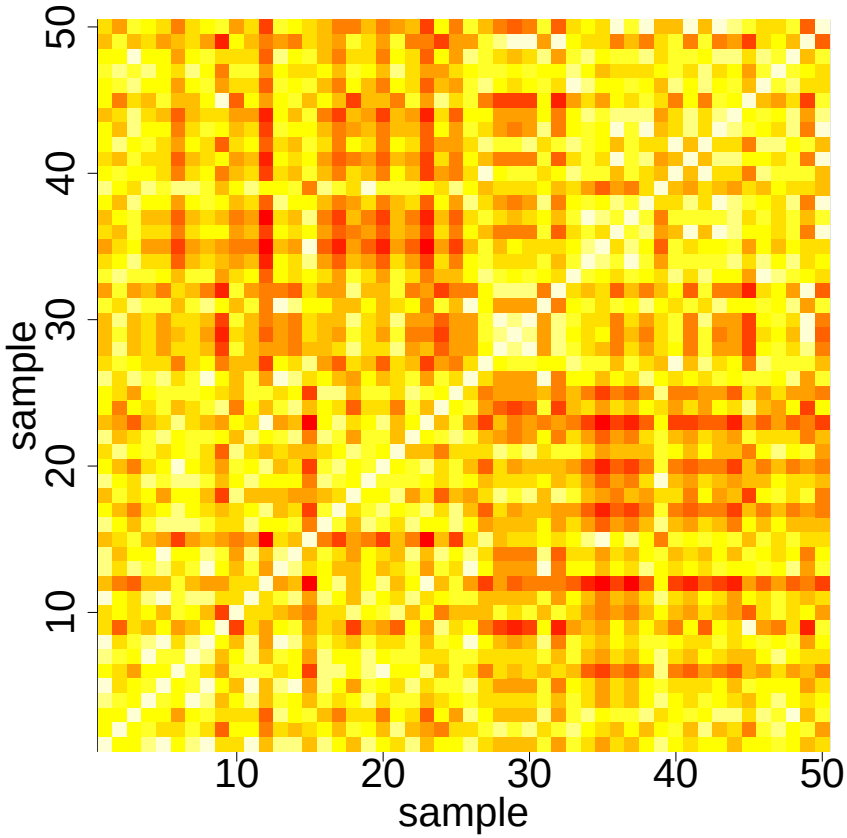
(b) $K_{C_{\mathrm{ds}}}$

カーネル関数のブースティング

Observation of boosting w.r.t. Gram matrix



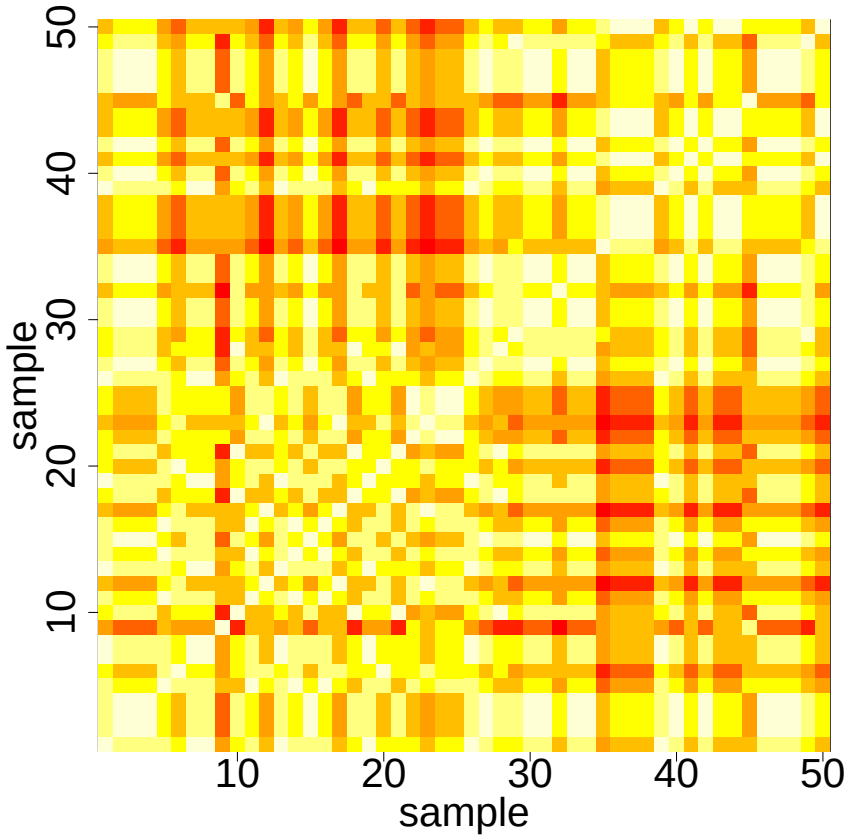
(a) $K_{C_{\mathrm{ds}}}$ **(T=6)**



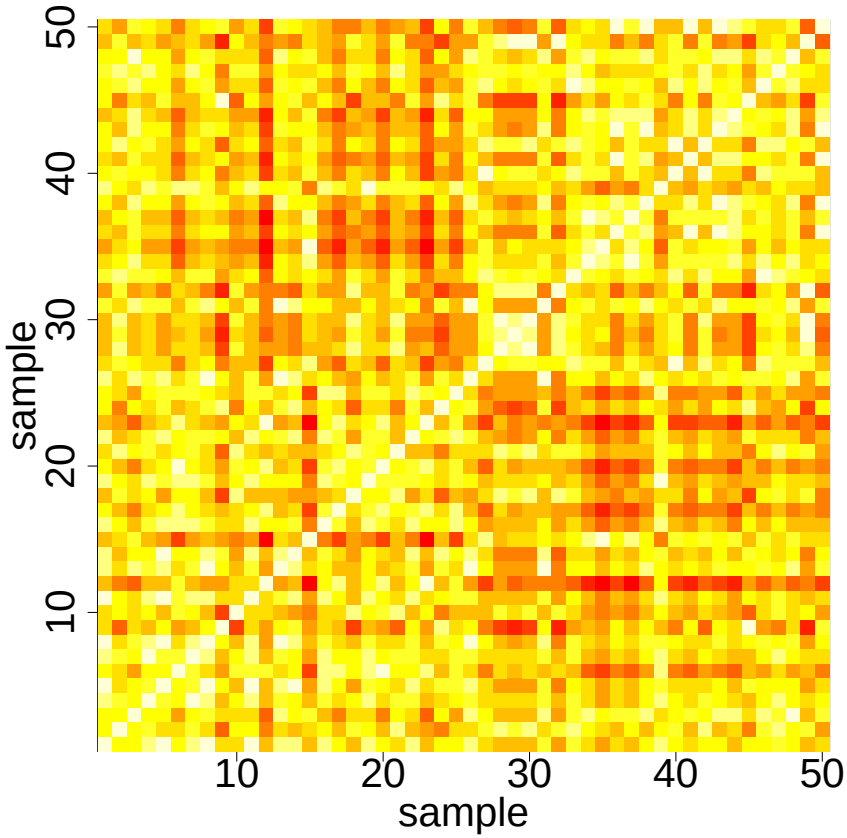
(b) $K_{C_{\mathrm{ds}}}$

カーネル関数のブースティング

Observation of boosting w.r.t. Gram matrix



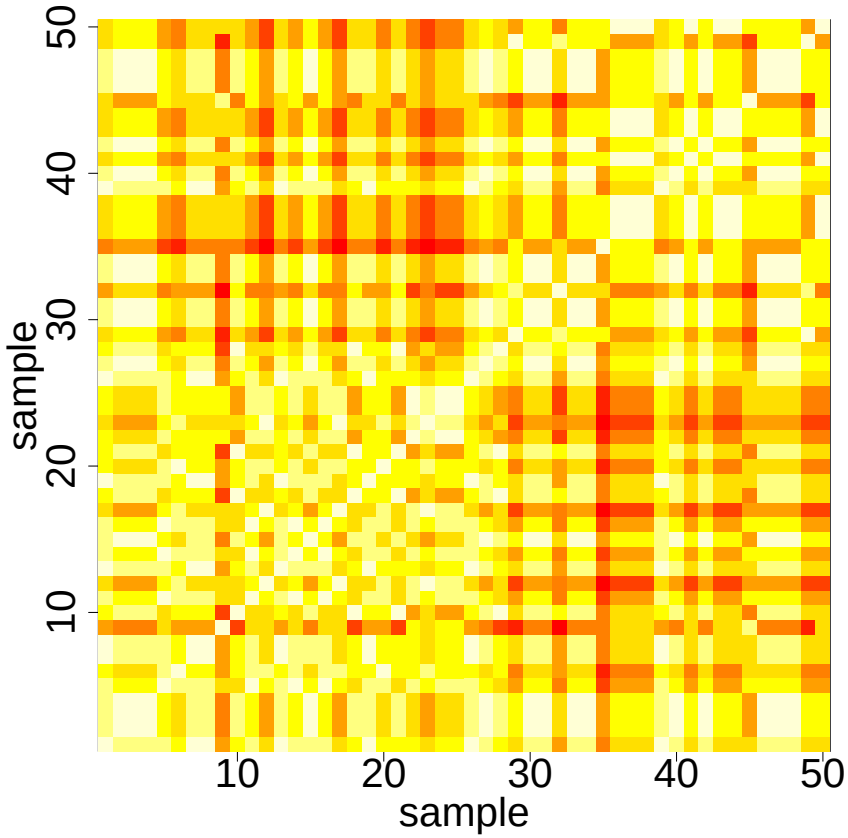
(a) $K_{C_{\mathrm{ds}}}$ **(T=7)**



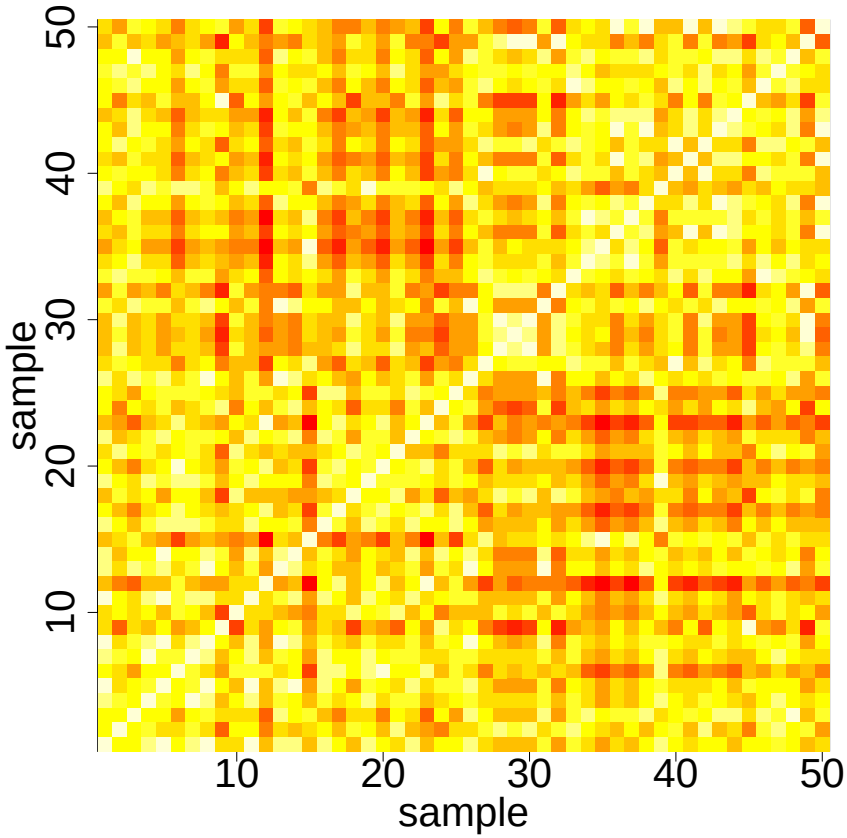
(b) $K_{C_{\mathrm{ds}}}$

カーネル関数のブースティング

Observation of boosting w.r.t. Gram matrix



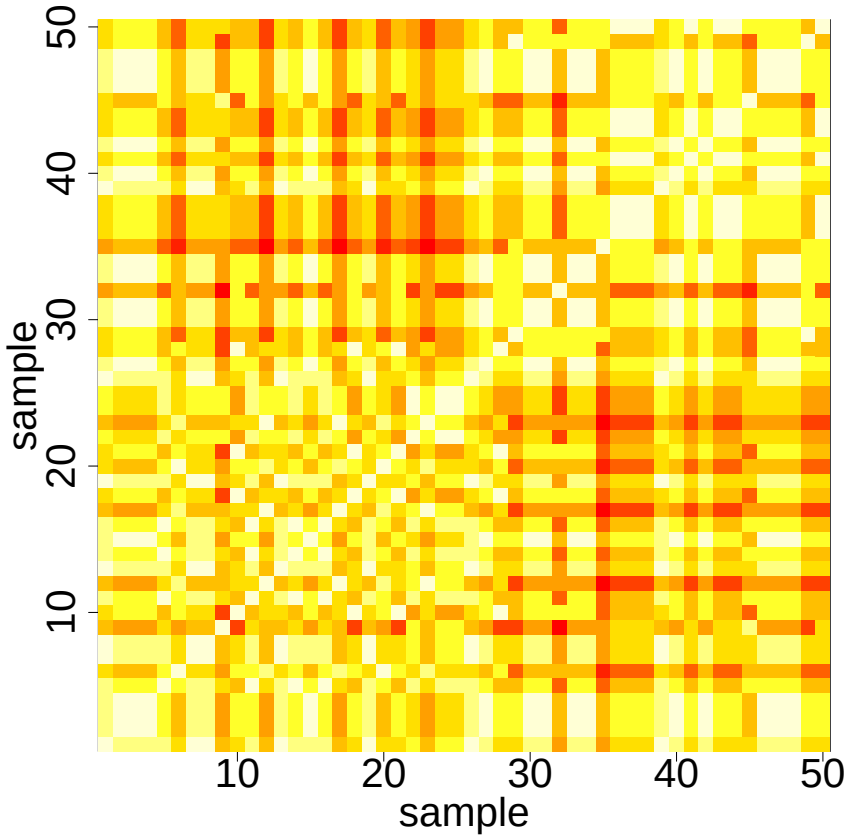
(a) $K_{C_{\mathrm{ds}}}$ **(T=8)**



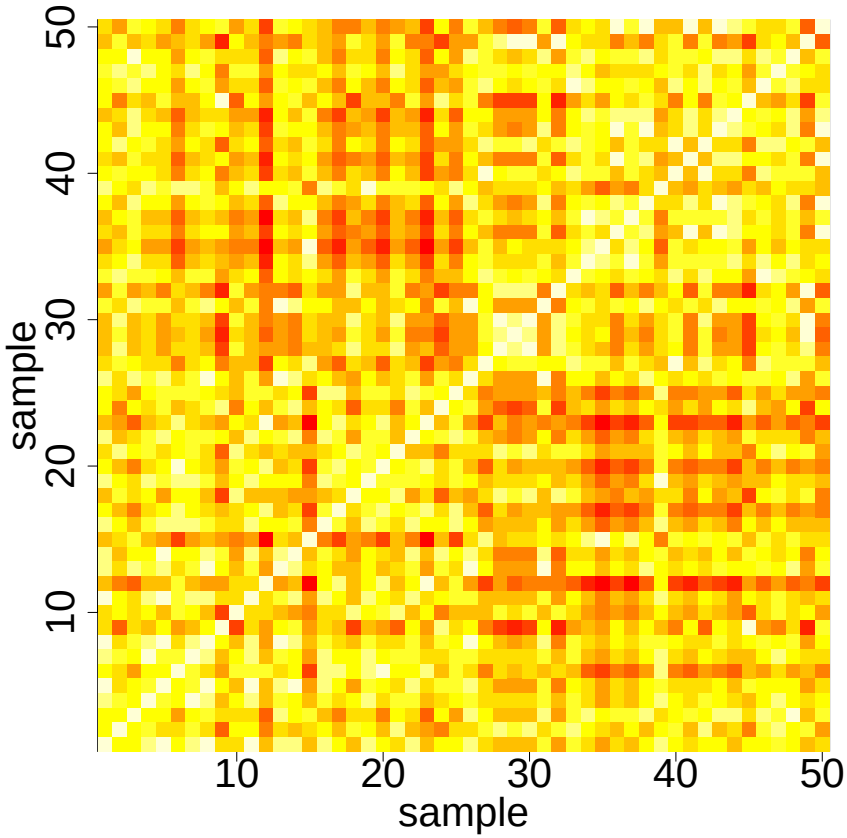
(b) $K_{C_{\mathrm{ds}}}$

カーネル関数のブースティング

Observation of boosting w.r.t. Gram matrix



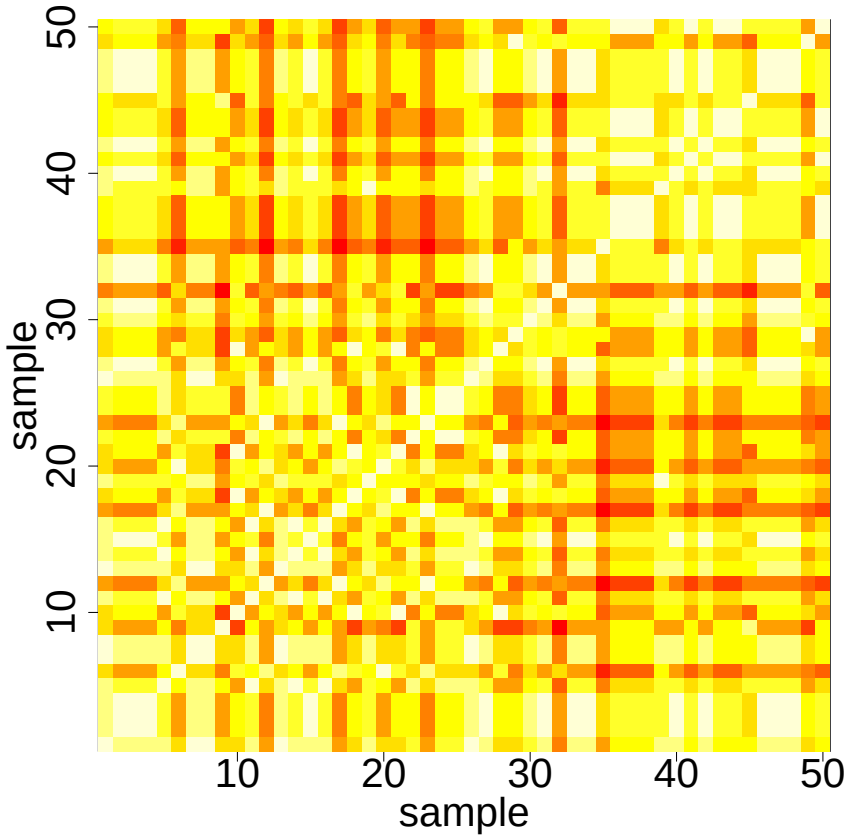
(a) $K_{C_{\mathrm{ds}}}$ **(T=9)**



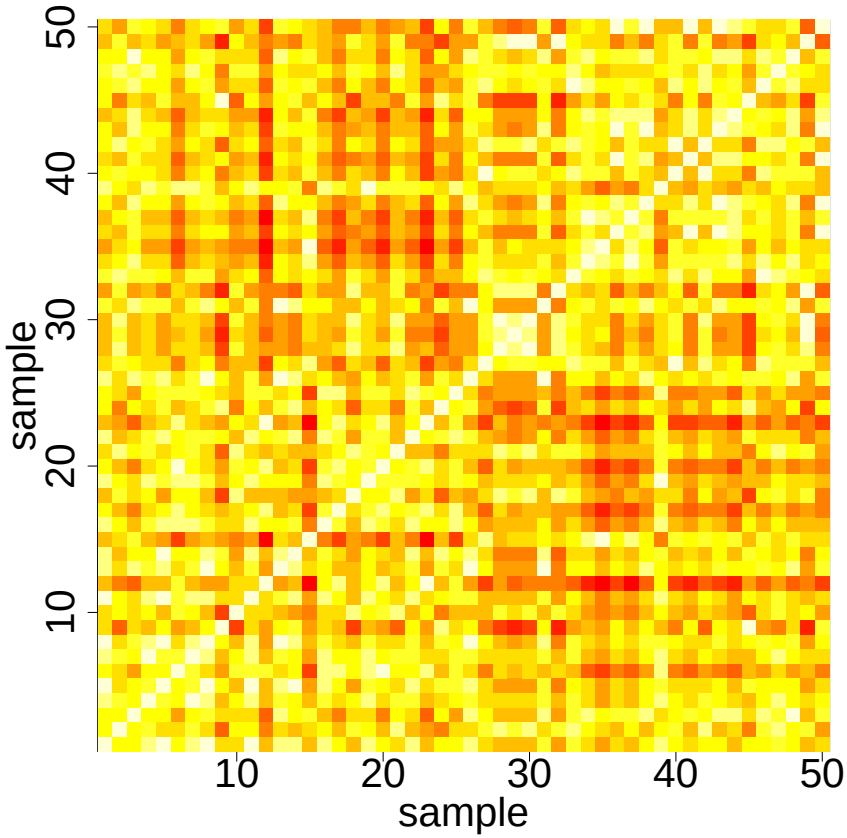
(b) $K_{C_{\mathrm{ds}}}$

カーネル関数のブースティング

Observation of boosting w.r.t. Gram matrix



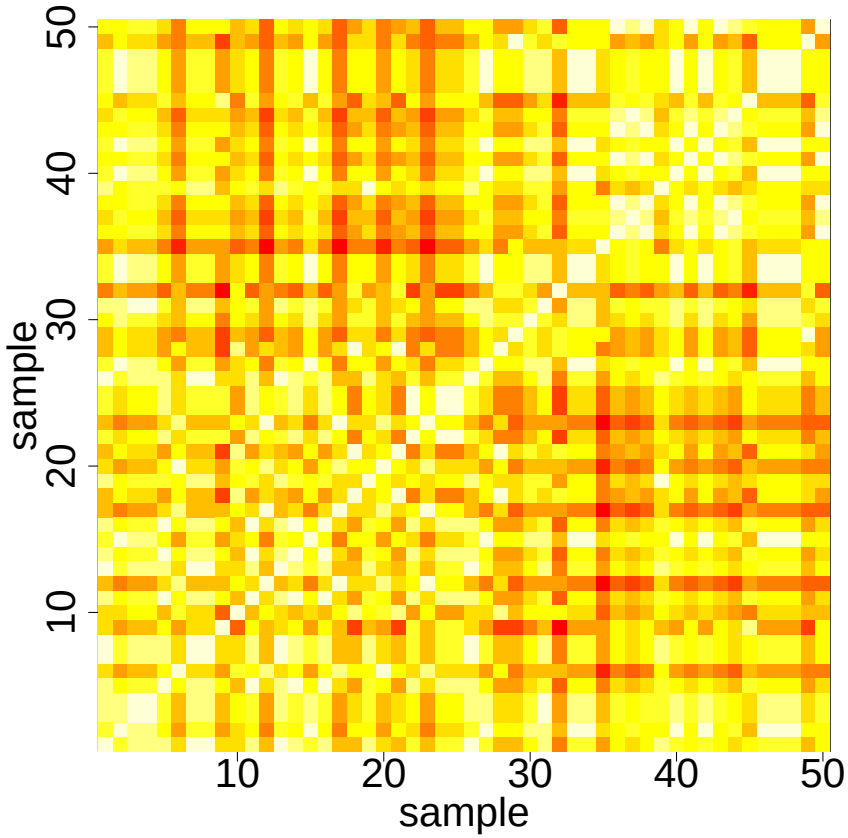
(a) $K_{C_{\mathrm{ds}}}$ **(T=10)**



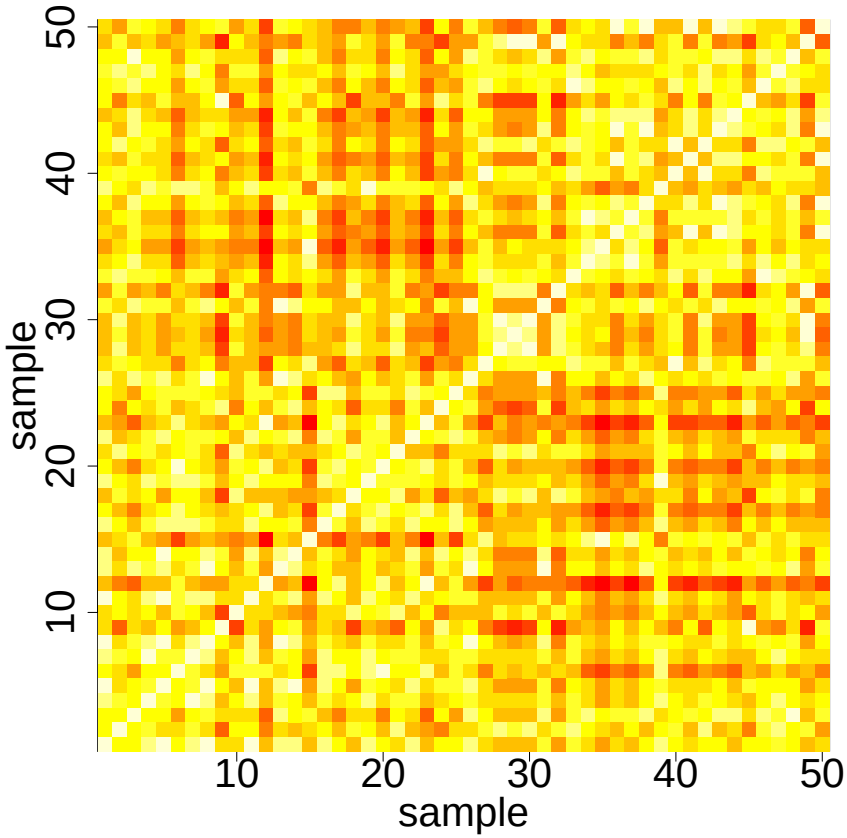
(b) $K_{C_{\mathrm{ds}}}$

カーネル関数のブースティング

Observation of boosting w.r.t. Gram matrix



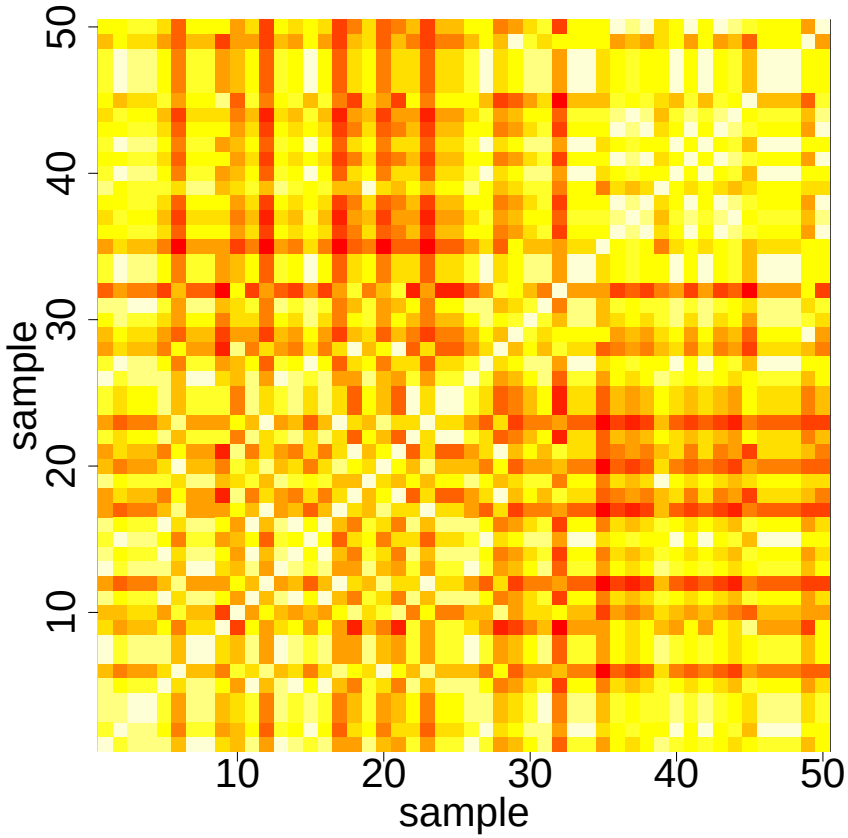
(a) $K_{C_{\mathrm{ds}}}$ **(T=11)**



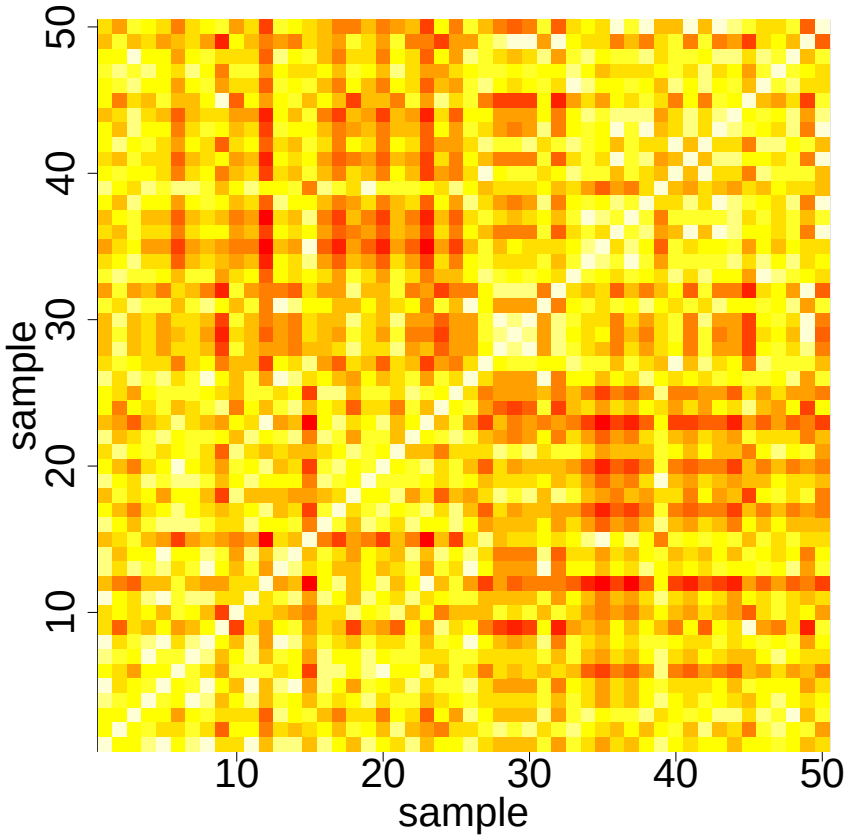
(b) $K_{C_{\mathrm{ds}}}$

カーネル関数のブースティング

Observation of boosting w.r.t. Gram matrix



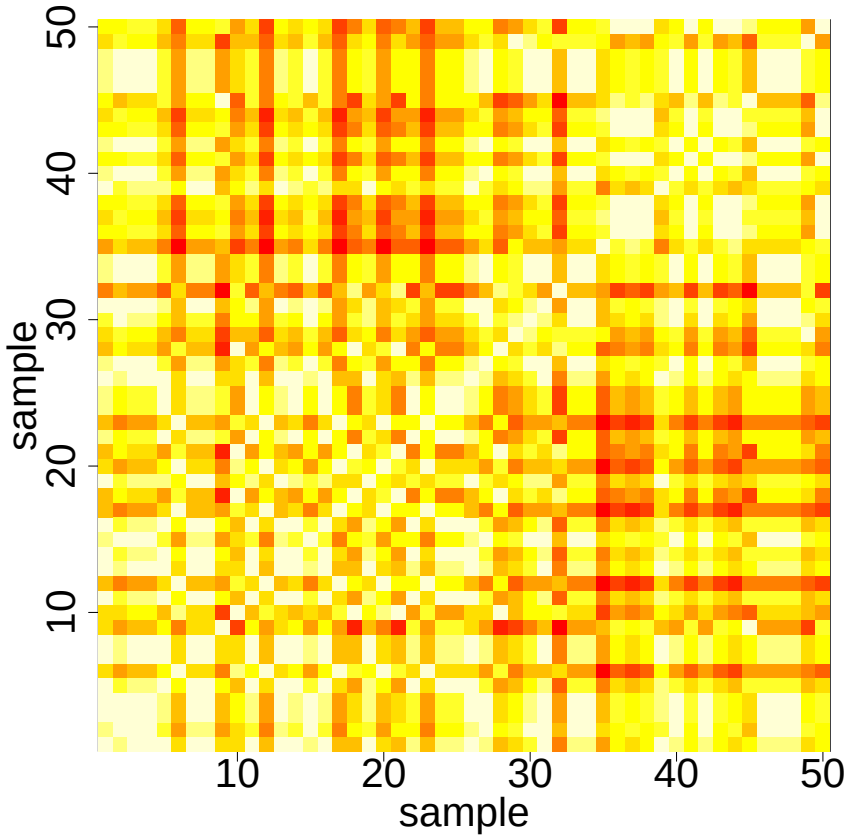
(a) $K_{C_{\mathrm{ds}}}$ **(T=12)**



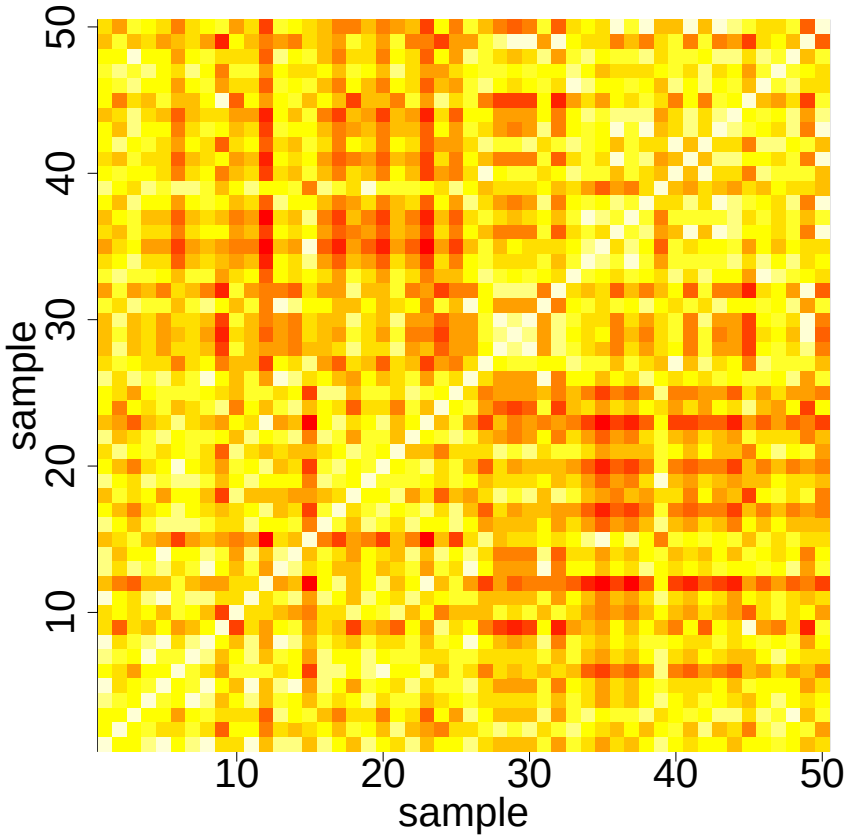
(b) $K_{C_{\mathrm{ds}}}$

カーネル関数のブースティング

Observation of boosting w.r.t. Gram matrix



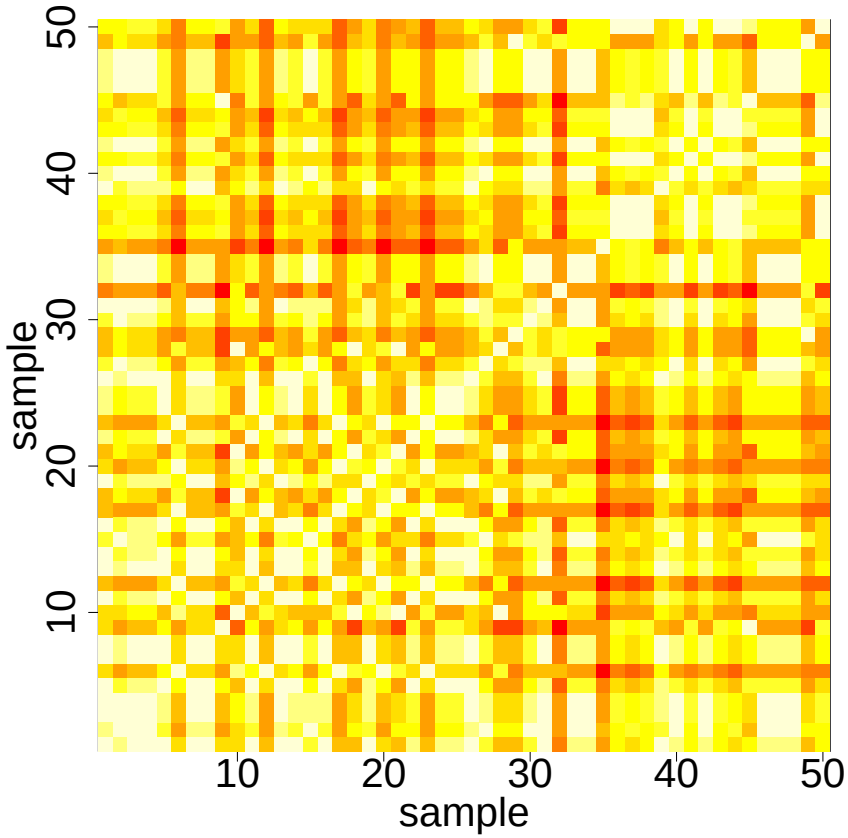
(a) $K_{C_{\mathrm{ds}}}$ (T=13)



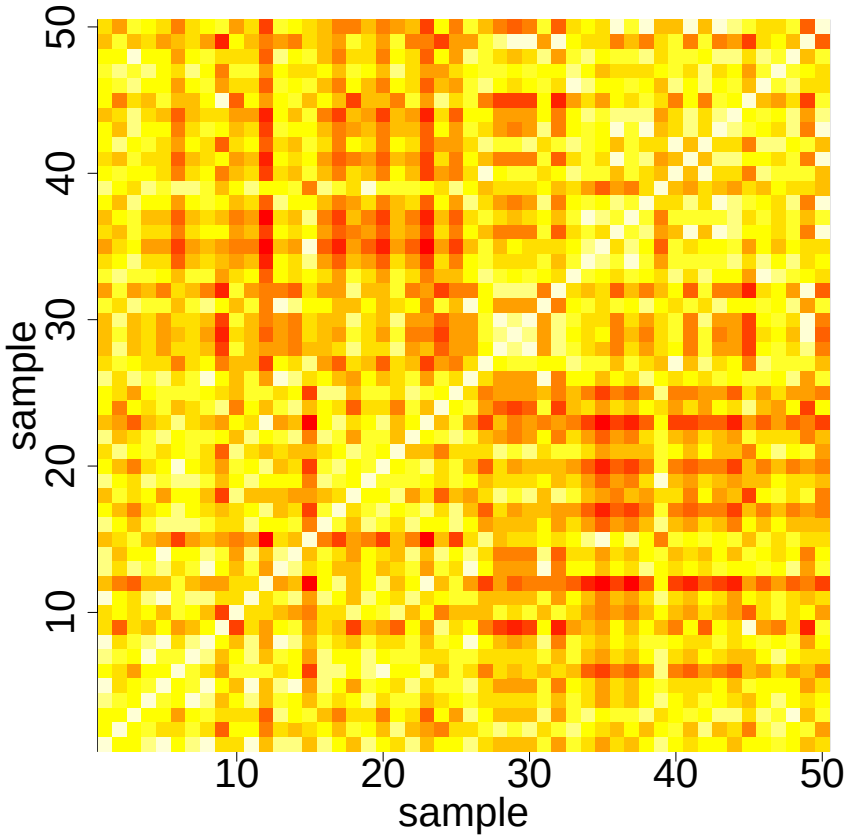
(b) $K_{C_{\mathrm{ds}}}$

カーネル関数のブースティング

Observation of boosting w.r.t. Gram matrix



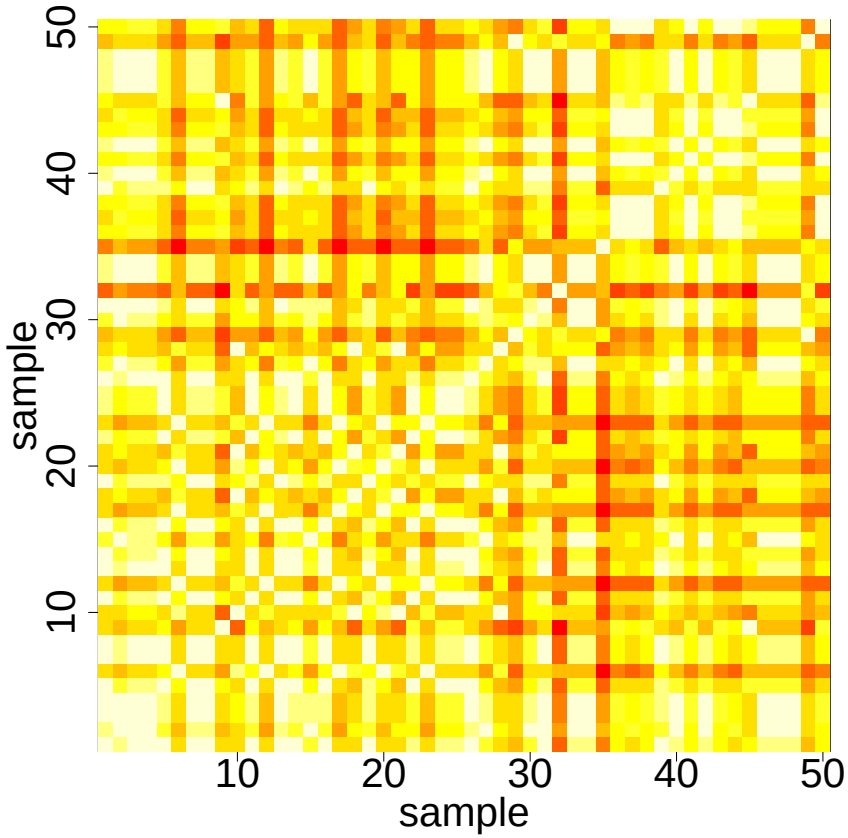
(a) $K_{C_{\mathrm{ds}}}$ **(T=14)**



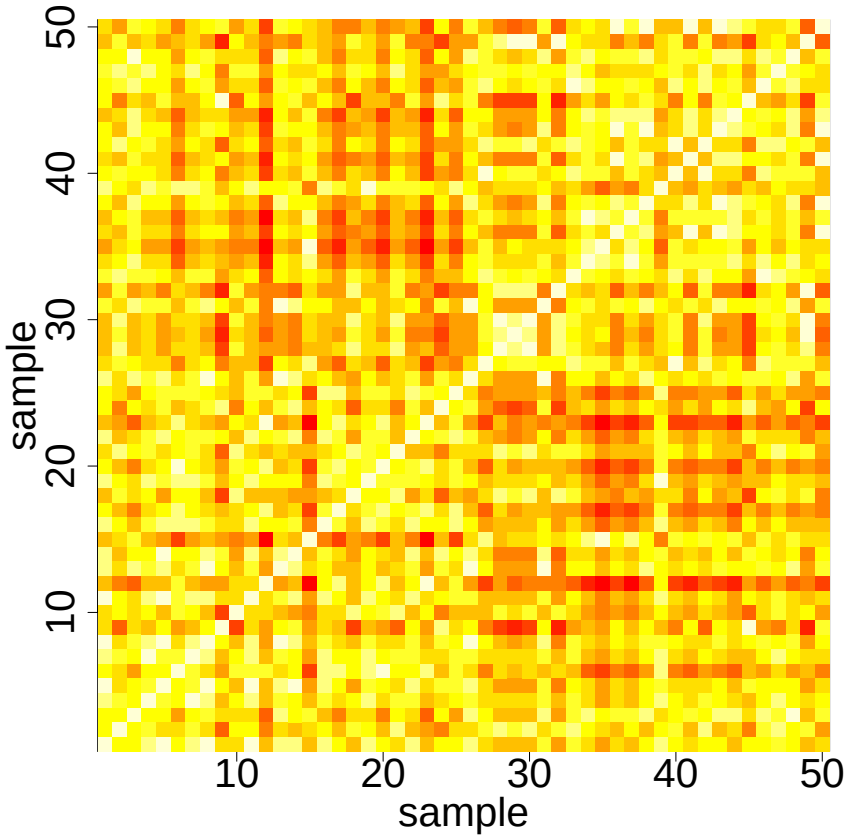
(b) $K_{C_{\mathrm{ds}}}$

カーネル関数のブースティング

Observation of boosting w.r.t. Gram matrix



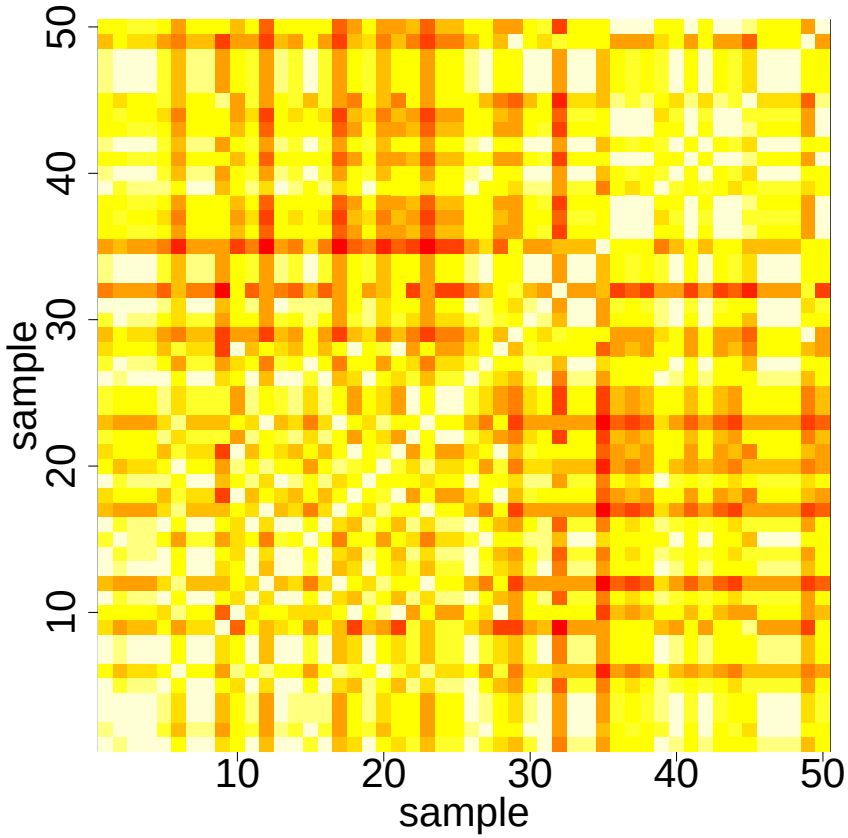
(a) $K_{C_{\mathrm{ds}}}$ **(T=15)**



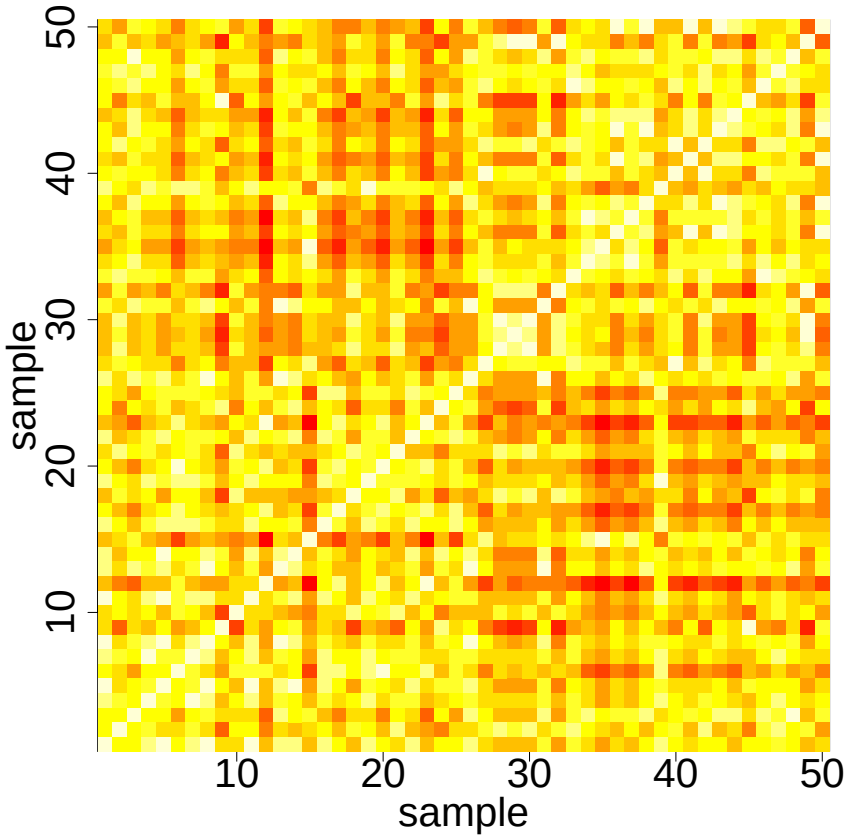
(b) $K_{C_{\mathrm{ds}}}$

カーネル関数のブースティング

Observation of boosting w.r.t. Gram matrix



(a) $K_{C_{\mathrm{ds}}}$ **(T=16)**



(b) $K_{C_{\mathrm{ds}}}$

Boosting Kernel Machine

1. **Fix a smoothing parameter $\lambda > 0$, set $\eta^0 = 0$.**
2. **For $t = 1, 2, \cdots, \tau$, repeat the following process.**
   (a) **Find f by boosting such that**

$$\begin{aligned} f_{j(t)} &= \operatorname*{argmin}_{f \in C} A(F(\cdot;\eta^{t-1}) + \alpha f) \\ \alpha_t &= \operatorname*{argmin}_{\alpha \in R} A(F(\cdot;\eta^{t-1}) + \alpha f_{j(t)}). \end{aligned}$$

   (b) **Set parameter distribution as**

$$\begin{aligned} \pi^t_{t'} &= \sum_{\ell=1}^{n} \pi^{t-1} f_{j(t')}(X_\ell) \eta^{t-1}_\ell / Z \quad (t' < t) \\ &= \alpha_t / Z \quad (t' = t), \end{aligned}$$

   where Z is a normalization constant such that $\sum_{t'} \pi^t_{t'} = 1$.
   (c) **Update kernel function as**

$$k_t(x, x') = \sum_{t'=1}^{t} \pi^t_{t'} f_{j(t')}(x) f_{j(t')}(x').$$

   (d) **Find η^t by regboosting with k_t.**
3. **Finally, we obtain a resultant classifier $g(x) = \mathbf{sign}(\overline{F}(x;\eta^\tau))$.**

Table of Contents

3. ベイズリスク一致性

ベイズリスク一致性の定義

ベイズ判別機 ベイズ判別機 $g^*(x) := \mathbf{argmax}_{y \in \mathcal{Y}} P(Y = y \mid x)$.

リスク $L(g) := P(Y \neq g(X))$.

ベイズリスク $L^* := \inf_g L(g)$. ただし $L^* = L(g^*)$.

ベイズリスク一致性 n 個のデータ D から推定された g_n が

$$L(g_n) \to L^* \ (n \to \infty)$$

ならベイズリスク一致性を持つという.

文献 多くの研究者はある困難のために AdaBoost 自身の一致性は示せなかった (Koltchinskii and Panchenko, 2002; Lugosi and Vayatis, 2004; Breiman, 2004; Jiang, 2004; Bickel et al. (2006)).

Bartlett and Traskin (2007) は自身のテクニック (Bartlett et al., 2006) と Bickel et al. (2006) の結果を用いて「AdaBoost は適切な停止条件を用いればベイズリスク一致性を持つ」ことを示した.

証明の準備

ϕ リスク リスクは凸関数でないため最適化が困難. 大抵狭義凸増加関数 ϕ を用いた ϕ リスクを最小化する.

経験 ϕ リスク $R_{\phi,n}(f) := \frac{1}{n}\sum_{i=1}^{n}\phi(Y_i f(X_i))$, **$\phi$ リスク** $R_\phi(f) := E[\phi(Yf(X)]$

制限された判別関数のクラス

$$\mathcal{F}_\lambda := \{f = \sum \lambda_j f_j \,|\, f_j \in C, \lambda_j \geq 0, \sum_j \lambda_j = \lambda\},$$

$$\mathcal{F}^t := \{f = \sum_{j=1}^{t} \lambda_j f_j \,|\, f_j \in C, \lambda_j \in R\}.$$

t_n ステップ後のブースティングの推定した判別関数を f_{t_n} と書くと $f_{t_n} \in \mathcal{F}^{t_n}$.

証明のための十分条件

ちゅう密性 任意の $p(x,y)$ に対して C が有限の VC 次元を持ち，
次を満たすと仮定 (そのような C は実際存在する)

$$\lim_{\lambda\to\infty} \inf_{f\in\mathcal{F}_\lambda} R_\phi(f) = R_\phi^* \quad \text{ただし } R_\phi^* := \inf_f R_\phi(f)$$

ϕ リスクによる十分条件 任意の判別関数の列 $\{f_n\}$ に対して

$$R_\phi(f_n) \to R_\phi^* \quad \text{ならば} \quad L(f_n) \to L^*$$

よって $R_\phi(f_{t_n}) \to R_\phi^*$ を示せば十分

ϕ リスクバイアス分解 十分に遅い増加列 λ_n $(\lambda_n \to \infty)$ に対して

$$R_\phi(f_n) - R_\phi^* = \underbrace{R_\phi(f_n) - \inf_{f\in\mathcal{F}_{\lambda_n}} R_\phi(f)}_{\text{推定誤差}} + \underbrace{\inf_{f\in\mathcal{F}_{\lambda_n}} R_\phi(f) - R_\phi^*}_{\text{近似誤差}}$$

近似誤差は 0 に概収束するので推定誤差の 0 収束をいえば十分.

従来の推定誤差の収束証明 (理想)

必要な条件 $\exists \{\zeta_n\}, \{t_n\}$ **s.t.** $\zeta_n \to \infty, t_n \to \infty$ **かつ以下を満たす**

参照関数列の存在 $R_\phi(\overline{f}_n) \to R_\phi^*$ **となる** $\overline{f}_n$ **が存在. 具体的には**

$$\overline{f}_n = \operatorname*{argmin}_{f \in \mathcal{F}_{\lambda_n}} R_\phi(f)$$

一様収束 **経験** ϕ **リスクが** ϕ **リスクに収束**

$$|R_\phi(f_{t_n}) - R_{\phi,n}(f_{t_n})| \to^{a.s.} 0.$$

アルゴリズム的収束 **Bickel et al. (2006) により証明**

$$|R_{\phi,n}(f_{t_n}) - R_{\phi,n}(\overline{f}_n)| \to^{a.s.} 0.$$

$\{\overline{f}_n\}$ の ϕ リスクの収束 **経験** ϕ **リスクが** ϕ **リスクに収束**

$$|R_\phi(\overline{f}_n) - R_{\phi,n}(\overline{f}_n)| \to^{a.s.} 0.$$

従来の推定誤差の収束証明失敗

必要な条件 $\exists \{\zeta_n\}, \{t_n\}$ **s.t.** $\zeta_n \to \infty, t_n \to \infty$ **かつ**

参照関数列の存在 $R_\phi(\overline{f}_n) \to R^*_\phi$ **となる** $\overline{f}_n$ **が存在. 具体的には**

$$\overline{f}_n = \underset{f \in \mathcal{F}_{\lambda_n}}{\operatorname{argmin}} R_\phi(f)$$

一様収束しない **経験** ϕ **リスクが** ϕ **リスクに収束しない**

$$|R_\phi(f_{t_n}) - R_{\phi,n}(f_{t_n})| \to^{a.s.} 0.$$

アルゴリズム的収束 **Bickel et al. (2006) により証明**

$$|R_{\phi,n}(f_{t_n}) - R_{\phi,n}(\overline{f}_n)| \to^{a.s.} 0.$$

$\{\overline{f}_n\}$ の ϕ リスクの収束 **経験** ϕ **リスクが** ϕ **リスクに収束**

$$|R_\phi(\overline{f}_n) - R_{\phi,n}(\overline{f}_n)| \to^{a.s.} 0.$$

証明に用いる重要な補題

Hoeffding-Aduma(McDiarmid) の不等式 **Let** $\{X_i\}_{i=1}^n$ **be random variables. Assume that a function** g **satisfies the bounded difference condition. Then, for all** $\epsilon > 0$,

$$P(g(X_1, X_2, \cdots, X_n) - E[g(X_1, X_2, \cdots, X_n)] \leq \epsilon) \leq \exp(-2\epsilon^2 / \sum_{i=1}^n c_i^2),$$

$$P(g(X_1, X_2, \cdots, X_n) - E[g(X_1, X_2, \cdots, X_n)] \geq \epsilon) \leq \exp(-2\epsilon^2 / \sum_{i=1}^n c_i^2).$$

Contraction principle **[Ledoux and Talagrand 1991]** **Let** $F : R_+ \to R_+$ **be convex and increasing. Let further** $\psi_i : R \to R$ $(i = 1, 2, \cdots, n)$ **be contractions such that** $\psi_i(0) = 0$. **Then, for any bounded subset** $T \subset R^n$,

$$EF\left(\frac{1}{2}\sup_{t\in T}\left|\sum_{i=1}^n \sigma_i\psi_i(t_i)\right|\right) \leq EF\left(\sup_{t\in T}\left|\sum_{i=1}^n \sigma_i t_i\right|\right).$$

補題のための定義

Definition 2 (Bounded difference condition). *Let A be some set and g be a function mapping A^n to R. We say that g satisfies the bounded difference condition if there exists $\{c_i\}_{i=1}^n$ such that*

$$\sup_{x_1,x_2,\cdots,x_n,x_i' \in A} |g(x_1, x_2, \cdots, x_n) - g(x_1, \cdots, x_{i-1}, x_i', x_{i+1}, \cdots, x_n)| \leq c_i.$$

Definition 3 (Contraction). *Let ψ be a function mapping from R to R. We say that ψ is a contraction if $|\psi(s) - \psi(t) \leq |s - t|$ for any $s, t \in R$.*

従来の推定誤差の収束証明 (少し妥協)

必要な条件 $\exists \{\zeta_n\}, \{t_n\}$ **s.t.** $\zeta_n \to \infty,\ t_n \to \infty$ **かつ**
参照関数列の存在 $R_\phi(\overline{f}_n) \to R_\phi^*$ **となる** $\overline{f}_n$ **が存在. 具体的には**

$$\overline{f}_n = \underset{\|f\|_* \le \lambda_n}{\mathbf{argmin}}\, R_\phi(f)$$

一様収束 $f_{t_n} := \mathbf{argmin}_{f \in \mathcal{F}_{\lambda_n}} R_{\phi,n}(f)$

$$|R_\phi(f_{t_n}) - R_{\phi,n}(f_{t_n})| \to^{a.s.} 0.$$

アルゴリズム的収束 **Bickel et al. (2006) により証明**

$$|R_{\phi,n}(f_{t_n}) - R_{\phi,n}(\overline{f}_n)| \to^{a.s.} 0.$$

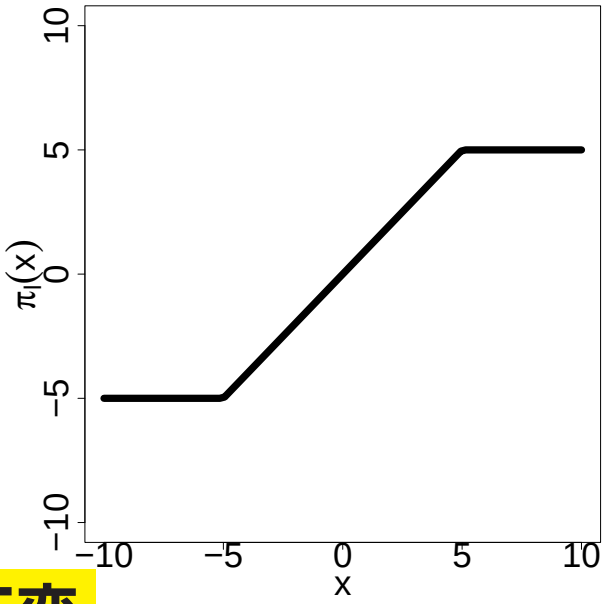
$\{\overline{f}_n\}$ **の** ϕ **リスクの収束** **経験** ϕ **リスクが** ϕ **リスクに収束**

$$|R_\phi(\overline{f}_n) - R_{\phi,n}(\overline{f}_n)| \to^{a.s.} 0.$$

妥協しない為のキーアイデア

クリッピングの導入

$$\pi_\ell(x) = \begin{cases} \ell & (\ell < x) \\ x & (-\ell \le x \le \ell) \\ -\ell & (x < -\ell) \end{cases}$$



リスクはクリッピングについて不変

- 任意の判別関数 $f : \mathcal{X} \to R$ について当然 $\pi_\ell \circ f$ は有界
- 任意の $\ell > 0$ と判別関数について $\mathbf{sign}(\pi_\ell \circ f) = \mathbf{sign}(f)$ より

$$L(\pi \circ f) = P(\mathbf{sign}(\pi_\ell \circ f(X)) \neq Y) = P(\mathbf{sign}(f(X)) \neq Y) = L(f).$$

新 推定誤差の収束証明

必要な条件 $\exists\{\zeta_n\}, \{t_n\}$ **s.t.** $\zeta_n \to \infty,\ t_n \to \infty$ **かつ**

参照関数列の存在 $R_\phi(\bar{f}_n) \to R^*_\phi$ **となる** $\overline{f}_n$ **が存在. 具体的には**

$$\overline{f}_n = \operatorname*{argmin}_{\|f\|_* \le \lambda_n} R_\phi(f)$$

一様収束 **経験** ϕ **リスクと期待** ϕ **リスク**

$$\sup_{f \in \pi_{\zeta_n} \circ \mathcal{F}^{t_n}} |R_\phi(f) - R_{\phi,n}(f)| \to^{a.s.} 0.$$

この収束を保証するため十分遅い増加列 $\{t_n\}$ **が必要**

アルゴリズム的収束 **Bickel et al. (2006) により証明**

$$|R_{\phi,n}(f_{t_n}) - R_{\phi,n}(\overline{f}_n)| \to^{a.s.} 0.$$

$\{\overline{f}_n\}$ **の** ϕ **リスクの収束** **経験** ϕ **リスクと期待** ϕ **リスク**

$$|R_\phi(\overline{f}_n) - R_{\phi,n}(\overline{f}_n)| \to^{a.s.} 0.$$

ベイズリスク一致性の証明

定理 今までの条件に加えて ϕ の凸性を仮定. $\phi_\lambda := \inf_{x\in[-\lambda,\lambda]} \phi(x)$ と定義すると一般性を失わずに $\lim_{\lambda\to\infty} \phi_\lambda = 0$ と仮定.

証明 ある数列 ϵ_n^j $(j = 1, 2, 3)$ が存在して $\epsilon_n^j \to 0$ かつ次を満たす

$$\begin{aligned} R_\phi(\pi_{\zeta_n} \circ f_{t_n}) &\leq R_{\phi,n}(\pi_{\zeta_n} \circ f_{t_n}) + \epsilon_n^1 \\ &\leq R_{\phi,n}(f_{t_n}) + \phi_{\zeta_n} + \epsilon_n^1 \\ &\leq R_{\phi,n}(\overline{f}_n) + \phi_{\zeta_n} + \epsilon_n^1 + \epsilon_n^2 \\ &\leq R_\phi(\overline{f}_n) + \phi_{\zeta_n} + \epsilon_n^1 + \epsilon_n^2 + \epsilon_n^3. \end{aligned}$$

故に $R_\phi(\pi_{\zeta_n} \circ f_{t_n}) \to R_\phi^*$. よって $L(\pi_{\zeta_n} \circ f_{t_n}) \to L^*$. これは $L(f_{t_n}) \to L^*$ を表している.

Table of Contents

4. 近似誤差の改善について

背景

近似誤差が 0 にならなければベイズリスク一致性は得られない.
近似誤差の改善のためには複雑な弱学習機の使用が考えられる.

しかし必要以上に複雑な弱学習機の使用は解釈性の低下を招くだけでなく, 判別精度の意味でも望ましくない.

研究紹介

4.1 局所ブースティング法の提案とベイズリスク一致性の証明.
局所的に解釈性を保持しつつ弱学習機を変更無しに近似誤差を改善　(今回は省略)

4.2 弱学習機モデルが既に真の分布を含む場合はブースティングすると却って改悪する事を示す

4.2 ブースティングによる改悪について

背景

- モデルが正しいとき plug-in 分布は m 埋め込み曲率方向へ曲率分だけシフトすることで期待 KL を改良できる. ベイズ予測分布はそれを達成する (Komaki, 1996).

研究成果 ブースティングを予測分布構成法とみなして解析

- 拡張空間 (規格化されていない) で曲 u モデルの plug-in 分布を U 曲率方向に U 測地線に沿って U 埋め込み曲率に応じてシフトするのが期待 U-divergence を最も改良する.
- 曲指数型分布族を弱学習機モデルとしたとき, **弱学習機モデルが正しい仮定の下**ではブースティングは plug-in 分布を平均的には「最適なシフトの指数型分布の空間への U 射影」と同じ分だけ逆向きに**改悪**している.

AdaBoost (Freund and Schapire, 1997)

幾何的解釈 (Lebanon and Lafferty, 2002)

---

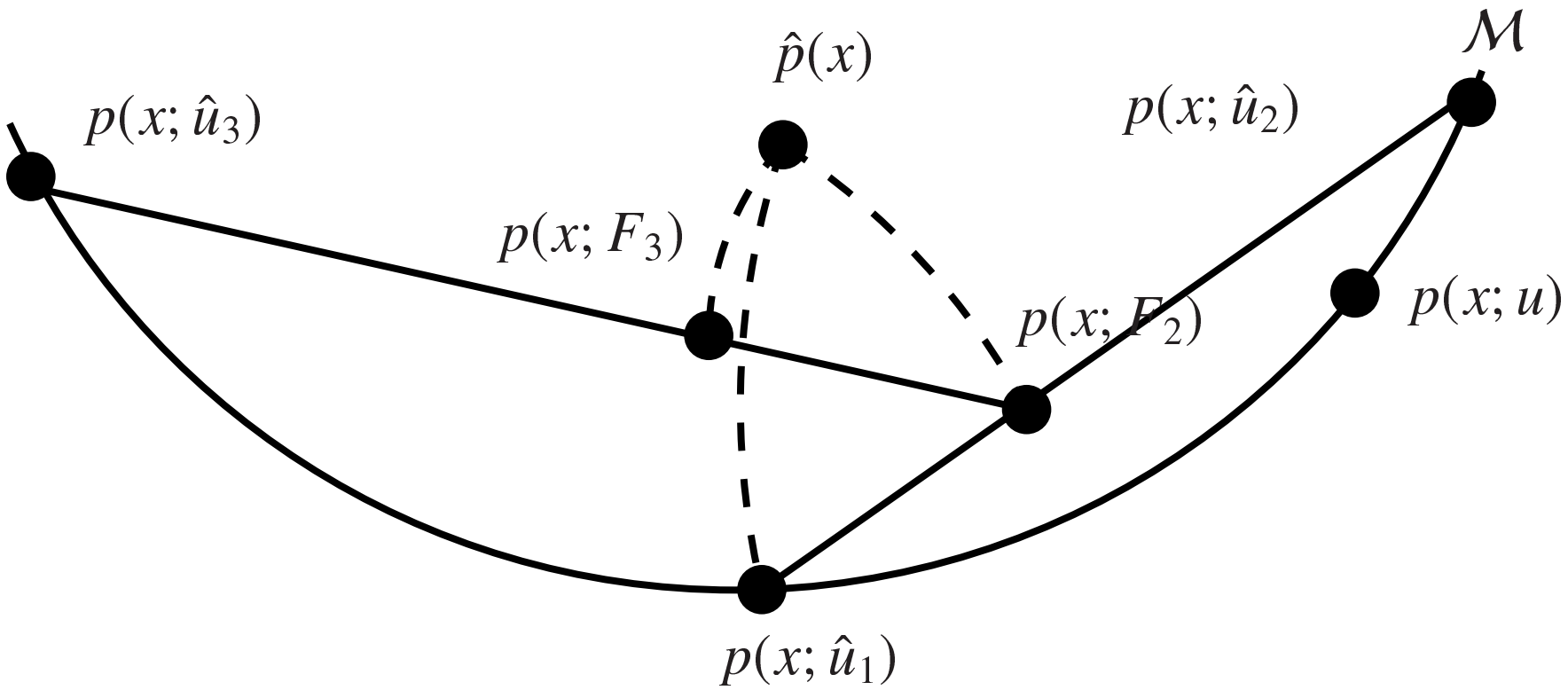
1. **Initialize a discriminant function** F **as** $F_0(x) \equiv 0$.
2. **For** $t = 1, 2, \cdots, T$**, do the followings:**
   (a) **Find** $\tau_t = \mathbf{argmin}_\tau \mathcal{D}(\hat{p}_D, p(\cdot; F_{t-1} + \theta f(\cdot, \tau)))$.
   (b) **Find** $\theta_t = \mathbf{argmin}_{\theta \geq 0} \mathcal{D}(\hat{p}_D, p(\cdot; F_{t-1} + \theta f(\cdot, \tau_t)))$
   (c) **Update** F_{t-1} **as** $F_t = F_{t-1}(x) + \theta_t f(x; \tau_t)$.

ここで拡張 KL-divergence $\mathcal{D}(p, q)$ **は任意の正値測度** p, q **について下記のように定義する**

$$\mathcal{D}(p, q) := \int_x \sum_{y \in \mathcal{Y}} p(x) \left\{ q(y \mid x) - p(y \mid x) - p(y \mid x) \frac{\log q(y \mid x)}{\log p(y \mid x)} \right\} dx.$$

ブースティングの幾何学的解釈

Murata et al. (2004) の幾何学的解釈



拡張空間 (Murata et al., 2004)

U-divergence 以下では $u := U', \xi = u^{-1}$ と記す.

$$\mathcal{D}_U(\rho_1, \rho_2) := \int_{\mathcal{X}} U(\xi(\rho_2(x))) - U(\xi(\rho_1(x))) - \rho_1(x)(\xi(\rho_2(x)) - \xi(\rho_1(x)))dx.$$

拡張空間 $\mathcal{F}$

$$\mathcal{F} := \left\{ \rho(x) \,|\, \forall x, \rho(x) \geq 0, 0 < Z(\rho) < \infty \text{ and } , \int_{\mathcal{X}} p(x)\xi(\rho(x))dx = 0. \right\}$$

ただし $Z(\rho) := \int \rho(x)dx$.

接空間の定義

$$T(\rho) := \left\{ v(x) \,|\, \int p(x)\xi'(\rho(x))v(x)dx = 0 \right\}.$$

拡張空間の幾何量 江口トリック (Eguchi, 1992)

幾何量	U 表現
$\rho(x;\tau)$	$\rho(x;\tau)$
接空間	$\mathcal{T}^U(\rho(x;\tau)) := \{v(x) \in \mathcal{F} \mid \int p(x)\xi'(\rho(x;\tau))v(x)dx = 0\}$
内積	$< v_1, v_2 >_U := \int \xi'(\rho)v_1(x)v_2(x)dx$
U 共変微分	$\nabla_{\partial_a\rho}{}^U \partial_b\rho := \partial_a\partial_b\rho(x)$
U^* 共変微分	$\nabla_{\partial_a\rho}{}^{U^*} \partial_b\rho := \frac{\partial_a\partial_b\xi(\rho)}{\xi'(\rho)}$
U 接続	$\Gamma_{abc}{}^U := < \nabla_{ab}{}^U, \partial_c\rho >_U$
U^* 接続	$\Gamma_{abc}{}^{U^*} := < \nabla_{\partial_a\rho}{}^{U^*} \partial_b\rho, \partial_c\rho >_U$

拡張空間の幾何量 江口トリック (Eguchi, 1992)

幾何量	U^* 表現
$\rho(x;\tau)$	$\xi(\rho(x;\tau))$
接空間	$\mathcal{T}^{U^*}(\rho(x;\tau)) := \{v(x) \in \mathcal{F} \mid \int p(x)v(x)dx = 0\}$
内積	$< v_1, v_2 >_{U^*} := \int u'(\xi(\rho))v_1(x)v_2(x)dx$
U 共変微分	$\nabla_{\partial_a \xi(\rho)}{}^{U} \partial_b \xi(\rho) := \frac{\partial_a \partial_b \rho}{u'(\xi(\rho))}$
U^* 共変微分	$\nabla_{\partial_a \xi(\rho)}{}^{U^*} \partial_b \xi(\rho) := \partial_a \partial_b \xi(\rho(x))$
U 接続	$\Gamma_{abc}{}^{U} :=< \nabla_{\partial_a \xi(\rho)}{}^{U} \partial_b \xi(\rho), \partial_c \xi(\rho) >_{U^*}$
U^* 接続	$\Gamma_{abc}{}^{U^*} :=< \nabla_{\partial_a \xi(\rho)}{}^{U^*} \partial_b \xi(\rho), \partial_c \xi(\rho) >_{U^*}$

考慮するモデル

u モデル $\mathcal{E}$ $\mathcal{D}_U$ **双対平坦 (双対座標系** θ **と** η**)**

$$\mathcal{E} := \left\{ \bar{\rho}(x;\theta) = u(\theta^i T_i(x) - \phi(\theta)) \,,\ \phi(\theta) := \theta^i \int p(x) T_i(x) dx \right\}.$$

曲 u モデル $\mathcal{P} \subset \mathcal{E}$ 推定に用いるモデル

$$\mathcal{P} := \left\{ \rho(x;\tau) = u(\theta^i(\tau) T_i(x) - \phi(\theta(\tau))) \,,\ \theta^i(\tau) : R^m \to R^n \right\}.$$

推定量 推定量は $\hat{\tau} := \mathbf{argmin}_{\tau'} \mathcal{D}_U(\hat{p}(x), \rho(x,\tau'))$ **を用いる. このとき推定部分多様体** $A(\tau) := \{\eta \,|\, \hat{\tau}(\eta) = \tau\}$ **はモデル** $\mathcal{P}$ **に直交する**

(証明略) 推定部分多様体の座標系を v^κ **として** $w^\alpha = (\tau^a, v^\kappa)$ **を** $\mathcal{E}$ **の新たな座標系と定義する.**

U 埋め込み曲率方向へのシフト

U 埋め込み曲率方向 h_{ab}

$$h_{ab}(x;\tau) := \partial_a\partial_b\rho(x;\tau) - \overset{U}{\Gamma}_{abc}\, g^{cd}\partial_d\rho(x;\tau)$$

拡張モデル $h_I(x;\tau)(I = m+1, \cdots, M)$ **を任意の** $T(\rho(x;\tau))$ **の元で** h_{ab} **が張る空間を含む空間を張る関数とする. このとき拡張モデルを**

$$\tilde{\rho}(x;\tau,n) := u\left(\xi(\rho(x;\tau)) + \frac{n^I h_I(x;\tau)}{\rho(x;\tau)} - \sum_I \left(\frac{n^I h_I(x;\tau)}{\rho(x;\tau)}\right)^2 - \psi_\tau(n)\right).$$

と定める. 後で示されるように $n = O(1/N)$ **より,** $o_p(1/N)$ **を無視すると** m **測地線に沿って** $h_I \in T(\rho)$ **方向に** $\mathcal{P}$ **から飛び出したモデル**

plug-in 分布の最適な改良

Theorem 4. 真の分布を $\rho(x;\tau_0)$ と仮定する. $\mathcal{P}$ の *plug-in* 分布を期待 *U-divergence* の意味で最も改良 *(*二次のオーダーで*)* するのは以下のようにシフトすれば十分

$$\hat{n}^I_{opt} = \frac{Z(\rho(x;\tau_0))^2}{N} \overset{U}{H}_{ab}{}^{I} B^j_{a'} B^k_{b'} \tilde{g}_{jk} g^{aa'} g^{bb'}$$

$$\text{ただし } \tilde{g}_{jk} := \int p(x)\partial_j\xi(\rho)\partial_k\xi(\rho)dx\,,\ B^j_\alpha := \partial_\alpha\theta^j$$

$$= \frac{Z}{N} \overset{U}{H}_{ab}{}^{I} g^{ab} \quad (when\ U = \exp).$$

- **ただし U 共変微分は常に m 共変微分なのでシフト方向は常に m 埋め込み曲率方向で十分. しかし計量は U に応じて変わる為, 接続係数の値などは見かけ上 U に依存する.**

ブースティングによるシフト

幾何学的考察 ブースティングも弱学習機モデルを U^* ミクスチャによりシフトしているとみなせる. 前スライドで求めた最適なシフトとの関係は?

- 最適なシフトは経験分布 $\hat{p}$ に依存せずに定まる. 一方でブースティングのシフトは $\hat{p}$ にまっしぐらなので依存する
- U 埋め込み曲率方向は一般に $\mathcal{E}$ に含まれるとは限らない. 一方で曲 u モデルを弱学習機とするとブースティングは $\mathcal{E}$ から飛び出せないため**一般にブースティングは最適なシフトは絶対に達成できない.**
- では全く関連がないのか? **そうではない**

ブースティングによるシフト

幾何学的考察 ブースティングも弱学習機モデルを U^* ミクスチャによりシフトしているとみなせる. 前スライドで求めた最適なシフトとの関係は?

- 最適なシフトは経験分布 $\hat{p}$ に依存せずに定まる. 一方でブースティングのシフトは $\hat{p}$ にまっしぐらなので依存する
- U 埋め込み曲率方向は一般に $\mathcal{E}$ に含まれるとは限らない. 一方で曲 u モデルを弱学習機とするとブースティングは $\mathcal{E}$ から飛び出せないため**一般にブースティングは最適なシフトは絶対に達成できない.**
- では全く関連がないのか? **そうではない** (んですが...)

ブースティングと最適シフトの比較 (1/2)

定理

Theorem 5. U を exp に制限する. 前定理より曲 u モデル $\mathcal{P}$ の *plug-in* 分布の最適なシフトが定まる. このシフトを $\mathcal{E}$ に U 射影した量とブースティングによるシフトの期待値は丁度反対になる *(*ただし規格化されてない空間への拡張によるおつりを除く*)*.

- **今回の設定ではブースティングの推定量はモデル $\mathcal{E}$ の最尤推定値だからブースティング予測分布と plug-in 分布の差は $\mathcal{E}$ と $\mathcal{P}$ のパラメータのバイアスの差となっている. その差は U 埋め込み曲率分である.**
- **一方で最適なシフトは全空間に対する $\mathcal{P}$ の U 埋め込み曲率に対応している. これは $\mathcal{E}$ に U 射影すれば $\mathcal{E}$ に対する U 埋め込み曲率に帰着される**

ブースティングと最適シフトの比較 (2/2)

最適なシフトの $\mathcal{E}$ への U 射影 射影先を $\check{\theta}$ と書く.

$$\begin{aligned}
\eta_i(\check{\theta}) &= \int \tilde{\rho}(x;\hat{\tau},n)\partial_i\xi(\bar{\rho})dx, \\
&= \eta_i(\theta(\hat{\tau})) + n^I \int h_I(x;\hat{\tau})\partial_i\xi(\bar{\rho})dx + o_p(1/N), \\
\eta_i(\check{\theta}) - \eta_i(\hat{\tau}) &= \frac{Z}{2} g^{ab}\partial_i w^\kappa \overset{U}{H}_{ab\kappa} + o_p(1/N).
\end{aligned}$$

ブースティングのシフトの期待値

$$\eta(\hat{\theta}) - \eta(\hat{\tau}) = -\frac{1}{N} g^{ab}\partial_a\eta_i\partial_b Z - \frac{Z}{2N}\overset{U}{H}_{ab\kappa}\, g^{ab}\partial_i w^\kappa + o_p(1/N).$$

ブースティングと最適シフトの比較 (2/2)

最適なシフトの $\mathcal{E}$ への U 射影 射影先を $\check{\theta}$ と書く.

$$\begin{aligned}
\eta_i(\check{\theta}) &= \int \tilde{\rho}(x;\hat{\tau},n)\partial_i\xi(\bar{\rho})dx, \\
&= \eta_i(\theta(\hat{\tau})) + n^I \int h_I(x;\hat{\tau})\partial_i\xi(\bar{\rho})dx + o_p(1/N), \\
\eta_i(\check{\theta}) - \eta_i(\hat{\tau}) &= \frac{Z}{2} g^{ab}\partial_i w^\kappa \overset{U}{H}_{ab\kappa} + o_p(1/N).
\end{aligned}$$

ブースティングのシフトの期待値

$$\eta(\hat{\theta}) - \eta(\hat{\tau}) = -\frac{1}{N} g^{ab}\partial_a\eta_i\partial_b Z - \frac{Z}{2N} \overset{U}{H}_{ab\kappa}\, g^{ab}\partial_i w^\kappa + o_p(1/N).$$

考察

- モデルが正しい事を知っているならブースティングのシフトを逆転すればよい. または凸結合ブースティングが望ましい. しかし有意義ではない.
- 確かにブースティングは最適な飛び出し方向へは完全には飛び出せない. しかし U 埋め込み曲率方向も弱学習機として加えれば解決できる.
- 同様に逆向きの $\mathcal{D}_U(\rho, \hat{p})$ を最も改良するシフトなら少なくともブースティングも最適なシフトを達成するポテンシャルはある.

今後の展望

- 曲 u モデルから一般のモデルへの拡張が望ましい 局所指数族バンドルならぬ局所 u 族バンドル
- 現在得られている最適なシフトは経験分布に依らずにシフトしている. 何らかの方法で真からの距離 (または曲率の大きさ) に応じてシフト量を調節できないか?
- ブースティングは通常弱学習機モデルが間違っている場合を想定しているので, この場合の評価が目標. ブースティングとベイズ, バギング予測分布との比較
- カーネルマシンについて最適な改良を考える. Kawakita et al. (2006) で示されたように連続弱学習機カーネルを用いたカーネルマシンはブースティングと同様モデルからの飛び出しを行なっている. 最適な飛び出しがカーネルのパラメータを決定する基準となりうる可能性がある.

まとめ

- ブースティングは提案されて以来様々な解釈と拡張が行なわれてきた. その結果統計的な解釈, 幾何学的な解釈, カーネルマシンとの関連性などが明らかにされた. またベイズリスク一致性が証明されている.

	カーネル法	ブースティング
ロスの選択	任意	任意
統計モデルの選択	カーネルの選択	弱学習機の選択
アルゴリズムの恣意性	カーネルの選択	弱学習機と学習停止条件
パラメータ推定	二次計画法など	関数勾配降下法
解釈性	やや弱い	適切な弱学習機のもと強い
実用性	ややプロ向け	アマチュアでも使い安い

まとめ (2/2)

- 発表者の感覚ではブースティングの良い性質は弱いモデルからスタートして, 強力な判別機を作る過程が利用可能である, すなわち分解能が高いことに由来しているように考えられる. この性質を活かす為にはアルゴリズムだけではなく停止条件と弱学習機の選択が重要である. 弱学習機が必要以上に複雑であれば解釈性が失われるだけでなく性能もブースティングしない方がましであることがわかった.

参考文献

References

Bartlett, P., Traskin, M., 2007. Adaboost is consistent. *Journal of Machine Learning Research* **8, 2347–2368.**

Bartlett, P. L., Jordan, M. I., McAuliffe, J. D., 2006. Convexity, classification, and risk bounds. *Journal of American Statistical Association* **101, 138–156.**

Bickel, P. J., Ritov, Y., Zakai, A., 2006. Some theory for generalized boosting algorithms. *Journal of Machine Learning Research* **7, 705–732.**

Breiman, L., 1998. Arcing classifiers. *The Annals of Statistics* **26 (3), 801–849.**

Breiman, L., 1999. Prediction games and arcing algorithms. *Neural Computation* **11 (7), 1493–1518.**

Breiman, L., 2004. Population theory for predictor ensembles. *The Annals of Statistics* **32 (1), 1–11.**

Bühlmann, P., 2006. Boosting for high-dimensional linear models. *The Annals of Statistics* **34 (2), 559–583.**

Canu, S., Smola, A. J., 2006. Kernel methods and the exponential family. *Neurocomputing* **69, 714–720.**

Collins, M., Schapire, R., Singer, Y., 2000. Logistic regression, adaboost and bregman distances. In *Proceedings of the Thirteenth Annual Conference on Computational Learning theory*, **158–169.**

Collins, M., Schapire, R., Singer, Y., 2002. Logistic regression, adaboost and bregman distances. In *Machine Learning*, **253–285.**

Eguchi, S., 1992. Geometry of minimum contrast. *Hiroshima Mathematical Journal* **22, 631–647.**

Freund, Y., Schapire, R. E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55 (1), 119–139.

Friedman, J. H., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: A statistical view of boosting. *The Annals of Statistics* **28, 337–407.**

Jiang, W., 2004. Process consistency for AdaBoost. *The Annals of Statistics* **32, 13–29.**

Kawakita, M., Eguchi, S., 2008. Boosting method for local learning in statistical pattern recognition. Neural Computation.

Kawakita, M., Ikeda, S., Eguchi, S., 9 2006. A bridge between boosting and a kernel machine. In: ISM Research Memorandum. Vol. 1006. Institute of Statistical Mathematics.

Kearns, M., Valiant, L. G., 1988. Learning boolean formulae or finite

automata is as hard as factoring. Technical Report TR-14-88, Harvard University Aiken Computation Laboratory.

Koltchinskii, V., Panchenko, D., 2002. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics* **30 (1), 1–30.**

Komaki, F., 1996. On asymptotic properties of predictive distributions. *Biometrika* **83 (2), 299–313.**

Kotsiantis, S. B., Kanellopoulos, D., Pintelas, P. E., 2006. Local boosting of decision stumps for regression and classification problems. *Journal of Computers* **1 (4), 30–37.**

Krishnapuram, B., Carin, L., Figueiredo, M. A. T., Hartemink, A. J., 2005. Sparse multinomial logistic regression: fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27 (6), 957–968.**

Lebanon, G., Lafferty, J., 2002. Boosting and maximum likelihood for exponential models. *Advances in Neural Information Processing Systems* **14.**

Ledoux, M., Talagrand, M., 1991. Probability in Banach Spaces: *isoperimetry and processes*. **Springer-Verlag, New York.**

Lugosi, G., Vayatis, N., 2004. On the Bayes-risk consistency of regularized boosting methods. *The Annals of Statistics* **32, 30–55.**

Mason, L., Baxter, J., Bartlett, P. L., Frean, M., 2000. Functional gradient techniques for combining hypotheses. In A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans editors, *Advances in Large Margin Classifiers*, **MIT Press, Cambridge, MA, 221–247.**

Murata, N., Takenouchi, T., Kanamori, T., Eguchi, S., 2004. Information geometry of U-Boost and Bregman divergence. *Neural Computation* **16, 1437–1481.**

Rätsch, G., Mika, S., Schölkopf, B., Müller, K. R., 2002. Constructing boosting algorithms from svms: an application to one-class classification. *IEEE trans. on pattern analysis and machine intelligence* **24 (9), 1184–1197.**

Rätsch, G., Onoda, T., Müller, K. R., 1999. Regularizing AdaBoost. *Neural Information Processing System***.**

Rätsch, G., Schölkopf, B., Mika, S., Müller, K. R., 2000. SVM and Boosting: One Class. *GMD FIRST* **119.**

Rosset, S., Segal, E., 2002. Boosting density estimation. Advances in Neural Processing Systems.

Schapire, R. E., Freund, Y., Bartlett, P., Lee, W. S., 1998. Boosting the margin: a new explantion for the effectiveness of voting methods. *The Annals of Statistics* **26, 1651–1686.**

Zhang, H., 2004. The optimality of naive bayes. *American Association for Artificial Intelligence***.**

Zhang, T., Yu, B., 2005. Boosting with early stopping: Convergence and consistency. *Annals of statistics* **33, 1538–1579.**