統計数理, *vol.58, no.2, pp.141–155, 2010.*
Proceedings of the Institute of Statistical Mathematics, vol.58, no.2, pp.141–155, 2010.

密度比に基づく機械学習の新たなアプローチ

A New Approach to Machine Learning Based on Density Ratios

杉山将 (sugi@cs.titech.ac.jp)
東京工業大学 大学院情報理工学研究科 計算工学専攻
〒 152-8552 東京都目黒区大岡山 2-12-1

Masashi Sugiyama (sugi@cs.titech.ac.jp)
Department of Computer Science, Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552, Japan.

概 要

本論文では，我々が最近導入した統計的機械学習の新しい枠組みを紹介する．この枠組みの特徴は，様々な機械学習問題を確率密度関数の比の推定問題に帰着させるところにある．そして，困難な確率密度関数の推定を経由せずに，確率密度比を直接推定することにより，精度良く学習を行なう．この密度比推定の枠組みには，非定常環境適応，異常値検出，次元削減，独立成分分析，因果推定，条件付き確率推定など様々な機械学習の問題が含まれるため，極めて汎用的である．

キーワード

密度比，非定常環境適応，異常値検出，次元削減，独立成分分析，因果推定，条件付き確率推定

Abstract

This paper reviews a new framework for statistical machine learning that we introduced recently. A distinctive feature of this framework is that various machine learning problems are formulated as a problem of estimating the ratio of probability densities in a unified way. Then the density ratio is estimated without going through the hard task of density estimation, which results in accurate estimation. This density ratio framework includes various machine learning tasks such as non-stationarity adaptation, outlier detection, dimensionality reduction, independent component analysis, and conditional density estimation. Thus, density ratio estimation is a highly versatile tool for machine learning.

Keywords

Density ratio, non-stationarity adaptation, outlier detection, dimensionality reduction, independent component analysis, conditional density estimation.

1 はじめに

確率密度の推定は困難な問題と知られており，これを回避することが統計的機械学習において非常に重要である．サポートベクトルマシン (SVM)(Vapnik, 1998) は，この原理を踏襲し成功した典型的な例であろう．すなわち，SVM はデータ生成の確率分布を推定するという一般的かつ困難な問題を解くことなく，パターン認識に必要最低限な決定境界のみを学習する．

この考え方に従い，確率密度でなく確率密度の**比**を推定する新たな統計的機械学習の枠組みが近年提案された (Sugiyama et al., 2009)．本論文では，まず密度比を推定するための様々な手法を紹介する．具体的には，カーネル密度推定 (Härdle et al., 2004) に基づく方法，カーネル平均適合法 (Huang et al., 2007)，ロジスティック回帰に基づく方法 (Qin, 1998; Cheng and Chu, 2004; Bickel et al., 2007)，カルバック・ライブラー重要度推定法 (Sugiyama et al., 2008)，最小二乗重要度適合法 (Kanamori et al., 2009a)，拘束無し最小二乗重要度適合法 (Kanamori et al., 2009a) を紹介する．**重要度**とは，重点サンプリング (Fishman, 1996) の文脈における確率密度比のことである．

そして次に，様々な機械学習問題が密度比推定の問題として定式化できることを説明する．具体的には，まず非定常環境適応 (Shimodaira, 2000; Zadrozny, 2004; Sugiyama and Müller, 2005; Sugiyama et al., 2007; Quiñonero-Candela et al., 2009)，異常値検出 (Hido et al., 2010; Smola et al., 2009; Kawahara and Sugiyama, 2009)，条件付き確率推定 (Sugiyama et al., 2010c; Sugiyama, 2009) の各問題が，密度比推定の問題として定式化できることを示す．次に，情報理論で重要な働きをする**相互情報量**が，密度比推定法によって近似できることを説明する (Suzuki et al., 2008; Suzuki et al., 2009b)．相互情報量は確率変数の独立性を表す量であるため，密度比推定法による相互情報推定量を用いて，特徴選択 (Suzuki et al., 2009a)，特徴抽出 (Suzuki and Sugiyama, 2010)，独立成分分析 (Suzuki and Sugiyama, 2009)，因果推定 (Yamada and Sugiyama, 2010) などを行うことができる．

2 密度比推定

本節では，密度比推定の問題を定式化すると共に，様々な密度比推定法を紹介する．

2.1 定式化

データの定義域を $\mathcal{D}$ ($\subset \mathbb{R}^d$) で表し，確率密度 $q(\boldsymbol{x})$ を持つ確率分布に独立に従う i.i.d. 標本 $\{\boldsymbol{x}_i\}_{i=1}^n$，および，確率密度 $q'(\boldsymbol{x})$ を持つ確率分布に独立に従う i.i.d. 標本 $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$ が与えられる場合を考える．ただし，$q(\boldsymbol{x})$ は

$$q(\boldsymbol{x}) > 0 \text{ for all } \boldsymbol{x} \in \mathcal{D}$$

を満たすと仮定する．本節では，2 組の標本 $\{\boldsymbol{x}_i\}_{i=1}^n$ と $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$ から確率密度比

$$r(\boldsymbol{x}) := \frac{q'(\boldsymbol{x})}{q(\boldsymbol{x})}$$

を推定する問題を論じる．

2.2 カーネル密度推定 (KDE) に基づく方法

カーネル密度推定法 (kernel density estimation; KDE) は，i.i.d. 標本 $\{\boldsymbol{x}_i\}_{i=1}^n$ からその標本を生成した確率密度関数 $p(\boldsymbol{x})$ を推定するノンパラメトリック法の一つである．ガウスカーネル

$$K_\sigma(\boldsymbol{x}, \boldsymbol{x}') := \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\sigma^2}\right), \tag{1}$$

を用いたとき，KDE によって得られる推定量は

$$\widehat{p}(\boldsymbol{x}) = \frac{1}{n(2\pi\sigma^2)^{d/2}} \sum_{i=1}^n K_\sigma(\boldsymbol{x}, \boldsymbol{x}_i)$$

で与えられる．

KDE の推定精度は，カーネル幅 σ の選び方に依存する．カーネル幅 σ は，尤度交差確認法 (Härdle et al., 2004) によって最適化できる．まず，標本 $\{\boldsymbol{x}_i\}_{i=1}^n$ を k 個の重なりのない（ほぼ）同じ大きさの部分集合 $\{\mathcal{X}_j\}_{j=1}^k$ に分割する．そして，$\{\mathcal{X}_j\}_{j\neq\ell}$ から確率密度関数の推定量 $\widehat{p}_{\mathcal{X}_\ell}(\boldsymbol{x})$ を求め，$\mathcal{X}_\ell$ に対する対数尤度を計算する：

$$\frac{1}{|\mathcal{X}_\ell|} \sum_{\boldsymbol{x}\in\mathcal{X}_\ell} \log \widehat{p}_{\mathcal{X}_\ell}(\boldsymbol{x})$$

この計算を $\ell = 1, 2, \ldots, k$ に対して行い，対数尤度の平均値を最大にする σ の値を選ぶ．

KDE を用いれば，次のようにして確率密度比を推定することができる．まず，確率密度関数 $q(\boldsymbol{x})$ と $q'(\boldsymbol{x})$ の推定量 $\widehat{q}(\boldsymbol{x})$ と $\widehat{q}'(\boldsymbol{x})$ を，$\{\boldsymbol{x}_i\}_{i=1}^n$ と $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$ からそれぞれ求める．そして，推定した確率密度関数の比をとる：

$$\widehat{r}(\boldsymbol{x}) = \frac{\widehat{q}'(\boldsymbol{x})}{\widehat{q}(\boldsymbol{x})}$$

この方法は非常に単純であるが，データの次元が高い場合に推定精度が良くない．

2.3 カーネル平均適合法 (KMM)

カーネル平均適合法 (kernel mean matching; KMM) は，確率密度関数の推定を行うことなく確率密度比を推定する方法である (Huang et al., 2007)．KMM では，確率密度関数 $q(\boldsymbol{x})$ と $q'(\boldsymbol{x})$ に従う標本を普遍再生核関数 (Steinwart, 2001) を用いて非線形変換し，それらの平均の差を最小にする関数 $\widehat{r}(\boldsymbol{x})$ を求める．ガウスカーネル (1) は普遍再生核であり，以下の最適化問題の解は真の密度比と一致することが知られている．

$$\begin{aligned} &\min_{r(\boldsymbol{x})} \left\| \int K_\sigma(\boldsymbol{x},\cdot) q'(\boldsymbol{x}) d\boldsymbol{x} - \int K_\sigma(\boldsymbol{x},\cdot) r(\boldsymbol{x}) q(\boldsymbol{x}) d\boldsymbol{x} \right\|_{\mathcal{H}}^2 \\ &\text{subject to } \int r(\boldsymbol{x}) q(\boldsymbol{x}) d\boldsymbol{x} = 1 \ \text{ and } \ r(\boldsymbol{x}) \geq 0 \end{aligned} \tag{2}$$

ただし，$\|\cdot\|_{\mathcal{H}}$ はガウス再生核ヒルベルト空間のノルムを表し，$K_\sigma(\boldsymbol{x},\boldsymbol{x}')$ はガウスカーネル (1) を表す．

上記の最適化問題に含まれる期待値を標本平均で近似すれば，次の凸二次計画問題が得られる．

$$\begin{aligned} &\min_{\{r_i\}_{i=1}^n} \left[\frac{1}{2} \sum_{i,i'=1}^{n} r_i r_{i'} K_\sigma(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) - \frac{n}{n'} \sum_{i=1}^{n} r_i \sum_{j=1}^{n'} K_\sigma(\boldsymbol{x}_i, \boldsymbol{x}'_j) \right] \\ &\text{subject to } \left| \frac{1}{n} \sum_{i=1}^{n} r_i - 1 \right| \leq \epsilon \ \text{ and } \ 0 \leq r_1, r_2, \ldots, r_n \leq B \end{aligned}$$

B (≥ 0) と ϵ (≥ 0) はユーザが設定するチューニングパラメータであり，正則化効果の強さを調整する．上記の最適化問題の解 $\{\widehat{r}_i\}_{i=1}^n$ は，標本 $\{\boldsymbol{x}_i\}_{i=1}^n$ における密度比の一致推定値になっている (Huang et al., 2007)．

KMM は確率密度推定を含まないため高次元でも精度が良いと期待されるが，KMM の近似精度はチューニングパラメータ B，ϵ，σ に依存するため，これらの値を適切に決定する必要がある．しかし，KMM では密度比関数全体でなく標本 $\{\boldsymbol{x}_i\}_{i=1}^n$ における密度比の推定値しか得られないため，これらのチューニングパラメータの値を交差確認法で決定することはできない．この問題は，密度比関数全体を学習できるように KMM の最適化問題 (2) を少し改変することにより解決できるが (Kanamori et al., 2009b)，そうであってもカーネル幅を交差確認法で決定することは妥当ではない．なぜならば，KMM の学習規準 (2) は普遍再生核空間のノルムで定義されており，カーネル幅を変えると学習規準の意味が変わってしまうからである．

カーネル法では，標本間の距離の中央値をガウス幅 σ として採用するヒューリスティックがよく用いられる (Schölkopf and Smola, 2002)．KMM では，このヒューリスティックを用いてガウス幅を決定するのが現実的であろうが，必ずしも妥当な結果が得られるとは限らない．

2.4 ロジスティック回帰 (LR) に基づく方法

確率的な分類器を用いても密度比を推定することができる．確率変数 η を考え，$q(\boldsymbol{x})$ から生成された標本には $\eta = 1$ を，$q'(\boldsymbol{x})$ から生成された標本には $\eta = -1$ をそれぞれ割り当てることにする．すなわち，$q(\boldsymbol{x})$ と $q'(\boldsymbol{x})$ は

$$q(\boldsymbol{x}) = p(\boldsymbol{x}|\eta = -1)$$
$$q'(\boldsymbol{x}) = p(\boldsymbol{x}|\eta = 1)$$

と表される．ベイズの定理より，密度比は η を用いて次式で表される (Qin, 1998; Cheng and Chu, 2004; Bickel et al., 2007):

$$r(\boldsymbol{x}) = \frac{p(\eta = -1)}{p(\eta = 1)} \frac{p(\eta = 1|\boldsymbol{x})}{p(\eta = -1|\boldsymbol{x})}$$

比 $p(\eta = -1)/p(\eta = 1)$ は単純に標本数の比 n/n' で近似する．条件付き確率 $p(\eta|\boldsymbol{x})$ は，ロジスティック回帰 (logistic regression; LR) によって $\{\boldsymbol{x}_i\}_{i=1}^n$ と $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$ を分離することにより近似できる．

LR は，条件付き確率 $p(\eta|\boldsymbol{x})$ を次のモデルで近似する確率的分類法である．

$$\widehat{p}(\eta|\boldsymbol{x}) = \left(1 + \exp\left(-\eta \sum_{\ell=1}^{m} \zeta_\ell \phi_\ell(\boldsymbol{x})\right)\right)^{-1}$$

ここで $\{\phi_\ell(\boldsymbol{x})\}_{\ell=1}^m$ は基底関数であり，m は基底関数の数を表す．パラメータ $\boldsymbol{\zeta}$ は，負の罰則付き対数尤度を最小にするように学習する．

$$\widehat{\boldsymbol{\zeta}} := \operatorname*{argmin}_{\boldsymbol{\zeta}} \left[\sum_{i=1}^{n} \log\left(1 + \exp\left(\sum_{\ell=1}^{m} \zeta_\ell \phi_\ell(\boldsymbol{x}_i)\right)\right) + \sum_{j=1}^{n'} \log\left(1 + \exp\left(-\sum_{\ell=1}^{m} \zeta_\ell \phi_\ell(\boldsymbol{x}'_j)\right)\right) + \lambda \boldsymbol{\zeta}^\top \boldsymbol{\zeta} \right]$$

上記の目的関数は下に凸であるため，勾配法や（準）ニュートン法などの標準的な最適化手法によって大域的最適解を求めることができる (Minka, 2007)．最終的に，密度比推定量は次式で与えられる．

$$\widehat{r}(\boldsymbol{x}) = \frac{n}{n'} \exp\left(\sum_{\ell=1}^{m} \widehat{\zeta}_\ell \phi_\ell(\boldsymbol{x})\right)$$

LR を用いた密度比推定法では通常の教師付き分類問題を解いているため，モデル選択（すなわち，基底関数 $\{\phi_\ell(\boldsymbol{x})\}_{\ell=1}^m$ と正則化パラメータ λ の選択）が標準的な交差確認法によって行えるという長所がある．

多クラスの LR を用いれば，3 つ以上の確率密度間の密度比を同時に求めることができる (Bickel et al., 2008).

2.5 カルバック・ライブラー重要度推定法 (KLIEP)

カルバック・ライブラー重要度推定法 (Kullback-Leibler importance estimation procedure; KLIEP)(Sugiyama et al., 2008) も，確率密度推定を介することなく密度比を直接推定できる手法である．

KLIEP では，密度比 $r(\boldsymbol{x})$ を次の線形モデルでモデル化する．

$$\widehat{r}(\boldsymbol{x}) = \sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}) \tag{3}$$

$\{\alpha_\ell\}_{\ell=1}^b$ はパラメータであり，$\{\varphi_\ell(\boldsymbol{x})\}_{\ell=1}^b$ は非負の値をとる基底関数である．

密度比モデル $\widehat{r}(\boldsymbol{x})$ を用いれば，$q'(\boldsymbol{x})$ を $\widehat{q}'(\boldsymbol{x}) = \widehat{r}(\boldsymbol{x})q(\boldsymbol{x})$ で推定することができる．KLIEP では，パラメータ $\{\alpha_\ell\}_{\ell=1}^b$ を，$q'(\boldsymbol{x})$ から $\widehat{q}'(\boldsymbol{x})$ へのカルバック・ライブラー情報量を最小にするように学習する．

$$\mathrm{KL}[q'(\boldsymbol{x})\|\widehat{q}'(\boldsymbol{x})] = \int_{\mathcal{D}} q'(\boldsymbol{x}) \log \frac{q'(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x} - \int_{\mathcal{D}} q'(\boldsymbol{x}) \log \widehat{r}(\boldsymbol{x}) d\boldsymbol{x} \tag{4}$$

第 1 項目は定数なので無視できる．$\widehat{q}'(\boldsymbol{x})$ は確率密度関数であることから，次の拘束条件を満たす必要がある．

$$1 = \int_{\mathcal{D}} \widehat{q}'(\boldsymbol{x}) d\boldsymbol{x} = \int_{\mathcal{D}} \widehat{r}(\boldsymbol{x}) q(\boldsymbol{x}) d\boldsymbol{x} \tag{5}$$

式 (4) の第 2 項目，および，式 (5) の期待値を標本平均で近似すれば，以下の最適化問題が得られる．

$$\max_{\{\alpha_\ell\}_{\ell=1}^b} \left[\sum_{j=1}^{n'} \log \left(\sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}'_j) \right) \right]$$
$$\text{subject to } \frac{1}{n} \sum_{\ell=1}^{b} \alpha_\ell \sum_{i=1}^{n} \varphi_\ell(\boldsymbol{x}_i) = 1 \quad \text{and} \quad \alpha_1, \alpha_2, \ldots, \alpha_b \geq 0$$

これは凸最適化問題であるため，勾配上昇と制約充足を繰り返せば大域的な最適解を求めることができる．また，最適解は疎になる傾向がある．KLIEP の擬似コードを図 1 に示す．

KLIEP の近似性能は，基底関数 $\{\varphi_\ell(\boldsymbol{x})\}_{\ell=1}^b$ の選び方に依存する．密度比関数は，分子の確率分布からの標本が多いところで大きな値をとる傾向があり，それ以外の場所ではゼロに近い値をとる傾向があるため，以下のモデルを用いるのが妥当であろう．

$$\widehat{r}(\boldsymbol{x}) = \sum_{\ell=1}^{n'} \alpha_\ell K_\sigma(\boldsymbol{x}, \boldsymbol{x}'_\ell)$$

ここで $K_\sigma(\boldsymbol{x}, \boldsymbol{x}')$ は，式 (1) で定義されるガウスカーネルである．KLIEP のモデル選択（すなわち，ガウスカーネルの幅 σ の選択）は，交差確認法で行うことができる．KLIEP に対する交差確認法の擬似コードを図 2 に示す．KLIEP の MATLAB® での実装は，

Input: $m = \{\varphi_\ell(\boldsymbol{x})\}_{\ell=1}^{b}$, $\{\boldsymbol{x}_i\}_{i=1}^{n}$, and $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$
Output: $\widehat{r}(\boldsymbol{x})$

$A_{j,\ell} \longleftarrow \varphi_\ell(\boldsymbol{x}'_j)$ for $j = 1, 2, \ldots, n'$ and $\ell = 1, 2, \ldots, b$;
$\xi_\ell \longleftarrow \frac{1}{n}\sum_{i=1}^{n} \varphi_\ell(\boldsymbol{x}_i)$ for $\ell = 1, 2, \ldots, b$;
Initialize $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_b)^\top$ $(> \mathbf{0}_b)$ and ε $(0 < \varepsilon \ll 1)$;
Repeat until convergence
    $\boldsymbol{\alpha} \longleftarrow \boldsymbol{\alpha} + \varepsilon \boldsymbol{A}^\top(\mathbf{1}_{n'}./\boldsymbol{A}\boldsymbol{\alpha})$; % Gradient ascent
    $\boldsymbol{\alpha} \longleftarrow \boldsymbol{\alpha} + (1 - \boldsymbol{\xi}^\top\boldsymbol{\alpha})\boldsymbol{\xi}/(\boldsymbol{\xi}^\top\boldsymbol{\xi})$; % Constraint satisfaction
    $\boldsymbol{\alpha} \longleftarrow \max(\mathbf{0}_b, \boldsymbol{\alpha})$; % Constraint satisfaction
    $\boldsymbol{\alpha} \longleftarrow \boldsymbol{\alpha}/(\boldsymbol{\xi}^\top\boldsymbol{\alpha})$; % Constraint satisfaction
end
$\widehat{r}(\boldsymbol{x}) \longleftarrow \sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x})$;

図 1: KLIEP の擬似コード．$\mathbf{0}_b$ は b 次元の 0 ベクトル，$\mathbf{1}_{n'}$ は n' 次元の 1 ベクトルを表す．'./' は要素毎の除算であり，$^\top$ は転置を表す．ベクトルに対する不等式と 'max' 作用素は，要素毎に適用される．

Input: $\mathcal{M} = \{m \mid m = \{\varphi_\ell(\boldsymbol{x})\}_{\ell=1}^{b}\}$, $\{\boldsymbol{x}_i\}_{i=1}^{n}$, and $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$
Output: $\widehat{r}(\boldsymbol{x})$

Split $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$ into k disjoint subsets $\{\mathcal{X}'_j\}_{j=1}^{k}$;
for each model $m \in \mathcal{M}$
    for each split $t = 1, 2, \ldots, k$
        $\widehat{r}_t(\boldsymbol{x}) \longleftarrow \mathrm{KLIEP}(m, \{\boldsymbol{x}_i\}_{i=1}^{n}, \{\mathcal{X}'_j\}_{j\neq t})$;
        $\widehat{L}_t(m) \longleftarrow \frac{1}{|\mathcal{X}'_t|}\sum_{\boldsymbol{x}\in\mathcal{X}'_t} \log \widehat{r}_t(\boldsymbol{x})$;
    end
    $\widehat{L}(m) \longleftarrow \frac{1}{k}\sum_{t=1}^{k} \widehat{L}_t(m)$;
end
$\widehat{m} \longleftarrow \mathrm{argmax}_{m\in\mathcal{M}}\, \widehat{L}(m)$;
$\widehat{r}(\boldsymbol{x}) \longleftarrow \mathrm{KLIEP}(\widehat{m}, \{\boldsymbol{x}_i\}_{i=1}^{n}, \{\boldsymbol{x}'_j\}_{j=1}^{n'})$;

図 2: KLIEP に対する交差確認法の擬似コード.

http://sugiyama-www.cs.titech.ac.jp/~sugi/software/KLIEP/

で公開されている．

2.6 最小二乗重要度適合法 (LSIF)

KLIEP では，カルバック・ライブラー情報量を用いて 2 つの確率密度の差異を評価したが，最小二乗重要度適合法 (least-squares importance fitting; LSIF)(Kanamori et al., 2009a) では，二乗損失のもとで密度比の適合を行う．密度比関数 $r(\boldsymbol{x})$ は，KLIEP と同様に線形モデル (3) でモデル化する．パラメータ $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_b)^\top$ は，次の二乗誤差 J_0 を最小にするように学習する．

$$\begin{aligned} J_0(\boldsymbol{\alpha}) &:= \frac{1}{2}\int (\widehat{r}(\boldsymbol{x}) - r(\boldsymbol{x}))^2 q(\boldsymbol{x}) d\boldsymbol{x} \\ &= \frac{1}{2}\int \widehat{r}(\boldsymbol{x})^2 q(\boldsymbol{x}) d\boldsymbol{x} - \int \widehat{r}(\boldsymbol{x}) q'(\boldsymbol{x}) d\boldsymbol{x} + \frac{1}{2}\int r(\boldsymbol{x}) q'(\boldsymbol{x}) d\boldsymbol{x} \end{aligned}$$

最後の項は定数のため無視できる．最初の 2 項を J と定義する．

$$J(\boldsymbol{\alpha}) := \frac{1}{2}\int \widehat{r}(\boldsymbol{x})^2 q(\boldsymbol{x}) d\boldsymbol{x} - \int \widehat{r}(\boldsymbol{x}) q'(\boldsymbol{x}) d\boldsymbol{x}$$

J に含まれる期待値を標本平均で近似すれば，次式が得られる．

$$\begin{aligned} \widehat{J}(\boldsymbol{\alpha}) &:= \frac{1}{2n}\sum_{i=1}^{n} \widehat{r}(\boldsymbol{x}_i)^2 - \frac{1}{n'}\sum_{j=1}^{n'} \widehat{r}(\boldsymbol{x}'_j) \\ &= \frac{1}{2}\sum_{\ell,\ell'=1}^{b} \alpha_\ell \alpha_{\ell'} \widehat{H}_{\ell,\ell'} - \sum_{\ell=1}^{b} \alpha_\ell \widehat{h}_\ell \end{aligned}$$

ただし，

$$\begin{aligned} \widehat{H}_{\ell,\ell'} &:= \frac{1}{n}\sum_{i=1}^{n} \varphi_\ell(\boldsymbol{x}_i)\varphi_{\ell'}(\boldsymbol{x}_i) \\ \widehat{h}_\ell &:= \frac{1}{n'}\sum_{j=1}^{n'} \varphi_\ell(\boldsymbol{x}'_j) \end{aligned}$$

である．密度比関数の非負性を考慮し，更に一次の正則化項を導入すれば，次の最適化問題が得られる．

$$\begin{aligned} &\min_{\{\alpha_\ell\}_{\ell=1}^{b}} \left[\frac{1}{2}\sum_{\ell,\ell'=1}^{b} \alpha_\ell \alpha_{\ell'} \widehat{H}_{\ell,\ell'} - \sum_{\ell=1}^{b} \alpha_\ell \widehat{h}_\ell + \lambda \sum_{\ell=1}^{b} \alpha_\ell\right] \\ &\text{subject to } \alpha_1, \alpha_2, \ldots, \alpha_b \geq 0 \end{aligned} \tag{6}$$

ただし，λ は非負の正則化パラメータである．式 (6) は凸二次計画問題であるので，標準的な最適化ソフトウェアによって大域的最適解を効率良く求めることができる．基底関数は，KLIEP と同様に設計すればよいであろう．ガウス幅 σ と正則化パラメータ λ は，交差確認法によって定めればよい．ただし KLIEP の場合と異なり，LSIF の交差確認法では $\{\boldsymbol{x}_i\}_{i=1}^n$ と $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$ を共に分割すべきである．

LSIF の解 $\widehat{\boldsymbol{\alpha}}$ は，正則化パラメータ λ に対して区分線形関数になることが知られている (Kanamori et al., 2009a)．従って，パラメトリック最適化の手法を用いれば，正則化パス（すなわち，全ての λ に対する解）を効率良く計算することができる (Best, 1982; Efron et al., 2004; Hastie et al., 2004)．LSIF の正則化パス追跡の擬似コードを図 3 に示す．この正則化パス追跡のアルゴリズムを用いれば，二次計画ソルバーはもはや不要であり，線形方程式を解くだけで解を得ることができる．これにより，LSIF の計算速度は大幅に高速化される．LSIF の R での実装は，

`http://www.math.cm.is.nagoya-u.ac.jp/~kanamori/software/LSIF/`

で公開されている．

2.7 拘束無し最小二乗重要度適合法 (uLSIF)

正則化パス追跡を用いることにより，LSIF の計算時間は大幅に短縮される．しかし，正則化パス追跡はしばしば数値的に不安定になり，実用上問題となることがある．この問題を解決すべく，拘束無し最小二乗重要度適合法 (unconstrained LSIF; uLSIF) が提案された (Kanamori et al., 2009a)．

uLSIF の考え方は非常に単純で，LSIF の最適化問題 (6) に含まれる非負拘束を無視するだけである．これにより，以下の拘束無し最適化問題が得られる．

$$\min_{\{\alpha_\ell\}_{\ell=1}^b} \left[\frac{1}{2} \sum_{\ell,\ell'=1}^{b} \alpha_\ell \alpha_{\ell'} \widehat{H}_{\ell,\ell'} - \sum_{\ell=1}^{b} \alpha_\ell \widehat{h}_\ell + \frac{\lambda}{2} \sum_{\ell=1}^{b} \alpha_\ell^2 \right] \tag{7}$$

ただし，非負拘束を取り除いたことに伴い，正則化項が二次に変更されていることに注意せよ．式 (7) の解は，解析的に次式で与えられる．

$$\widetilde{\boldsymbol{\alpha}} = (\widetilde{\alpha}_1, \widetilde{\alpha}_2, \ldots, \widetilde{\alpha}_b)^\top = (\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I}_b)^{-1} \widehat{\boldsymbol{h}}$$

ただし，$\boldsymbol{I}_b$ は b 次元の単位行列である．非負拘束を無視したため，いくつかのパラメータ値は負になる可能性がある．そこで，解を次式で補正する．

$$\widehat{\alpha}_\ell = \max(0, \widetilde{\alpha}_\ell) \quad \text{for } \ell = 1, 2, \ldots, b$$

uLSIF は，収束性やアルゴリズムの数値的な安定性において優れた性質を持っていることが理論的に示されている (Kanamori et al., 2009b)．

Input: $\widehat{\boldsymbol{H}}$ and $\widehat{\boldsymbol{h}}$
Output: entire regularization path $\widehat{\boldsymbol{\alpha}}(\lambda)$ for $\lambda \geq 0$

$\tau \longleftarrow 0$; $\quad k \longleftarrow \mathrm{argmax}_i\{\widehat{h}_i \mid i = 1, 2, \ldots, b\}$;
$\lambda_\tau \longleftarrow \widehat{h}_k$; $\quad \widehat{\mathcal{A}} \longleftarrow \{1, 2, \ldots, b\} \backslash \{k\}$;
$\widehat{\boldsymbol{\alpha}}(\lambda_\tau) \longleftarrow \boldsymbol{0}_b$; $\quad$ % the vector with all zeros
While $\lambda_\tau > 0$
$\quad \widehat{\boldsymbol{E}} \longleftarrow \boldsymbol{O}_{|\widehat{\mathcal{A}}| \times b}$; $\quad$ % the matrix with all zeros
$\quad$ **For** $i = 1, 2, \ldots, |\widehat{\mathcal{A}}|$
$\quad\quad \widehat{E}_{i,j_i} \longleftarrow 1$; % $\widehat{\mathcal{A}} = \{j_1, j_2, \ldots, j_{|\widehat{\mathcal{A}}|} \mid j_1 < j_2 < \cdots < j_{|\widehat{\mathcal{A}}|}\}$
$\quad$ **end**
$\quad \widehat{\boldsymbol{G}} \longleftarrow \begin{pmatrix} \widehat{\boldsymbol{H}} & -\widehat{\boldsymbol{E}}^\top \\ -\widehat{\boldsymbol{E}} & \boldsymbol{O}_{|\widehat{\mathcal{A}}| \times |\widehat{\mathcal{A}}|} \end{pmatrix}$;
$\quad \boldsymbol{u} \longleftarrow \widehat{\boldsymbol{G}}^{-1} \begin{pmatrix} \widehat{\boldsymbol{h}} \\ \boldsymbol{0}_{|\widehat{\mathcal{A}}|} \end{pmatrix}$; $\quad \boldsymbol{v} \longleftarrow \widehat{\boldsymbol{G}}^{-1} \begin{pmatrix} \boldsymbol{1}_b \\ \boldsymbol{0}_{|\widehat{\mathcal{A}}|} \end{pmatrix}$;
$\quad$ **If** $\boldsymbol{v} \leq \boldsymbol{0}_{b+|\widehat{\mathcal{A}}|}$ % the final interval
$\quad\quad \lambda_{\tau+1} \longleftarrow 0$; $\quad \widehat{\boldsymbol{\alpha}}(\lambda_{\tau+1}) \longleftarrow (u_1, u_2, \ldots, u_b)^\top$;
$\quad$ **else** % an intermediate interval
$\quad\quad k \longleftarrow \mathrm{argmax}_i\{u_i / v_i \mid v_i > 0,\ i = 1, 2, \ldots, b + |\widehat{\mathcal{A}}|\}$;
$\quad\quad \lambda_{\tau+1} \longleftarrow \max\{0, u_k / v_k\}$;
$\quad\quad \widehat{\boldsymbol{\alpha}}(\lambda_{\tau+1}) \longleftarrow (u_1, u_2, \ldots, u_b)^\top - \lambda_{\tau+1}(v_1, v_2, \ldots, v_b)^\top$;
$\quad\quad$ **If** $1 \leq k \leq b$
$\quad\quad\quad \widehat{\mathcal{A}} \longleftarrow \widehat{\mathcal{A}} \cup \{k\}$;
$\quad\quad$ **else**
$\quad\quad\quad \widehat{\mathcal{A}} \longleftarrow \widehat{\mathcal{A}} \backslash \{j_{k-b}\}$;
$\quad\quad$ **end**
$\quad$ **end**
$\quad \tau \longleftarrow \tau + 1$;
end

$$\widehat{\boldsymbol{\alpha}}(\lambda) \longleftarrow \begin{cases} \boldsymbol{0}_b & \text{if } \lambda \geq \lambda_0 \\ \frac{\lambda_{\tau+1} - \lambda}{\lambda_{\tau+1} - \lambda_\tau} \widehat{\boldsymbol{\alpha}}(\lambda_\tau) + \frac{\lambda - \lambda_\tau}{\lambda_{\tau+1} - \lambda_\tau} \widehat{\boldsymbol{\alpha}}(\lambda_{\tau+1}) & \text{if } \lambda_{\tau+1} \leq \lambda \leq \lambda_\tau \end{cases}$$

図 3: LSIF の正則化パス追跡の擬似コード. $\widehat{\boldsymbol{G}}^{-1}$ の計算が不安定な場合は, $\widehat{\boldsymbol{H}}$ の対角成分に小さな正数を加えて安定化させると良い.

uLSIF の重要な特徴は，一つ抜き交差確認 (leave-one-out cross-validation; LOOCV) のスコアを解析的に計算できることである．これにより，解を一度計算するのと同じ計算量で LOOCV のスコアを求めることができる．LOOCV を計算するための擬似コードを図 4 に示す．uLSIF の MATLAB® および R の実装は，

http://sugiyama-www.cs.titech.ac.jp/~sugi/software/uLSIF/
http://www.math.cm.is.nagoya-u.ac.jp/~kanamori/software/LSIF/

で公開されている．

2.8 密度比推定法のまとめ

KDE は最適化プロセスを含まないため計算が非常に高速であり，尤度交差確認によってモデル選択も可能である．しかし，高次元データに対するノンパラメトリック密度推定は精度が良くない (Vapnik, 1998; Härdle et al., 2004)．

KMM は密度比を直接推定することにより，KDE の弱点を克服しようとしている．しかし，モデル選択法がないため，カーネル幅などのチューニングパラメータを適切に設定するのは実用上難しい．また，解を求めるためには二次計画問題を解く必要があるため，計算にやや時間がかかる．次節で示すように，密度比の応用例には分母と分子の確率密度の定義域が異なる場合がある．普遍再生核ヒルベルト空間内で分母と分子の分布から生成された標本の平均を適合させるという定式化のため，分母と分子の確率密度の定義域が異なる場面に KMM を適用するのは困難であると考えられる．

LR と KLIEP も密度推定を含まない密度比推定法であり，交差確認法によりモデル選択を行うことができるという特徴を持っている．しかし，解を求めるためには非線形最適化問題を解く必要があるため，計算にやや時間がかかる．KLIEP は，分母と分子の確率密度の定義域が異なる場合の密度比推定に用いることができるが，LR は分母と分子の分布から生成された標本を分類するという定式化のため，そのような場面に適用するのは困難である．

LSIF は LR や KLIEP と同様に，密度推定を含まず交差確認によるモデル選択が可能な密度比推定法である．LSIF も，KLIEP と同様に分母と分子の確率密度の定義域が異なる場合の密度比推定に用いることができる．更に，LSIF では正則化パス追跡アルゴリズムが利用できるため，LR や KLIEP よりも計算効率が良い．しかし，正則化パス追跡アルゴリズムは数値的にやや不安定である．

uLSIF は密度比推定を含まず，交差確認によるモデル選択が可能であり，更に解が解析的に求まるという特徴を持つ．そのため，uLSIF の解は高速かつ安定に計算することができる．また，uLSIF は分母と分子の確率密度の定義域が異なる場合の密度比推定に用いることもできる．更に，一つ抜き交差確認のスコアを解析的に計算できるため，uLSIF ではモデル選択を含めた全体の計算時間が大幅に短縮される．

以上より，uLSIF が実用上最も優れた密度比推定法であると考えられる．

Input: $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$
Output: $\widehat{r}(\boldsymbol{x})$

$b \longleftarrow \min(100, n')$; $\quad \overline{n} \longleftarrow \min(n, n')$;
Randomly choose b centers $\{\boldsymbol{c}_\ell\}_{\ell=1}^b$ from $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$ without replacement;
For each candidate of Gaussian width σ

$$\widehat{H}_{\ell,\ell'} \longleftarrow \frac{1}{n}\sum_{i=1}^{n} \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{c}_\ell\|^2 + \|\boldsymbol{x}_i - \boldsymbol{c}_{\ell'}\|^2}{2\sigma^2}\right) \text{ for } \ell, \ell' = 1, 2, \ldots, b;$$

$$\widehat{h}_\ell \longleftarrow \frac{1}{n'}\sum_{j=1}^{n'} \exp\left(-\frac{\|\boldsymbol{x}'_j - \boldsymbol{c}_\ell\|^2}{2\sigma^2}\right) \text{ for } \ell = 1, 2, \ldots, b;$$

$$X_{\ell,i} \longleftarrow \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{c}_\ell\|^2}{2\sigma^2}\right) \text{ for } i = 1, 2, \ldots, \overline{n} \text{ and } \ell = 1, 2, \ldots, b;$$

$$X'_{\ell,i} \longleftarrow \exp\left(-\frac{\|\boldsymbol{x}'_i - \boldsymbol{c}_\ell\|^2}{2\sigma^2}\right) \text{ for } i = 1, 2, \ldots, \overline{n} \text{ and } \ell = 1, 2, \ldots, b;$$

For each candidate of regularization parameter λ

$$\widehat{\boldsymbol{B}} \longleftarrow \widehat{\boldsymbol{H}} + \frac{\lambda(n-1)}{n}\boldsymbol{I}_b;$$

$$\boldsymbol{B}_0 \longleftarrow \widehat{\boldsymbol{B}}^{-1}\widehat{\boldsymbol{h}}\boldsymbol{1}_{\overline{n}}^\top + \widehat{\boldsymbol{B}}^{-1}\boldsymbol{X}\,\mathrm{diag}\left(\frac{\widehat{\boldsymbol{h}}^\top\widehat{\boldsymbol{B}}^{-1}\boldsymbol{X}}{n\boldsymbol{1}_{\overline{n}}^\top - \boldsymbol{1}_b^\top(\boldsymbol{X} * \widehat{\boldsymbol{B}}^{-1}\boldsymbol{X})}\right);$$

$$\boldsymbol{B}_1 \longleftarrow \widehat{\boldsymbol{B}}^{-1}\boldsymbol{X}' + \widehat{\boldsymbol{B}}^{-1}\boldsymbol{X}\,\mathrm{diag}\left(\frac{\boldsymbol{1}_b^\top(\boldsymbol{X}' * \widehat{\boldsymbol{B}}^{-1}\boldsymbol{X})}{n\boldsymbol{1}_{\overline{n}}^\top - \boldsymbol{1}_b^\top(\boldsymbol{X} * \widehat{\boldsymbol{B}}^{-1}\boldsymbol{X})}\right);$$

$$\boldsymbol{B}_2 \longleftarrow \max\left(\boldsymbol{O}_{b\times\overline{n}}, \frac{n-1}{n(n'-1)}(n'\boldsymbol{B}_0 - \boldsymbol{B}_1)\right);$$

$$\boldsymbol{r} \longleftarrow (\boldsymbol{1}_b^\top(\boldsymbol{X} * \boldsymbol{B}_2))^\top; \quad \boldsymbol{r}' \longleftarrow (\boldsymbol{1}_b^\top(\boldsymbol{X}' * \boldsymbol{B}_2))^\top;$$

$$\mathrm{LOOCV}(\sigma, \lambda) \longleftarrow \frac{\boldsymbol{r}^\top\boldsymbol{r}}{2\overline{n}} - \frac{\boldsymbol{1}_{\overline{n}}^\top\boldsymbol{r}'}{\overline{n}};$$

end
end

$$(\widehat{\sigma}, \widehat{\lambda}) \longleftarrow \mathrm{argmin}_{(\sigma,\lambda)}\,\mathrm{LOOCV}(\sigma, \lambda);$$

$$\widetilde{H}_{\ell,\ell'} \longleftarrow \frac{1}{n}\sum_{i=1}^{n} \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{c}_\ell\|^2 + \|\boldsymbol{x}_i - \boldsymbol{c}_{\ell'}\|^2}{2\widehat{\sigma}^2}\right) \text{ for } \ell, \ell' = 1, 2, \ldots, b;$$

$$\widetilde{h}_\ell \longleftarrow \frac{1}{n'}\sum_{j=1}^{n'} \exp\left(-\frac{\|\boldsymbol{x}'_j - \boldsymbol{c}_\ell\|^2}{2\widehat{\sigma}^2}\right) \text{ for } \ell = 1, 2, \ldots, b;$$

$$\widehat{\boldsymbol{\alpha}} \longleftarrow \max(\boldsymbol{0}_b, (\widetilde{\boldsymbol{H}} + \widehat{\lambda}\boldsymbol{I}_b)^{-1}\widetilde{\boldsymbol{h}});$$

$$\widehat{r}(\boldsymbol{x}) \longleftarrow \sum_{\ell=1}^{b} \widehat{\alpha}_\ell \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{c}_\ell\|^2}{2\widehat{\sigma}^2}\right);$$

図 4: uLSIF の擬似コード．モデル選択は一つ抜き交差確認で行う．$\boldsymbol{B} * \boldsymbol{B}'$ は要素毎の乗算を表す．n 次元ベクトル $\boldsymbol{b}, \boldsymbol{b}'$ に対して，$\mathrm{diag}\left(\frac{\boldsymbol{b}}{\boldsymbol{b}'}\right)$ は i 番目の対角要素が b_i/b'_i である対角行列を表す．

3 密度比推定を用いた機械学習

本節では，前節で紹介した密度比推定法を用いることにより，様々な機械学習問題が解決できることを示す．

3.1 共変量シフト適応

共変量シフト(Shimodaira, 2000) とは，教師付き学習において訓練標本とテスト標本の入力分布が $q(\boldsymbol{x})$ から $q'(\boldsymbol{x})$ に変化するが，入出力関係 $p(y|\boldsymbol{x})$ は変化しないという状況のことを指す．共変量シフト下では，最尤推定などの標準的な学習法はバイアスを持つ．このバイアスは，損失関数（対数尤度）を**重要度**によって重み付けすることにより漸近的に打ち消すことができる．

$$\mathop{\mathbb{E}}_{q'(\boldsymbol{x})}[\mathrm{loss}(\boldsymbol{x})] = \int \mathrm{loss}(\boldsymbol{x})q'(\boldsymbol{x})d\boldsymbol{x} = \int \mathrm{loss}(\boldsymbol{x})r(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x} = \mathop{\mathbb{E}}_{q(\boldsymbol{x})}[r(\boldsymbol{x})\mathrm{loss}(\boldsymbol{x})]$$

すなわち，損失関数 $\mathrm{loss}(\boldsymbol{x})$ の $q'(\boldsymbol{x})$ に関する期待値は，その $q(\boldsymbol{x})$ に関する重要度重み付き期待値により計算できる．

交差確認規準や赤池情報量規準などのモデル選択規準も，共変量シフト下では不偏性を失う．しかし，同様に重要度重み付けを行うことにより，不偏性が回復できる (Shimodaira, 2000; Zadrozny, 2004; Sugiyama and Müller, 2005; Sugiyama et al., 2007)．また，モデルが正しくない場合の能動学習においても共変量シフトの影響を明示的に考慮する必要があり，重要度重み付けはバイアスを軽減するために重要である (Wiens, 2000; Kanamori and Shimodaira, 2003; Sugiyama, 2006; Sugiyama and Rubens, 2008; Sugiyama and Nakajima, 2009)．

このように，密度比推定法は共変量シフト適応に不可欠な技術である．共変量シフトの実応用例は，ブレイン・コンピュータインターフェース (Sugiyama et al., 2007)，ロボット制御 (Hachiya et al., 2009a; Hachiya et al., 2009b)，話者識別 (Yamada et al., 2010)，自然言語処理 (Tsuboi et al., 2009)，顔画像認識 (Ueki et al., 2010) など多岐にわたる．同様な重要度重み付けは，マルチタスク学習 (Bickel et al., 2008) に用いることもできる．

3.2 正常値に基づく異常値検出

正常標本集合に基づいて，評価標本集合に含まれる異常値を検出する問題を考える．これら 2 つの標本集合の密度比を考えれば，正常値に対する密度比の値は 1 に近く，異常値に対する密度比の値は 1 から大きく離れることがわかる．従って，密度比は異常値検出の規準として用いることができる (Hido et al., 2010; Smola et al., 2009)．

正常標本集合を用いない教師なしの異常値検出法では，正常値と異常値の定義が不明確なことが多い (Breunig et al., 2000; Schölkopf et al., 2001)．それに対して，正常値に基づ

く異常値検出では正常値を明確に定義していることから，妥当な異常値検出結果が得られると考えられる．

uLSIF に基づく異常値検出法は，精密機器の異常診断に適用されている (Takimoto et al., 2009)．また，正常値に基づく異常値検出と同様な考え方は，時系列の変化点検出 (Kawahara and Sugiyama, 2009) に応用することもできる．

3.3 相互情報量推定

確率密度 $p(\boldsymbol{x}, \boldsymbol{y})$ を持つ分布に従う n 個の i.i.d. 標本 $\{(\boldsymbol{x}_k, \boldsymbol{y}_k)\}_{k=1}^n$ から，$\boldsymbol{x}$ と $\boldsymbol{y}$ の相互情報量を推定する問題を考える．

$$\iint p(\boldsymbol{x}, \boldsymbol{y}) \log \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})} d\boldsymbol{x} d\boldsymbol{y}$$

密度比推定の文脈において，$\{(\boldsymbol{x}_k, \boldsymbol{y}_k)\}_{k=1}^n$ を分子の確率分布からの標本とみなし，$\{(\boldsymbol{x}_k, \boldsymbol{y}_{k'})\}_{k,k'=1}^n$ を分母の確率分布からの標本とみなせば，密度比推定法により相互情報量を推定することができる．すなわち，密度比

$$r(\boldsymbol{x}, \boldsymbol{y}) := \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})}$$

の推定量 $\widehat{r}(\boldsymbol{x}, \boldsymbol{y})$ を求め，相互情報量を

$$\frac{1}{n} \sum_{k=1}^n \log \widehat{r}(\boldsymbol{x}_k, \boldsymbol{y}_k)$$

と近似する (Suzuki et al., 2008; Suzuki et al., 2009b)．

また，相互情報量の二乗損失版

$$\frac{1}{2} \int \left(\frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})} - 1 \right)^2 p(\boldsymbol{x})p(\boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{y}$$

も同様に推定することができる (Suzuki et al., 2009a)．

相互情報量は確率変数間の独立性を表す指標であるため，その推定量は，特徴選択 (Suzuki et al., 2009a)，特徴抽出 (Suzuki and Sugiyama, 2010)，独立成分分析 (Suzuki and Sugiyama, 2009)，因果推定 (Yamada and Sugiyama, 2010) などに応用することができる．

3.4 条件付き確率推定

確率密度 $p(\boldsymbol{x}, \boldsymbol{y})$ を持つ分布に従う n 個の i.i.d. 標本 $\{(\boldsymbol{x}_k, \boldsymbol{y}_k)\}_{k=1}^n$ から，条件付き確率 $p(\boldsymbol{y}|\boldsymbol{x})$ を推定する問題を考える．条件変数 $\boldsymbol{x}$ が連続の場合は単純な経験近似ができないため，条件付き確率の推定は自明でない (Bishop, 2006; Takeuchi et al., 2009)．

密度比推定の文脈において，$\{(\boldsymbol{x}_k, \boldsymbol{y}_k)\}_{k=1}^n$ を分子の確率分布からの標本とみなし，$\{\boldsymbol{x}_k\}_{k=1}^n$ を分母の確率分布からの標本とみなせば，密度比推定法により条件付き確率を直接推定することができる．すなわち，$p(\boldsymbol{y}|\boldsymbol{x})$ の等価表現である

$$r(\boldsymbol{x}, \boldsymbol{y}) := \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})}$$

を推定することに対応する．$\boldsymbol{y}$ が連続変数の場合の手法 (Sugiyama et al., 2010c)，$\boldsymbol{y}$ がカテゴリ変数の場合の手法 (Sugiyama, 2009) がそれぞれ提案されている．

4 まとめ

本論文では，密度比を用いた新しい機械学習の枠組みを紹介した．この枠組みには，非定常環境適応，異常値検出，次元削減，独立成分分析，因果推定，条件付き確率推定など様々な機械学習の問題が含まれるため，極めて汎用的である．また，密度比推定のための様々な手法を紹介した．今後，より大規模・高次元のデータに対応できるロバストな密度比推定法の開発が望まれる．高次元の密度比に対しては，次元削減と密度比推定を組み合わせた手法が提案されている (Sugiyama et al., 2010b; Sugiyama et al., 2010a)．

本研究は，AOARD, SCAT, JST さきがけの支援を受けて行われた．

References

Best, M. J. (1982). *An algorithm for the solution of the parametric quadratic programming problem* (Technical Report 82-24). Faculty of Mathematics, University of Waterloo, Waterloo, Canada.

Bickel, S., Brückner, M. and Scheffer, T. (2007). Discriminative learning for differing training and test distributions. *Proceedings of the 24th International Conference on Machine Learning (ICML2007)* (pp. 81–88).

Bickel, S., Bogojeska, J., Lengauer, T. and Scheffer, T. (2008). Multi-task learning for HIV therapy screening. *Proceedings of 25th Annual International Conference on Machine Learning (ICML2008)* (pp. 56–63).

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer, New York, NY, USA.

Breunig, M. M., Kriegel, H.-P., Ng, R. T. and Sander, J. (2000). LOF: Identifying density-based local outliers. *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD2000)* (pp. 93–104).

Cheng, K. F. and Chu, C. K. (2004). Semiparametric density estimation under a two-sample density ratio model. *Bernoulli*, *10*, 583–604.

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

Fishman, G. S. (1996). *Monte Carlo: Concepts, algorithms and applications*. Springer-Verlag, Berlin, Germany.

Hachiya, H., Akiyama, T., Sugiyama, M. and Peters, J. (2009a). Adaptive importance sampling for value function approximation in off-policy reinforcement learning. *Neural Networks*, *22*, 1399–1410.

Hachiya, H., Peters, J. and Sugiyama, M. (2009b). Efficient sample reuse in EM-based policy search. *Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD2009)* (pp. 469–484).

Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004). *Nonparametric and semiparametric models*. Springer, Berlin, Germany.

Hastie, T., Rosset, S., Tibshirani, R. and Zhu, J. (2004). The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, *5*, 1391–1415.

Hido, S., Tsuboi, Y., Kashima, H., Sugiyama, M. and Kanamori, T. (2010). Statistical outlier detection using direct density ratio estimation. *Knowledge and Information Systems*, to appear.

Huang, J., Smola, A., Gretton, A., Borgwardt, K. M. and Schölkopf, B. (2007). Correcting sample selection bias by unlabeled data. In B. Schölkopf, J. Platt and T. Hoffman (Eds.), *Advances in neural information processing systems 19*, 601–608. MIT Press, Cambridge, MA, USA.

Kanamori, T. and Shimodaira, H. (2003). Active learning algorithm using the maximum weighted log-likelihood estimator. *Journal of Statistical Planning and Inference*, *116*, 149–162.

Kanamori, T., Hido, S. and Sugiyama, M. (2009a). A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, *10*, 1391–1445.

Kanamori, T., Suzuki, T. and Sugiyama, M. (2009b). *Condition number analysis of kernel-based density ratio estimation* (arXiv Technical Report). http://www.citebase.org/abstract?id=oai:arXiv.org:0912.2800.

Kawahara, Y. and Sugiyama, M. (2009). Change-point detection in time-series data by direct density-ratio estimation. *Proceedings of 2009 SIAM International Conference on Data Mining (SDM2009)* (pp. 389–400).

Minka, T. P. (2007). *A comparison of numerical optimizers for logistic regression* (Technical Report). Microsoft Research, USA. http://research.microsoft.com/˜minka/papers/logreg/minka-logreg.pdf

Qin, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, *85*, 619–639.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A. and Lawrence, N. (Eds.). (2009). *Dataset shift in machine learning*. MIT Press, Cambridge, MA, USA.

Schölkopf, B. and Smola, A. J. (2002). *Learning with kernels*. MIT Press, Cambridge, MA, USA.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J. and Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation*, *13*, 1443–1471.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, *90*, 227–244.

Smola, A., Song, L. and Teo, C. H. (2009). Relative novelty detection. *Proceedings of Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS2009)* (pp. 536–543).

Steinwart, I. (2001). On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, *2*, 67–93.

Sugiyama, M. (2006). Active learning in approximately linear regression based on conditional expectation of generalization error. *Journal of Machine Learning Research*, *7*, 141–166.

Sugiyama, M. (2009). *Superfast-trainable multi-class probabilistic classifier by least-squares posterior fitting* (Technical Report TR09-0011). Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan.

Sugiyama, M. and Müller, K.-R. (2005). Input-dependent estimation of generalization error under covariate shift. *Statistics and Decisions*, *23*, 249–279.

Sugiyama, M. and Nakajima, S. (2009). Pool-based active learning in approximate linear regression. *Machine Learning*, *75*, 249–274.

Sugiyama, M. and Rubens, N. (2008). A batch ensemble approach to active learning with model selection. *Neural Networks*, *21*, 1287–1286.

Sugiyama, M., Krauledat, M. and Müller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, *8*, 985–1005.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P. and Kawanabe, M. (2008). Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, *60*, 699–746.

Sugiyama, M., Kanamori, T., Suzuki, T., Hido, S., Sese, J., Takeuchi, I. and Wang, L. (2009). A density-ratio framework for statistical data processing. *IPSJ Transactions on Computer Vision and Applications*, *1*, 183–208.

Sugiyama, M., Hara, S., von Bünau, P., Suzuki, T., Kanamori, T. and Kawanabe, M. (2010a). Direct density ratio estimation with dimensionality reduction, *Proceedings of 2010 SIAM International Conference on Data Mining (SDM2010)*, to appear.

Sugiyama, M., Kawanabe, M. and Chui, P. L. (2010b). Dimensionality reduction for density ratio estimation in high-dimensional spaces. *Neural Networks*, *23*, 44–59.

Sugiyama, M., Takeuchi, I., Suzuki, T., Kanamori, T., Hachiya, H. and Okanohara, D. (2010c). Least-squares conditional density estimation. *IEICE Transactions on Information and Systems*, *E93-D*, 583–594.

Suzuki, T. and Sugiyama, M. (2009). Estimating squared-loss mutual information for independent component analysis. *Proceedings of 8th International Conference on Independent Component Analysis and Signal Separation (ICA2009)* (pp. 130–137).

Suzuki, T. and Sugiyama, M. (2010). Sufficient dimension reduction via squared-loss mutual information estimation, *Proceedings of Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS2010)*, to appear.

Suzuki, T., Sugiyama, M., Sese, J. and Kanamori, T. (2008). Approximating mutual information by maximum likelihood density ratio estimation. *Proceedings of ECML-PKDD2008 Workshop on New Challenges for Feature Selection in Data Mining and Knowledge Discovery 2008 (FSDM2008)* (pp. 5–20).

Suzuki, T., Sugiyama, M., Kanamori, T. and Sese, J. (2009a). Mutual information estimation reveals global associations between stimuli and biological processes. *BMC Bioinformatics*, *10*, S52.

Suzuki, T., Sugiyama, M. and Tanaka, T. (2009b). Mutual information approximation via maximum likelihood estimation of density ratio. *Proceedings of 2009 IEEE International Symposium on Information Theory (ISIT2009)* (pp. 463–467).

Takeuchi, I., Nomura, K. and Kanamori, T. (2009). Nonparametric conditional density estimation using piecewise-linear solution path of kernel quantile regression. *Neural Computation*, *21*, 533–559.

Takimoto, M., Matsugu, M. and Sugiyama, M. (2009). Visual inspection of precision instruments by least-squares outlier detection. *Proceedings of Fourth International Workshop on Data-Mining and Statistical Science (DMSS2009)* (pp. 22–26).

Tsuboi, Y., Kashima, H., Hido, S., Bickel, S. and Sugiyama, M. (2009). Direct density ratio estimation for large-scale covariate shift adaptation. *Journal of Information Processing*, *17*, 138–155.

Ueki, K., Sugiyama, M. and Ihara, Y. (2010). Perceived age estimation under lighting condition change by covariate shift adaptation. submitted.

Vapnik, V. N. (1998). *Statistical learning theory*. Wiley, New York, NY, USA.

Wiens, D. P. (2000). Robust weights and designs for biased regression models: Least squares and generalized M-estimation. *Journal of Statistical Planning and Inference*, *83*, 395–412.

Yamada, M. and Sugiyama, M. (2010). Dependence minimizing regression with model selection for non-linear causal inference under non-Gaussian noise. submitted.

Yamada, M., Sugiyama, M. and Matsui, T. (2010). Semi-supervised speaker identification under covariate shift. *Signal Processing*, to appear.

Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. *Proceedings of Twenty-First International Conference on Machine Learning (ICML2004)* (pp. 903–910). ACM Press, New York, NY, USA.