オペレーションズ・リサーチ, *vol.57, no.7, pp.353-359, 2012.*

機械学習入門

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

概 要

本稿では，これまでに筆者が携わってきた機械学習の基礎技術と応用事例を紹介する．具体的には，非定常環境適応学習，異常値検出，二標本検定，特徴選択，条件付き確率推定などの基本的な原理とそれらの応用例を紹介する．最後に，これらの機械学習技術はすべて密度比とよばれる量に基づいていることを述べ，密度比推定の手法を紹介する．

キーワード

密度比推定，非定常環境適応，確率分布比較，相互情報量推定，条件付き確率推定

1 はじめに

近年，インターネットやセンサーを通して膨大な量のデータが容易に入手できるようになり，大量のデータからいかにして有用な知識を得るかが重要な研究課題となっている．このような背景のもと，データの統計的な性質を活用する**統計的機械学習**が，有望な情報処理パラダイムの一つとして注目されている [1, 2, 3].

本稿では，これまでに筆者が携わってきた機械学習の基礎技術と応用事例を紹介する．

2 非定常環境適応学習

入力と出力が対になったデータの背後に潜んでいる入出力関係を推定する問題を，**教師付き学習**とよぶ（図 1）．この名称は，入力が生徒の質問，出力が教師の答えに例えられることによる．未知の入出力関係を学習することができれば，学習に用いていない新しい入力に対する出力を予測できるようになる．未知の状況に対して一般化できるということから，これを**汎化能力**とよぶ．この汎化能力の獲得こそが，教師付き学習の目的である．

汎化能力の獲得を保証するために，学習に用いる訓練データと将来予測を行いたいテストデータが同じ規則に基づいて生成されているという条件が一般的に仮定される．し

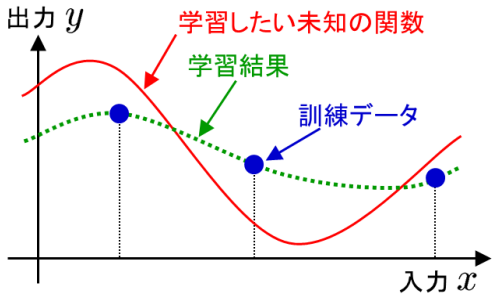


図 1: 教師付き学習．入力と出力が対になった訓練データから，その背後に潜んでいる入出力関係を推定する．

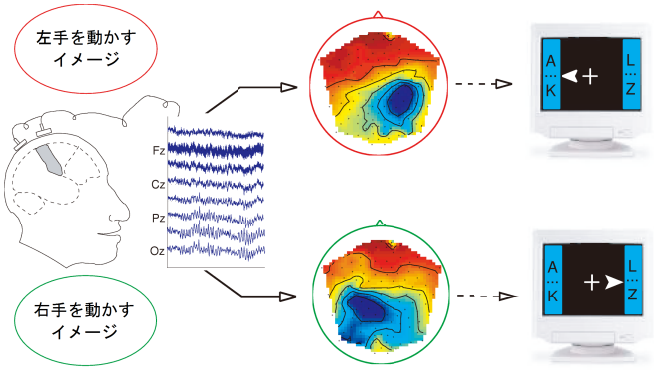
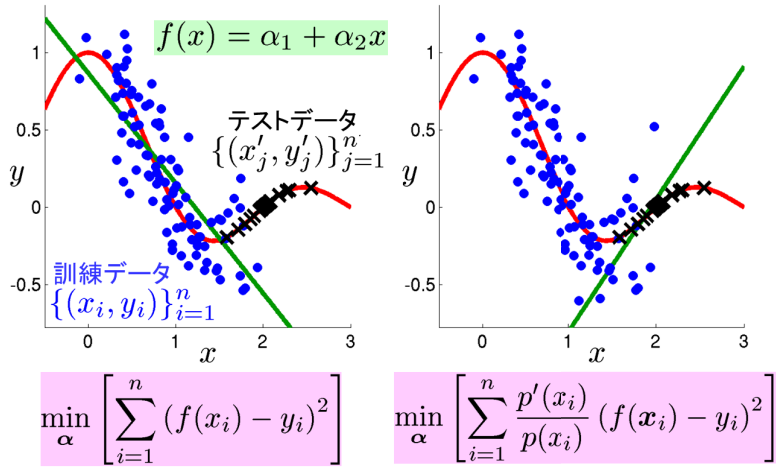


図 2: ブレイン・コンピュータ・インターフェース．脳波でコンピュータを直接操作する．

かし，近年の機械学習の多くの応用分野では，この基本的な仮定が成り立たないことが多い．例えば，脳波解析では脳の振る舞いが時間と共に変化をするため，訓練データとテストデータの傾向が異なる．一方，訓練データとテストデータが全く別の規則に基づいて生成されると，訓練データからテストデータの情報を予測することは原理的に不可能である．従って，訓練データとテストデータをつなぐ何らかの仮定は必ず必要である．

共変量シフトは，そのような仮定の一つである [23]．共変量とは入力データの別称であり，共変量シフトとは，入力データの生成規則が訓練時とテスト時で変化するが，入出力関係は変化しないという状況を指す．以下では，ブレイン・コンピュータ・インターフェース (BCI) を例に，共変量シフトに対処するための適応学習技術を紹介する．

BCI とは，脳波によって計算機に意志を伝える技術であり [5]，手足を動かすことのできない人でもコンピュータを操作できるようにするための重要な技術である (図 2)．ここでは，脳波でマウスカーソルを左右に動かすタスクを考えよう．脳波パターンを実ベクトル $\boldsymbol{x}$ で表し，その脳波によって伝えようとしている意志（左か右か）を $y \in \{+1, -1\}$ で表す．学習の目標は，訓練データ $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ からその背後に潜む入出力関係 $y = f(\boldsymbol{x})$ を獲得し，将来与えられるテスト入力 $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$ に対する出力 $\{y'_j\}_{j=1}^{n'}$ を正しく予測することである．ただし脳の非定常性のため，訓練入力 $\{\boldsymbol{x}_i\}_{i=1}^{n}$ とテスト入力 $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$ は一般に異なる確率分布に従う．



(a) 通常の最小二乗法　　(b) 重要度重み付き最小二乗法

図 3: 重要度重み付き最小二乗法による共変量シフト適応．(a) 通常の最小二乗法ではモデルを訓練データに適合させるため，テストデータが訓練データと異なる分布に従う場合はテストデータをうまく予想できない．(b) テストデータに近い訓練データに強い重みをつけることにより，テスト出力に適合させる．

学習に最小二乗法を用いるならば，関数 f を

$$\min_{f} \sum_{i=1}^{n} (f(\boldsymbol{x}_i) - y_i)^2$$

によって求めるが，共変量シフト下では入力分布の変化のため適切な解が得られない．共変量シフト下では，訓練入力の確率密度関数 $p(\boldsymbol{x})$ とテスト入力の確率密度関数 $p'(\boldsymbol{x})$ の比 $p'(\boldsymbol{x})/p(\boldsymbol{x})$ で重みをつけた最小二乗法を用いるのが良い（図 3）：

$$\min_{f} \sum_{i=1}^{n} \frac{p'(\boldsymbol{x}_i)}{p(\boldsymbol{x}_i)} (f(\boldsymbol{x}_i) - y_i)^2$$

このような重要度重み付けによる共変量シフト適応は，最小二乗法だけでなく損失関数を用いたあらゆる学習法に適用できる．

共変量シフト適応は，BCI の性能向上に有用であることが示されている [25, 16, 13]．他にも，ロボット制御における標本再利用 [7, 8] や最適データ収集 [4]，自然言語処理におけるドメイン適合 [39]，顔画像からの年齢予測における照明環境適合 [40]，加速度センサーからの行動識別におけるユーザ適合 [9]，話者識別における声質適合 [47]，半導体露光装置の位置合わせ [26] など，共変量シフト適応は様々な実問題に応用されている．

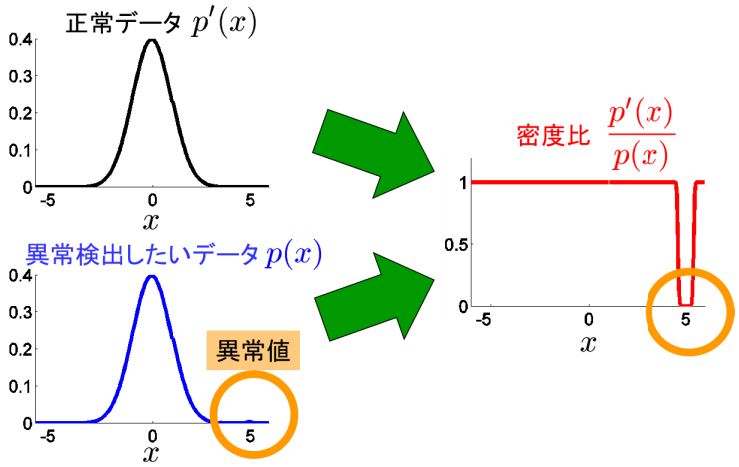


図 4: 密度比に基づく異常値検出．$p(\boldsymbol{x})$ の異常値を直接検出するのは困難だが（左下），正常データの密度（左上）との比を取ることにより，異常値が強調され検出が容易になる（右）．

3 確率分布比較

データ集合 $\{\boldsymbol{x}_i\}_{i=1}^n$ に含まれる異常値を見つける問題を，**異常値検出**とよぶ．このような入力データだけからの機械学習問題は，前述の教師付き学習と対比して**教師なし学習**とよばれる．一般に教師なし学習では，データ解析の目的があいまいなことが多い．異常値検出も例外ではなく，どういうデータを異常とみなすかを決めないと主観的な議論におちいってしまう．しかしながら，異常には様々なパターンが存在し得るため，あらかじめ異常とは何かを厳密に定義することは難しい．そこで，逆に正常とは何かを定義し，正常でないものを異常とみなすことにする．具体的には，異常を発見したいデータ集合 $\{\boldsymbol{x}_i\}_{i=1}^n$ 以外に，正常データの集合 $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$ が与えられると仮定し，$\{\boldsymbol{x}_i\}_{i=1}^n$ のうち $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$ から外れたものを異常値とみなす．この考え方は，異常値検出したいデータ集合 $\{\boldsymbol{x}_i\}_{i=1}^n$ の確率密度関数 $p(\boldsymbol{x})$ と正常データ $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$ の確率密度関数 $p'(\boldsymbol{x})$ の比 $p'(\boldsymbol{x})/p(\boldsymbol{x})$ を考え，この比の値が 1 から大きく離れたデータを異常値とみなすことにより実現できる（図 4）．このような方式に基づく異常値検出は，光学部品の異常検出 [38] やローン顧客の審査 [10] などに応用されている．

異常値検出は，二つの確率分布の一点を比較することに対応するが，二つの確率分布の全体を比較することも重要である．これは，二つのデータ集合 $\{\boldsymbol{x}_i\}_{i=1}^n$，$\{\boldsymbol{x}'_j\}_{j=1}^{n'}$ が同じ確率分布から生成されたかどうかを判定する問題に対応し，**二標本検定**とよぶ [28]．二標本検定は，二つの確率分布間の距離，例えば，カルバック距離

$$\int p'(\boldsymbol{x}) \log \frac{p'(\boldsymbol{x})}{p(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}$$

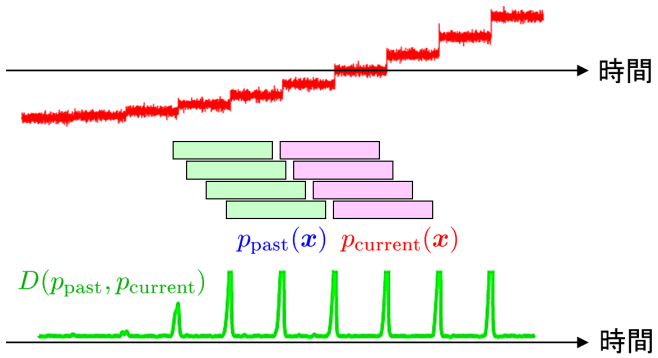


図 5: 密度比に基づく変化検知．過去の時系列データと現在の時系列データの分布間の距離 $D(p_{\mathrm{past}}, p_{\mathrm{current}})$ を推定することにより，時系列の傾向の変化を捉えることができる．

や，ピアソン距離

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

がある閾値より大きいかどうかを判定することにより実現できる[1]．二標本検定は，共変量シフトが起こっているかどうかの判定や，異なる状況で採取されたデータを合併して処理して良いかどうかの判定などに用いることができる．

また，過去の時系列データと現在の時系列データが従う分布間の距離を推定することにより（図 5），時系列の傾向の変化検出を行うこともできる [14, 17]．このような変化検出手法は，生体信号からの状態推定 [14]，画像中の注目領域の抽出 [49]，動画からのイベント抽出 [18]，ツイッターからのイベント抽出 [17] などに応用されている．

4 相互情報量推定

入出力データ $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ が与えられた時，入力 $\boldsymbol{x}$ と出力 $\boldsymbol{y}$ に依存性があるかどうかを判定することにより，様々なデータ解析が可能となる．例えば，入力ベクトルの一部の要素が出力と独立であることがわかると，そのような要素は教師付き学習においては無視することができる．これは，出力 $\boldsymbol{y}$ の予測に役立つ入力変数ベクトル $\boldsymbol{x}$ の部分集合を求めることに対応し，**特徴選択**とよぶ．特徴選択によりデータの解釈性が高まるため，例えば遺伝子解析に応用することができる [37]．一方，出力 $\boldsymbol{y}$ の予測の精度を向上させるために，入力ベクトル $\boldsymbol{x}$ を低次元表現に変換することを**特徴抽出**とよぶ．特徴抽出は，出力 $\boldsymbol{y}$ との依存性が最大の低次元表現を求める事により実現できる [35, 42]．

入力データ $\{\boldsymbol{x}_i\}_{i=1}^n$ だけが与えられる場合でも，$\{\boldsymbol{x}_i\}_{i=1}^n$ と最も依存性が高い出力 $\{y_i \in \{1, \ldots, k\}\}_{i=1}^n$ を求めることにより，データの**クラスタリング**を行うことができる [33, 15]．

[1] カルバック距離やピアソン距離は非対称で三角不等式を満たさないため，数学的な意味での距離ではない．しかし，常に非負の値を取り，ゼロになるのは二つの分布が等しい場合に限られるため，二つの分布の何らかの近さの尺度として用いることはできる．

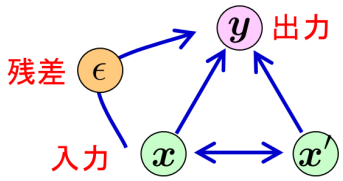
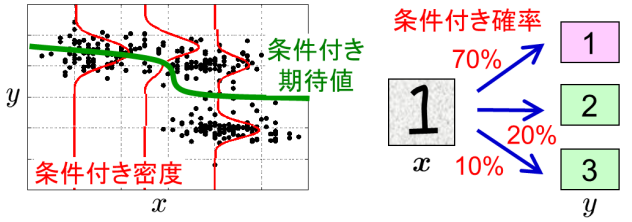


図 6: 相互情報量に基づく独立性判定．入出力間の独立性判定により，特徴選択，特徴抽出，クラスタリングが行え，入力間の独立性判定により，独立成分分析やオブジェクト適合が行える．また，入力と残差の間の独立性判定により，因果推論が行える．



(a) 条件付き密度推定　　(b) 確率的パターン認識

図 7: 条件付き確率推定．(a) 出力変数 $\boldsymbol{y}$ が連続値を取るとき，条件付き密度の推定に対応する．これは，条件付き期待値を推定する回帰分析の一般化になっており，出力の条件付き分布が多峰性や非対称性を持つときに有用である．(b) 出力変数 y がカテゴリ値を取るとき，確率的パターン認識とよばれ，カテゴリの予測だけでなく予測の信頼度も同時に得ることができる．

他にも依存性の推定により，**ブラインド信号源分離** [36]，**異ドメイン間オブジェクト適合** [45]，**独立性検定** [27]，**因果解析** [44] など様々なデータ解析を行うことができる（図 6）．

確率変数 $\boldsymbol{x}$ と $\boldsymbol{y}$ の依存性（独立性）は，$\boldsymbol{x}$ と $\boldsymbol{y}$ の同時確率密度 $p(\boldsymbol{x},\boldsymbol{y})$ から $\boldsymbol{x}$ と $\boldsymbol{y}$ の周辺確率密度の積 $p(\boldsymbol{x})p(\boldsymbol{y})$ までの距離によって見積ることができる．例えば，カルバック距離を用いた**相互情報量**

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

や，ピアソン距離を用いた**二乗損失相互情報量**

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

がよく用いられる．

5　条件付き確率推定

回帰とよばれる教師付き学習では，連続値をとる出力変数 $\boldsymbol{y}$ の，入力 $\boldsymbol{x}$ が与えられたもとでの**条件付き期待値**を推定する．しかし，出力 $\boldsymbol{y}$ の条件付き分布が多峰性や非対称

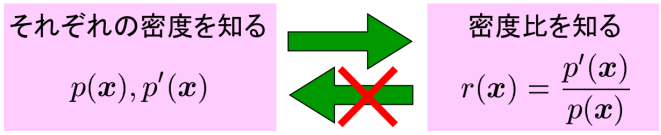


図 8: 密度比推定．分母と分子の密度 $p(\boldsymbol{x})$，$p'(\boldsymbol{x})$ がわかればそれらの比 $r(\boldsymbol{x})$ もわかるが，密度比 $r(\boldsymbol{x})$ がわかったとしてもそれぞれの密度はわからない．従って，分子と分母の密度を個別に推定するよりも，密度比を直接推定する方が易しいと考えられる．

性を持つときは，回帰分析では十分な情報が得られないため，**条件付き密度** $p(\boldsymbol{y}|\boldsymbol{x})$ そのものを推定することが重要である（図 7(a)）．このような条件付き密度の推定は，データの可視化や，移動ロボットの状態遷移確率 [32] などに応用できる．

一方，出力がカテゴリ値 $y \in \{1, \ldots, k\}$ を取るとき，条件付き確率 $p(y|\boldsymbol{x})$ はカテゴリの事後確率を表し，これを最大にするカテゴリを選ぶことにより**パターン認識**を行うことができる（図 7(b)）．このようなパターン認識法には，カテゴリの予測だけでなく予測の信頼度も同時に得られるという特徴があり，顔画像からの年齢予測 [41] や加速度センサーからの行動識別 [9] などに応用されている．

条件付き確率は，その定義から

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

と確率密度比の形で表すことができ，この形式を利用することにより精度よく推定できる [32, 22]．

6 密度比推定

ここまで，筆者が携わってきた機械学習の基礎的な技術とその応用例をいくつか紹介してきたが，これらの技術は全て確率密度の比の推定に基づいている [29]．そこで本節では，**密度比推定**の手法を紹介することにする．

確率密度 $p(\boldsymbol{x})$ を持つ確率分布に独立に従う標本 $\{\boldsymbol{x}_i\}_{i=1}^{n}$ と，確率密度 $p'(\boldsymbol{x})$ を持つ確率分布に独立に従う標本 $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$ から，確率密度比

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

を推定する問題を考える．

$\{\boldsymbol{x}_i\}_{i=1}^{n}$ と $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$ から $p(\boldsymbol{x})$ と $p'(\boldsymbol{x})$ をそれぞれ推定し，それらの比とれば密度比を推定することができる．しかし，このような素朴な方法では，必ずしも精度よく密度比を推定できるとは限らない（図 8）．以下では，密度比を直接推定する手法を紹介する．

6.1 確率的分類法

確率的分類法では，$p(\boldsymbol{x})$ と $p'(\boldsymbol{x})$ から生成された標本に，ラベル $y=+1$，-1 をそれぞれ割り当てる [20]．このとき，$p(\boldsymbol{x})$ と $p'(\boldsymbol{x})$ を

$$p(\boldsymbol{x}) = p(\boldsymbol{x}|y=+1), \quad p'(\boldsymbol{x}) = p(\boldsymbol{x}|y=-1)$$

と表すことができ，ベイズの定理より，密度比を

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

と表現できる．ここで，ラベルの事前確率 $p(y)$ の比を標本数の比で近似し，ラベルの事後確率 $p(y|\boldsymbol{x})$ を $\{\boldsymbol{x}_i\}_{i=1}^{n}$ と $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$ に対する確率的分類器 $\widehat{p}(y|\boldsymbol{x})$（例えば，ロジスティック回帰や最小二乗確率的分類により求める）で近似すれば，密度比の近似を求めることができる．

6.2 積率適合法

積率適合法では，密度比のモデル $g(\boldsymbol{x})$ を用いて，$g(\boldsymbol{x})p(\boldsymbol{x})$ の積率を $p'(\boldsymbol{x})$ の積率に最小二乗適合させる [6, 12]．例えば一次の積率（すなわち期待値）を適合させる場合は，次式を解く：

$$\min_{g} \|\mathbb{E}_{p}[\boldsymbol{x}g(\boldsymbol{x})] - \mathbb{E}_{p'}[\boldsymbol{x}]\|^2$$

ただし，$\|\cdot\|$ はユークリッドノルム，$\mathbb{E}$ は期待値を表す．真の密度比を正しく求めるためには全ての次数の積率を適合させる必要がある．ガウス核などの普遍再生核 $K(\boldsymbol{x}, \boldsymbol{x}')$ を用いれば，これを効率よく実現することができる：

$$\min_{g} \|\mathbb{E}_{p}[K(\boldsymbol{x}, \cdot)g(\boldsymbol{x})] - \mathbb{E}_{p'}[K(\boldsymbol{x}, \cdot)]\|_{\mathcal{H}}^2$$

ただし，$\|\cdot\|_{\mathcal{H}}$ は $K(\boldsymbol{x}, \boldsymbol{x}')$ が属するヒルベルト空間のノルムを表す．実際には，期待値を標本平均で近似した規準を最小化することにより解を求める．

6.3 密度適合法

密度適合法では，一般化カルバック距離のもとで $p'(\boldsymbol{x})$ に $g(\boldsymbol{x})p(\boldsymbol{x})$ を適合させる [31, 19]：

$$\min_{g} \mathbb{E}_{p'}\left[\log \frac{p'(\boldsymbol{x})}{g(\boldsymbol{x})p(\boldsymbol{x})}\right] + \mathbb{E}_{p}\left[g(\boldsymbol{x})\right]$$

ただし，実際の推定には期待値を標本平均で近似した規準を用いる．$g(\boldsymbol{x})$ として，線形モデル [31, 19]，対数線形モデル [39]，混合モデル [43, 48] を用いた手法が提案されている．

6.4 密度比適合法

密度比適合法では，密度比モデル $g(\boldsymbol{x})$ を真の密度比 $r(\boldsymbol{x})$ に最小二乗適合させる [11]：

$$\min_{g} \mathbb{E}_p \left[\left(g(\boldsymbol{x}) - r(\boldsymbol{x}) \right)^2 \right]$$

ただし，実際の推定には期待値を標本平均で近似した規準を用いる．$g(\boldsymbol{x})$ として線形モデルを用いれば，密度比適合法の解は解析的に求められる．更に非負拘束と ℓ_1 正則化項を加えた場合は，全ての正則化パラメータに対する解が効率よく計算できる．

6.5 統一的枠組み

上記の最小二乗密度比適合法を一般化し，ブレグマン距離のもとで $g(\boldsymbol{x})$ を $r(\boldsymbol{x})$ に適合させる [30]：

$$\min_{g} \mathbb{E}_p [f(r(\boldsymbol{x})) - f(g(\boldsymbol{x})) - \nabla f(g(\boldsymbol{x}))(r(\boldsymbol{x}) - g(\boldsymbol{x}))]$$

ただし，$f(t)$ は微分可能な強凸関数であり，$\nabla f(t)$ はその微分を表す．$f(t)$ を変えることにより，様々な密度比推定法が表現できる．

- ロジスティック回帰：$t \log t - (1+t) \log(1+t)$
- 再生核積率適合：$(t-1)^2/2$
- カルバック密度適合：$t \log t - t$
- 最小二乗密度比適合：$(t-1)^2/2$
- ロバスト密度比適合：$(t^{1+\alpha} - t)/\alpha$, $(\alpha > 0)$

6.6 次元削減付き密度比推定

ベクトル $\boldsymbol{x}$ を線形射影により $\boldsymbol{u}$ と $\boldsymbol{v}$ に分解したときに，$\boldsymbol{v}$ 成分が $p(\boldsymbol{x})$ と $p'(\boldsymbol{x})$ で共通，すなわち，ある共通の $p(\boldsymbol{v}|\boldsymbol{u})$ を用いて，$p(\boldsymbol{x})$ と $p'(\boldsymbol{x})$ が

$$p(\boldsymbol{x}) = p(\boldsymbol{v}|\boldsymbol{u})p(\boldsymbol{u}), \quad p'(\boldsymbol{x}) = p(\boldsymbol{v}|\boldsymbol{u})p'(\boldsymbol{u})$$

と表現できるならば，密度比 $r(\boldsymbol{x})$ を $p'(\boldsymbol{u})/p(\boldsymbol{u})$ と簡略化することができる．従って，$\boldsymbol{u}$ が属する部分空間（異分布部分空間とよぶ）を特定すれば，高次元の密度比推定問題を低次元の問題に還元できる．異分布部分空間の探索は，局所フィッシャー判別分析 [21] などの教師付き次元削減手法により $\{\boldsymbol{x}_i\}_{i=1}^n$ と $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$ を最もよく分離する部分空間を求める [24]，あるいは，$p'(\boldsymbol{u})$ から $p(\boldsymbol{u})$ へのピアソン距離を最大にする部分空間を求める [34, 46] ことにより行う．

7 まとめ

本稿では，これまでに筆者が携わってきた機械学習の基礎技術とその応用例を紹介した．そして，これら様々な機械学習タスクが，密度比推定により統一的に解決できることを示した．密度比推定の精度や計算効率を向上させれば，密度比推定に基づく全ての機械学習アルゴリズムの性能を改善できるため，密度比推定技術の今後の更なる発展が望まれる．また，密度比推定により解決できる新たな機械学習タスクを開拓することも重要な研究課題である．

密度比推定に関する論文やソフトウェアが，著者のホームページ

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

からダウンロードできる．また，密度比推定に関するより詳細な説明は，文献 [29] にまとめられている．興味を持って下さった方は，ご覧いただければ幸いである．

参考文献

[1] 元田浩, 栗田多喜夫, 樋口知之, 松本裕治, 村田昇（編）．パターン認識と機械学習（上）：ベイズ理論による統計的予測. シュプリンガー・ジャパン, 東京, 2007.

[2] 元田浩, 栗田多喜夫, 樋口知之, 松本裕治, 村田昇（編）．パターン認識と機械学習（下）：ベイズ理論による統計的予測. シュプリンガー・ジャパン, 東京, 2008.

[3] 杉山将. 統計的機械学習—生成モデルに基づくパターン認識. オーム社, 東京, 2009.

[4] T. Akiyama, H. Hachiya, and M. Sugiyama. Efficient exploration through active learning for value function approximation in reinforcement learning. *Neural Networks*, Vol. 23, No. 5, pp. 639–648, 2010.

[5] G. Dornhege, J. d. R. Millán, T. Hinterberger, D. McFarland, and K.-R. Müller, editors. *Toward Brain Computer Interfacing*. MIT Press, Cambridge, MA, USA, 2007.

[6] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf. Covariate shift by kernel mean matching. In *Dataset Shift in Machine Learning*, chapter 8, pp. 131–160. MIT Press, Cambridge, MA, USA, 2009.

[7] H. Hachiya, T. Akiyama, M. Sugiyama, and J. Peters. Adaptive importance sampling for value function approximation in off-policy reinforcement learning. *Neural Networks*, Vol. 22, No. 10, pp. 1399–1410, 2009.

[8] H. Hachiya, J. Peters, and M. Sugiyama. Reward weighted regression with sample reuse. *Neural Computation*, Vol. 11, No. 23, pp. 2798–2832, 2011.

[9] H. Hachiya, M. Sugiyama, and N. Ueda. Importance-weighted least-squares probabilistic classifier for covariate shift adaptation with application to human activity recognition. *Neurocomputing*, Vol. 80, pp. 93–101, 2012.

[10] S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, and T. Kanamori. Statistical outlier detection using direct density ratio estimation. *Knowledge and Information Systems*, Vol. 26, No. 2, pp. 309–336, 2011.

[11] T. Kanamori, S. Hido, and M. Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, Vol. 10, pp. 1391–1445, Jul. 2009.

[12] T. Kanamori, T. Suzuki, and M. Sugiyama. Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, Vol. 86, No. 3, pp. 335–367, 2012.

[13] M. Karasuyama, N. Harada, M. Sugiyama, and I. Takeuchi. Multi-parametric solution-path algorithm for instance-weighted support vector machines. *Machine Learning*, 2012. to appear.

[14] Y. Kawahara and M. Sugiyama. Sequential change-point detection based on direct density-ratio estimation. *Statistical Analysis and Data Mining*, Vol. 5, No. 2, pp. 114–127, 2012.

[15] M. Kimura and M. Sugiyama. Dependence-maximization clustering with least-squares mutual information. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 15, No. 7, pp. 800–805, 2011.

[16] Y. Li, H. Kambara, Y. Koike, and M. Sugiyama. Application of covariate shift adaptation techniques in brain computer interfaces. *IEEE Transactions on Biomedical Engineering*, Vol. 57, No. 6, pp. 1318–1324, 2010.

[17] S. Liu, M. Yamada, N. Collier, and M. Sugiyama. Change-point detection in time-series data by relative density-ratio estimation. Technical Report 1203.0453, arXiv, 2012.

[18] M. Matsugu, M. Yamanaka, and M. Sugiyama. Detection of activities and events without explicit categorization. In *Proceedings of the 3rd International Workshop on Video Event Categorization, Tagging and Retrieval for Real-World Applications (VECTaR2011)*, pp. 1532–1539, Barcelona, Spain, Nov. 13 2011.

[19] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, Vol. 56, No. 11, pp. 5847–5861, 2010.

[20] J. Qin. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, Vol. 85, No. 3, pp. 619–630, 1998.

[21] M. Sugiyama. Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis. *Journal of Machine Learning Research*, Vol. 8, pp. 1027–1061, May 2007.

[22] M. Sugiyama. Superfast-trainable multi-class probabilistic classifier by least-squares posterior fitting. *IEICE Transactions on Information and Systems*, Vol. E93-D, No. 10, pp. 2690–2701, 2010.

[23] M. Sugiyama and M. Kawanabe. *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*. MIT Press, Cambridge, MA, USA, 2012.

[24] M. Sugiyama, M. Kawanabe, and P. L. Chui. Dimensionality reduction for density ratio estimation in high-dimensional spaces. *Neural Networks*, Vol. 23, No. 1, pp. 44–59, 2010.

[25] M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, Vol. 8, pp. 985–1005, May 2007.

[26] M. Sugiyama and S. Nakajima. Pool-based active learning in approximate linear regression. *Machine Learning*, Vol. 75, No. 3, pp. 249–274, 2009.

[27] M. Sugiyama and T. Suzuki. Least-squares independence test. *IEICE Transactions on Information and Systems*, Vol. E94-D, No. 6, pp. 1333–1336, 2011.

[28] M. Sugiyama, T. Suzuki, Y. Itoh, T. Kanamori, and M. Kimura. Least-squares two-sample test. *Neural Networks*, Vol. 24, No. 7, pp. 735–751, 2011.

[29] M. Sugiyama, T. Suzuki, and T. Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, Cambridge, UK, 2012.

[30] M. Sugiyama, T. Suzuki, and T. Kanamori. Density ratio matching under the Bregman divergence: A unified framework of density ratio estimation. *Annals of the Institute of Statistical Mathematics*, 2012. to appear.

[31] M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, Vol. 60, No. 4, pp. 699–746, 2008.

[32] M. Sugiyama, I. Takeuchi, T. Suzuki, T. Kanamori, H. Hachiya, and D. Okanohara. Least-squares conditional density estimation. *IEICE Transactions on Information and Systems*, Vol. E93-D, No. 3, pp. 583–594, 2010.

[33] M. Sugiyama, M. Yamada, M. Kimura, and H. Hachiya. Information-maximization clustering based on squared-loss mutual information. Technical Report 1112.0611, arXiv, 2011.

[34] M. Sugiyama, M. Yamada, P. von Bünau, T. Suzuki, T. Kanamori, and M. Kawanabe. Direct density-ratio estimation with dimensionality reduction via least-squares hetero-distributional subspace search. *Neural Networks*, Vol. 24, No. 2, pp. 183–198, 2011.

[35] T. Suzuki and M. Sugiyama. Sufficient dimension reduction via squared-loss mutual information estimation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS2010)*, pp. 804–811, Sardinia, Italy, May 13-15 2010.

[36] T. Suzuki and M. Sugiyama. Least-squares independent component analysis. *Neural Computation*, Vol. 23, No. 1, pp. 284–301, 2011.

[37] T. Suzuki, M. Sugiyama, T. Kanamori, and J. Sese. Mutual information estimation reveals global associations between stimuli and biological processes. *BMC Bioinformatics*, Vol. 10, No. 1, p. S52, 2009.

[38] M. Takimoto, M. Matsugu, and M. Sugiyama. Visual inspection of precision instruments by least-squares outlier detection. In *Proceedings of the Fourth International Workshop on Data-Mining and Statistical Science (DMSS2009)*, pp. 22–26, Kyoto, Japan, Jul. 7–8 2009.

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

[40] K. Ueki, M. Sugiyama, and Y. Ihara. Lighting condition adaptation for perceived age estimation. *IEICE Transactions on Information and Systems*, Vol. E94-D, No. 2, pp. 392–395, 2011.

[41] K. Ueki, M. Sugiyama, Y. Ihara, and M. Fujita. Multi-race age estimation based on the combination of multiple classifiers. In *Proceedings of the First Asian Conference on Pattern Recognition (ACPR2011)*, pp. 633–637, Beijing, China, Nov. 28–30 2011.

[42] M. Yamada, G. Niu, J. Takagi, and M. Sugiyama. Computationally efficient sufficient dimension reduction via squared-loss mutual information. In C.-N. Hsu and W. S. Lee, editors, *Proceedings of the Third Asian Conference on Machine Learning (ACML2011)*, pp. 247–262, Taoyuan, Taiwan, Nov. 13-15 2011.

[43] M. Yamada and M. Sugiyama. Direct importance estimation with Gaussian mixture models. *IEICE Transactions on Information and Systems*, Vol. E92-D, No. 10, pp. 2159–2162, 2009.

[44] M. Yamada and M. Sugiyama. Dependence minimizing regression with model selection for non-linear causal inference under non-Gaussian noise. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI2010)*, pp. 643–648, Atlanta, Georgia, USA, Jul. 11–15 2010. The AAAI Press.

[45] M. Yamada and M. Sugiyama. Cross-domain object matching with model selection. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS2011)*, pp. 807–815, Fort Lauderdale, Florida, USA, Apr. 11-13 2011.

[46] M. Yamada and M. Sugiyama. Direct density-ratio estimation with dimensionality reduction via hetero-distributional subspace analysis. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence (AAAI2011)*, pp. 549–554, San Francisco, California, USA, Aug. 7–11 2011. The AAAI Press.

[47] M. Yamada, M. Sugiyama, and T. Matsui. Semi-supervised speaker identification under covariate shift. *Signal Processing*, Vol. 90, No. 8, pp. 2353–2361, 2010.

[48] M. Yamada, M. Sugiyama, G. Wichern, and J. Simm. Direct importance estimation with a mixture of probabilistic principal component analyzers. *IEICE Transactions on Information and Systems*, Vol. E93-D, No. 10, pp. 2846–2849, 2010.

[49] M. Yamanaka, M. Matsugu, and M. Sugiyama. Automatic detection of regions of interest based on density ratio estimation. In *Proceedings of 2011 Annual Conference of IEE of Japan*, pp. 143–149, Okinawa, Japan, Sep. 6–8 2011.