密度比推定
Density Ratio Estimation

杉山　将
東京工業大学　計算工学専攻
sugi@cs.titech.ac.jp
http://sugiyama-www.cs.titech.ac.jp/˜sugi

統計的機械学習のほとんど全てのタスクは，データの背後に潜む確率分布を推定することにより解決できる．しかし，確率分布の推定は困難であることが知られているため，これを回避しつつ所望のデータ処理を実現することが望ましい．ここでは，確率分布でなく確率密度の比を通して様々なデータ処理タスクを解決する密度比推定の枠組みを紹介する [1].

1　密度比推定手法

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

$$r^*(\boldsymbol{x}) = p^*_{\rm nu}(\boldsymbol{x})/p^*_{\rm de}(\boldsymbol{x})$$

を推定する問題を考える．‘nu’ と ‘de’ は，分子 (numerator) と分母 (denominator) の頭文字である．

1.1　確率的分類法 (Probabilistic Classification)

確率的分類法では，$p^*_{\rm nu}(\boldsymbol{x})$ と $p^*_{\rm de}(\boldsymbol{x})$ から生成された標本に，ラベル $y =$ ‘nu’ と ‘de’ をそれぞれ割り当てる．このとき，$p^*_{\rm nu}(\boldsymbol{x})$ と $p^*_{\rm de}(\boldsymbol{x})$ を

$$p^*_{\rm nu}(\boldsymbol{x}) = p^*(\boldsymbol{x}|y = \text{‘nu’}),\quad p^*_{\rm de}(\boldsymbol{x}) = p^*(\boldsymbol{x}|y = \text{‘de’})$$

と表すことができ，ベイズの定理より，密度比を

$$r^*(\boldsymbol{x}) = \frac{p^*(y = \text{‘de’})}{p^*(y = \text{‘nu’})}\frac{p^*(y = \text{‘nu’}|\boldsymbol{x})}{p^*(y = \text{‘de’}|\boldsymbol{x})}$$

と表現することができる．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 $\{\boldsymbol{x}^{\rm nu}_i\}_{i=1}^{n_{\rm nu}}$ と $\{\boldsymbol{x}^{\rm de}_j\}_{j=1}^{n_{\rm de}}$ 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 $\widehat{r}(\boldsymbol{x})$ を次式で求めることができる：

$$\widehat{r}(\boldsymbol{x}) = \frac{n_{\rm de}}{n_{\rm nu}}\frac{\widehat{p}(y = \text{‘nu’}|\boldsymbol{x})}{\widehat{p}(y = \text{‘de’}|\boldsymbol{x})}$$

1.2 積率適合法 (Moment Matching)

積率適合法では，密度比のモデル $r(\boldsymbol{x})$ を用いて，$r(\boldsymbol{x})p^*_{\mathrm{de}}(\boldsymbol{x})$ の積率を $p^*_{\mathrm{nu}}(\boldsymbol{x})$ の積率に最小二乗適合させる．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：

$$\min_r \left\| \mathbb{E}_{p^*_{\mathrm{de}}}[\boldsymbol{x} r(\boldsymbol{x})] - \mathbb{E}_{p^*_{\mathrm{nu}}}[\boldsymbol{x}] \right\|^2$$

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．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：

$$\min_r \left\| \mathbb{E}_{p^*_{\mathrm{de}}}[K(\boldsymbol{x}, \cdot) r(\boldsymbol{x})] - \mathbb{E}_{p^*_{\mathrm{nu}}}[K(\boldsymbol{x}, \cdot)] \right\|^2_{\mathcal{H}}$$

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．

1.3 密度適合法 (Density Fitting)

密度適合法では，一般化カルバック距離のもとで $p^*_{\mathrm{nu}}(\boldsymbol{x})$ に $r(\boldsymbol{x})p^*_{\mathrm{de}}(\boldsymbol{x})$ を適合させる：

$$\min_r \mathbb{E}_{p^*_{\mathrm{nu}}} \left[\log \frac{p^*_{\mathrm{nu}}(\boldsymbol{x})}{r(\boldsymbol{x}) p^*_{\mathrm{de}}(\boldsymbol{x})} \right] + \mathbb{E}_{p^*_{\mathrm{de}}} [r(\boldsymbol{x})]$$

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．$r(\boldsymbol{x})$ として，線形モデル，対数線形モデル，混合モデルを用いた手法が提案されている．

1.4 密度比適合法 (Density-Ratio Fitting)

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

$$\min_r \mathbb{E}_{p^*_{\mathrm{de}}} \left[\left(r(\boldsymbol{x}) - r^*(\boldsymbol{x}) \right)^2 \right]$$

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．$r(\boldsymbol{x})$ 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．

1.5 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 $r(\boldsymbol{x})$ を $r^*(\boldsymbol{x})$ に適合させる：

$$\begin{aligned} \min_r \mathbb{E}_{p^*_{\mathrm{de}}} [& f(r^*(\boldsymbol{x})) - f(r(\boldsymbol{x})) \\ & - f'(r(\boldsymbol{x}))(r^*(\boldsymbol{x}) - r(\boldsymbol{x}))] \end{aligned}$$

ただし，$f(t)$ は微分可能な強凸関数であり，$f'(t)$ 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：$(t^{1+\alpha}-t)/\alpha$, $(\alpha>0)$

1.6 次元削減付き密度比推定

$\boldsymbol{x}$ を線形射影により $\begin{bmatrix}\boldsymbol{u}\\ \boldsymbol{v}\end{bmatrix}$ と分解したときに，$\boldsymbol{v}$ 成分が $p^*_{\mathrm{nu}}(\boldsymbol{x})$ と $p^*_{\mathrm{de}}(\boldsymbol{x})$ で共通，すなわち，ある共通の $p^*(\boldsymbol{v}|\boldsymbol{u})$ を用いて $p^*_{\mathrm{nu}}(\boldsymbol{x})$ と $p^*_{\mathrm{de}}(\boldsymbol{x})$ が

$$p^*_{\mathrm{nu}}(\boldsymbol{x})=p^*(\boldsymbol{v}|\boldsymbol{u})p^*_{\mathrm{nu}}(\boldsymbol{u}),\ p^*_{\mathrm{de}}(\boldsymbol{x})=p^*(\boldsymbol{v}|\boldsymbol{u})p^*_{\mathrm{de}}(\boldsymbol{u})$$

と表現できるならば，密度比 $r^*(\boldsymbol{x})$ を $p^*_{\mathrm{nu}}(\boldsymbol{u})/p^*_{\mathrm{de}}(\boldsymbol{u})$ 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などの教師付き次元削減手法により $\{\boldsymbol{x}^{\mathrm{nu}}_i\}_{i=1}^{n_{\mathrm{nu}}}$ と $\{\boldsymbol{x}^{\mathrm{de}}_j\}_{j=1}^{n_{\mathrm{de}}}$ を最もよく分離する部分空間を求める，あるいは，$p^*_{\mathrm{nu}}(\boldsymbol{u})$ から $p^*_{\mathrm{de}}(\boldsymbol{u})$ へのピアソン距離

$$\mathbb{E}_{p^*_{\mathrm{de}}}[(p^*_{\mathrm{nu}}(\boldsymbol{u})/p^*_{\mathrm{de}}(\boldsymbol{u})-1)^2]$$

を最大にする部分空間を求めることにより行う．

2 密度比に基づく機械学習

2.1 重点標本化 (Importance Sampling)

入力 $\boldsymbol{x}$ から出力 y への変換規則を学習する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 $p^*_{\mathrm{tr}}(\boldsymbol{x})$ から $p^*_{\mathrm{te}}(\boldsymbol{x})$ に変化するが，入出力関係 $p^*(y|\boldsymbol{x})$ は変化しない状況を共変量シフトとよぶ [2]．共変量シフト下では最尤推定などの学習法はバイアスを持つが，このバイアスは損失関数を重要度 $p^*_{\mathrm{te}}(\boldsymbol{x})/p^*_{\mathrm{tr}}(\boldsymbol{x})$ に従って重み付けすることにより打ち消すことができる．すなわち，損失関数 $\ell(\boldsymbol{x})$ の $p^*_{\mathrm{te}}(\boldsymbol{x})$ に関する期待値は，損失関数の $p^*_{\mathrm{tr}}(\boldsymbol{x})$ に関する重要度重み付き期待値により計算できる：

$$\begin{aligned}\mathbb{E}_{p^*_{\mathrm{te}}}[\ell(\boldsymbol{x})]&=\int\ell(\boldsymbol{x})p^*_{\mathrm{te}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\&=\int\ell(\boldsymbol{x})\frac{p^*_{\mathrm{te}}(\boldsymbol{x})}{p^*_{\mathrm{tr}}(\boldsymbol{x})}p^*_{\mathrm{tr}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=\mathbb{E}_{p^*_{\mathrm{tr}}}\left[\ell(\boldsymbol{x})\frac{p^*_{\mathrm{te}}(\boldsymbol{x})}{p^*_{\mathrm{tr}}(\boldsymbol{x})}\right]\end{aligned}$$

交差確認などのモデル選択法も共変量シフト下では不偏性を失うが，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．

2.2 分布比較 (Distribution Comparison)

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．これら二つの標本集合に対する密度比を考えれば，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．従って，密度比の値を評価基準とすることにより異常値を検出することができる．

また，密度比推定量$\widehat{r}(\boldsymbol{x})$を用いることにより，分布間の距離を精度よく推定することができる：

- カルバック距離： $\frac{1}{n_{\mathrm{nu}}}\sum_{i=1}^{n_{\mathrm{nu}}}\log\widehat{r}(\boldsymbol{x}_i^{\mathrm{nu}})$
- ピアソン距離: $\frac{2}{n_{\mathrm{nu}}}\sum_{i=1}^{n_{\mathrm{nu}}}\widehat{r}(\boldsymbol{x}_i^{\mathrm{nu}})$
  $-\frac{1}{n_{\mathrm{de}}}\sum_{j=1}^{n_{\mathrm{de}}}\widehat{r}(\boldsymbol{x}_j^{\mathrm{de}})^2-1$

これらの距離推定量を用いれば，並べ替え検定により二つの分布の同一性を検定することができる．

2.3 相互情報量推定 (Mutual Information)

確率密度$p^*_{\mathrm{x,y}}(\boldsymbol{x},\boldsymbol{y})$を持つ分布に独立に従う$n$個の標本$\{(\boldsymbol{x}_k,\boldsymbol{y}_k)\}_{k=1}^n$から，$\boldsymbol{x}$と$\boldsymbol{y}$の相互情報量

$$\mathbb{E}_{p^*_{\mathrm{x,y}}}\left[\log\frac{p^*_{\mathrm{x,y}}(\boldsymbol{x},\boldsymbol{y})}{p^*_{\mathrm{x}}(\boldsymbol{x})p^*_{\mathrm{y}}(\boldsymbol{y})}\right]$$

を推定する問題を考える．ただし，$p^*_{\mathrm{x}}(\boldsymbol{x})$と$p^*_{\mathrm{y}}(\boldsymbol{y})$は$\boldsymbol{x}$と$\boldsymbol{y}$の周辺密度である．$\{(\boldsymbol{x}_k,\boldsymbol{y}_k)\}_{k=1}^n$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，密度比推定により相互情報量が推定できる．同様に，二乗損失版の相互情報量

$$\mathbb{E}_{p^*_{\mathrm{x}}(\boldsymbol{x})}\mathbb{E}_{p^*_{\mathrm{y}}(\boldsymbol{y})}\left[\left(\frac{p^*_{\mathrm{x,y}}(\boldsymbol{x},\boldsymbol{y})}{p^*_{\mathrm{x}}(\boldsymbol{x})p^*_{\mathrm{y}}(\boldsymbol{y})}-1\right)^2\right]$$

も推定できる．相互情報量は確率変数間の独立性を表す指標であり，その推定量は，独立性検定，特徴選択，特徴抽出，クラスタリング，独立成分分析，オブジェクト適合，因果推定など，様々な機械学習タスクに応用することができる．

2.4 条件付き確率 (Conditional Probability) 推定

確率密度$p^*_{\mathrm{x,y}}(\boldsymbol{x},\boldsymbol{y})$を持つ分布に独立に従う$n$個の標本$\{(\boldsymbol{x}_k,\boldsymbol{y}_k)\}_{k=1}^n$から，条件付き確率

$$p^*_{\mathrm{y|x}}(\boldsymbol{y}|\boldsymbol{x})=\frac{p^*_{\mathrm{x,y}}(\boldsymbol{x},\boldsymbol{y})}{p^*_{\mathrm{x}}(\boldsymbol{x})}$$

を推定する問題を考える．$\{(\boldsymbol{x}_k,\boldsymbol{y}_k)\}_{k=1}^n$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，密度比推定により条件付き確率が推定できる．$\boldsymbol{y}$が連続変数の場合，これは条件付き密度推定に対応し，$\boldsymbol{y}$がカテゴリ変数の場合は確率的パターン認識となる．

3　まとめ

様々な機械学習タスクを統一的に解決できる密度比推定の枠組みを紹介した．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，密度比推定手法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である．

References

[1] M. Sugiyama, T. Suzuki & T. Kanamori: *Density Ratio Estimation in Machine Learning*, Cambridge University Press, 2012.

[2] M. Sugiyama & M. Kawanabe: *Machine Learning in Non-Stationary Environments*, MIT Press, 2012.