データ・サイエンティスト育成のためのビッグデータの利活用術,
pp. 81–87, エヌ・ティー・エス, 東京, 2014.

最小二乗法によるビッグデータ解析

東京工業大学　杉山将

本節では，最小二乗法を用いた教師付きのビッグデータ解析手法を概説する [1].

１．教師付き学習

教師付き学習とは，入力$\boldsymbol{x}$と出力yとをつなぐ関数$y = f(\boldsymbol{x})$を，入出力が対になったn個の訓練標本$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$から学習する問題である．入力$\boldsymbol{x}$は一般に$d$次元の実数値ベクトルであり，出力$y$は実数値，あるいは，$\pm 1$のようなカテゴリ値を取る．出力$y$が実数値を取る場合は回帰問題とよび（図１），カテゴリ値を取る場合は分類問題とよぶ（図２）．

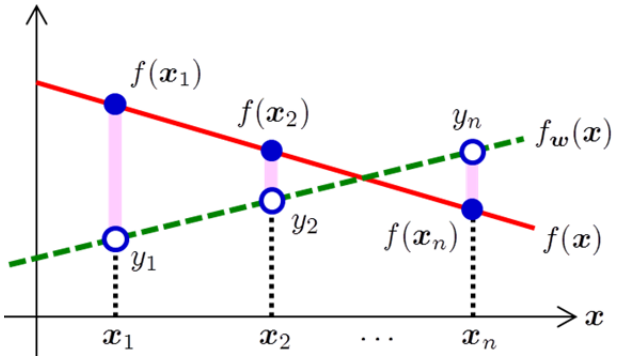

図１：教師付き回帰.

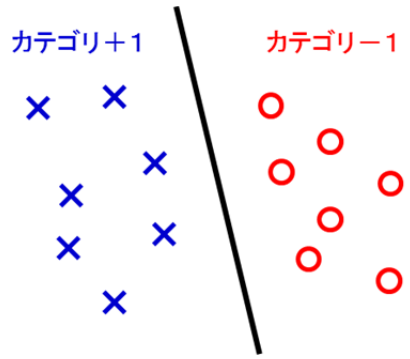

図２：教師付き分類

ビッグデータとは，入力$\boldsymbol{x}$の次元数dと訓練標本数nが非常に大きい状況を指す．ビッグデータ解析においては，以下の単純な線形モデル$f_{\boldsymbol{w}}$を用いて関数fをモデル化する事が多い．

$$f_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{w}$$

ここで$\boldsymbol{w}$はモデルのパラメータでありd次元の実数値ベクトルである．モデル$f_{\boldsymbol{w}}$が真の関数fにできるだけ近くなるようにパラメータ$\boldsymbol{w}$を学習することが，教師付き学習の目的である．

以下では，まずは出力yが実数値を取る回帰問題に対する学習法を紹介し，続いて出力yがカテゴリ値を取る分類問題に対する学習法を紹介する．最後に，最小二乗法の更なる応用について述べる．

２．最小二乗回帰とその確率的勾配法による実装

回帰問題では，パラメータ$\boldsymbol{w}$を次の二乗損失を最小化するように学習する．

$$\min_{\boldsymbol{w}} \sum_{i}^{n} \left(y_i - \boldsymbol{x}_i^T \boldsymbol{w}\right)^2$$

この最小解$\hat{\boldsymbol{w}}$は以下のように解析的に求められる．

$$\hat{\boldsymbol{w}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

ただし，$\boldsymbol{X}$は訓練入力標本を並べた$n \times d$行列であり，$\boldsymbol{y}$は訓練出力標本を並べたn次元の縦ベクトルである．

$$\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T, \qquad \mathbf{y} = (y_1, \ldots, y_n)^T$$

しかし，入力$\boldsymbol{x}$の次元数dと訓練標本数nが非常に大きいとき，訓練入力標本行列$\boldsymbol{X}$をメモリに保持することは困難である．また，$d \times d$行列である$\boldsymbol{X}^T \boldsymbol{X}$の逆を計算することも困難である．

そこで，次の確率的勾配法を用いて逐次的に解を求めることにする（図３）．

（１） パラメータ$\boldsymbol{w}$を適当に初期化する．

（２） n個の訓練標本$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$からランダムに一つ選ぶ．

（３） 選んだ訓練標本$(\boldsymbol{x}, y)$に対する二乗誤差$(y - \boldsymbol{x}^T \boldsymbol{w})^2$を小さくするようにパラメータ$\boldsymbol{w}$を更新する．

$$\boldsymbol{w} \leftarrow \boldsymbol{w} + \epsilon \cdot (y - \boldsymbol{x}^T \boldsymbol{w}) \boldsymbol{x}$$

ただし，ϵは勾配降下の歩幅を表す小さい正のスカラーである．

（４） 収束するまで（２）～（３）を繰り返す．

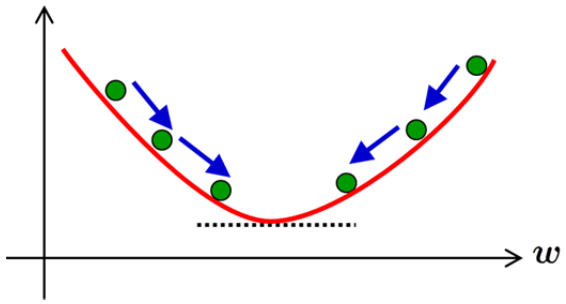

図３：確率的勾配法．誤差規準を小さくするように勾配を降下していく．

３．受動攻撃回帰

確率的勾配法はビッグデータ解析の有効な学習アルゴリズムであるが，勾配降下の歩幅ϵの決め方が難しいという問題がある．すなわち，ϵが大きすぎると収束せず（図４），ϵが小さすぎると収束に時間がかかる（図５）．最初はϵを大きめに設定し，学習が進むにつれて徐々にϵを小さくしていくのが妥当と考えられるが，どのくらいの速さでϵを減らすかを決めるのもまた困難である．

一方，最小二乗回帰では一気に谷底までジャンプするようにパラメータを更新することが可能である．そうすると勾配降下の歩幅ϵを決める必要がなく，実用上非常に便利である．

しかし，このような急激な学習を行うとパラメータの値が大きく変化するため，これまでの学習で得られていた知識がたった一つのデータの追加によって失われてしまう恐れがある．

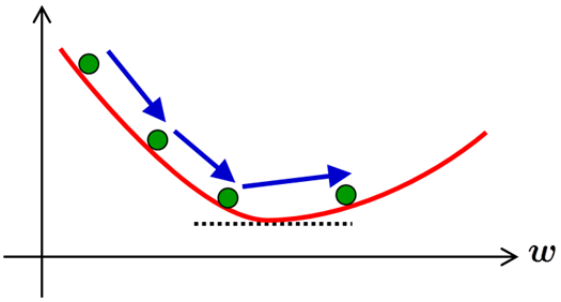

図４：確率的勾配法の問題点１．
歩幅が大きすぎると収束しない．

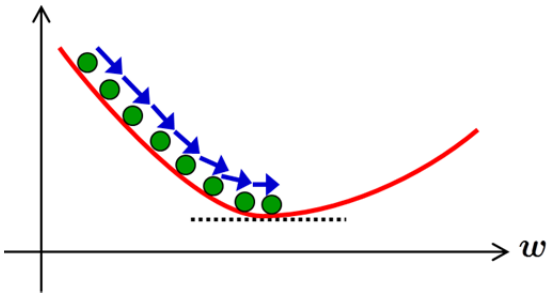

図５：確率的勾配法の問題点２．
歩幅が小さすぎると収束が遅い．

この問題を避けるためには，現在の解からの移動量を考慮しながら勾配降下を行う受動攻撃回帰が有効である．受動攻撃回帰では，次式を最小にするようにパラメータ$\boldsymbol{w}$を更新する．

$$\min_{\boldsymbol{w}}[(y-\boldsymbol{x}^T\boldsymbol{w})^2+\gamma\|\boldsymbol{w}-\tilde{\boldsymbol{w}}\|^2]$$

ただし，$\tilde{\boldsymbol{w}}$は現在のパラメータ値，$\|\cdot\|$はユークリッドノルム，γは現在の解からの移動量を調整する正のスカラーをそれぞれ表す．この最小化問題は解析的に解くことができ，次のパラメータ更新式が得られる．

$$\boldsymbol{w}\leftarrow\boldsymbol{w}+\frac{y-\boldsymbol{x}^T\boldsymbol{w}}{\|\boldsymbol{x}\|^2+\gamma}\boldsymbol{x}$$

これは，確率的勾配法において歩幅を$\epsilon=1/(\|\boldsymbol{x}\|^2+\gamma)$とおくことに対応しており，受動攻撃回帰では$\boldsymbol{x}$のノルムに合わせて適応的に歩幅を調整していることがわかる．

４．適応正則化回帰

統計的な枠組みのもとでパラメータ$\boldsymbol{w}$をデータから学習する限りは，推定に対する不確定性が必ず伴う．すなわち，パラメータ$\boldsymbol{w}$も確率分布を持つ．受動攻撃回帰では，現在のパラメータ値$\tilde{\boldsymbol{w}}$からの移動量をユークリッドノルムで測ったが，パラメータが確率分布を持つ場合は，例えばカルバック・ライブラー距離など，確率分布間の距離尺度を用いる方が自然である．確率密度関数pからqへのカルバック・ライブラー距離は，次式で定義される．

$$\mathrm{KL}(p||q)=\int p(\boldsymbol{x})\log\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}$$

パラメータの分布として，ガウス分布を用いることにする．ガウス分布の確率密度関数は次式で定義される（図６）．

$$N_{\boldsymbol{w}}(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}\det(\boldsymbol{\Sigma})^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{w}-\boldsymbol{\mu})\right)$$

ただし，$\boldsymbol{\mu},\boldsymbol{\Sigma}$はガウス分布の期待値ベクトルと分散共分散行列を表し，$\det(\cdot)$は行列式を表す．

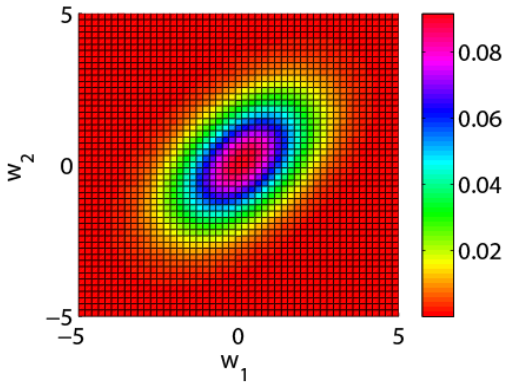

図６：ガウス分布．楕円形の等高線を持つ．

適応正則化回帰では，パラメータ$\boldsymbol{w}$そのものでなくその期待値と分散共分散行列$\boldsymbol{\mu},\boldsymbol{\Sigma}$を考え，次式を最小にするように$\boldsymbol{\mu},\boldsymbol{\Sigma}$を更新する．

$$\min_{\boldsymbol{\mu},\boldsymbol{\Sigma}}\left[(y-\boldsymbol{x}^T\boldsymbol{\mu})^2 + 2\gamma\cdot\mathrm{KL}\left(N(\boldsymbol{\mu},\boldsymbol{\Sigma})||N(\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}})\right) + \boldsymbol{x}^T\boldsymbol{\Sigma}\boldsymbol{x}\right]$$

ただし，$\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}}$は現在のパラメータ$\tilde{\boldsymbol{w}}$の期待値と分散共分散行列を表す．第３項目の$\boldsymbol{x}^T\boldsymbol{\Sigma}\boldsymbol{x}$は，分散共分散行列$\boldsymbol{\Sigma}$を標本$\boldsymbol{x}$に合わせて適応的に正則化する役割を果たす．この最小化問題は解析的に解くことができ，次の更新式が得られる．

$$\boldsymbol{\mu}\leftarrow\boldsymbol{\mu}+\frac{y-\boldsymbol{x}^T\boldsymbol{\mu}}{\boldsymbol{x}^T\boldsymbol{\Sigma}\boldsymbol{x}+\gamma}\boldsymbol{\Sigma}\boldsymbol{x},\quad \boldsymbol{\Sigma}\leftarrow\boldsymbol{\Sigma}-\frac{1}{\boldsymbol{x}^T\boldsymbol{\Sigma}\boldsymbol{x}+\gamma}\boldsymbol{\Sigma}\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{\Sigma}$$

$\boldsymbol{\mu}$の更新式を受動攻撃回帰の$\boldsymbol{w}$の更新式と比べると，分散共分散行列$\boldsymbol{\Sigma}$が表す楕円形状（図６）を考慮した形になっていることがわかる．$\boldsymbol{\Sigma}$が単位行列のとき（すなわち楕円が単位円になるとき），$\boldsymbol{\mu}$の更新式は受動攻撃回帰の$\boldsymbol{w}$の更新式と一致する．

分散共分散行列$\boldsymbol{\Sigma}$の計算が煩雑な場合は，対角成分のみを残して非対角成分をゼロにするという近似を行うことにより，計算時間とメモリ消費量を大幅に削減できる．

５．二値分類問題への拡張

ここまで，出力yが実数値を取る回帰問題に対する学習アルゴリズムを紹介してきた．二値分類問題では出力yは±1のカテゴリ値を取るが，これを実数とみなせば上記の回帰アルゴリズムをそのまま分類学習にも適用できる．そして，テスト入力$\boldsymbol{x}$に対するモデル$f_{\boldsymbol{w}}(\boldsymbol{x})$の出力をその正負に従って$\pm1$のカテゴリ値に変換すれば，テスト入力の分類が行なえる．

しかし，分類問題では学習した関数の符号のみが重要であることから，分類学習に二乗

誤差を用いるのはやや不自然である．分類学習では，誤分類数に対応する0/1損失の最小化が理想的である．

$$\min_{\boldsymbol{w}} \sum_{i}^{n} I\left(y_i \boldsymbol{x}_i^T \boldsymbol{w} > 0\right)$$

ここで，$I(\cdot)$は条件が真なら0ば，偽ならば1を出力する指示関数である．しかし，0/1損失の最小化はn個の標本のカテゴリ割り当てを求める離散最適化問題に帰着されるため，2^nの組み合わせから最適なカテゴリの割り当てを求めることは一般に困難である．

そこで，次の二乗ヒンジ損失を代理損失として用いることにする．

$$\min_{\boldsymbol{w}} \sum_{i}^{n} \left(\max(0,1 - y_i \boldsymbol{x}_i^T \boldsymbol{w})\right)^2$$

図７に示すように，二乗ヒンジ損失は0/1損失の滑らかな上界となっている．一方，出力yが± 1のカテゴリ値を取るとき，二乗損失の最小化は

$$\min_{\boldsymbol{w}} \sum_{i}^{n} \left(1 - y_i \boldsymbol{x}_i^T \boldsymbol{w}\right)^2$$

と表現できる．図７に示すように，二乗損失も0/1損失の上界となっていることがわかるが，右側で損失が増加するため0/1損失と形状が大きく異なる．

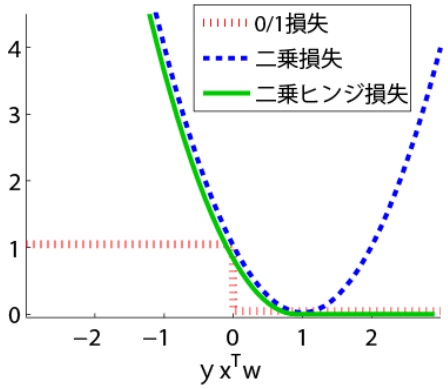

図７：分類問題の損失関数．

この二乗ヒンジ損失を用いた分類学習規準に対しても，回帰の場合と同様にして学習アルゴリズムを導出できる．

- 確率的勾配分類のパラメータ更新式：

$$\boldsymbol{w} \leftarrow \boldsymbol{w} + \epsilon \cdot \max(0,1 - y\boldsymbol{x}^T\boldsymbol{w}) y\boldsymbol{x}$$

- 受動攻撃分類のパラメータ更新式：

$$\boldsymbol{w} \leftarrow \boldsymbol{w} + \frac{\max(0,1 - y\boldsymbol{x}^T\boldsymbol{w}) y}{\|\boldsymbol{x}\|^2 + \gamma} \boldsymbol{x}$$

- 適応正則化分類のパラメータ更新式：

$$\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} + \frac{\max(0,1 - y\boldsymbol{x}^T\boldsymbol{\mu})y}{\boldsymbol{x}^T\boldsymbol{\Sigma}\boldsymbol{x} + \gamma}\boldsymbol{\Sigma}\boldsymbol{x}, \quad \boldsymbol{\Sigma} \leftarrow \boldsymbol{\Sigma} - \frac{1}{\boldsymbol{x}^T\boldsymbol{\Sigma}\boldsymbol{x} + \gamma}\boldsymbol{\Sigma}\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{\Sigma}$$

これらは，回帰学習のパラメータ更新式の$y - \boldsymbol{x}^T\boldsymbol{w}$を$\max(0,1 - y\boldsymbol{x}^T\boldsymbol{w})y$に置き換えた形になっている．

６．まとめ

本節では，最小二乗法に基づく教師付きのデータ解析手法を紹介した．これらの学習法はオンライン学習とよばれる手法であり，わずか数行のプログラムによって簡単に実装できるという優れた特徴を持っている．ビッグデータ解析にはこのような単純な学習法が特に有効であり，今後益々重要性が高まってくるものと思われる．

このように回帰と分類に対してはビッグデータ解析に適した優れた学習アルゴリズムが開発されているが，回帰と分類以外にも，クラスタリング，異常検知，特徴選択など様々なデータタスクが存在する．機械学習の最も汎用的なアプローチは，データを生成する確率分布を推定することである[2]．なぜならば，データの生成分布を知ることは，そのデータに関する全ての知識を得ることと本質的に等価だからである．従って，データの生成分布がうまく推定できれば，あらゆるデータ解析タスクを精度良く解決できる．しかし，高次元の確率分布をうまく推定することは一般に困難であり，データの生成分布に基づくアプローチでは必ずしも精度良くデータ解析を行えない．

一方，クラスタリング，異常検知，特徴選択などの様々なデータ解析タスクそれぞれに対して，ビッグデータ解析に適した優れたアルゴリズムを個別に開発することも現実的には困難である．実際，回帰や分類の基礎理論は既に１９６０年ころには研究されており，半世紀にも及ぶ更なる研究開発を経て，様々な実データ解析に応用されるようになった．ビッグデータ時代には大量のデータを前に迅速な意思決定を行うことが望まれており，様々なデータ解析タスクに対して長期間に及ぶ研究開発を行うことは現実的でない．

この状況を打破するために，生成分布推定アプローチとタスク特化アプローチの中間的なアプローチが近年提案された[3,4]．それは，ある条件を満たすデータ解析タスクのクラスに対して学習アルゴリズムを開発するというアプローチである（図８）．具体的には，複数の確率分布が含まれるデータ解析タスクのうち，それらの確率分布そのものは必要なく，確率密度関数の比さえわかればデータ解析を行えるというクラスを考える．このクラスには，非定常環境適応学習[5,6]，特徴選択，クラスタリング[7]，条件付き確率推定，独立成分分析，異常値検出，変化検知[8]など，多くの重要なデータ解析タスクが含まれる．そして，この確率密度関数の比を，それぞれの確率密度関数を推定することなく直接推定する．密度比の推定は確率密度の推定よりも一般に容易であるため（図９），密度比推定によって上記の全てのデータ解析タスクを統一的かつ優れた精度で解決できる．そして，密度比は最小二乗法によって推定できるため，本節で紹介した学習アルゴリズムと同等の考え方を密度比推定にも適用できる．

密度比推定によって様々な機械学習タスクを統一的に解決できるため，密度比推定の精度や計算効率を更に向上させることにより，様々な機械学習アルゴリズムの性能を一挙に

改善できる．今後は，密度比推定の基礎技術を更に発展させていくとともに，密度比推定によって解決できる新たなデータ解析タスクを開拓し，それらの機械学習技術を様々な実世界問題の解決に活用してくことが期待される．

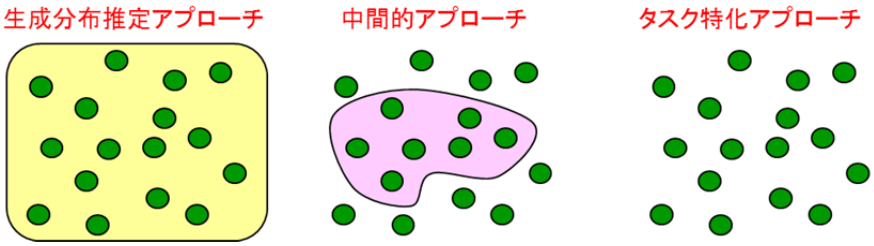

図８：機械学習のアプローチ．各●印は回帰，分類などのデータ解析タスクを表す．

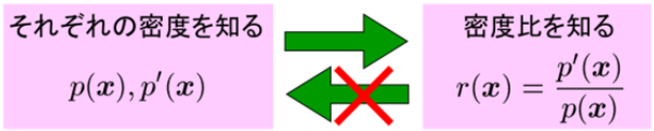

図９：密度比推定．それぞれの密度を推定するよりも密度比を推定するほうが容易である．

参考文献

1) 杉山 将．イラストで学ぶ機械学習：最小二乗法による識別モデル学習を中心に，講談社, 2013.

2) 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, 2009.

3) M. Sugiyama et al. Density Ratio Estimation in Machine Learning, Cambridge University Press, 2012.

4) 杉山 将．機械学習入門．オペレーションズ・リサーチ, vol.57, no.7, pp.353-359, 2012.

5) M. Sugiyama et al. Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation, MIT Press, 2012.

6) 杉山 将．非定常環境下での教師付き学習：データの入力分布が変化する場合．画像ラボ, vol.18, no.10, pp.1-6, 2007.

7) 杉山 将．機械学習によるデータの自動クラスタリング．シミュレーション, vol.31, no.2, pp.36-40, 2012.

8) 杉山 将．確率分布間の距離推定：機械学習分野における最新動向．日本応用数理学会論文誌, vol.23, no.3, pp.439-452, 2013.