適応正則化オンライン学習における特徴選択問題

大阪府立大学大学院工学研究科　野崎 俊貴
電気通信大学大学院情報システム学研究科　川野 秀一

1. はじめに

オンライン学習とは近年機械学習の分野において注目を集めている学習法であり，その利点は，逐次的にデータを学習しモデルを構成することにある．これまで多くのオンライン学習法が提案されてきているが，その中でも Crammer *et al.* (2009) により提案された適応正則化学習に着目する．適応正則化学習はデータの外れ値に強く，また信頼度を用いることで更新回数の少ない重みベクトルをより大きく更新することが特徴的である．しかし，適応正則化学習では，将来の予測に寄与しない不要な変数をモデルから除去する特徴選択を実行することは困難である．

本報告では，このような問題点を解決するために，適応正則化学習の最適化問題内に新たな正則化項として lasso 項を加えたモデリング手法を提案する．また，提案手法の有効性を数値実験を通して検証する．

2. スパース適応正則化オンライン学習

ラベル変数 $y \in \{-1, 1\}$ と d 次元説明変数 $\boldsymbol{x} = (x_1, \cdots, x_d)^T$ に関して，逐次的にデータが得られることを想定し，識別・判別問題を考える．このとき，t $(t = 1, 2, \cdots)$ 番目に得られるデータを $(y_t, \boldsymbol{x}_t)$ とする．適応正則化オンライン学習では，重みベクトル $\boldsymbol{w} \in \mathbb{R}^d$ が，平均ベクトル $\boldsymbol{\mu} \in \mathbb{R}^d$，分散共分散行列 $\Sigma \in \mathbb{R}^{d\times d}$ の正規分布 $N(\boldsymbol{\mu}, \Sigma)$ に従っていると仮定し，パラメータ $\boldsymbol{\mu}, \Sigma$ を更新する．また，このときの分散共分散行列は各特徴量の信頼度を表していることに注意されたい．

いま，パラメータ $\boldsymbol{\mu}, \Sigma$ の $t-1$ 回目の推定値を μ_{t-1}, Σ_{t-1} とする．このとき，次の最適化問題を考える．

$$(\boldsymbol{\mu}_t, \Sigma_t) = \mathrm{argmin}_{\boldsymbol{\mu},\Sigma}\Bigg\{ D_{KL}(N(\boldsymbol{\mu}, \Sigma)||N(\boldsymbol{\mu}_{t-1}, \Sigma_{t-1})) + \lambda_1\{\max(0, 1 - y_t\boldsymbol{\mu}^T\boldsymbol{x}_t)\}^2 + \lambda_2\boldsymbol{x}_t^T\Sigma\boldsymbol{x}_t + \lambda_3||\boldsymbol{\mu}||_1\Bigg\}. \tag{1}$$

ここで，$\lambda_1, \lambda_2, \lambda_3$ はすべて正の値をとる正則化パラメータである．また，$D_{KL}(p||q)$ は確率分布 p, q 間の Kullback-Leibler divergence，$\{\max(0, 1 - y_t\boldsymbol{\mu}^T\boldsymbol{x}_t)\}^2$ は 2 乗ヒンジ損失と呼ばれ，モデルと t 番目のデータの当てはまりの良さを表し，$\boldsymbol{x}_t^T\Sigma\boldsymbol{x}_t$ はモデルの分散を小さくする項である．つまり，$\boldsymbol{x}_t^T\Sigma\boldsymbol{x}_t$ は各特徴量に対する信頼度を大きくする役割を担っている．右辺の $||\boldsymbol{\mu}||_1$ で表される lasso 項を除いた部分が，Crammer *et al.* (2009) により提案された適応正則化学習の最適化問題であり，lasso 項を加えることによってパラメータ $\boldsymbol{\mu}$ のいくつかを 0 にすることができ特徴選択が可能となる．また，(1) 式の推定値を得るために，coordinate descent アルゴリズム (Friedman *et al.*, 2010) に基づく推定アルゴリズムを開発する．推定アルゴリズムの詳細ならびに数値実験については当日紹介する．

参考文献

Crammer, K., Kulesza, A. and Dredze, M. (2009) Adaptive regularization of weight vectors. *Advances in Neural Information Processing Systems*, **23**, 414–422.

Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**, 1–22.