Model Selection and Active Learning for High-Dimensional Data Analysis

梅津, 佑太

https://doi.org/10.15017/1654664

出版情報：九州大学, 2015, 博士（機能数理学）, 課程博士
バージョン：

（様式３）

氏　　名　：梅津　佑太

論 文 名　：Model Selection and Active Learning for High-Dimensional Data Analysis
（高次元データ解析のためのモデル選択と能動学習）
区　　分　：甲

論　文　内　容　の　要　旨

1990 年代以降，高次元データ解析の重要性はますます高まっている．高次元データは標本数に対して変数の次元が比較的大きいという特徴をもっており，例えば画像データや次世代シーケンサによるゲノム情報データなどが典型的なものである．高次元データの抱える問題として，古典的な統計手法の適用が困難あるいは適用できたとしても推定精度が不安定になるということが挙げられる．このような問題を解決するために，近年では数理統計や機械学習，量子物理など様々な分野で研究が行われている．

本論文では，高次元データに対しても適用可能な統計解析手法を用いた，モデル選択および能動学習について得られた 4 つの結果を報告する．

第 2 章では，工業製品などの製造過程における不良率解析を目的とした，判別境界を効率よく推定するための能動学習について述べる．不良率の解析のためには，不良品の分布をできるだけ精度よく推定する必要があり，そのためには多くの標本が求められる．したがって，実験を繰り返すことで十分な標本数になるまでサンプリングしなければならない．しかしながら，実際には実験にかかるコストや時間などにより，実験回数に制約があることがある．さらに，変数の次元が大きな場合，完全ランダムサンプリングでは不良品を見つけることが困難であるため，能動学習によって積極的に推定精度を向上させるようなサンプリングを行うことを考える．不良率解析におけるデータの持つ構造として，不良品の分布はいくつかのクラスターに分かれており，非線形性を伴うと考えられている．このようなデータに対しては，active learning SVM (Tong & Koller, 2002)のような最新の手法であっても上手くいかないことがある． そこで，これらの問題を解決する能動学習法を提案し，数値実験と実データ解析によってその有用性を確認する．また，インポータンスサンプリングを用いて不良率を行う．

アンケート調査によって得られるデータに対してのモデリングと Cp 型の情報量規準について第 3 章述べる．Amazon や Netflix などのように大量の商品に対するアンケート調査を行う場合，数多くの商品すべてについて回答が要求されることは稀であり，実際にはそのうちのいくつかについての評価値が得られる．このとき，評価されなかった対象は欠損値となる．欠損を補完するために，機械学習では低ランク行列補完や協調フィルタリングなどの手法が用いられることが多い(Candès & Recht, 2009; Koltchinskii et. al., 20101)．しかしながら，このような手法では予測量の分散を評価することが難しいという問題がある．そこで，欠損の有無を表すダミー行列を用いて，GMANOVA モデル (Potthoff & Roy, 1964)の 3 次元配列への拡張として解釈できるモデルを考える．そして，推定量の安定性を保証するため Ridge 型の推定量に対して基準化した平均 2 乗誤差の不偏推定量として Cp 基準を提案する．

第 4 章では,一般化線形モデルに対するスパース正則化法を考える.高次元データを解析する際,実際に現象を説明する有意な変数は少数であるということがある．このような場合，ある程度標本数が少なくても正則化法によって推定精度を安定させることができることもある.その中でも特に,Lasso (Tibshirani, 1996)などのスパース正則化法は，モデルに含まれるパラメータの推定と有意な変数の選択を同時に実行できる手法として広く知られている．スパース正則化法は基本的に，推定関数に原点で微分不可能な罰則を付加する．罰則はパラメータの振る舞いを制限するものであると解釈できる．しかしながら，調整パラメータの値により推定結果が大きく異なってしまうため，その選択は一般に重要な問題として知られている．通常，調整パラメータの選択には cross validation (Stone, 1974)などの計算機的手法が用いられるが，一般には計算負荷が高いことが知られている．Ninomiya & Kawano (2014)では，Lasso のための AIC を提案しているが推定量の一致性が保証されないという問題がある．そこで，Ninomiya & Kawano (2014)の問題点を解決するとともに，Bridge (Frank & Friedman, 1993)や SCAD (Fan & LI, 2001), MCP (Zhang, 2010)などの非凸罰則を用いたスパース正則化法に対して，調整パラメータを AIC 最小化により選択することを提案する．

SCAD や MCP に対する AIC はバイアス補正項に評価の難しい期待値が現れるため，バイアス補正項の一致推定量の導出が困難となる．また，Bridge に対しては漸近分布にバイアスが生じる．そこで，第 5 章では Fan & Li による oracle property に着目することで，バイアス補正項の一致推定量を構築するための条件について考察する．