カーネル回帰による最適判別曲線の推定

杉崎朱菜　(指導教員：吉田裕亮)

1　はじめに

統計学上のデータ解析手法のひとつに, 判別分析がある. いくつかのグループに分かれているデータを元に, それらが「どういう基準で分けられているのか」という関係を解析することで, 分類されていないサンプルがどのグループに属するかを予測する手法である.

判別器の代表格として, カーネル法を用いたSVM(サポートベクターマシン) が挙げられる. SVM とは, 2 クラスのパターン判別器を構成するひとつの手法である. しかし, この手法は, 学習データが増えると, 計算量も膨大になってしまうというデメリットがある.

本研究では, カーネル回帰を用いて, SVM よりも比較的簡易的に判別曲線を推定する手法を提案する.

2　カーネル回帰

2.1　カーネル法

カーネル法とは, カーネル関数を利用し, 観測データを高次元のベクトル空間に写像にし, 変換後のデータに線形の手法を用いることで, 非線形の関係を実現することができるものである.

2.2　カーネル関数

カーネル法にとって最も重要なのはカーネル関数と呼ばれる, 内積演算に相当する関数である. 本研究では, カーネル法でよく使用されるガウスカーネル

$$k(x, x') = \exp(-\beta \|x - x'\|^2)$$

を用いる. これは $x = x'$ のとき最大値 1 をとり, 直感的には x と x' の近さを表す量になっている.

ここで, β はあらかじめ適当に決めておくパラメータ (本研究ではカーネルパラメータと呼ぶ) である.

2.3　正則化

一般にパラメータの次元が高くなると, 関数の表現能力が指数関数的に増大するため汎化能力が落ちる. 対してカーネル法では, 高次元に保ったまま関数の表現を抑える, 正則化という方法を用いる.

正則化は, サンプルに対する誤差のほかに負荷項を付け加えた

$$R_{k,\lambda}(\alpha) = (y - K\alpha)^T (y - K\alpha) + \lambda \alpha^T K \alpha, \quad \lambda > 0$$

を最小化することによって, カーネル関数の表現能力を落とすという方法である. 正則化の際に加えた $\lambda\alpha^T K\alpha$ を正則化項と呼び, ここでは, その強さを調節している λ を正則化パラメータと呼ぶ.

2.4　カーネル回帰

線形回帰の内積をカーネルにおきかえ, 非線形化することにより, カーネル回帰は与えられる.

3　提案手法

2 群データのそれぞれのラベルを, カーネル回帰を用いて 1,-1 に分け, 回帰曲面でつなぐ. 1, -1 の中点である 0 での等高線を判別曲線とする. カーネル関数のカーネルパラメータ (β) は非線形性の強度, 正則化パラメータ (λ) はカーネル曲面の平滑度に対応する. 各パラメータを調節し, データを 2 群に判別するための最適な判別曲線を推定するのが本研究の手法である.

4　実験

4.1　画像の重ね合わせによる判別

参考文献 [2] に記載があったデータを基に, SVM による判別曲線と比較し, パラメータを調節して形状を近づけていく. この手法では $\beta = 0.025$, $\lambda = 2.0$ が最も近いと判断した.

4.2　マージン内のデータ混入率による判別

SVM では, マージンを分類の良さの基準としている. マージンとは, 判別面と, それぞれのクラスのサンプル集合との最小距離である. カーネル回帰でも同様にマージンをとり, その中に存在するデータの個数を比較したい. しかし, 今回使用したデータにおいて, 等高線によるマージン設定は, 2 次元データのマージン設定には適さない.

本研究では, 図 1 のように, グラフを 4 方向にずらし重ね合わせ, その内部をマージンとする.

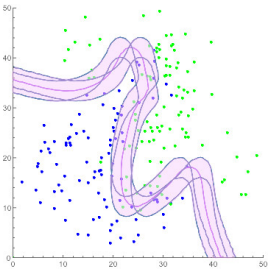


図 1: グラフを 4 方向にずらしたもの

各パラメータを調節した結果, それぞれのマージン内のデータの個数は図 2 のグラフのようになった.

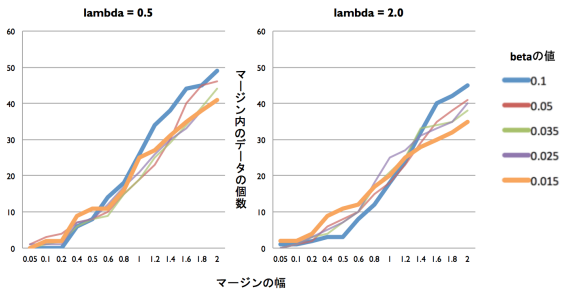


図 2: 実験 2 の結果

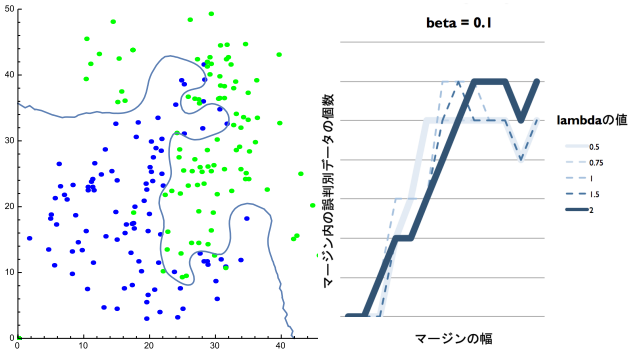
λ の値によらず, β の値とマージン内のデータの個数は比例していた.

また, 実験 4.1 で最適と判断した $\beta = 0.025$ を固定し, λ を増減させると, λ の増加による個数の変化は確認されなかった. データの個数の増加率で比較すると, 比較に用いた 5 つのパラメータの中で, $\lambda = 0.75$ の時が最も増加率が低かった. マージンは大きいほど良いとされるため, 増加率が低いほど良い. そのため, この値を本実験の最適解とした.

4.3 マージン内の誤判別率判別

本実験では, マージン内のデータの誤判別率を比較する. マージン内のデータは無視されるものと考えるため, マージンの中により誤判別データが集まるほうがよい. そのため, 本実験では誤判別率が高い方が最適であると判断する.

β が大きすぎる場合は非線形性が低いため, 図 3[1] のように平滑でない形になる. その場合のマージン内のデータの誤判別率は図 3[2] のグラフのようになった.

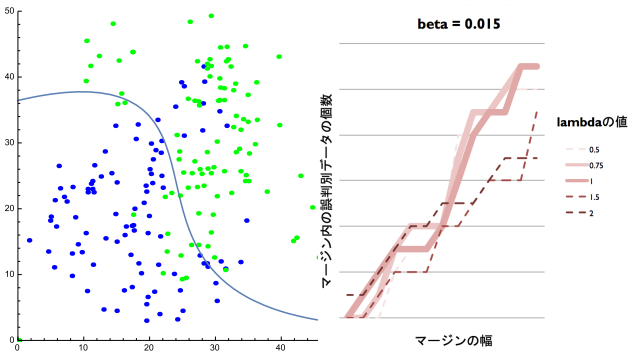


[1] 判別曲線　　[2] 誤判別のデータ個数

図 3: β が大きすぎる場合

この場合は, λ が比較した値の中で大きい, または小さい場合の誤判別率が高い傾向が見られる.

一方, β が小さい場合は非線形性が低いため, 図 4[1] のように平滑すぎる曲線になってしまう. その場合のマージン内のデータの誤判別率は図 4[2] のグラフのようになった.



[1] 判別曲線　　[2] 誤判別のデータ個数

図 4: β が小さすぎる場合

この場合, λ が比較した値の中で, 中間の場合の誤判別率が高い傾向が見られる.

実験 4.2 と同様, 実験 4.1 で最適と判断した $\beta = 0.025$ を固定し, λ を増減させると, 比較に用いた 5 つのパラメータの中で, $\lambda = 0.75$ の時が最も誤判別率が高かったため, この値を本実験の最適解とした.

実験 4.2 と同様の結果になったため, 本研究で用いたデータ上における判別曲線のパラメータは, $\beta = 0.025$, $\lambda = 0.75$ を最適解とした.

5 実験結果のまとめ

実験 1 で SVM の判別曲線を基にしてパラメータを調整した結果, $\beta = 0.025$, $\lambda = 2.0$ のときの曲線が近くなった. 対し, 実験 2 と 3 でカーネル回帰における最適なパラメータを推定した結果は、$\lambda = 0.75$ と, SVM と比べて小さい値となった. それぞれのパラメータを用いた判別曲線を比べた結果が図 5 のようになる.

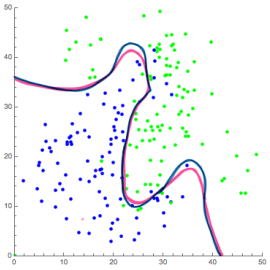


図 5: それぞれの判別曲線の重ね合わせ

赤色が SVM を基にしてパラメータを調整した曲線、紺がカーネル回帰によって最適と推定されたパラメータを用いた曲線である.

大体重なっているが, SVM を基にした曲線の方が、孤立したデータを無視している傾向がある. 一方でカーネル回帰で最適と推定された判別曲線は, 孤立したデータも取り込もうとする傾向があることが確認できた.

6 今後の課題

本研究では, カーネルパラメータの値を固定して数値実験を行ったが, カーネルパラメータの値自体を更に吟味することを今後の課題としたい.

また、本研究ではカーネル回帰を用いたが, ロジスティック回帰などを用いると更に簡易的に判別曲線が作ることができる可能性もあるので, 様々な手法での検討も行いたい.

参考文献

[1] 赤穂昭太郎 (2008) 『カーネル多変量解析 - 非線形データ解析の新しい展開 -』: 岩波書店

[2] Gareth James, Daniela Witten, Trevor Hastie Robert Tibshirani (2013) 『An Introduction to Statistical Learning』: Springer