カーネル回帰分析のモンテカルロ拡張による判別分析

川戸翔子　　(指導教員：吉田裕亮)

1　はじめに

判別分析とは, いくつかのグループに分かれているデータを基に, 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ようにグループに属するかを予測する手法である.

本研究では, カーネル回帰を用いた判別分析を, 元データからランダムに抽出されたデータにより行い, その弱い学習を重ね合わせることによって判別境界を求める手法を考察する.

2　カーネル回帰

2．1　カーネル法

カーネル法とは, データ x, x' が与えられたとき, それらの間の関係を $k(x, x')$ という実数値関数であるカーネル関数によって要約し, 全てを数値に置き換えて処理する方法である.

2．2　カーネル関数

カーネル関数は, 特徴量で見たときの x と x' の類似度 (直感的には x と x' の近さ) を表していると考えることもでき, 2 つの要素 x, x' に対し, それぞれの特徴ベクトルどうしの内積として定義される. すなわち, $\phi(x), \phi(x')$ を高次元空間の特徴ベクトルとして

$$k(x, x') = \phi(x)^T \phi(x')$$

と表される. 本研究では, カーネル関数としては、Gauss カーネル

$$k(x, x') = \exp(-\beta ||x - x'||^2), \quad \beta > 0$$

を用いた. β は非線形性を調整するパラメータの一種と考えられる.

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. 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, λ はその強さを調整しているパラメータである.

3　提案手法

2 群データのそれぞれのラベル 1 と -1 を, カーネル回帰を用いて, 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. これを, モンテカルロ的に拡張するのが本研究の手法である. 具体的には次の手順を行う.

1. データをランダムに 10%程度抽出する.
2. それらのデータでカーネル回帰により回帰曲面を求める.
3. 1,2 を複数回繰り返し, 回帰曲面の平均をとる.
4. 平均曲面の 0 の等高線を判別境界とする.

4　実験例

4．1　実験 1

図 1 のような, 判別境界を $2\sin(x) + 2\sin(3x)$ とし, 正規乱数 $N(0, 1)$ で境界の縦軸方向にノイズを加えた 2 つの群からなる各 500 個のデータを用意する.

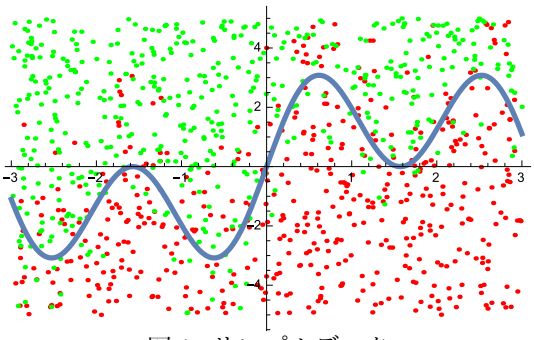

図 1: サンプルデータ

このサンプルデータの判別境界を提案手法の 1, 2 を 5 回, 10 回, 15 回と繰り返して求めると, それぞれ以下の図のような結果が得られた.

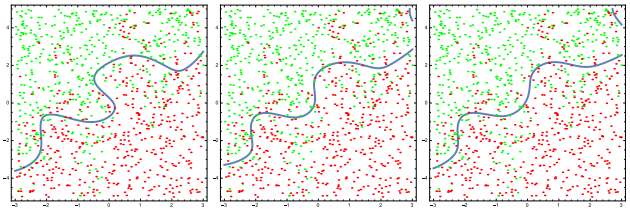

図 1 - A: 5 回　　図 1 - B: 10 回　　図 1 - C: 15 回

全データを用いて判別分析を行ったときと, データを抽出して判別分析を行ったときの結果を比較すると以下のようになる.

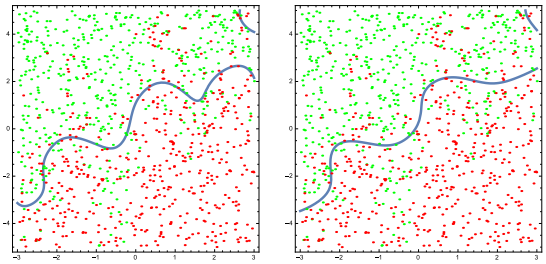

図 1 - D: 全データ　　図 1 - E: 抽出（15 回）

図 1 - D と図 1 - E を比較すると, かなり近い判別境界が得られていることがわかる.

また, 抽出したデータでの判別分析を 10 回繰り返して誤判別率を比較すると, 表 1 のようになる.

全データ	抽出 5 回	抽出10回	抽出15回
0.186	0.212	0.195	0.201
	0.212	0.202	0.200
	0.214	0.206	0.196
	0.206	0.202	0.209
	0.220	0.204	0.204
	0.206	0.194	0.203
	0.210	0.189	0.195
	0.204	0.206	0.211
	0.202	0.199	0.196
	0.193	0.187	0.186

表 1 :誤判別率

誤判別率の平均は, 5 回で 0.208, 10 回で 0.198, 15 回で 0.200 になり, 5 回, 10 回, 15 回と重ね合わせる回数を増やしても, あまり変わらない結果となった. また, 全データを用いて判別分析したときの誤判別率が, 0.186 だったので, 重ね合わせ 10 回で全データでの判別分析に近い結果が得られている.

4．2　実験２

図 2 のような, 判別境界を $x^2+y^2=1$ の円とし, 正規乱数 $N(0,1)$ で境界の半径方向にノイズを加えた 2 つの群からなる各 500 個のデータを用意する.

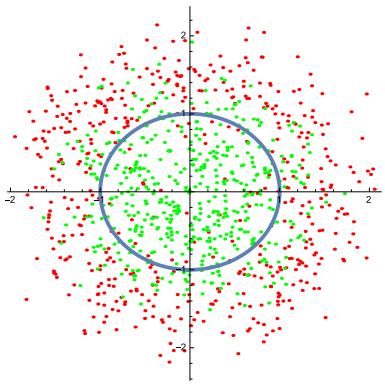

図 2 :サンプルデータ

このサンプルデータの判別境界を提案手法の 1, 2 を 5 回, 10 回, 15 回と繰り返して求めると, それぞれ以下の図のような結果が得られた.

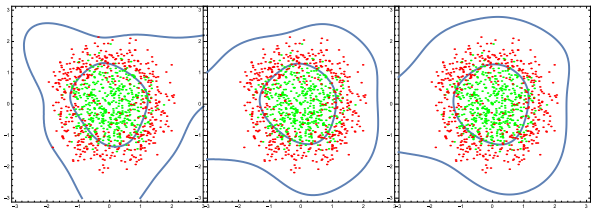

図 2 - A: 5 回　図 2 - B: 10 回　図 2 - C: 15 回

内側の円はうまく描けているが, 外側の円はどれも描けていないように見える. そこで, この図の x, y の範囲をそれぞれ ± 10 に広げてみると, 図 2 - D , E , F のようになる.

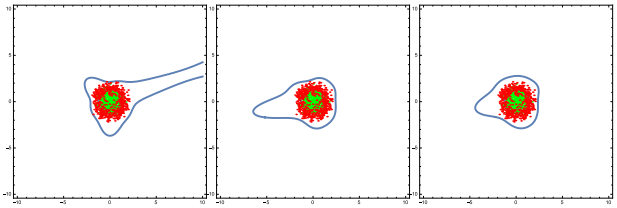

図 2 - D: 5 回　図 2 - E: 10 回　図 2 - F: 15 回

上図を見ると, 外側の円はどれもうまく描けてはいないが, 重ね合わせの回数を増やすと徐々に判別境界として良くなっていっていることがわかる.

全データを用いて判別分析を行ったときと, データを抽出して判別分析を行ったときの結果を比較すると以下のようになる.

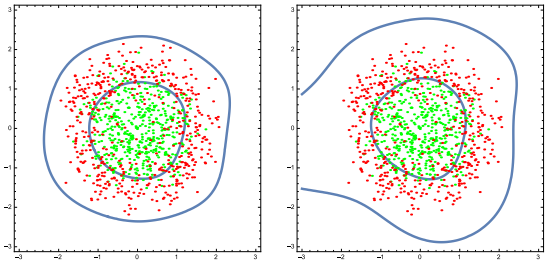

図 2 - G: 全データ　図 2 - H: 抽出（15 回）

また, 抽出したデータでの判別分析を 10 回繰り返して誤判別率を比較すると, 以下のようになる.

全データ	抽出 5 回	抽出10回	抽出15回
0.186	0.184	0.187	0.190
	0.183	0.187	0.182
	0.193	0.195	0.184
	0.190	0.187	0.188
	0.197	0.186	0.183
	0.190	0.190	0.182
	0.199	0.185	0.188
	0.195	0.185	0.190
	0.205	0.189	0.188
	0.197	0.197	0.193

表 2 :誤判別率

誤判別率の平均は, 5 回で 0.193, 10 回で 0.189, 15 回で 0.187 になり, 5 回, 10 回, 15 回と重ね合わせる回数を増やすと, 少しずつ誤判別が減っていることがわかる. また, 全データを用いて判別分析したときの誤判別率が, 0.186 だったので, 重ね合わせ 15 回で全データでの判別分析とほぼ同じ結果が得られている.

5　まとめ

円の外側の判別境界以外は, 判別分析をデータを抽出して行っても, 全データを用いて判別分析を行ったときとかなり近い判別境界を描けることがわかった. 誤判別率についても, 全データを用いて判別分析したときとかなり近い結果が得られた.

参考文献

[1] 赤穂昭太郎, 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 , 東京 , 2008.