分類の観点からの判別分析とサポートベクターマシンの比較

2017SS028 柄澤俊也
指導教員：阿部俊弘

1 導入

線形判別分析 (LDA) は, R. A. Fisher によって基本的な考え方が示された手法である [3]. また, V. N Vapnik, A. Y. Chervonenkis が発表したサポートベクターマシン (SVM)[1] は, 現在最も広く利用されているパターン認識学習アルゴリズムの一つであり, マージン最大化に基づき, 主に 2 クラス問題に用いられる [2]. ラベル付きの分類には統計的手法の判別分析と機械学習の SVM の 2 種類があり, 本研究では分類の観点からこれらの比較を行う.

2 識別関数

2 クラスに分ける線形識別関数を考える. 係数ベクトルはバイアスを含めてパラメータ $\boldsymbol{\omega} = (\omega_0, \omega_1, \ldots, \omega_d)^T$, i 番目の学習用入力ベクトルは, $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{id})^T$ と定義する. 線形識別関数の識別境界は, 入力データの次元を d とすれば $d-1$ 次元の超平面となる. 識別関数は,

$$y = f(\boldsymbol{x}; \boldsymbol{\omega}) = \omega_0 + \omega_1 x_1 + \cdots + \omega_d x_d = \boldsymbol{\omega}^T \boldsymbol{x} \quad (1)$$

である.

3 フィッシャーの線形判別分析

2 クラス (C_1, C_2) 問題について考える. 各クラスの学習データの大きさを N_1, N_2 とする. $k = 1, 2$ に対して変換前の平均ベクトルは, $\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{i \in C_k} \boldsymbol{x}_i$ となる. 式 (1) より変換後のクラスの平均を $m_k = \boldsymbol{\omega}^T \boldsymbol{\mu}_k$ と表現できる. クラス間の差の 2 乗は,

$$(m_1 - m_2)^2 = (\boldsymbol{\omega}^T (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))^2 \quad (2)$$

となり, これをクラス間変動という. 値が大きいほどクラスの分離がうまくできている. また, 変換後の各クラスの散らばりは,

$$S_k^2 = \sum_{i \in C_k} (\boldsymbol{\omega}^T \boldsymbol{x}_i - m_k)^2 \quad (3)$$

で, これをクラス内変動という. 値が小さいほど各クラスのデータが密集できている. 式 (2) を変形していくと,

$$(m_1 - m_2)^2 = \boldsymbol{\omega}^T (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\omega} = \boldsymbol{\omega}^T \boldsymbol{S}_B \boldsymbol{\omega}$$

ここで $\boldsymbol{S}_B = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T$ は変換前のクラス間変動行列という. 式 (3) を変形していくと,

$$S_k^2 = \boldsymbol{\omega}^T \left(\sum_{i \in C_k} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \right) \boldsymbol{\omega}$$

変換後のクラス内変動は, 変換前のクラス内変動行列 $\boldsymbol{S}_W = \sum_{i \in C_1} (\boldsymbol{x}_i - \boldsymbol{\mu}_1)^2 + \sum_{i \in C_2} (\boldsymbol{x}_i - \boldsymbol{\mu}_2)^2$ より, $S_1^2 + S_2^2 = \boldsymbol{\omega}^T \boldsymbol{S}_W \boldsymbol{\omega}$ となる. クラス間変動とクラス内変動の比

$$J(\boldsymbol{\omega}) = \frac{(m_1 - m_2)^2}{S_1^2 + S_2^2} = \frac{\boldsymbol{\omega}^T \boldsymbol{S}_B \boldsymbol{\omega}}{\boldsymbol{\omega}^T \boldsymbol{S}_W \boldsymbol{\omega}}$$

を最大化する解は, 一般化固有値問題をラグランジュの未定乗数法で解く. λ を未定乗数と見立てて, $\boldsymbol{\omega}$ で微分すると, $\boldsymbol{S}_B \boldsymbol{\omega} = \lambda \boldsymbol{S}_W \boldsymbol{\omega}$ となる. $\boldsymbol{S}_W$ が正則であれば, $\boldsymbol{S}_W^{-1} \boldsymbol{S}_B \boldsymbol{\omega} = \lambda \boldsymbol{\omega}$ と書ける. $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\omega}$ はスカラーだから, $\boldsymbol{S}_B \boldsymbol{\omega} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\omega} \propto (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ となり, $\boldsymbol{S}_B \boldsymbol{\omega}$ が $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ に比例することを用いて,

$$\boldsymbol{\omega} \propto \boldsymbol{S}_W^{-1} \boldsymbol{S}_B \boldsymbol{\omega} \propto \boldsymbol{S}_W^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

が最適な $\boldsymbol{\omega}$ となる. また, $\hat{\boldsymbol{\omega}} = \boldsymbol{S}_W^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ とすると, 判別点を y 軸上で C_1, C_2 の中点として, 線形識別関数は,

$$f(\boldsymbol{x}) = \frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{S}_W^{-1} \boldsymbol{x} - \frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{S}_W^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$$

となる.

4 サポートベクターマシン

4.1 ハードマージン

2 クラス (C_1, C_2) 問題を考える. クラスラベル付き学習データの集合を $C_L = \{(t_i, \boldsymbol{x}_i)\} (i = 1, \ldots, N)$ とする. $t_i = \{-1, +1\}$ は教師データとする. 式 (1) とパラメータ c を用いて, 分離超平面の式は $\boldsymbol{\omega}^T \boldsymbol{x}_i + c = 0$ と表現できる. また, ラベル変数 t_i を用いると, $t_i(\boldsymbol{\omega}^T \boldsymbol{x}_i + c) \geq 1$ とまとまる. M を最大化する条件式は,

$$\max_{\boldsymbol{\omega}, c} \ M, \quad \frac{t_i(\boldsymbol{\omega}^T \boldsymbol{x}_i + c)}{\|\boldsymbol{\omega}\|} \geq M \quad (4)$$

となる. 式 (4) を変形すると最適化問題は,

$$\max_{\tilde{\boldsymbol{\omega}}, \tilde{c}} \ \frac{1}{\|\tilde{\boldsymbol{\omega}}\|}, \quad t_i(\tilde{\boldsymbol{\omega}}^T \boldsymbol{x}_i + \tilde{c}) \geq 1$$

となる. 最適化問題の主問題としてまとめると,

$$\begin{aligned} &\arg\min \ \frac{1}{2} \|\boldsymbol{\omega}\|^2 \\ &\text{subject to } t_i(\boldsymbol{\omega}^T \boldsymbol{x}_i + c) \geq 1 \end{aligned}$$

この問題はラグランジュ未定乗数法で解く. λ を未定乗数と見立てて, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)^T$ $(\lambda_i \geq 0)$ とし,

$$\frac{1}{2} \|\boldsymbol{\omega}\|^2 - \sum_{i=1}^{n} \lambda_i (t_i(\boldsymbol{\omega}^T \boldsymbol{x}_i + c) - 1) \quad (5)$$

となる. $\boldsymbol{\omega}$ と c についてそれぞれ微分し, そこから得られた条件を式 (5) に代入すると次の双対問題が得られる.

$$\arg\max \sum_{i=1}^{n}\lambda_i - \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j t_i t_j \boldsymbol{x}_i^T\boldsymbol{x}_j$$
$$\text{subject to } \sum_{i=1}^{n}\lambda_i t_i = 0$$

4.2 ソフトマージン

線形分離不可能なデータを前提とし, 誤判別を許容する方法である. 分離超平面の反対側に入ることを許し, 入り込んだ距離をスラック変数 ξ_i で表し, $t_i(\boldsymbol{\omega}^T\boldsymbol{x}_i+c)-1+\xi_i \geq 0$ と表現する. ソフトマージンの主問題を定義する.

$$\arg\min \frac{1}{2}\|\boldsymbol{\omega}\|^2 + C\sum_{i=1}^{n}\xi_i$$
$$\text{subject to } t_i(\boldsymbol{\omega}^T\boldsymbol{x}_i + c) - 1 + \xi_i \geq 0,\ \xi_i \geq 0$$

ハイパーパラメータ C は, 誤識別数に対するペナルティの強さを表し, 大きければ大きいほど $\|\boldsymbol{\omega}\|$ の最小化より誤識別数を小さくすることを優先する解が得られる. 適切な C は交差確認法などで実験的に選ぶ必要がある. 未定乗数を $\tau_i \geq 0$ とし, ハードマージン同様に主問題に関するラグランジュ関数を求める. 主問題に対する双対問題は, ハードマージンと同様に,

$$\arg\max \sum_{i=1}^{n}\lambda_i - \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j t_i t_j \boldsymbol{x}_i^T\boldsymbol{x}_j$$
$$\text{subject to } \sum_{i=1}^{n}\lambda_i t_i = 0,\ 0 \leq \lambda_i \leq C$$

となる. 解ベクトルは $\hat{\boldsymbol{\omega}} = \sum_{i=1}^{n}\lambda_i t_i \boldsymbol{x}_i$ で得られる.

5 判別分析とサポートベクターマシンの比較

LDA と SVM における境界近くのデータに対する対応を検証した. データは 2 変量 2 群のデータで, クラス 1 の平均 $(\mu_1,\mu_2) = (6,6)$, 分散 $(\sigma_1^2,\sigma_2^2) = (1,1)$, クラス 2 の平均 $(\mu_1,\mu_2) = (0,0)$, 分散 $(\sigma_1^2,\sigma_2^2) = (1,1)$, 相関係数 $\rho = -0.5$ とする 2 変量正規分布に従う乱数を用いた. これらのデータセットを 2 つ用意し, 片方には境界付近の点として $(4,4)$ を打ち込み, クラス 1 に混ぜたものを用意した. 試行回数 3000 回行い, LDA と SVM の傾きと切片の平均と分散を比較した. 各群 30 個の結果を示す.

表 1 $(4,4)$ なしの場合

	傾き-平均	傾き-分散	切片-平均	切片-分散
LDA	−1.0115	0.0240	6.0373	0.2349
SVM	−1.0393	0.0669	6.1243	0.7427

表 2 $(4,4)$ ありの場合

	傾き-平均	傾き-分散	切片-平均	切片-分散
LDA	−1.0110	0.0254	5.9644	0.2399
SVM	−1.0518	0.1059	5.0573	0.7278

データの大きさを 30 個にすると視覚的に図からも LDA と SVM の精度の差が確認できる. LDA は分散が小さいことにより直線がまとまって見える. $(4,4)$ の有無による LDA と SVM の変化も LDA の方が安定している.

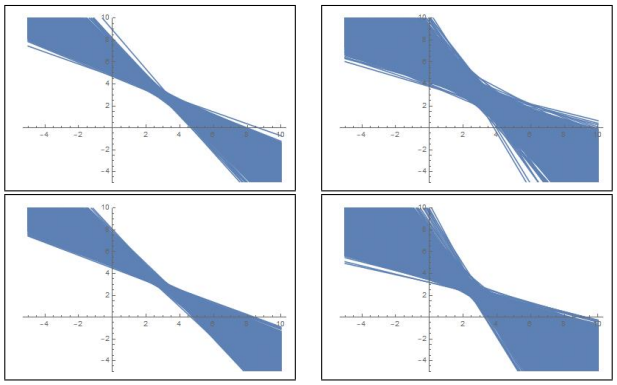

図 1 3000 本の識別境界直線. 左上図:LDA $(4,4)$ なし, 右上図:SVM $(4,4)$ なし, 左下図 LDA $(4,4)$ あり, 右下図 SVM $(4,4)$ あり

境界付近のデータに対して, LDA はデータの大きさが大きくなるほど, 傾きと切片の分散は小さくなり精度が向上する. これは LDA がデータの平均や分散から識別境界線を求めているためである. 一方 SVM はデータの大きさが小さいと, 傾きの分散が大きくなり切片の分散は小さくなった. これは境界付近のデータの影響を受けているためと考えられる. 境界付近のデータがサポートベクトルとして選ばれ, それ以外のデータは識別境界線に反映されないため, 今回の検証では LDA より精度が劣る結果となった.

6 まとめ

分類の観点から統計学での代表的な手法である判別分析と機械学習での代表的な手法であるサポートベクターマシンの比較を行った. 多群化や線形分離不可能な場合, 境界付近のデータへの対応などそれぞれに一長一短があり, 分類したいデータに合わせて分類法も検討することが重要であると考える. 判別分析は正規分布を基にしていることから, 一般化がしやすい. 一方で, サポートベクターマシンは状況に応じて拡張を考えなければならず, 少し複雑な拡張を考えるととたんに困難が生じる. 今後の課題として, 多群化や非線形化の数値的検証をしていきたい.

参考文献

[1] Corinna Cortes and Vladimir Vapnik: "Support-vector networks." Machine learning, Vol. 20, No. 3, pp. 273-297, 1995.

[2] 平井有三:『はじめてのバターン認識』. 森北出版, 東京 2012.

[3] 小西貞則:『多変量解析入門』. 岩波書店, 2010.