サポートベクターマシンとその応用

T02KF020　児島佑樹

1.初めに

サポートベクターマシン(以下 svm)はパターン認識性能の優れた学習モデルのひとつである.今回は svm を理解したうえで,svm の特徴を利用した実験を行い svm の性能を見ることを目的としている.

2.線形 svm

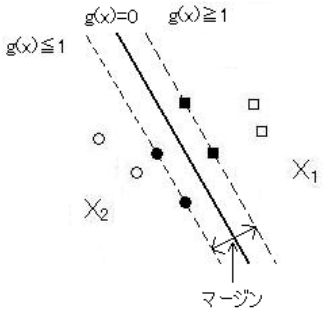
線形識別関数 f(x)を

$$f(x) = sign(g(x)) = sign(w^t x + b) \cdots ①$$

とおく.ここで n 個の学習データ $x_i (i = 1...n)$ が与えられていて,二つのクラス X1,X2 に分離することを考えるので満たすべき条件を

$$g(x) = w^t x_i + b \begin{cases} \geq 1 & x_i \in X1 \\ \leq -1 & x_i \in X2 \end{cases} \cdots ②$$

とする.ところで,点 x_i から分離境界 g(x)=0 との距離は $|g(x_i)| = \|w\|$ となる.



ということは②を満たす識別関数において学習データは識別境界線から $1/\|w\|$ の領域には存在しない.この領域をマージン領域と呼び,svm ではこの領域を最大にする識別境界を最良とする.つまり②の条件の下でマージン領域 $2/\|w\|$ を最大にするような w および b を考えればよい.ここで学習データ x_i に関する教師信号を y_i とし,次のように定義する.

$$y_i = \begin{cases} 1 & if \quad x_i \in X_1 \\ -1 & if \quad x_i \in X_2 \end{cases} \cdots ③$$

これにより②の式を書き直すと

$$y_i (w^t x_i + b) - 1 \geq 0 \cdots ④$$

この条件でマージン領域 $2/\|w\|$ を最大化する問題を考える.式の扱い上 $2/\|w\|$ のままだと後々大変なので

$$G(w) = \frac{1}{2}\|w\|^2 \cdots ⑤$$

の最小化問題に置き換える.

この最小化問題をそのまま解くのは困難なので,ラグランジュ未定乗数法を使って問題を置き換える.

x_i に対応するラグランジュ未定乗数 λ_i $(\lambda_i \geq 0, i = 1..N)$ を要素とするベクトル λ を定義すると,ラグランジュ関数 L_p は次のようになる.

$$L_p(w,b,\lambda) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} \lambda_i (y_i (w^t x_i - b) - 1) \cdots ⑥$$

w および b に関する偏微分より

$$w = \sum_{i=1}^{N} \lambda_i y_i x_i , \quad 0 = \sum_{i=1}^{N} \lambda_i y_i \cdots ⑦$$

という関係式が成り立つ.この式を⑥に代入すると

$$L_D(\lambda) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{N} \lambda_i \lambda_j y_i y_j x_i^t x_j \cdots ⑧$$

が得られる.つまり $0 = \sum_{i=1}^{N} \lambda_i y_i$, $\lambda_i \geq 0$ の条件の下で⑧を最大にする問題となる.

解 λ_i が 0 でない学習データ x_i のことを「サポートベクター」と呼び, この x_i はマージン領域を最大とするようなマージン領域の最端に存在する.

3.svm とその応用

svm は未知のデータに対して非常に精度が高い識別ができ,また大量のデータを識別するのに有効である.これらの特徴を生かせる実験として,svm を利用したメールフィルタを作成しテストを行うことにした.

3.1.作成するメールフィルタの概要

メールの読み込み作業を簡単にするために日本語のメールのみを取り扱うことにする.

メールの本文より漢字のみを抜き出し,出てきた回数により svm に適用する.ただし,漢字の

数が膨大になることが予想されるので一文字の漢字は除くことにした.

漢字抜き出しの例：

私は山梨の出身です.他に山梨出身の人は居ますか？

山梨→１,出身→１,山梨出身→１

3.2.作成する svm の概要

全ての迷惑メールについて出てくる単語の回数を調べ,出てくる回数の多いものについてはメール一つに何回ずつ出てくるかを調べる.svm の入力は出てくる回数で学習用メールはランダムに選び,また学習用メールによって出力も変わってくるので 10 回ごとの平均を四捨五入して表すことにした.

4.実験

A,B,C,D,E,F でそれぞれ条件を変えた場合の識別の結果を表 1,2 に表した.通常は通常メールを迷惑は迷惑メールを○は識別したことを×は間違って識別をしたことをそれぞれ表す.

A：svm を利用せず出てくる単語のうち,出てくる頻度が上から 10 番目までの単語が出てきたら迷惑メールと判別する.

B：A の条件で使用する単語のうち「通常メールのうち一回でも出てくる単語については除外する.」という条件を加える.

C：svm を利用する.入力する単語数は 10,学習用のメールは通常メール 10 通,迷惑メール 10 通とする.

表 1:A,B,C の識別結果

	通常○	通常×	迷惑○	迷惑×
A	335	89	230	15
B	424	0	147	88
C	423	1	160	75

D：C の条件の入力の単語数を 20 に変更する.

E：C の条件の学習用メールを通常メール 20 通,迷惑メール 20 通に変更する.

F：C の条件の入力の単語数を 20 に,学習用メールを通常メール 20 通,迷惑メール 20 通に変更する.

表 3:C,D,E,F の識別結果

	通常○	通常×	迷惑○	迷惑×
C	423	1	160	75
D	424	0	184	51
E	424	0	185	50
F	424	0	201	34

5,考察

入力の単語数,学習用のメールを増やすほど,識別度が高くなる.

識別が上手くできなかったメールを見てみるとメールの本文が短い場合が多かった.識別に単語の頻度を使用するので短ければ通常メールと単語自体が少なくなり誤識別してしまうと予想される.

これらの考察を踏まえたうえでの改良点として挙げられるのは

・メールの読み込み方法の改良.

漢字だけでの識別ではなく,ひらがなをまじえた単語や英単語を含む単語での識別ができると良い.

・識別方法の改良.

今回は線形 svm を利用して識別をしたが,非線形 svm を利用しての識別でも可能である.非線形 svm の方が細かく分けることができるので,良い結果を出すと思われる.

6.参考文献

(1)前田英作:“痛快！サポートベクトルマシン～古くて新しいパターン認識手法”、情報処理学会誌 Vol42, No. 7, (2001)

(2)青葉雅人：Support Vector Machine ってなに？

(online), http://www.neuro.sfc.keio.ac.jp/~masato/study/SVM/index.htm

(3)堀内亮介：サポートベクトルマシンを使った 6 パリティ問題の解決(2005)