ExcelによるSVMの実装とPythonによる応用例

**Support Vector Machine Implementation in Excel and application examples in Python**

高橋　湧汰

Yuta Takahashi

法政大学情報科学部コンピューター科学科

*E-mail: yuta.takahashi.5t@stu.hosei.ac.jp*

# **1. SVMとは**

SVM(Support Vector Machine)とは教師あり学習を用いる機械学習モデルの1つである。特徴空間上で学習データを分離する識別長平面(識別関数)を求めることで2クラス分類を行う。

手法としてはマージン(判定する境界線と各データとの距離)の最大化を使用し、サポートベクトル(識別関数と最も近くにあるデータ)と識別関数のマージンを最大化することを方針としている。マージンは、

で表すことが出来、これを最大化すなわちを最小化すればよいことになる。

ここで、識別関数をと置き、データを表す点と教師データを利用すると、

と問題設定ができる。(1.2)式は今後の計算過程を見やすいものにするため(1.1)式を少し改変したものだが、意味は同じである。

さらに、双対問題を導入してラグランジュの未定乗数法を用いると以下のように書き直せる：

(1.4)式は可視性が高いものとは言えないが、データを入力する際には便利である。

## **2.Excelによる実装**

下の表は、男性A,B,Cと女性D,E,Fを対象に製品X,Yの好感度x,yを調べたものである。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No. | 名前 | 好感度 |  | 性別 |
|  |  | x | y |  |
| 1 | A | 0 | 0 | 男 |
| 2 | B | 0 | 1 | 男 |
| 3 | C | 1 | 1 | 男 |
| 4 | D | 1 | 0 | 女 |
| 5 | E | 2 | 0 | 女 |
| 6 | F | 2 | 1 | 女 |

この表とExcelを用いて男女を区別する識別関数を求める。

まず、男性を負例,女性を正例とし、正解ラベルを正例に対して1,負例に対して-1とすると、

と示せる。また、ラグランジュ定数をそれぞれとおき(Excelファイルの方では、教材[7]に合わせて定数をとしている。)、(1.4),(1.5)式に当てはめ最適解を求めるととなる。

故に求める識別関数はとなり、グラフで図示すると以下のようになる：

グラフ, 折れ線グラフ

自動的に生成された説明

## **3. SVMの応用例**

　SVMは応用例としてスパムメールの検出に用いられるテキスト分類, 郵便番号の追跡などに使用される手書き数字認識, 写真を分析し、それが誰であるかを判定する顔検出などが挙げられる。

# **3-1. ハードマージンSVMと**ソフトマージンSVM

　これまでに実装してきた方法は、クラスを完全に線形分離することが出来、マージン最大化を完全に満たす場合でハードマージンSVMと言うものであった。しかし、データが増えてしまうと前述の2つの条件を満たさなくなってしまう。そこで、

・マージンが最大である必要がないようにする

・分離に失敗するデータがあったとしても許容する

と条件を緩めたソフトマージンSVMと言う手法があり、どれだけ誤分類を許容するかのパラメータCと言うものを導入している。またパラメータと言うものがあり、この値が大きいほど複雑な決定境界となる。

# **3-2. Pythonでの手書き数字認識**

ここで追加実験として、Pythonによる手書き数字認識を実装する。数字認識をするデータとして、Scikit-learnのライブラリに含まれているデータセットを用いる。実行環境はそれぞれ

Python3.8

Jupyter Notebook

とし、使用するモジュールはmatplotlib, sklearnとした。

Scikit-learnのサンプルデータとして、以下のようなものが含まれている：

テキスト, クロスワードパズル, 挿絵, 時計 が含まれている画像

自動的に生成された説明

また、今回は読み込んだデータセットの2/3を訓練データ,　1/3をテストデータとし、パラメータ(誤分類は一切許容しない、決定境界は単純)とした。

実行結果は以下の表のようになった：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数字 | 適合率 | 再現率 | F値 | 個数 |
| 0 | 1.00 | 0.98 | 0.99 | 59 |
| 1 | 0.97 | 1.00 | 0.98 | 62 |
| 2 | 1.00 | 0.98 | 0.99 | 60 |
| 3 | 0.96 | 0.85 | 0.91 | 62 |
| 4 | 0.98 | 0.95 | 0.97 | 62 |
| 5 | 0.95 | 0.98 | 0.97 | 59 |
| 6 | 0.98 | 0.98 | 0.98 | 61 |
| 7 | 0.98 | 1.00 | 0.99 | 61 |
| 8 | 0.90 | 0.98 | 0.94 | 55 |
| 9 | 0.95 | 0.97 | 0.96 | 58 |

また、各数字ごとの認識率は以下の表になった：

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 0 | 58 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 62 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 59 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 53 | 0 | 2 | 0 | 1 | 6 | 0 |
| 4 | 0 | 0 | 0 | 0 | 59 | 0 | 0 | 0 | 0 | 3 |
| 5 | 0 | 0 | 0 | 0 | 0 | 58 | 1 | 0 | 0 | 0 |
| 6 | 0 | 1 | 0 | 0 | 0 | 0 | 60 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 61 | 0 | 0 |
| 8 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 54 | 0 |
| 9 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 56 |

全体的に認識率は9割を超えており精度が良いが、テストデータの400番目周辺の誤検出が目立つので、可視化すると以下のようになる。例えば1行3列目では、「0」を「7」,また2行1列目では「2」を「3」と誤認識しているが、これらは人間の目でも間違えかねない数字の汚さが目立っており、致し方がない事がわかる。

テキスト, スコアボード, クロスワードパズル が含まれている画像

自動的に生成された説明

# **4.まとめ**

本レポートでは、配布された教材を使いつつ他の教材や学習サイトにおいてグループ学習を行い、一定の機械学習の知識を学ぶことが出来た。しかし、あくまで初歩的な部分を触っただけである。以上の事から、今後の自分がしたい研究のことを踏まえながら、機械学習の学習・研究を進めていきたいと考えている。また。この場を借りて佐藤裕二教授、TAの田浦さんと土屋さんにお礼を申し上げたい。

# **参　考　文　献**

1. http://www.kana-lab.c.titech.ac.jp/lecture/lec\_2018\_osaka/note\_03-svm.pdf
2. http://ibisforest.org/index.php?F%E5%80%A4
3. https://www.slideshare.net/mknh1122/svm-13623887
4. https://qiita.com/yhyhyhjp/items/ebda34f46369b7d3ac8e
5. https://aiacademy.jp/media/?p=248
6. https://logics-of-blue.com/svm-concept/
7. Excelでわかる機械学習超入門 涌井 良幸 (著) 技術評論社 (講義資料第4章.pdf)
8. 詳細!Python3入門ノート 大重美幸(著) ソーテックス社