

# 平成 28 年度 3 回生後期実験 (エージェント)

## 課題 2 SVM の評価

村田 叡

2016/10/21

### 1 プログラム概要

前回の課題では 1 と -1 の 2 クラスのサンプル点集合を分類する識別器を生成した。この課題ではその交差検定を行った。ガウスクーネル または 多項式クーネルの最良パラメータを推定できるようになった。以下にその詳細を述べる。

### 2 外部仕様

#### 2.1 svm.py

今回のコードは 前回のコードの `svm.py` にて追加で実装した。このコードは、引数としてサンプル点集合のファイル名、クーネル名、及びオプションをとる。サンプル点集合の形式については、実験のページに書かれてあるものに従った。クーネル名は、`gauss`, `polynomial`, `sigmoid`, `linear` のいずれかをとる。デフォルトのクーネル名は `gauss` である。`-cross` オプションを使うことで、交差検定ができる。そのオプションに数字を指定することで、交差検定のデータの分割数を指定できる。この交差検定は、ガウスクーネルまたは多項式クーネルについて適応できる。指定しなかった場合は分割数は 10 である。`-plot` オプションを使うと結果をプロットしたものを `images` フォルダに保存していく。

#### 2.2 実行例と実行結果

```
# 必要なライブラリの導入
$pip3 install -r requirements.txt
# サンプル点集合 sample_circle.dat のガウスクーネルの SVM の交差検定をデータ分割数 5 で行う。
$python3 svm.py sample_circle.dat -m gauss -c 5
>> 2 ** -13.0000 : 67.0%
>> 2 ** -11.6000 : 67.0%
...
...
>> 2 ** -1.5060 : 96.0%
```

```
>> p : 0.34868591658760156 | 96.0%
```

上記のように、徐々に範囲を狭めて適切な値を探索する。最後の出力には適切なパラメータとその精度を表示する。

### 3 内部仕様

以下では `svm.py` で、交差検定のために新たに実装した関数について述べる。

#### 3.1 `cross_validation(x,y,kernel,div)`

この関数では交差検定を行う。引数は、サンプル点ベクトル集合 `x`, クラス集合 `y`, カーネル関数 `kernel`, 分割数 `div` である。

#### 3.2 `search_parameter(x, y, kernel, param_ranges, div, do_plot=False, eps=0.001)`

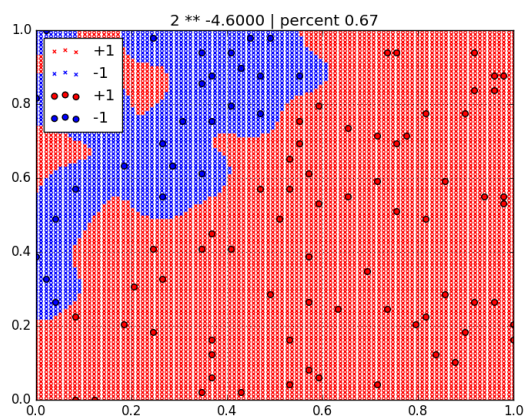
この関数で、交差検定を用いることでパラメータを推定する。引数は、サンプル点ベクトル集合 `x`, クラス集合 `y`, カーネル関数 `kernel`, 探索するパラメータの範囲 `param_ranges`, 分割数 `div`, プロットするか否かの `do_plot`, 探索を更に続けるかの `eps` をとる。探索するパラメータの範囲は、`i, offset` の2つのパラメータからなり、 $2^{i-offset}$  から  $2^{i+offset}$  の範囲を探す。探索範囲をしばめていって精度を上げていく。`eps` を上げるほど、細部を探索しなくなる。

### 4 考察

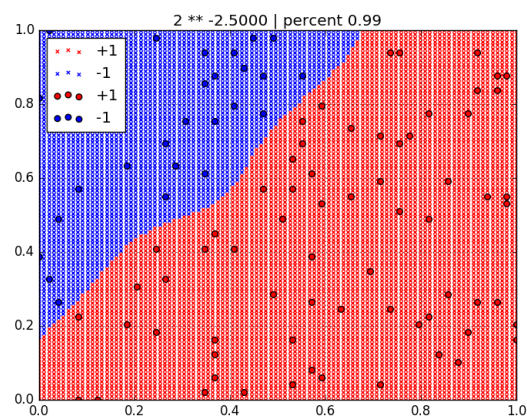
以下では実際に上記のプログラムを用いて、分割数 10 の交差検定を行った予測精度の違いを述べる。データは線形分離可能なデータと線形分離不可能なデータの2つについて、カーネルはガウスカーネルと多項式カーネルの2つについてをそれぞれ述べる。線形不可能なデータについては、やや湾曲した図形の自作のデータを用いた。(画像を二値化して、実験データと同様のフォーマットのサンプルデータに変換するプログラム `image2scatter.py` を書いてデータを作成した。(注: これの実行には `pillow` が必要である))

## 4.1 ガウスカーネル

ガウスカーネルは、パラメータ  $\sigma$  に依存して、 $K(x_k, x_l) = \exp(-||x_k - x_l||^2 / 2\sigma^2)$  と書ける。下図のように、 $\sigma$  の値に応じて、教師の点の周りにのみ分類されやすくなるかどうかが変わる。そういうカーネルなので、線形分離が可能な集合に対しては、本質的に弱点があるように思われる。

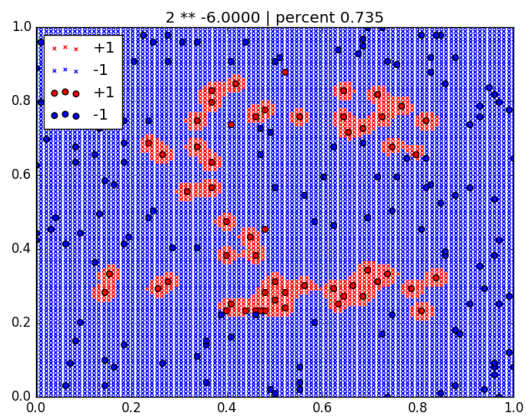


(a) 精度 67% ,  $\sigma = 2^{-4.6}$

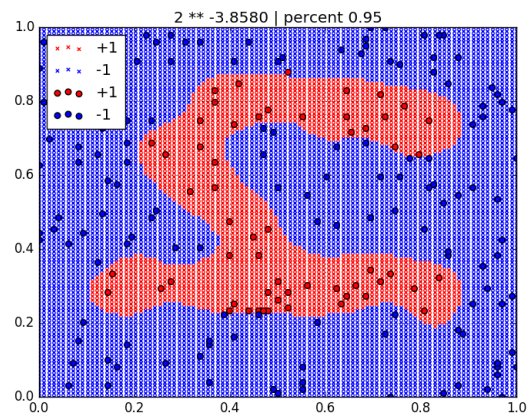


(b) 精度 99% ,  $\sigma = 2^{-2.5}$

図 1: 線形分離可能習合のガウスカーネルの識別器



(a) 精度 73.5% ,  $\sigma = 2^{-6}$

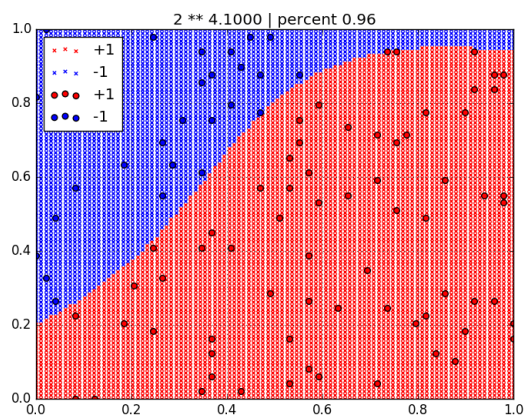


(b) 精度 95% ,  $\sigma = 2^{-3.858}$

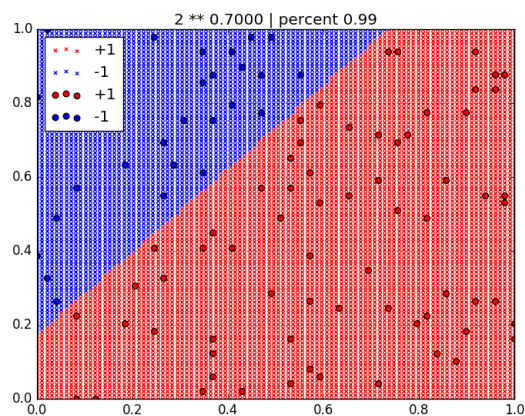
図 2: 線形分離不可能習合のガウスカーネルの識別器

## 4.2 多項式カーネル

多項式カーネルは、パラメータ  $d$  に依存して、 $K(x_k, x_l) = (1 + (x_k, x_l))^d$  と書ける。ガウスカーネルと違って、教師の周りのサンプルをとるというよりは、境界をなめらかに決定する用に

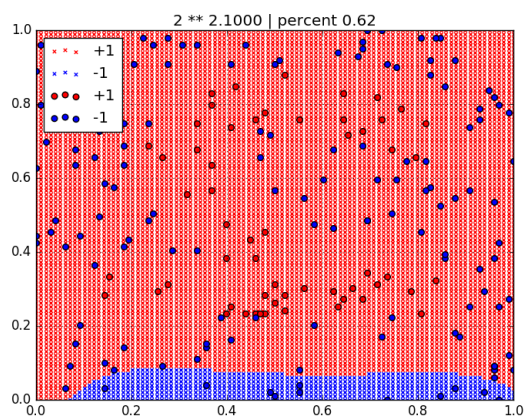


(a) 精度 96% ,  $d = 2^{4.1}$

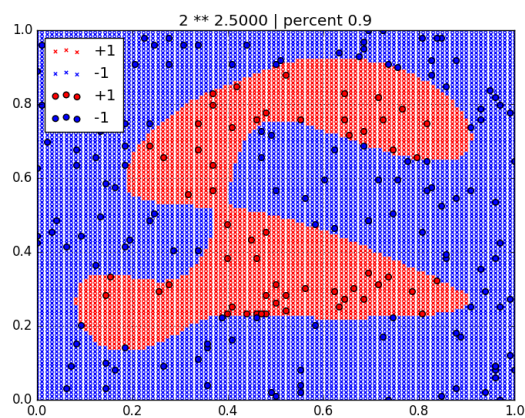


(b) 精度 99% ,  $d = 2^{0.7}$

図 3: 線形分離可能習合の多項式カーネルの識別器



(a) 精度 62% ,  $d = 2^{2.1}$



(b) 精度 90% ,  $d = 2^{2.5}$

図 4: 線形分離不可能習合の多項式カーネルの識別器