Corpus Statistics
Baayen (2008: 160–163)
2009/02/05

Toru Hitomi

### 5.2.3 Support vector machine

- 概説

サポート・ベクター・マシン（support vector machine, SVM）は、比較的新しい分類（classification）手法で、優れた結果を示すことができる。

2つへの分類問題においては、すべての項目が、理想的に乗るような超平面（hyperplane）[1]を多次元中から探す。さらに、その超平面の周りに辺縁（margin）を見出し、超平面からちょうどその辺縁分はなれたところに位置する点を示し、これをサポート・ベクター（support vector）と呼ぶ。

コメント [T1]: ある分類の項目が片面に、もう一つの分類項目がその裏面に乗る。

- e.g.① コレスポンデンス分析時に使用した中世フランス語のデータ←tag trigram における、ジャンル間（prose vs. poetry）の相違
  - パッケージ e1071 を呼び込む。

```
> library (e1071)
```

  - SVM 関数 svm()の式を立てる。

```
> genre.svm = svm (oldFrench, oldFrenchMeta$Genre)     # svm (頻度, ジャンル)
> genre.svm

 Call:
 svm.default(x = oldFrench, y = oldFrenchMeta$Genre)

 Parameters:
    SVM-Type:  C-classification
  SVM-Kernel:  radial
        cost:  1
       gamma:  0.02857143

 Number of Support Vectors:  158                # support vector の数
```

  - 作図する（Figure 5.19, p. 161）。 # 図中の＋が support vector を示す。

SVM を直接視覚化する方法はないので、多次元尺度法（multidimensional scaling）を利用する。

```
> plot (cmdscale (dist(oldFrench)),
+ col = c ("black", "darkgrey")[as.integer(oldFrenchMeta$Genre)],
+ pch = c ("o", "+")[1:nrow(oldFrenchMeta) %in% genre.svm$index + 1])
```

コメント [T2]: ジャンルごとに 1 か 2 を返す。

コメント [T3]: 行番号を返す。

  - 予測された分類と実際の分類を比較する。

```
> xtabs ( ~ oldFrenchMeta$Genre + predict (genre.svm))
                        predict(genre.svm)
oldFrenchMeta$Genre poetry prose
             poetry    198     0
             prose       1   143
```

---

[1] 0 でない $n$ 変数の $K$ 係数一次多項式 $a_1x_1 + a_2x_2 + \cdots + a_nx_n$ ($a_i \in K$) に対し、

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = b \quad (b \in K)$$

つまり、一本の一次方程式の解空間として定義される $K^n$ の $n-1$ 次の部分線型空間

→異なっているのは、1 つのみ。しかし、データを過剰に適合させようとしている可能性があるので、10 回の cross-validation を行う。

- SVM の式の中に cross-validation を行うように指示を与える。

```
> genre.svm = svm ( oldFrench, oldFrenchMeta$Genre, cross = 10)
10-fold cross-validation on training data:

Total Accuracy: 97.36842          # 平均成功率
Single Accuracies:
 97.05882 100 100 97.05882 97.14286 97.05882 94.11765 97.05882 94.11765 100
```

→平均成功率 0.97 で、ジャンルは、筆者の統語的な慣習により予測することができる。

e.g.② 中世フランス語データの地域(Region; R1, R2, & R3)による分類

- 式を立てる。

```
> region.svm = svm (oldFrench, oldFrenchMeta$Region, cross = 10)
> xtab = xtabs ( ~ oldFrenchMeta$Region + predict (region.svm))
> xtab
                          predict(region.svm)
oldFrenchMeta$Region   R1   R2   R3
                  R1   86   32    1
                  R2    1  152    0
                  R3    6   18   46
```

- 正しく分類できた割合を計算するために、diag ()で対角線上の数値を抽出

```
> diag (xtab)
 R1  R2  R3
 86 152  46
```

- 成功率を算出するために、対角線上の数値の合計をすべての数値の合計で割る。
  →成功率 0.83 しかし、これは、過剰適合をされてしまっていることが、cross-validation によりわかる。

```
> summary (region.svm)
10-fold cross-validation on training data:

Total Accuracy: 62.5731
Single Accuracies:
 52.94118 61.7647 64.70588 55.88235 71.42857 58.82353 58.82353 70.58824 67.64706
62.85714
```

- 平均正確率の正当性を見るために、大多数の地域(R2)が全体に占める割合と比べてみる。

```
> max (xtabs ( ~ oldFrenchMeta$Region)) / nrow (oldFrench)
[1] 0.4473684
```

- R2に属するデータの割合と、cross-validationを行った後の正確率との間に有意差があるかどうかを検定する。

① 分類に成功している数と誤っている数でχ2乗検定をする。

```
> chisq.test(cbind (c (153, 342-153), c (202, 342-202)))

        Pearson's Chi-squared test with Yates' continuity correction

data:  cbind(c(153, 342 - 153), c(202, 342 - 202))
X-squared = 13.4932, df = 1, p-value = 0.0002394
```

② proportions testを行う。

```
> prop.test(cbind (c (153, 342-153), c (202, 342-202)))

        2-sample test for equality of proportions with continuity
        correction

data:  cbind(c(153, 342 - 153), c(202, 342 - 202))
X-squared = 13.4932, df = 1, p-value = 0.0002394
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.22062631 -0.06633803
sample estimates:
    prop 1     prop 2
0.4309859 0.5744681
```

→2つの統計的な検定より、有意差が見られた。成功率は有意に高いということが示された。