# 情報意味論 (課題3)

慶應義塾大学理工学部
櫻井 彰人



## レポート課題3

- レポートは、前回同様、web を通じて行ってください。
- 言うまでもありませんが、決して、他人の著作物をコピーしないで下さい。
- レポートは電子的に作成してください。TeX, MsWord で作成して結構です。提出は pdf 形式でも結構です。
- 書くべき内容に関しては特には述べません。すべて常識的に判断してください。
- 締め切りは、1/31 23:59 とします。万が一遅れる場合には、メールにてご連絡下さい(特に、B4, M2の方)。
- 万が一の緊急の連絡のため、電子メールアドレスを書いて下さい。
- 課題1、課題2について修正・追記されたい方は、再提出して下さい。



## 3.1 SVM

データは csv ファイルとして用意してあります。プログラムはスライド No. 4, 5 のものを用いて下さい(独自に作ってもよい)

UCI ML repository の Heart Disease dataset を対象とし、SVMを用い、パラメータによる分類精度の違いを調べて下さい。なお、クラス属性(病気の疑いの有無)は、2種類に簡単化してあります(http://archive.ics.uci.edu/ml/machine-learning-databases )。参考とする

(1) まず、全データを学習データとし、scale, kernel, cost, type を替えて accuracyの違いを調べて下さい。scaleは、説明変数値を平均0分散1に正規化するか否かを指定するものです。kernelは線形とRBFを、typeは分類と$\nu$-regressionを、costは必要に応じて 0.1, 1, 10, 100 (またはそれ以外)を試してください。
(2) 次に 10-fold cv で(1)と同じ(類似の)ことを試してみて下さい
(3) 次にデータを Pima Indians Diabetes データ
https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes
を対象に、(1)(2) を行って下さい。
(4) 考察を忘れないように。

注 scaleは普通は行った方がよい結果が得られます。

```
library(e1071)
library(bootstrap)

setwd("E:/D/data/")

df <- na.omit( read.csv( "HW03HeartDiseaseCleveland.csv", header=T) )
# df <- na.omit( read.csv( "HW03PimaIndiansDiabetes.csv", header=T) )
target <- df["Class"]
x <- df[-which(colnames(df)=="Class")]
x.scaled <- scale( x )

kernel <- "r"          # "r" for radial basis function and "l" for linear kernel
cost   <- 1
type   <- "nu-r"       # "C" for classification and "nu-r" for regression
ifscale <- F           # to scale or not to scale

  m <- svm( Class~., cbind(x,Class=target[,1]), scale=ifscale, kernel=kernel, type=type, cost=cost)
  targetPred <- predict( m )

  if ( type=="C" )  cm <- table( targetPred, target[,1])      else
                    cm <- table( (sign(targetPred-0.5)+1)/2, target[,1])

  print( cm )
  print( sum(diag(cm))/sum(cm) )
```

```
#### 10-fold cv

kernel <- "r"           # "r" for radial basis function and "l" for linear
cost   <- 1
type   <- "nu-r"        # "C" for classification and "nu-r" for nu-regression
ifscale <- F            # to scale or not to scale

  ngroup <- 10          # the number of folds for cv

  theta.fit <- function ( x, y )
                 svm( Class~., data.frame( x, Class=y),
                              scale=ifscale, kernel=kernel, type=type, cost=cost)
  theta.predict <- function( fit, x ) predict( fit, x )

  results <- crossval( x.scaled, target[,1], theta.fit, theta.predict, ngroup=ngroup)
  targetPred <- results$cv.fit

  if ( type=="C" )  cm <- table( targetPred, target[,1])      else
                    cm <- table( (sign(targetPred-0.5)+1)/2, target[,1])

  print( cm )
  print( sum(diag(cm))/sum(cm) )
```

5

## 3.2a Adaboost 簡単な証明

- Adaboost の訓練誤差の上限を証明してください。すなわち、次の式を証明してください。

$$\text{training error}(H_{\text{final}}) \le \exp\left(-2\sum_t \gamma_t^2\right)$$

- 講義のスライドの式変形の部分を埋めれば結構です。当然ながら、式の個数があう必要は全くありません。
  - ヒントは、スライド中にある式です

6

## 3.2b Adaboost 簡単な実例

- Adaboost で、次のデータの学習を、手で、行って下さい。ただし、仮説は軸並行な線によるもの（例えば、x>1 や y<5）とし、2個まで用いるとします。

| id | x | y | Label |
|---|---|---|---|
| 1 | 5 | 2 | 0 |
| 2 | 4 | 11 | 0 |
| 3 | 5 | 4 | 0 |
| 4 | 2 | 10 | 0 |
| 5 | 6 | 10 | 1 |
| 6 | 9 | 7 | 1 |
| 7 | 9 | 8 | 0 |
| 8 | 8 | 7 | 1 |
| 9 | 11 | 12 | 1 |
| 10 | 6 | 6 | 1 |

- $h_1, D_1, \varepsilon_1, \alpha_1, h_2, D_2, \varepsilon_2, \alpha_2$ を回答して下さい。

7

## 3.3 frequent itemsets

講義が進まなかったら、無し

- 次のデータに対し、Aprioriアルゴリズムを適用し 1～5-frequent itemsets を得て下さい。
  - Minimum support は 40% として下さい。
  - 5-frequent itemsets は空になるはずです。

| 取引ID | トロ | 酢 | 醤油 | 茶 | 米 |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 0 |
| 2 | 1 | 1 | 1 | 1 | 1 |
| 3 | 1 | 0 | 1 | 1 | 0 |
| 4 | 1 | 0 | 1 | 1 | 1 |
| 5 | 1 | 1 | 1 | 1 | 0 |

## 3.4

3.3 がなしとなった場合のみ

(1) SVM（線形カーネル）において、マージンが最大となる超平面が、その他の超平面より、汎化能力がよくなりそうな、直観的な理由を示して下さい。

(2) Bootstrap や bagging の要素学習器に、決定木がよく使われる理由は何でしょうか？（逆に、ニューラルネットワークがあまり使われない（もちろん、使えますよ。使う例もあります。けれどもあまり使われません）理由を述べて下さっても結構です）



## 3.4（続）

3.3 がなしとなった場合のみ

(3) 教師ラベルのないデータと教師ラベルのあるデータとが混在しているときに用いる学習手法は（講義ではほとんど説明しませんでしたが）、半教師あり学習と言われています。どのような課題に適用するのでしょうか？その課題の性質を述べて下さい。課題の例を挙げて説明して下さるとなお、よい。