



第三回 レポート課題

- これまで(皆さんが)学習した機械学習手法のうち、naive Bayes, 決定木, SVM(線形カーネルとRBFカーネル), とニューラルネットワークとを Iris データ, Heart Disease データ, Breast Cancerデータに適用し、その結果に基づき、手法の性質の違いを述べてください。
 - Heart Disease データの取扱い方法については、次のスライドに記すことを参考にして下さい。Breast Cancerについても同様です。
 - なお、svm にも、解の複雑度を制御するパラメータがあります。e1071のsvm では cost という引数です。default値は1です。今回はdefault で結構です。RBFカーネルには、このsvmでは、gammaがあります。defaultは1/ncol(x) で、BreastCancerのときは1/9になっています。その10倍と1/10倍を比較して下さい。
- 締め切りは、1/13(土曜日) 23:59 とします。
- 他の注意はこれまでと同じです。特に: **必ず、自分でやって下さい。**

```
library(e1071)

# Breast-cancer data
# 良性 ('benign') と悪性 ('malignant') を区別する。
# パッケージ mlbench にあるデータ BreastCancer
# nnet のプログラム例では、数値 0 と1 に直して回帰を行ったが、
# 本プログラム例では、もとのカテゴリデータのままだ。

setwd("D:/R/Sample")
hddata <- read.csv('HW03HeartDiseaseCleveland.csv', header=TRUE)

# 属性名とデータの最初の部分を見ると、不要な(というよりあるべきでない)
# 属性が分る。勿論、慎重に考えるべきであるが
# 属性数、データ数もみてみよう

names(hddata)
dim(hddata)
head(hddata)

# 測定値のない、すなわち、NA と書いてある行がある。
# アルゴリズムによっては、単に無視したり、誤動作したりする。
# 削除しておくとう便利である(安易な解決方法ではある)

hddata <- na.omit(hddata)
dim(hddata)
ndata <- dim(hddata)[1]
```

```
# subset() 関数を用いて、不要な属性の削除ができる
# Classは分離してみよう

hd <- subset(hddata, select=c(-Class))
hdclass <- subset(hddata, select=Class)

# CV ではなく、学習データとテストデータを分けて実験してみよう
# 訓練データ(全データの部分集合)を作る

ntdata <- round(ndata*0.75)
hdtrain <- hd[1:ntdata,]
hdtrainclass <- hdclass[1:ntdata,]

# テストデータ(全データの部分集合)を作る
hdtest <- hd[(ntdata+1):ndata,]
hdtestclass <- hdclass[(ntdata+1):ndata,]
```

svm での学習には、この二つのデータを用いて下さい。

```
# iris のデータから各クラス40個の学習データと
# 10個のテストデータは、例えば、次のようにして、
# 取り出すことができます
iristrain <- iris[c(1:40, 51:90, 101:140),]
iristest <- iris[c(41:50, 91:100, 141:150),]
```

テストデータを用いた予測は次のように行う

```
pred <- predict(model, hdtest)
(cm <- table(hdtestclass, pred))
(accuracy <- sum(diag(cm))/sum(cm))
```