Wk11-3 : 서포트벡터머신 III
(Support Vector Machine)

1. Breast Cancer 데이터 설명

- **Breast Cancer Wisconsin (Diagnostic) Data Set**
  - 세침흡인 세포검사를 통해 얻은 683개 유방조직의 9개 특성을 나타냄
  - 자료 출처: UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/index.php)

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	Y
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	benign
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10	9	7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	1	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign
1035283	1	1	1	1	1	1	3	1	1	benign
1036172	2	1	1	1	2	1	2	1	1	benign
1041801	5	3	3	3	2	3	4	4	1	malignant
1043999	1	1	1	1	2	3	3	1	1	benign
1044572	8	7	5	10	7	9	5	5	4	malignant
1047630	7	4	6	4	6	1	4	3	1	malignant
1048672	4	1	1	1	2	1	2	1	1	benign
1049815	4	1	1	1	2	1	3	1	1	benign

input변수(독립변수) output변수(종속변수, 타겟변수)

#	Attribute	Domain
1	샘플 코드 번호	ID number
2	종양 두께	1 - 10
3	조직 크기의 균등성	1 - 10
4	조직 모양의 균등성	1 - 10
5	가장자리 흡착	1 - 10
6	상피조직 크기	1 - 10
7	노출핵	1 - 10
8	순한염색질	1 - 10
9	정상 세포핵	1 - 10
10	유사분열	1 - 10
11	Class	Benign(양성, 정상) Malignant(악성)

2. 서포트벡터머신 패키지와 함수

- 서포트벡터머신을 수행하기 위한 패키지 : e1071
- 오분류율 교차표 생성을 위한 패키지 : caret

```
# load package for support vector machine
library(e1071) #svm model

# load package for Confusion matrix
library(caret)

# set working directory
setwd("D:/tempstore/moocr/wk11")

# read data
cancer<-read.csv("cancer.csv")
head(cancer, n=10)

# remover X1 column(ID number)
cancer<-cancer[, names(cancer) != "X1"]
attach(cancer)
```

(e1071, caret) 라이브러리 설정

데이터 불러오기, 첫번째 10줄을 데이터보기

첫번째 column인 ID number는 필요없는 feature이므로 제거

```
> # read data
> cancer<-read.csv("cancer.csv")
> head(cancer, n=10)
        X1 X2 X3 X4 X5 X6 X7 X8 X9 X10         Y
1  1000025  5  1  1  1  2  1  3  1   1    benign
2  1002945  5  4  4  5  7 10  3  2   1    benign
3  1015425  3  1  1  1  2  2  3  1   1    benign
```

ID

3. kernel 함수에 따른 결과비교

- Breast Cancer 데이터 (학습데이터와 검증데이터의 분할)

```
# training (455) & test set (228)
#set.seed(1000)
N=nrow(cancer)
set.seed(998)

# split train data and test data
tr.idx=sample(1:N, size=N*2/3, replace=FALSE)
train <- cancer[ tr.idx,]
test  <- cancer[-tr.idx,]
```

데이터분할 (학습데이터 2/3, 검증데이터 1/3)

train (455개의 데이터)
test (228개의 데이터)

3. kernel 함수에 따른 결과비교

• Kernel 함수에 따른 서포트벡터머신

```
#svm using kernel
m1<-svm(Y~., data = train)
summary(m1)
m2<-svm(Y~., data = train,kernel="polynomial")
summary(m2)
m3<-svm(Y~., data = train,kernel="sigmoid")
summary(m3)
m4<-svm(Y~., data = train,kernel="linear")
summary(m4)
```

m1-kernel: radial
m2-kernel: polynomial
m3-kernel: sigmoid
m4-kernel: linear

3. kernel 함수에 따른 결과비교

▪ 서포트벡터머신 결과(kernel-radial basis function)

```
> summary(m1)

Call:
svm(formula = Y ~ ., data = train)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.1

Number of Support Vectors:  85

 ( 56 29 )


Number of Classes:  2

Levels:
 benign malignant
```

```
#정확도 측정
pred11<-predict(m1,test)
confusionMatrix(pred11, test$Y)
```

```
> confusionMatrix(pred11, test$Y)
Confusion Matrix and Statistics

          Reference
Prediction  benign malignant
  benign       138         1
  malignant      4        85

               Accuracy : 0.9781
```

3. kernel 함수에 따른 결과비교

- 서포트벡터머신 결과(kernel-polynomial)

```
> summary(m2)

Call:
svm(formula = Y ~ ., data = train, kernel = "polynomial")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  polynomial
       cost:  1
     degree:  3
      gamma:  0.1
     coef.0:  0

Number of Support Vectors:  79

 ( 42 37 )


Number of Classes:  2

Levels:
 benign malignant
```

```
#정확도 측정
pred12<-predict(m2,test)
confusionMatrix(pred12, test$Y)
```

```
> pred12<-predict(m2,test) # polynomial
> confusionMatrix(pred12, test$Y)
Confusion Matrix and Statistics

          Reference
Prediction  benign malignant
  benign       142        12
  malignant      0        74

               Accuracy : 0.9474
```

Q : False positive와 False negative중 어느것이 더 위험할까?

3. kernel 함수에 따른 결과비교

- 서포트벡터머신 결과(kernel-sigmoid)

```
> summary(m3)

Call:
svm(formula = Y ~ ., data = train, kernel = "sigmoid")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  sigmoid
       cost:  1
      gamma:  0.1111111
     coef.0:  0

Number of Support Vectors:  30

 ( 15 15 )


Number of Classes:  2

Levels:
 benign malignant
```

```
#정확도 측정
pred13<-predict(m3,test)
confusionMatrix(pred13, test$Y)
```

```
> pred13<-predict(m3,test) # sigmoid
> confusionMatrix(pred13, test$Y)
Confusion Matrix and Statistics

          Reference
Prediction  benign malignant
  benign       137         1
  malignant      5        85

               Accuracy : 0.9737
```

3. kernel 함수에 따른 결과비교

- 서포트벡터머신 결과(kernel-linear)

```
> summary(m4)

Call:
svm(formula = Y ~ ., data = train, kernel = "linear")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  1
      gamma:  0.1111111

Number of Support Vectors:  41

 ( 21 20 )


Number of Classes:  2

Levels:
 benign malignant
```

```
#정확도 측정
pred14<-predict(m4,test)
confusionMatrix(pred14, test$Y)
```

```
> pred14<-predict(m4,test) # linear
> confusionMatrix(pred14, test$Y)
Confusion Matrix and Statistics

          Reference
Prediction  benign malignant
  benign       141         3
  malignant      1        83

               Accuracy : 0.9825
```