판별분석(Discriminant Analysis)

http://wolfpack.hannam.ac.kr/Stat_Notes/adv_stat/MDA/MDA_%ED%8C%90%EB%B3%84%EB%B6%84%EC%84%9D.pdf
http://contents.kocw.net/KOCW/document/2015/dongguk/shimkyubark1/10-2.pdf

판별 분석: 두개 이상의 모집단으로부터 표본이 섞였을 경우, 개별 경우에 대하여 그것이 어떤 모집단에 속하는지를 판별하기 위한 함수를 만들어서 데이터를 분류하는 방법
판별함수는 그룹 안 분산(Variance within group)에 비하여 그룹 간 분산(Variance between group)의 최대화로 얻어짐.

로직스틱 회귀분석과 많이 비교됨.

종류:

1) 선형(Linear)판별 분석: 정규 분포의 분산-공분산 행렬이 범주에 관계없이 동일한 경우 적용
2) 이차(Quadratic)판별 분석: 정규 분포의 분산-공분산 행렬이 범주별로 다른 경우 적용

1. 선형 판별 분석(Linear Discriminant Analysis)

선형 판별 분석: 데이터를 특정 축에 투영하여 데이터를 잘 구분할 수 있는 직선을 찾는 것을 목표로 하는 분석 방법

실습

==========

```
#패키지 설치
install.packages("caTools")
install.packages("MASS")

library(caTools)

#training, test set
```

```
set.seed(1000)
split <- sample.split(iris$Species, SplitRatio=.7)
train <- subset(iris, split == T)
test <- subset(iris,split == F)
test.y <- test[,5]

# LDA 실행
library(MASS)

# Species가 3종류이므로 prior 3개 설정
iris.lda <- lda(Species~., data=train, prior=c(1/3, 1/3, 1/3))
iris.lda
plot(iris.lda)
```

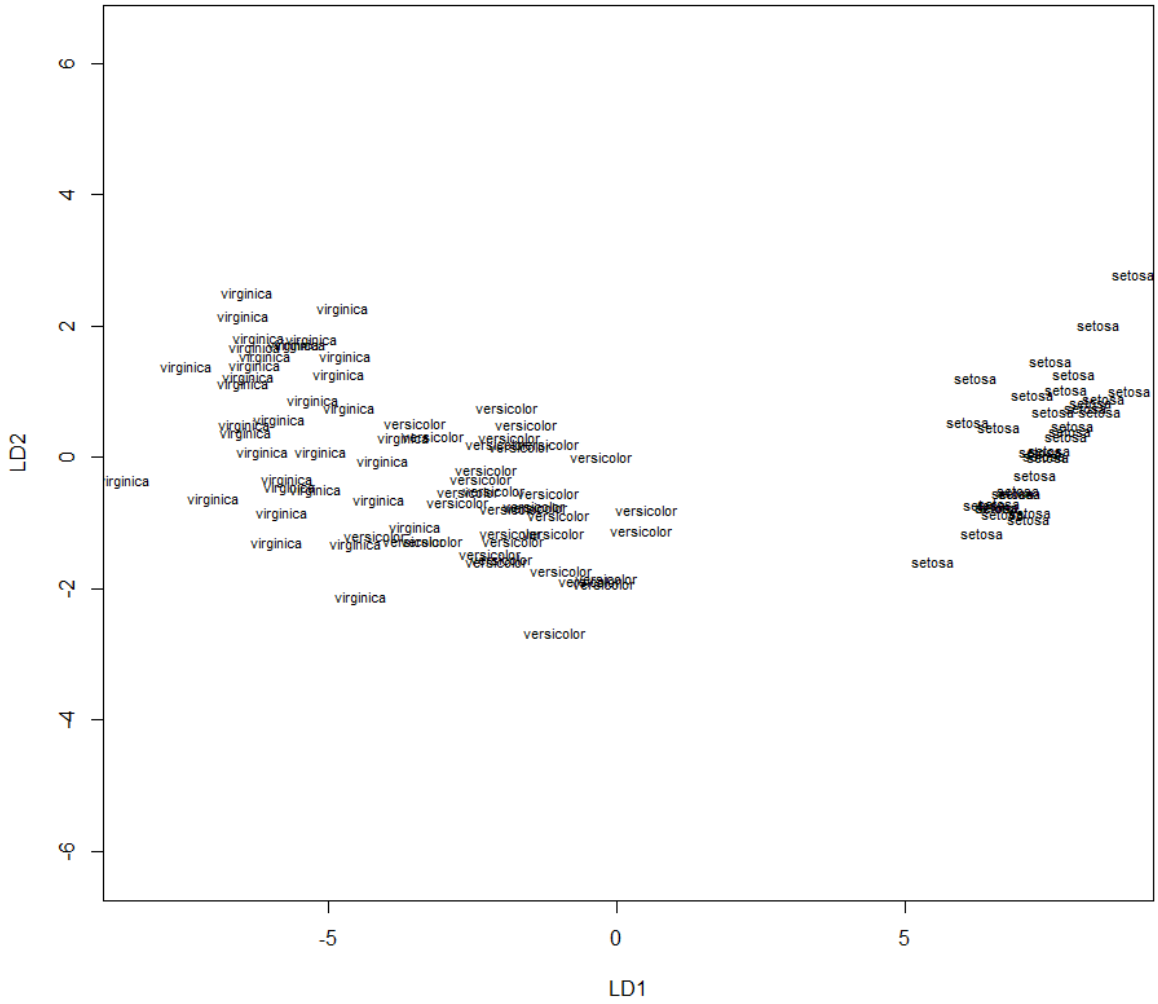
==========

```
> iris.lda
Call:
lda(Species ~ ., data = train, prior = c(1/3, 1/3, 1/3))

Prior probabilities of groups:
    setosa versicolor  virginica
 0.3333333  0.3333333  0.3333333

Group means:
           Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa         5.017143    3.405714     1.477143   0.2514286
versicolor     5.854286    2.742857     4.225714   1.3457143
virginica      6.642857    2.985714     5.602857   2.0885714

Coefficients of linear discriminants:
                    LD1        LD2
Sepal.Length  0.9125594  0.8643219
Sepal.Width   1.4603364  1.5815447
Petal.Length -2.0975009 -1.4019661
Petal.Width  -2.8104854  2.9102139

Proportion of trace:
   LD1    LD2
0.9897 0.0103
```

Prior probabilities of groups: 각 특성별 사전 부여 비율
Group means: 각 Species 별 변수의 평균 값
Coefficients of linear discriminants: 변수 특성

데이터가 LD1을 기준으로 확실히 분리되는 것을 볼 수 있음 (98.87%).

LD2로는 특별한 분류의 효과를 보지 못함.



LD1, LD2로 구성된 모델을 기반으로 주어진 데이터에 대한 예측

==========

```
testpred <- predict(iris.lda, test)
table(test.y, testpred$class)
```

==========

```
> testpred <- predict(iris.lda, test)
> table(test.y, testpred$class)

test.y       setosa versicolor virginica
  setosa         15          0         0
  versicolor      0         15         0
  virginica       0          0        15
> 
```

2. 이차판별분석 (Quadratic Discriminant Analysis)

실습.

```
=========================
# 패키지설치
install.packages("biotools")
library(biotools)

# 분산-공분산 행렬이 동일하지 않은 지 확인
boxM(iris[1:4], iris$Species)
# p-value가 0.05보다 작으면 분산-공분산 행렬이 동일하지 않음.
# p-value가 0.05보다 크면 QDA 적용 못함.

iris.qda <- qda(Species~., data=train, prior=c(1/3, 1/3, 1/3))
iris.qda
# QDA는 직선이 아니므로 그래프로 표현하기 어려움

#테스트 데이터 대상으로 예측
testqda <- predict(iris.qda, test)
table(test.y, testqda$class)

=======================
```

```
>
> library(biotools)
>
> boxM(iris[1:4], iris$Species)

        Box's M-test for Homogeneity of Covariance Matrices

data:  iris[1:4]
Chi-Sq (approx.) = 140.94, df = 20, p-value < 2.2e-16

>
> iris.qda <- qda(Species~., data=train, prior=c(1/3, 1/3, 1/3))
> iris.qda
Call:
qda(Species ~ ., data = train, prior = c(1/3, 1/3, 1/3))

Prior probabilities of groups:
    setosa versicolor  virginica
 0.3333333  0.3333333  0.3333333

Group means:
           Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa         5.017143    3.405714     1.477143   0.2514286
versicolor     5.854286    2.742857     4.225714   1.3457143
virginica      6.642857    2.985714     5.602857   2.0885714
>
>
> testqda <- predict(iris.qda, test)
> table(test.y, testqda$class)

test.y       setosa versicolor virginica
  setosa         15          0         0
  versicolor      0         15         0
  virginica       0          0        15
>
```

실습2

http://contents.kocw.net/KOCW/document/2015/chungbuk/najonghwa1/12.pdf