

## 1. 판별분석이란?

판별분석은 두 개 이상의 모 집단에서 추출된 표본들이 지닌 정보를 이용해 이 표본들이 어느 모 집단에서 추출된 것인지를 결정해 줄 수 있는 기준을 찾는 분석법입니다.

## 2. 판별분석의 예시

은행에서의 대출을 행하고자 할 때, 채무자가 대출금을 갚을 수 있는지 없는지의 여부를 과거에 대출금을 반환치 않는 사람의 정보를 참고하여 담보 신청 시 신청자의 정보 유형을 과거의 유형과 비교하여 가능성을 파악하거나 고등학교 학생의 다수를 랜덤하게 선택해 각 학생의 성취도 시험 점수, 동기 점수 및 현재 학습 과정을 기록하고 이를 분류해 이후 새로운 학생의 경우에 여러 진로 중 하나로 판별하는 등의 예가 있습니다.

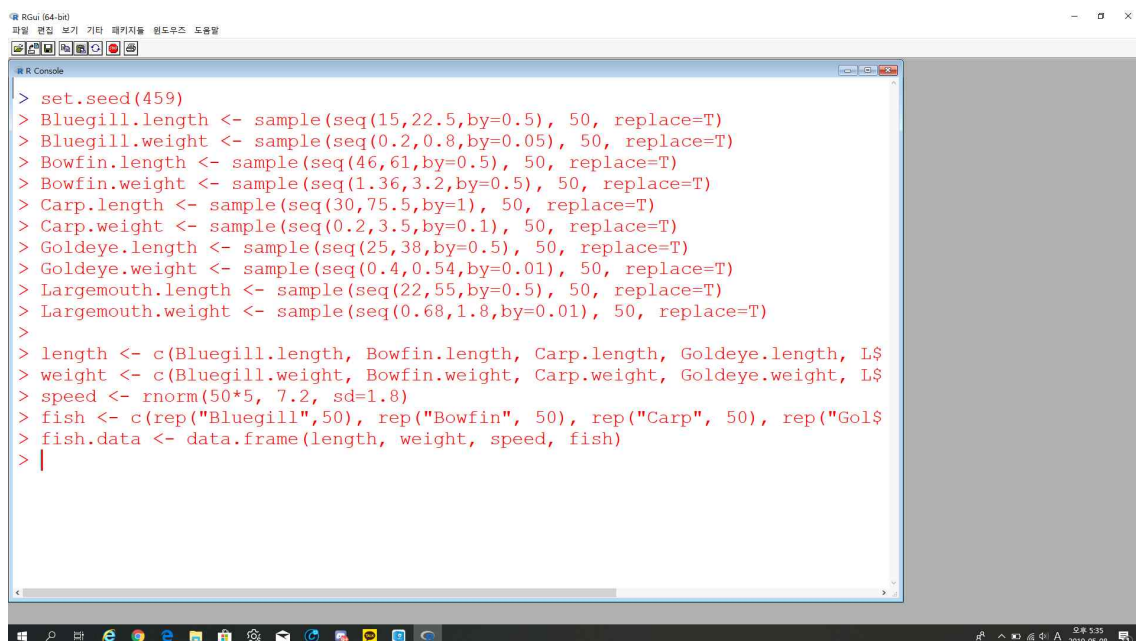
## 3. 판별분석의 단계

- 1) 케이스가 속한 집단을 구분하는데 기여할 수 있는 독립 변수를 찾습니다.
- 2) 집단을 구분하는 기준이 되는 독립 변수들의 선형 결합 즉, 판별 함수를 도출합니다.
- 3) 도출된 판별 함수에 의한 분류의 정확도를 파악합니다.
- 4) 판별 함수를 이용하여 새로운 케이스가 속하는 집단을 예측합니다

## 4. 판별분석 실습

온타리오 호수에 사는 어종에 대한 자료

- 블루길, 아미아, 잉어, 골드아이, 입큰송어의 데이터로 시뮬레이션 데이터셋 만들기



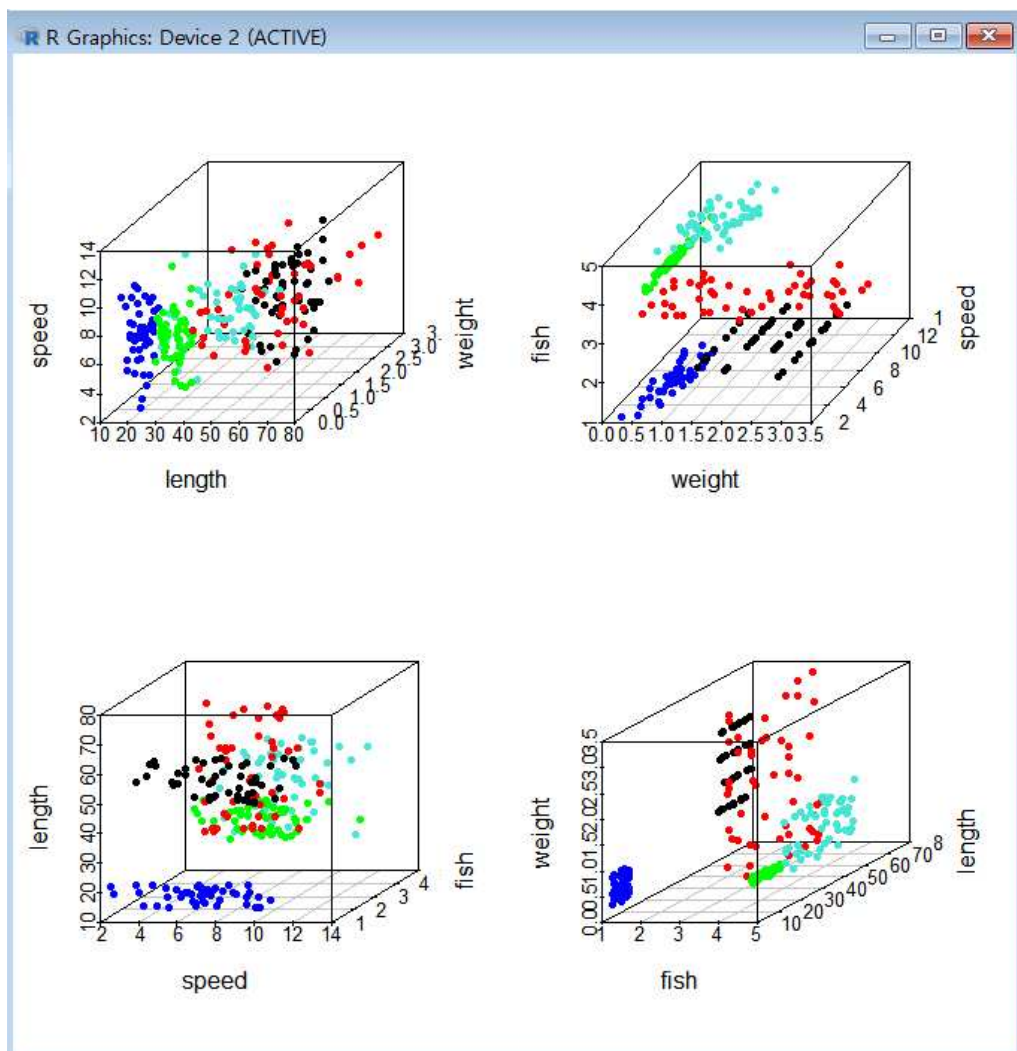
```
> set.seed(459)
> Bluegill.length <- sample(seq(15,22.5,by=0.5), 50, replace=T)
> Bluegill.weight <- sample(seq(0.2,0.8,by=0.05), 50, replace=T)
> Bowfin.length <- sample(seq(46,61,by=0.5), 50, replace=T)
> Bowfin.weight <- sample(seq(1.36,3.2,by=0.5), 50, replace=T)
> Carp.length <- sample(seq(30,75.5,by=1), 50, replace=T)
> Carp.weight <- sample(seq(0.2,3.5,by=0.1), 50, replace=T)
> Goldeye.length <- sample(seq(25,38,by=0.5), 50, replace=T)
> Goldeye.weight <- sample(seq(0.4,0.54,by=0.01), 50, replace=T)
> Largemouth.length <- sample(seq(22,55,by=0.5), 50, replace=T)
> Largemouth.weight <- sample(seq(0.68,1.8,by=0.01), 50, replace=T)
>
> length <- c(Bluegill.length, Bowfin.length, Carp.length, Goldeye.length, L$
> weight <- c(Bluegill.weight, Bowfin.weight, Carp.weight, Goldeye.weight, L$
> speed <- rnorm(50*5, 7.2, sd=1.8)
> fish <- c(rep("Bluegill",50), rep("Bowfin", 50), rep("Carp", 50), rep("Gol$
> fish.data <- data.frame(length, weight, speed, fish)
> |
```

산점도를 그려줄 함수 만들기

```
> plot3DfishData <- function(x, y, z, data=fish.data)
+ {
+   require("scatterplot3d")
+   fish.variable <- colnames(data)
+   scatterplot3d(data[,x], data[,y], data[,z], color=c("blue", "black", "red", "green", "turquoise")[data$fish]
+     , pch=19, xlab=fish.variable[x], ylab=fish.variable[y], zlab=fish.variable[z])
+ }
```

3차원 산점도 그리기

```
> par(mfrow=c(2,2))
> plot3DfishData(1,2,3)
> plot3DfishData(2,3,4)
> plot3DfishData(3,4,1)
> plot3DfishData(4,1,2)
```



LDA수행 lda()함수를 사용, fish.data에 대하여 LDA를 수행

```
> library("MASS")
> fish.lda <- lda(fish ~., data=fish.data, prior=c(1,1,1,1,1)/5)
> fish.lda
Call:
lda(fish ~ ., data = fish.data, prior = c(1, 1, 1, 1, 1)/5)
```

Prior probabilities of groups:

Bluegill	Bowfin	Carp	Goldeye	Largemouth
0.2	0.2	0.2	0.2	0.2

Group means:

	length	weight	speed
Bluegill	19.25	0.5280	7.222587
Bowfin	53.57	2.1600	7.243531
Carp	52.26	1.7320	7.588852
Goldeye	31.87	0.4748	7.122371
Largemouth	40.11	1.2428	7.163396

Coefficients of linear discriminants:

	LD1	LD2	LD3
length	0.10121664	-0.067114012	0.003296811
weight	1.17216477	1.335393654	0.011024527
speed	-0.02745768	-0.009936134	-0.555543270

Proportion of trace:

LD1	LD2	LD3
0.9730	0.0256	0.0014

새로운 데이터셋 분류 결과와 클래스별 사후확률

- fish.data에 100마리를 선택하여 선형판별방법을 학습

```
> set.seed(10)
> train100 <- sample(1:nrow(fish.data),100)
> table(fish.data$fish[train100])

Bluegill    Bowfin    Carp    Goldeye  Largemouth
    18         25         16         20         21
> fish100.lda <- lda(fish ~., data=fish.data, prior=c(1,1,1,1,1)/5, subset=train100)
```

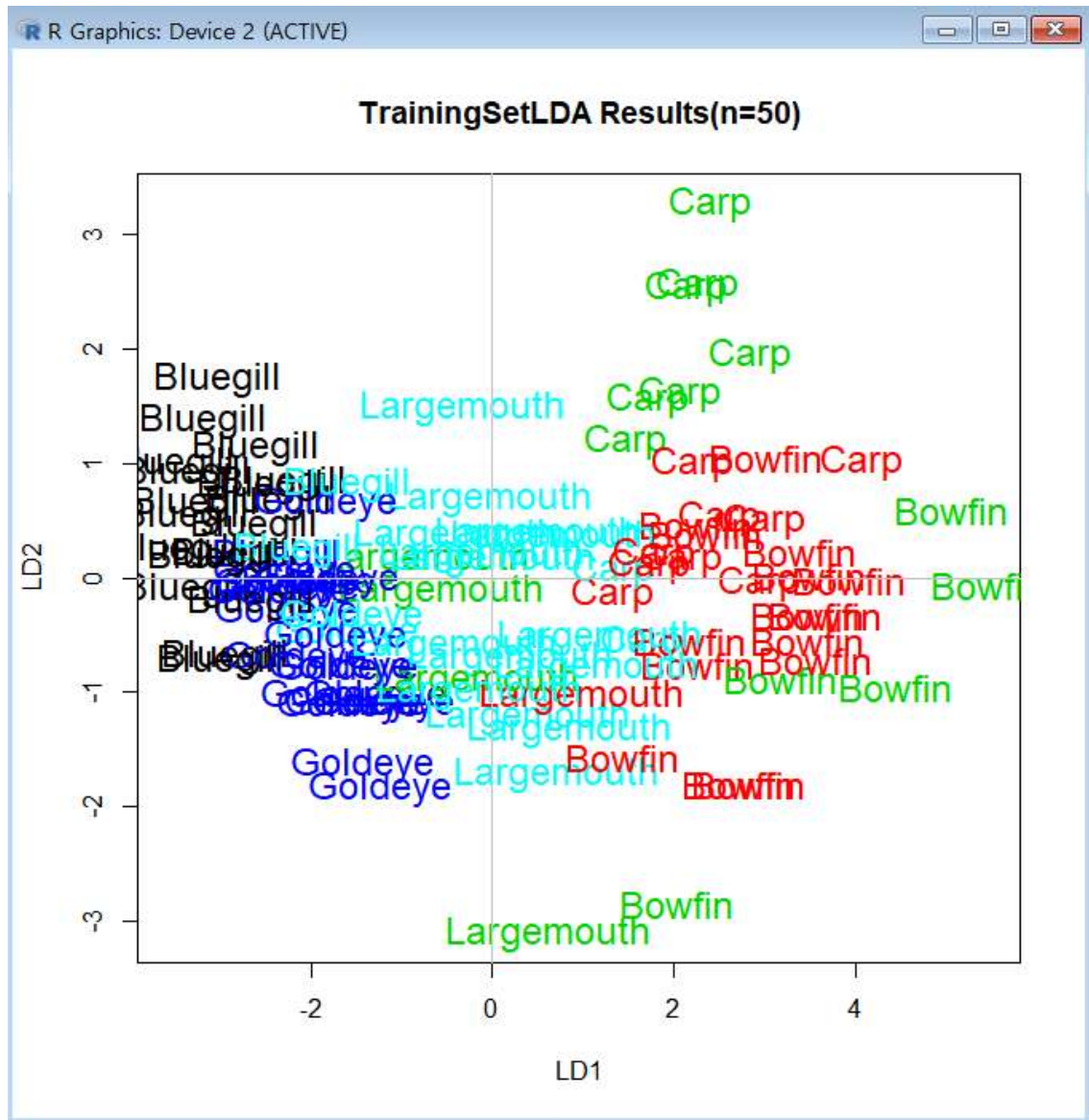
predict() 함수를 사용하여 표본에 포함된 100마리에 대한 분류 결과

```
> predict.fish100 <- predict(fish100.lda)
> table(fish.data$fish[train100], predict.fish100$class)
```

	Bluegill	Bowfin	Carp	Goldeye	Largemouth
Bluegill	18	0	0	0	0
Bowfin	0	15	9	0	1
Carp	0	5	7	0	4
Goldeye	1	0	0	19	0
Largemouth	2	0	2	2	15

몇몇은 분류가 잘못되었음. 시각화를 해본다.

```
> par(mfrow=c(1,1))
> plot(predict.fish100$x, type="n", xlab="LD1", ylab="LD2", main="TrainingSetLDA Results(n=50)")
> text(predict.fish100$x, as.character(predict.fish100$class), col=as.numeric(fish.data$fish[train100]), cex=1.5)
> abline(h=0, col="gray")
> abline(v=0, col="gray")
```



predict() 함수로 나머지 데이터를 분류하고, 분류결과, 사후확률, 오분류율을 계산함.

```
> predict.new <- predict(fish100.lda, newdata=fish.data[-train100,])  
> table(fish.data$fish[-train100], predict.new$class)
```

	Bluegill	Bowfin	Carp	Goldeye	Largemouth
Bluegill	32	0	0	0	0
Bowfin	0	11	13	0	1
Carp	0	16	5	3	10
Goldeye	0	0	0	30	0
Largemouth	3	0	3	5	18

```
> TAB <- table(fish.data$fish[-train100], predict.new$class)  
> mcrlad <- 1-sum(diag(TAB))/sum(TAB)  
> mcrlad  
[1] 0.36
```

## 5. 소감

전부 처음 접해보는 개념들 생소한 단어들에 머릿속에 혼란스럽고, 강의중에는 이해가 안 되고 집중하기 어려웠습니다. 그러나, 이렇게 R언어로 예제를 생각해보고 또 다른 예시로 실습을 해보면서 직접 데이터를 구하고 눈으로 보면서 이해를 더 잘 할 수 있었고, 앞으로의 공부에서도 도움이 될 유익한 시간이었습니다. 이러한 과제를 내어주신 심재창 교수님께 감사드리며, R언어에 더 흥미를 갖게 되었습니다! R언어 사랑해요~!