

채널 G

2016125077 최재혁

2016126049 박희재

데이터 사이언스

- Histogram, Scatter plot 분류예측



목 차

- 과제 설명

분류예측을 위해 Histogram에서 최적 cut-off, Scatter plot에서 부등식 구하기.

- 과제 목표

Histogram과 Scatter plot을 전반적으로 이해하고, 분류예측이 정확하도록 최적의 지점을 구한다.

- 사용 Tool

분석용 언어 : R

목 차

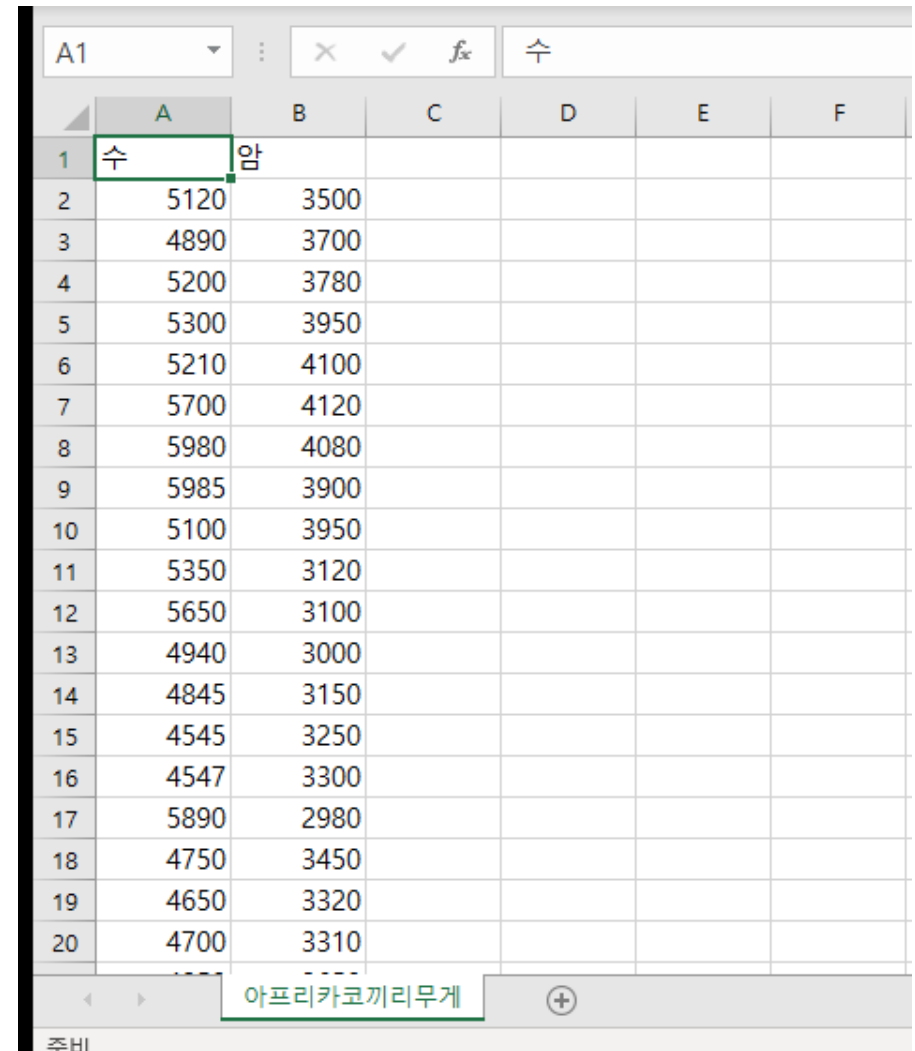
-
- 과제1 – Histogram

-
- 과제2 – Scatter Plot
-

과제1 – Histogram

- Data

- 아프리카 코끼리의 개체 별 무게와 성별이 나타난 데이터.
- 수컷의 몸무게는 평균 5500kg, 암컷의 몸무게는 평균 3500kg
- 수컷과 암컷의 비율 : (1 : 1)



	A	B	C	D	E	F
1	수	암				
2	5120	3500				
3	4890	3700				
4	5200	3780				
5	5300	3950				
6	5210	4100				
7	5700	4120				
8	5980	4080				
9	5985	3900				
10	5100	3950				
11	5350	3120				
12	5650	3100				
13	4940	3000				
14	4845	3150				
15	4545	3250				
16	4547	3300				
17	5890	2980				
18	4750	3450				
19	4650	3320				
20	4700	3310				

과제1 – Histogram

- 코드

```
scatter.R x histogram.R* x
← → | | | Source on Save | | | Run | | | Source | | |
1 df <- read.csv(file = "C:/Users/user/Desktop/3학년2학기/데이터사이언스
2 /과제/3주/아프리카코끼리무게.csv",header=TRUE, fileEncoding = "UCS-2LE")
3 df
4 head(df)
5 summary(df)
6 value_m <- df$수
7 value_f <- df$암
8 mhist<-hist(value_m, rnorm(49,5179,1),main="아프리카 코끼리 무게",
9 xlab="무게",ylab="마리수",col=rgb(0,0,1,0.5),breaks=c(seq(2800,6100,by=100)),xaxt="n")
10 fhist<-hist(value_f, xaxt='n',rnorm(49,3590,1),main="아프리카 코끼리 무게",
11 xlab="무게",ylab="마리수",col=rgb(1,0,0,0.5),breaks=c(seq(2800,6100,by=100)),xaxt="n",add=TRUE)
12 axis(side=1, at=seq(2800,6100,100), labels=seq(2800,6100,100))
13
13:1 (Top Level) R Script
```

과제1 - Histogram

- Histogram



- 개체의 성별이 수컷인 경우에 무게가 무겁고, 반대로 암컷의 경우에는 무게가 상대적으로 적게 나가는 경향이 있다
- 이를 통해 성별이 파악이 되지 않는 개체는 무게를 통해서 예측할 수 있을 것이다
- 분류 예측시, Error가 나타날 수 있는 구간은 4100 - 4400 의 구간으로 이 구간의 무게를 가진 코끼리의 성별을 예측 할 때, 잘못된 예측을 하게 될 수 있다.

과제1 - Histogram

- Cut-Off



우리 조에서 생각해 낸 Cut-Off는 히스토그램에서 성별에 따라 두개의 봉우리가 생긴다

두 봉우리에 중첩되는 구간은 4100kg~4400kg으로 확인이 된다

4200kg~4300kg 막대에 암컷의 개체 수가 더 많기 때문에 Cut-Off는 오른쪽으로 약간 치우쳐진다고 할 수 있다.

반면 4300kg~4400kg 막대에 수컷의 개체 수가 더 많기 때문에 Cut-off는 왼쪽으로 치우쳐진다.

따라서 약 4280kg정도에서 Cut-Off가 생긴다고 할 수 있다.

과제2 - Scatter Plot

- Data

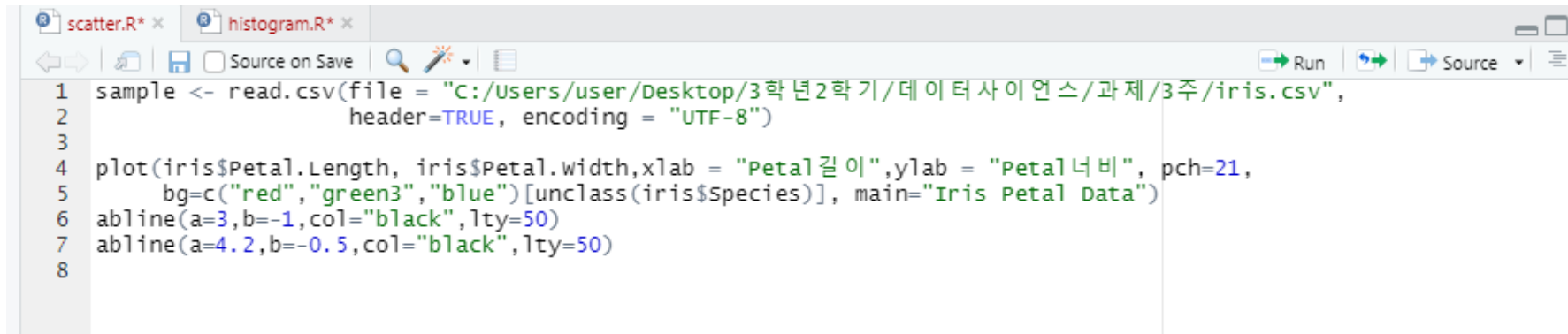
- PPT에 등장하는 Iyris 데이터이다.
- 목표변수 : 개체의 종류
- 예측변수 : 개체의 꽃잎의 길이, 꽃잎의 너비
- 개체 종의 비율 : setosa : 50, versicolor : 50, virginica : 50
으로 각 1:1:1의 비율이다.
- 데이터를 다룰 때 주의해야할 점 :
목표는 분류 예측이므로, 분류예측시에 Error이 되도록
발생하지 않도록 예측 변수를 잘 선택해야한다.

Id	SepalLeng	SepalWidth	PetalLeng	PetalWidth	Species
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa
4	4.6	3.1	1.5	0.2	Iris-setosa
5	5	3.6	1.4	0.2	Iris-setosa
6	5.4	3.9	1.7	0.4	Iris-setosa
7	4.6	3.4	1.4	0.3	Iris-setosa
8	5	3.4	1.5	0.2	Iris-setosa
9	4.4	2.9	1.4	0.2	Iris-setosa
10	4.9	3.1	1.5	0.1	Iris-setosa
11	5.4	3.7	1.5	0.2	Iris-setosa
12	4.8	3.4	1.6	0.2	Iris-setosa
13	4.8	3	1.4	0.1	Iris-setosa
14	4.3	3	1.1	0.1	Iris-setosa
15	5.8	4	1.2	0.2	Iris-setosa
16	5.7	4.4	1.5	0.4	Iris-setosa
17	5.4	3.9	1.3	0.4	Iris-setosa
18	5.1	3.5	1.4	0.3	Iris-setosa
19	5.7	3.8	1.7	0.3	Iris-setosa
20	5.1	3.8	1.5	0.3	Iris-setosa
21	5.4	3.4	1.7	0.2	Iris-setosa
22	5.1	3.7	1.5	0.4	Iris-setosa
23	4.6	3.6	1	0.2	Iris-setosa
24	5.1	3.3	1.7	0.5	Iris-setosa
25	4.8	3.4	1.9	0.2	Iris-setosa
26	5	3	1.6	0.2	Iris-setosa
27	5	3.4	1.6	0.4	Iris-setosa
28	5.2	3.5	1.5	0.2	Iris-setosa
29	5.2	3.4	1.4	0.2	Iris-setosa
30	4.7	3.2	1.6	0.2	Iris-setosa
31	4.8	3.1	1.6	0.2	Iris-setosa
32	5.4	3.4	1.5	0.4	Iris-setosa
33	5.2	4.1	1.5	0.1	Iris-setosa
34	5.5	4.2	1.4	0.2	Iris-setosa

<사진 1> - Iyris Data

과제2 - Scatter Plot

- 코드

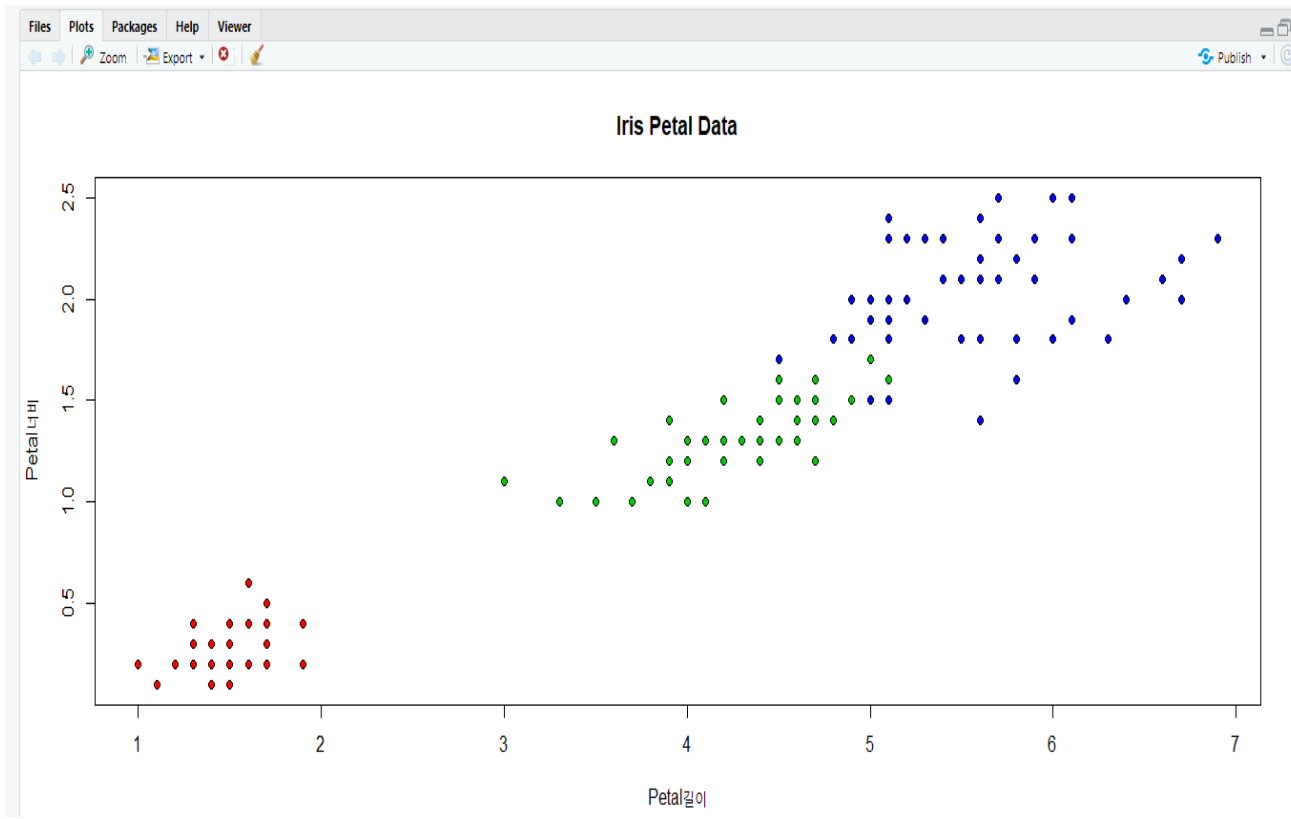


```
1 sample <- read.csv(file = "C:/Users/user/Desktop/3학년2학기/데이터사이언스/과제/3주/iris.csv",
2                     header=TRUE, encoding = "UTF-8")
3
4 plot(iris$Petal.Length, iris$Petal.width,xlab = "Petal 길이",ylab = "Petal 너비", pch=21,
5      bg=c("red","green3","blue")[unclass(iris$Species)], main="Iris Petal Data")
6 abline(a=3,b=-1,col="black",lty=50)
7 abline(a=4.2,b=-0.5,col="black",lty=50)
8
```

- sample에 iris데이터를 불러들인다.
- x축은 꽃잎의 길이, y축에는 꽃잎의 너비로 설정한다
- Scatter Plot로 각 데이터를 표현할 때에, 빨강, 파랑, 초록 색으로 iris의 종류별로 분류하여 나타낸다.

과제2 - Scatter Plot

- Scatter-Plot



- 빨간색은 setosa, 파란색은 versicolor, 초록색은 virginica 종이다

- 이를 통해 x,y좌표가 0에 가까운 개체의 종류는 setosa,

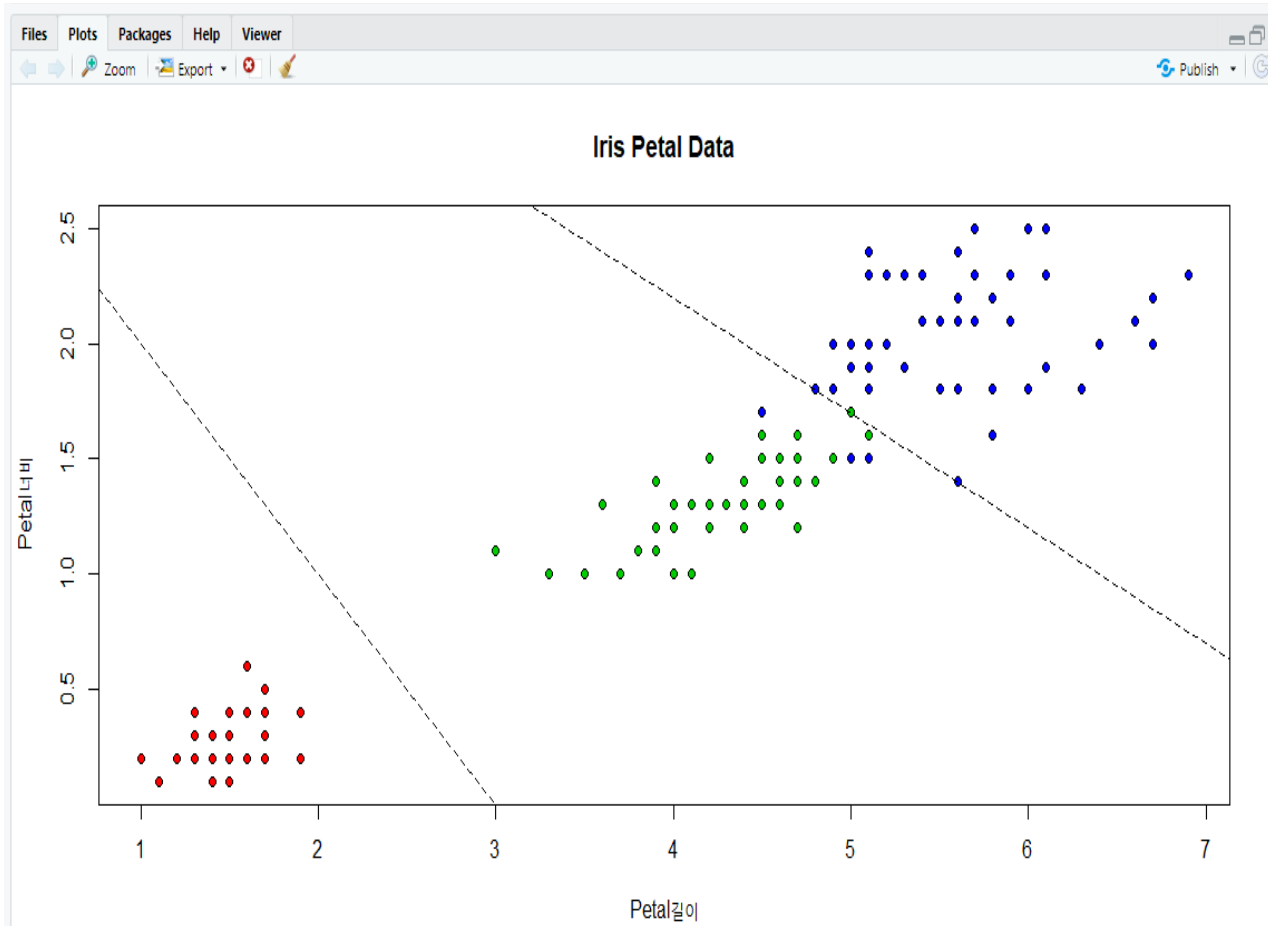
x좌표가 3~5 안에 있고, y좌표가 5.0~7.0안에 있으면 versicolor,

x좌표가 5이상, y좌표가 7.0이상이면 virginica종으로 대략적으로 예측할 수 있다.

- 분류 예측시, Error가 나타날 수 있는 구간은 파랑색과 초록색이 동시에 나타나는 구간으로 versicolor, virginica 종으로 분류예측시에 어려움이 예상되고 그렇기에 Cut-Off의 필요성이 존재한다

과제2 - Scatter Plot

- Cut-Off



- 우선 우리 조에서 Cut-Off를 구한 방식은 직접 계수 값을 정해가면서 최대한 분류가 잘 이루어질 수 있는 방향으로 선택하였다.

- abline()함수를 이용하여 직선의 방정식을 구하였다. (기본 틀 $y = a + bx$)

※ abline의 arguments들

a : y절편 , b : 기울기 , h : 수평선 , v : 수직선

왼쪽 방정식 : $y = -x + 3$

오른쪽 방정식 : $y = -0.5x + 4.2$

결론 :

$y < -x + 3$ 이면 setosa,

$y > -x + 3 \ \&\& \ y < -0.5x + 4.2$ 이면 virginica

$y > -0.5x + 4.2$ 이면 versicolor

즉, Petal의 길이와 너비를 각각 대입하면 해당 종을 예측할 수 있다.