```
rm(list=ls())
##### Chapter 5 단순 베이즈 분류


### 5.1 서론
# 단순 베이즈 분류 모형은 베이즈 정리에 기반한 방법으로. 사후 확률의 계산 시 조건부 독립을 가정하여 계산을 단순화한 방법이다.
# 사후 확률이 높은 집단으로 새로운 데이터를 분류하게 된다. 조건부 독립이라는 가정이 비현실적인 측면이 있지만, 계산이 간편하여 널리 사용


### 5.2 단순 베이즈 분류
# 단순 베이즈 분류기는 연속형 또는 이산형과 관계없이 임의의 크기의 예측변수를 다룰 수 있다.
# 일반적인 베이즈 분류에서는 사후 확률이 가장 큰 집단으로 개체의 대한 분류를 시작한다.
# 단순 베이즈 분류는 위의 사후 확률의 계산을 좀 더 편하게 할 수 있도록 예측 변수들 간의 독립을 가정
```

데이터가 $x = (x_1, x_2, \dots, x_d)$ 으로 주어질 때, 이 데이터가 C_j 집단으로부터 나왔을 사후확률은 베이즈 정리로 부터 다음과 같다.

$$p(C_j|x) = \frac{p(C_j)p(x|C_j)}{p(x)} = \frac{p(C_j)p(x|C_j)}{\sum_{j=1}^{k} p(C_j)p(x|C_j)}, j = 1, 2, \dots, k$$

$$= \frac{p(C_j)p(x_1|C_j)p(x_2|C_j)\dots p(x_d|C_j)}{p(x)}, \text{(계산을 편리하게 하기 위해 예측변수들 간의 독립을 가정)}$$

$$= \frac{p(C_j)p(x_1|C_j)p(x_2|C_j)\dots p(x_d|C_j)}{\sum_{j=1}^{k} p(C_j)p(x|C_j)}$$

```
######## 사례 1(이산형) ########
```

● 문서분류

문서번호	주요단어	문서분류
1	fun, couple, love, love	comedy
2	fast, furious, shoot	action
3	couple, fly, fast, fun, fun	comedy
4	furious, shoot, shoot, fun	action
5	fly, fast, shoot, love	action

```
## 사례 1-1

# 입력문서가 {fast, furious, fun}을 주요 단어로 가질 때, 이 문서는 어떤 문서로 분류될 것인가?
```

$$P(comedy|x) = \frac{P(comedy)P(x|comedy)}{P(x)}$$

$$P(action|x) = \frac{P(action)P(x|action)}{P(x)}$$

```
# 위에 두 사후 확률을 비교하는 것이므로 분자들의 계산 결과로 비교를 진행한다.
```

$$P(comedy|x) = P(comedy)P(x|comedy) = P(comedy)P(fast|comedy)P(furious|comedy)P(fun|comedy)$$

$$= \frac{2}{5} \times \frac{1}{9} \times \frac{0}{9} \times \frac{3}{9} = 0$$

$$P(action|x) = P(action)P(x|action) = P(action)P(fast|action)P(furious|action)P(fun|action)$$

$$= \frac{3}{5} \times \frac{2}{11} \times \frac{2}{11} \times \frac{1}{11} = \frac{12}{5 \times 11^3} = 0.0018$$

$$P(comedy)P(x|comedy) < P(action)P(x|action)$$

```
# 사후 확률을 비교했을 때, action의 사후 확률이 더 높으므로 action으로 분류될 것이다.

# 단순 베이즈 분류에서 '낮은-빈도 문제'에 주의할 필요가 있다.
# '낮은-빈도 문제' : 빈도가 0인 경우에 새로운 자료에 대한 사후 확률은 항상 0이 되는 문제점
# 문제점을 해결하기 위해 모든 속성값-군집 조합에 대한 빈도에 작은 수(아주 작은 수)를 더해주어 계산을 수행한다.
```

사례 1-2

입력문서가 {fast, love, fun}을 주요 단어로 가질 때, 이 문서는 어떤 문서로 분류될 것인가?

$$P(comedy|x) = \frac{P(comedy)P(x|comedy)}{P(x)}$$

$$P(action|x) = \frac{P(action)P(x|action)}{P(x)}$$

위에 두 사후 확률을 비교하는 것이므로 분자들의 계산 결과로 비교를 진행한다.

$$P(comedy)P(x|comedy) = P(comedy)P(fast|comedy)P(love|comedy)P(fun|comedy)$$

$$= \frac{2}{5}\times\frac{1}{9}\times\frac{2}{9}\times\frac{3}{9} = \frac{12}{5\times 9^3} = \frac{12}{3645} = 0.00329218$$

$$P(action)P(x|action) = P(action)P(fast|action)P(love|action)P(fun|action)$$

$$= \frac{3}{5}\times\frac{2}{11}\times\frac{1}{11}\times\frac{1}{11} = \frac{6}{5\times 11^3} = \frac{6}{6658} = 0.00090157$$

$$P(comedy)P(x|comedy) > P(action)P(x|action)$$

사후 확률을 비교했을 때, comedy의 사후 확률이 더 높으므로 comedy으로 분류될 것이다.

######## 사례 2(연속형 자료일 때) ########

3개의 변수(키, 몸무게, 발 크기)로서 데이터가 형성되었을 때, 세 변수 모두 독립이며, 정규분포를 따른다고 가정

데이터

성별	키 (feet)	몸무게 (lbs)	발크기 (inches)	성별	키 (feet)	몸무게 (lbs)	발크기 (inches)
남성	6	180	12	여자	5	100	6
남성	5.92	190	11	여자	5.5	150	8
남성	5.58	170	12	여자	5.42	130	7
남성	5.92	165	10	여자	5.75	150	9

성별	키		몸무게		발 크기	
	평균	분산	평균	분산	평균	분산
남	5.855	3.5033e-02	176.26	1.2292e+02	11.25	9.1667e-01
여	5.4175	9.7225e-02	132.5	5.5833e+02	7.5	1.6667e+00

남성과 여성 그룹에 속한 사전확률을 P(남성)=P(여성)=0.5 이라고 하자. 주어진 자료(키:6 , 몸무게:130 , 발크기:8)의 성별은 무엇인지 분류해보자.

$$P(\text{남성}|x) = \frac{P(\text{남성})P(x|\text{남성})}{P(x)}$$

$$P(\text{여성}|x) = \frac{P(\text{여성})P(x|\text{여성})}{P(x)}$$

따라서 분자끼리 비교를 통해 사후 확률을 분류한다.

$$P(\text{남성})P(x|\text{남성}) = P(\text{남성})P(\text{키}|\text{남성})P(\text{몸무게}|\text{남성})P(\text{발크기}|\text{남성})$$

$$= \frac{1}{2}\times N(6, 5.855, \sqrt{0.035}) \times N(130, 176.26, \sqrt{122.92}) \times N(8, 11,25, \sqrt{0.91667})$$

$$= \frac{1}{2}\times 1.578886 \times 5.965584e-6 \times 0.001311246$$

$$\fallingdotseq 6.175299\times 10^{-9}$$

$$P(\text{여성})P(x|\text{여성}) = P(\text{여성})P(\text{키}|\text{여성})P(\text{몸무게}|\text{여성})P(\text{발크기}|\text{여성})$$

$$= \frac{1}{2}\times N(6, 5.4175, \sqrt{0.097225}) \times N(130, 132.5, \sqrt{558.33}) \times N(8, 7.5, \sqrt{1.6667})$$

$$= \frac{1}{2}\times 0.2234587 \times 0.01678935 \times 0.2866883$$

$$\fallingdotseq 0.0005377879$$

$$P(\text{남성})P(x|\text{남성}) < P(\text{여성})P(x|\text{여성})$$

사후 확률을 비교했을 때, 여성의 사후 확률이 더 높으므로 여성으로 분류될 것이다.

```
## 남성일 경우의 사후확률
height1 <- dnorm(6, mean=5.855, sd=sqrt(0.035033)) # 1.578886
weight1 <- dnorm(130, mean=176.26, sd=sqrt(122.92)) # 5.965584e-06
feet1 <- dnorm(8, mean=11.25, sd=sqrt(0.91667)) # 0.001311246
height1; weight1; feet1
B_m.x <- 0.5*height1*weight1*feet1
B_m.x # 6.175299e-09

## 여성일 경우의 사후확률
height2 <- dnorm(6, mean=5.4175, sd=sqrt(0.097225)) # 0.2234587
weight2 <- dnorm(130, mean=132.5, sd=sqrt(558.33)) # 0.01678935
feet2 <- dnorm(8, mean=7.5, sd=sqrt(1.6667)) # 0.2866883
height2; weight2; feet2
B_f.x <- 0.5 * height2 * weight2 * feet2
B_f.x # 0.0005377879

## 결과적으로 여자인 경우의 사후 확률이 더 높으므로 주어진 자료는 여성으로 예측된다.
B_m.x # 6.175299e-09
B_f.x # 0.0005377879



### 예제 1번
## 단순 베이즈 분석을 위해 iris자료를 사용한다.
data(iris) # 데이터 선언
head(iris) # 해드 데이터 출력

## {e1071} 패키지의 naiveBayes()함수를 이용하여 단순 베이즈를 수행한다.
library(e1071)
m <- naiveBayes(Species ~ ., data = iris)
m # 단순 베이지 분류를 실행한 결과 대입

# predict()함수를 이용하여 단순베이즈분류를 한 기준을 가지고 iris 데이터에 대해 예측을 새로 실시, 정오분류표를 작성한다.
table(predict(m, iris), iris[,5])
# 결과
##              setosa versicolor virginica
##   setosa         50          0         0
##   versicolor      0         47         3
##   virginica       0          3        47
```

예제 2 {klaR} 패키지 사용
데이터 불러오기, 스팸 데이터 사용
# {klaR} 패키지의 NaiveBayes()함수를 이용하여 단순 베이즈 분류를 수행한다. (대문자)
# {klaR} 패키지는 분류 및 시각화를 위한 다양한 함수를 제공한다.
install.packages("https://cran.r-project.org/src/contrib/Archive/ElemStatLearn/ElemStatLearn_2015.6.26.tar.gz", repos = NULL, type = "source")
library(ElemStatLearn)
install.packages("klaR")
library(klaR)
data(spam) # spam데이터자료, 4061개 메일 중 1813개가 스팸메일이며 58개의 변수로 구성
str(spam) # 58번째 변수가 스팸메일의 여부(1: spam, 0: non-spam)이며, 전체의 39.4% 1813개가 스팸메일이다.

전체 자료의 2/3를 훈련용 자료로 하여 분석을 진행
train.ind <- sample(1:nrow(spam), ceiling(nrow(spam)*2/3), replace=FALSE)

{klaR}패키지의 NaiveBayes()함수를 이용하여 단순 베이즈 분류를 수행, (대문자)
nb.res <- NaiveBayes(spam ~ . , data=spam[train.ind,])
par(mfrow=c(2,3))
plot(nb.res) # 그림으로 표시

predict()함수를 이용하여 예측을 실시, 전체 자료의 1/3인 검증용 자료를 이용하여 모형의 정확도를 측정
nb.pred <- predict(nb.res, spam[-train.ind,])
confusion.mat <- table(nb.pred$class, spam[-train.ind,"spam"])
confusion.mat
결과
email spam
email   536   29
spam   393 575

정분류율은 0.7247228%로 나타난다. (536+575)/(536+29+575+393)=0.7247228
sum(diag(confusion.mat))/sum(confusion.mat) # 0.7247228

예제 2_1번 {kernlab} 사용하기
# 스팸 이메일
install.packages("kernlab")
library(kernlab)
library(klaR)
data(spam) # spam데이터자료, 4061개 메일 중 1813개가 스팸메일이며 58개의 변수로 구성
str(spam) # 58번째 변수가 스팸메일의 여부이며, 전체의 39.4%가 스팸메일이다.

전체 데이터의 2/3만 훈련용 데이터 셋 설정
train.ind <- sample(1:nrow(spam), ceiling(nrow(spam)*2/3), replace=FALSE)

{klaR}패키지의 NaiveBayes()함수를 이용하여 단순 베이즈 분류를 수행, (대문자)
nb.res <- NaiveBayes(type ~ ., data=spam[train.ind,])
par(mfrow=c(2,3))
plot(nb.res)

분석에서 제외된 검증용 자료를 이용하여 모형의 정확도 탐색
nb.pred <- predict(nb.res, spam[-train.ind,])
confusion.mat <- table(nb.pred$class, spam[-train.ind,"type"])
confusion.mat
nonspam spam
nonspam   521   45
spam   415 552

 ## 정분류율은 0.6999348%로 나타난다.
sum(diag(confusion.mat))/sum(confusion.mat) # 0.6999348

단순 베이즈 분류는 결측값을 포함하는 자료를 다음과 같이 처리한다.
훈련단계(구축단계) : 속성값-군집 조합에 대한 빈도 계산 시 결측값을 포함하는 케이스가 제외됨
분류단계 : 결측인 속성이 계산과정에서 생략됨

예제 3번
결측값이 있는 데이터 사용. 16개 안건에 대한 의원들의 의견 데이터
# 단순 베이즈 분류에서 결측값에 대한 처리가 매우 유연하게 이루어진다.
# 모형구축에서는 결측값을 포함하는 케이스를 제외하며 분류과정에서는 결측 속성에 대한 확률만 계산에서 제외되므로 수행과정에 문제가 없다.
install.packages("mlbench")
library(mlbench)
data (HouseVotes84) # 데이터 선언
head(HouseVotes84) # 데이터 해드 확인
summary(HouseVotes84) # 데이터 요약, 17개 변수(class, v1(결측 12개), v2(결측 48개) 등등)

{e1071} 패키지의 naiveBayes()함수를 이용하여 단순 베이즈분석 실시
library (e1071)
model <- naiveBayes(Class ~ . , data = HouseVotes84) # 단순 베이즈분석을 실시하여 분류 모델 생성

predict를 사용하여 새로운 데이터에 대해 예측을 진행한다. (공화 민주 둘 중 하나 예측)
pred <- predict(model, HouseVotes84[,-1]) # class(첫번째 열)를 뺀 원본 데이터로 베이즈분류 모델로 분류를 시행
tab <- table(pred, HouseVotes84$Class) # 분석 결과를 table로 정리하여 대입, class는 공화 민주 둘 중 하나로 나타난다.
tab # 결과
```
## pred           democrat republican
##   democrat          238         13
##   republican         29        155
```

정분류율 : (2389+155) / (238+13+29+155) = 0.90344 = 90.3%정도 잘 분류 하더라
sum(tab[row(tab)==col(tab)])/sum(tab)