

## 기술 통계량 개요

기술 통계(Descriptive Statistics)란 자료를 요약해주는 기본적인 통계량을 의미한다.

기술 통계량의 사용 목적

전체적인 데이터의 개략적인 분포와 통계적 수치를 제공한다.  
모집단의 특성을 유추하는데 사용될 수 있다.

명목 척도를 대상으로 하는 빈도 분석과 비율/등간 척도에 많이 사용 되는 기술 통계 분석 등이 있다.

### 빈도 분석

빈도 분석(Frequency Analysis)은 설문 조사 결과에 대한 가장 기초적인 정보를 제공해 주는 분석 방법이다.  
특히 성별이나 직급을 수치화하는 명목 척도나 서열 척도 같은 범주형 데이터를 대상으로 비율을 측정하는데 주로 이용된다.

빈도 분석

범주형 데이터에 대한 비율을 알아 보고자 하는 경우에 사용한다.  
table() 함수를 이용한다.

다음은 빈도 분석의 응용 사례이다.

빈도 분석의 응용 사례

특정 후보의 지지율은 50%이다.  
응답자 중에서 남자는 30%, 여자는 70%이다.  
연령대 별로 차지하는 비율 구하기

### 기술 통계 분석

각 척도에 따른 기술 통계량 분석은 다음 표를 이용하여 분석하면 된다.

비율 척도인 경우 빈도 수를 구하려면 코딩 변경을 해서 범주화 척도로 변경을 해야 한다.  
예를 들어서 "나이 → 연령대", "점수 → 학점" 등으로 변경해야 한다.

항목	명목	서열	등간(설문 조사)	비율
기술 통계량(summary)	X	X	0	0
빈도 수( table )	0	0	0	△(범주화 되면)
비율( prop.table )	0	0	0	△(범주화 되면)
그래프	막대	막대	막대, 파이	히스토그램, 산점도

### 기술 통계량 보고서 작성

빈도 분석과 기술 통계 분석을 통해서 구해진 기초 통계량 정보를 제시하기 위해서 다음과 같이 표본의 통계적 특성 결과표를 작성한다.

변수		빈도수	구성 비율(%)	
성별	남자	1,500	50	
	여자	1,500	50	
나이	청년층	2,065	68.83	
	중년층	935	31.17	
광고 유형	연예인 CF	1,489	49.63	
	일반인 CF	1,511	50.37	
관심도	관심 있음	2,460	82.00	
	관심 없음	540	18.00	

정리해야 할 패키지 목록

prettyR 패키지

Hmisc 패키지 : 홍보 이벤트 효과 분석.R 파일 참조 요망

# 'doBy' 패키지 설치

```
install.packages('doBy')
```

```
library(doBy)
```

# 성별로 그룹핑하여 brandA 에 대한 평균을 구해준다.

```
mytttest01 <- summaryBy(brandA ~ gender, ttest01)
```

```
mytttest01
```

```
# gender brandA.mean
```

```
# 1 남자 2.95
```

```
# 2 여자 3.20
```

### Hmisc 패키지

전체 데이터 셋에 포함된 모든 변수들을 대상으로 **기술 통계량**을 제공하며, 빈도와 비율 데이터를 일괄적으로 제공해준다.

describe() 함수를 자주 사용한다.

특히 데이터 내 결측치(NA)의 존재 및 서로 다른 값(unique)의 수를 알려주는 점이 편리하다.

describe() 함수는 변수의 척도에 따라서 서로 다른 통계량을 제공해준다.

항목	설명
명목, 서열, 등간 척도	빈도수와 비율 등을 제공한다.
비율 척도	mean, lowest, highest 등의 통계량을

	제공해준다.
--	--------

summary() 함수를 이용하여 기술 통계량을 구할 수 있지만, describe() 함수를 사용하면 좀더 유용한 통계량 등을 얻을 수 있다.

---

### prettyR 패키지

Hmisc 패키지에서 제공하는 describe() 함수와 유사한 freq( ) 함수를 제공한다.

freq( ) 함수는 변수별 빈도수, 결측치 비율을 제공해준다.

비율은 소수점까지 제공하며, 각 명목 척도의 변수를 대상으로 NA 의 비율까지 제공한다.

---

## 교차 분석과 카이 제곱 검정

교차 분석은 두 개 이상의 범주형 변수를 대상으로 교차 분할표를 작성하고, 변수 상호 간의 관련성 여부를 분석하는 방법이다. 교차 분석은 특히 빈도 분석 결과에 대한 보충 자료를 제시하는데 효과적으로 이용할 수 있다.

또한 카이 제곱 검정은 교차 분석으로 얻어진 교차 분할표를 대상으로 유의 확률(p-value)을 적용하여 변수들 간의 독립성 및 관련성 여부 등을 검정하는 분석 방법이다.

---

### 교차 분할표

교차 분할표를 만들기 위해서는 table() 함수를 사용하면 된다.

반환 값은 'table' 인스턴스의 분할표이다.

분할 표에는 값이 배열 형태로 저장되어 있다.

항목	설명
사용 형식	table(x, useNA="always") # useNA="always" : NA 도 포함하여 보여 주기

x

Factor 로 해석할 수 있는 하나 이상의 객체를 의미한다.(예시 : data.frame)

### 교차 분할 표 예시

#### 사용 예시

```
# table 함수를 이용하여 교차 분할표를 생성한다.
table(result) # 부모의 학력과 자녀의 대학 합격 여부에 대한 교차 분할표
#           Pass
# Level      fail pass
# 고졸       40   49
# 대졸       27   55
# 대학원졸   23   31
```

### 교차 분할표 관련 패키지

gmodels 패키지는 교차 분할표를 좀더 세밀하게 보여주는 함수가 있다.

#### 패키지 설치

```
from install.packages("gmodels")
library(gmodels)
```

### CrossTable() 함수 사용하기

교차 분석을 위하여 사용할 수 있는 함수이다.

chisq 옵션을 이용하여 카이 제곱 검정을 수행할 수 있다.

항목	설명
사용 형식	CrossTable(x, y, prop.chisq = FALSE, prop.t = FALSE, prop.r = FALSE, dnn = c('predicted', 'actual'))
x, y	교차하고자 하는 데이터를 지정한다. 예를 들어 분류 문제에서 x에는 정답을 의미하는 label 을, y에는 예측 값을 의미하는 prediction 을 지정하면 된다.
chisq	카이 제곱 검정을 수행하고자 하는 경우에는 True 을 사용하면 된다.
prop.t	True 이면 테이블 비율(proportion table)을 포함한다.
prop.r	True이면 행 비율(proportion row)을 포함한다.
dnn	dimnames names 을 사용자 정의 형식으로 부여하고 자 할 때 사용한다.

## CrossTable() 함수 사용 예시

	자녀의 대학 진학 여부			
부모 학력	실패	합격	Row Total	
고졸	40	49	89	① 관측치
	0.544	0.363		② 기대 비율
	0.449	⑥ 0.551	③ 0.396	행비율
	0.444	④ 0.363		열비율
	0.178	⑤ 0.218		셀비율
대졸	27	55	82	관측치
	1.026	0.684		기대 비율
	0.329	0.671	0.364	행비율
	0.3	0.407		열비율
	0.12	0.244		셀비율
대학원졸	23	31	54	관측치
	0.091	0.06		기대 비율
	0.426	0.574	0.24	행비율
	0.256	0.23		열비율
	0.102	0.138		셀비율
Column Total	90	135	225	전체 관측치
	0.4	0.6		열비율

## 교차 분할표 범례

① 관측된 데이터 수 = 40 + 49 = 89

② 기대 값 및 기대 비율

부모 학력이 고졸이고, 자녀가 합격인 경우

기대값 = (고졸 소계)\*(합격 소계)/(총계) = 89\*135/225 = 53.4

기대 비율 =  $\sum (\text{관측값} - \text{기대값})^2 / \text{기대값} = (49 - 53.4)^2 / 53.4 = 0.362546816$

부모 학력이 대졸이고, 자녀가 실패인 경우

기대값 = (대졸 소계)\*(실패 소계)/(총계) = 82\*90/225 = 32.8

기대 비율 =  $\sum (\text{관측값} - \text{기대값})^2 / \text{기대값} = (27 - 32.8)^2 / 32.8 = 1.025609756$

③ 현재 행의 비율 = 89 / ( 89 + 82 + 54 )

④ 현재 열의 비율 = 49 / ( 49 + 55 + 31 )

⑤ 전체 비율에서 현재 셀이 차지하는 비율 = 49 / ( 40 + 49 + 27 + 55 + 23 + 31 )

⑥ 고졸 합격율

**<논문/보고서에서 교차 분할표 해석>**

학력 수준에 상관없이 대학 진학 합격률이 평균 60.0%로 학력 수준별로 유사한 결과가 나타났다.

전체 응답자 225명을 대상으로 고졸 39.6%(89명) 중 55.1%가 진학에 성공하였고, 대졸 36.4%(82명) 중 68.4%가 성공했으며, 대학원졸은 24%(54명) 중 57.4%가 대학 진학에 성공하였다.

특히 대졸 부모의 대학 진학 합격율이 평균보다 조금 높고, 고졸 부모의 대학 진학 합격율이 평균보다 조금 낮은 것으로 분석이 된다.

**교차 분석(카이 제곱 검정)**

카이 제곱을 설명하기 전에 다음 항목을 살펴 보자.

주사위를 60 번 던졌는데, 다음과 같이 관측이 되었다고 가정하자.

눈금 5 의 기대치는 10 번인데 이번 테스트에서는 8 번이 나왔다.

주사위 눈금	1	2	3	4	5	6
관측도수	4	6	17	16	8	9
기대도수	10	10	10	10	10	10

이와 같이 범주별로 관측 빈도와 기대 빈도의 차이를 통하여 확률 모형이 데이터를 얼마나 잘 설명해줄 수 있는지를 검정하는 통계적 방법을 카이 제곱 검정이라고 한다.

범주형 자료를 대상으로 변수들에 대한 관련성을 알아 보기 위하여 교차 분할 표를 만든다.

이와 같이 범주형 자료를 대상으로 2 개 이상의 변수들에 대한 상호 관련 여부를 분석하는 방법이다.

교차 분석은 카이 제곱 검정 통계량을 사용하므로 카이 제곱 검정이라고 한다.

항목	설명
특징	빈도 분석의 특성별 차이를 분석하기 위해 수행하는 방법이다. 빈도 분석 결과에 대한 보충 자료를 제시하는 데 효과적이다. 참조 문서 : 기술 통계량, 척도의 분류
고려 사항	변수는 10 미만인 범주형 변수(명목, 서열 척도)이어야 한다. 비율 척도는 코딩 변경(리코딩)을 이용하여 범주형 자료로 변경해야 한다. 예시) 10~19(10 대), 20~29(20 대), 30~39(30 대) 등등
변수 모델링	분석할 속성(변수)을 선택하여 속성 간의 관계를 설정하는 작업이다. 사용 예시) 교육 수준(education, 독립 변수)이 흡연률(smoking, 종속 변수)에 연관성이 있는 가를 분석해본다. education → smoking 의 형식으로 표기 한다.

### 카이 제곱의 유형

카이 제곱 검정 유형은 교차 분할표를 사용하느냐에 따라 크게 일원 카이와 이원 카이 제곱 검정으로 분류가 된다.

	변수	교차 분할표 사용 여부	주용도
일원	1 개	no	적합도 검정, 선호도 분석
이원	2 개	yes	독립성 검정, 동질성 검정

#### (1) 일원 카이 제곱 검정

교차 분할표를 이용하지 않는 카이 제곱 검정으로 한 개의 변인(집단 또는 범주)을 대상으로 검정을 수행한다.  
관찰 도수가 기대 도수와 일치하는지를 검정하는 적합도 검정(test for goodness of fit)이 여기에 속한다.

항목	설명
적합도 검정	관측치와 기대치가 일치하는 지를 조사하는 것을 의미한다. 예) 주사위의 눈금의 확률은 각각 1/6 이다.(게임에 적합한가?)
선호도 분석	관측치와 기대치가 일치하는 지를 조사하는 것을 의미한다. 적합도와 차이점은 필요한 연구 환경과 자료이다. 예) 스포츠 음료에 대한 선호도에 차이가 없다.

#### (2) 이원 카이 제곱

교차 분할표를 이용하는 카이 제곱 검정으로 두 개 이상의 변인(집단 또는 범주) 대상으로 검정을 수행한다.  
분석 대상의 집단 수에 의해서 독립성 검정과 동질성 검정으로 나누어진다.

항목	설명
독립성 검정	관련성 검정이라고도 한다. 두 변수를 대상으로 관련성이 있는가? 없는가?를 검정하는 방법이다. 예시 ) 경제력과 대학 진학률과 관련이 있는가? 예시 ) 흡연과 폐암은 연관성이 있는가? 예시 ) 회사에서 나이와 직위가 연관성이 있는가?
동질성 검정	두 집단의 분포가 동일한 모집단에 추출된 것인지를 검정하는 방법이다. 즉, 동일한 분포를 갖는 모집단에서 추출된 것인지를 검정하는 방법이다. 예) 직업의 유형에 따라서 만족도에 차이가 있다.

### 카이 제곱 함수

변수간의 독립성 검정에는 카이 제곱 검정을 수행한다.

2 개의 데이터 간에 의존 관계가 있는 지를 검정하는 수단이다.

항목	설명
사용 형식	chisq.test(x, y=NULL, p)
x	숫자 벡터 또는 행렬, 또는 x, y 모두 Factor
y	숫자 벡터 또는 x 가 Factor 인 경우 Factor 으로 지정되어야 한다. x 와 같은 길이를 가질 확률, 이 값이 지정되어 있지 않으면 확률이 서로 같은지 테스트한다.
p	이 매개 변수가 입력이 되지 않으면 모든 확률이 동일하다고 가정한다. rep(1/length(x), length(x)) x 가 5 개라면 rep(0.2, 5) 즉, 5 개가 모두 확률이 0.2 라는 의미이다.

### 카이 제곱 함수의 결과 정보

카이 제곱 함수의 결과 정보는 다음과 같은 항목들이 존재한다.

항목	설명
X-squared	검정 통계량 $\chi^2 = \sum (\text{관측값} - \text{기대값})^2 / \text{기대값}$ <p>"카이 제곱 검정 통계량 &gt;= 분포표 결과 값"이면 귀무 가설을 기각한다.  "카이 제곱 검정 통계량"은 카이 제곱 검정 표를 참고하길 바란다.</p>
df	자유도(degree of freedom) 일원 카이 제곱에서는 N-1 개이다. 교차 분할표에서는 자유도(df) = (행수-1) * (열수-1)로 구해진다.
p-value	유의 확률(p-value) > 0.05 이면 귀무 가설을 채택하면 된다.

### 카이 제곱 검정 통계량 구하기 예시

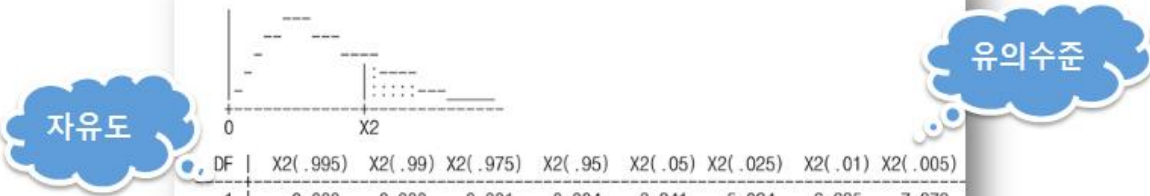
주사위 눈금	1	2	3	4	5	6
관측도수	4	6	17	16	8	9
기대도수	10	10	10	10	10	10
(기대치-관측치)^2	36	16	49	36	4	1
(기대치-관측치)^2/기대치	3.6	1.6	4.9	3.6	0.4	0.1
검정 통계량	14.2 = 3.6 + 1.6 + 4.9 + 3.6 + 0.4 + 0.1					
자유도	일원 카이 제곱 검정이므로 N-1=6-1=5 이다.					



## 카이 제곱 검정 통계표

참조 문서 : <https://math7.tistory.com/58>

CHI-SQUARE TABLE: VALUES OF CHI-SQUARE (ALPHA) OF THE CHI-SQUARE DISTRIBUTION



DF	X2(.995)	X2(.99)	X2(.975)	X2(.95)	X2(.90)	X2(.85)	X2(.80)	X2(.75)
1	0.000	0.000	0.001	0.004	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928

## 교차 분석과 검정 보고서 작성

카이 제곱 검정은 교차 분석으로 얻어진 교차 분할표를 대상으로 유의 확률을 적용하여 변수들 간의 독립성(관련성) 여부를 검정하기 때문에 논문이나 보고서에서는 다음과 같이 교차 분할표와 카이 제곱 검정 통계량을 함께 제시한다.

학력 수준		실패	합계	X-squared	유의 확률(p)
고졸	관찰 빈도(%)	40(44.9%)	49(55.1%)	2.766951	0.2507057
	기대 빈도	35	54		
대졸	관찰 빈도(%)	27(32.9%)	55(67.1%)		
	기대 빈도	30	52		
대학원졸	관찰 빈도(%)	23(42.6%)	31(57.4%)		
	기대 빈도	25	29		

---

### 집단간 차이 분석

집단 간 차이 분석은 모집단에서 추출한 표본의 정보를 이용하여 모집단의 다양한 특성을 과학적으로 추론하는 학문이다.

한 집단과 기존 집단과 **비율**과 **평균**에 대한 차이 검정에 대하여 알아 본다.

즉, 1개 또는 2개, 그 이상의 데이터에 차이가 있다고 볼 수 있는지를 검증하는 절차라고 보면 된다.

- 비율 검정
- T-test(T 검정)
- Anova(3 개 이상의 비교 검정) = 분산 분석 = F 검정

---

### 검정 대상

표본을 이용하여 모수를 검정하는 방법에는 비율과 평균을 검정하는 방법이 있다.

검정 방법	설명
비율	기술 통계량으로 <b>빈도 수에 대한 비율</b> 을 검정한다.
평균	<b>표본 평균</b> 에 의미를 두고 있다.

### 비교 대상에 다른 표본

비교 하는 집단이 한 집단인가 서로 다른 집단인가에 따라서 독립 표본과 대응 표본으로 나누어 진다

표본	설명
독립 표본	서로 <b>다른 두 집단</b> 의 비교를 말한다. 각 그룹에서의 표본이 상대 그룹에서의 표본과 아무런 상관 관계가 없는 표본을 말한다. 예를 들어서, 남자의 커피에 대한 만족도는 여자의 그것과 완전 독립적이다.
대응 표본	<b>한 집단의 이전과 이후에 대한 비교</b> 를 말한다. 한 쌍으로 되어 있기 때문에 paired sample 이라고 한다. 동일한 표본을 대상으로 측정된 두 변수 사이의 평균 차이를 검정하는 방법이다.

동질성 검정이란 **두 집단의 분포가 동일한 분포를 갖는 모집단에서 추출된 것인지를 검정하는 방법**이다.

동질성 검정의 귀무 가설 :
두 집단간 분포의 모양이 동질적이다. p-value 을 유의 확률이라고 하고, alpha 를 유의 수준이라고 한다. <b>p-value &gt; alpha 이면 귀무 가설을 채택</b> 한다.

### 집단 수에 따른 검정 방법

집단의 수와 검정 대상에 따른 함수는 다음과 같은 항목을 사용하면 된다.

집단	검정 대상	관련 함수	동질성 검정	정규 분포 검정
단일 집단	비율	binom.test()	-	-
단일 집단	평균	t.test()	-	shapiro.test()
두 집단	비율	prop.test()	-	-
두 집단	평균	t.test()	var.test()	-
세 집단	비율	prop.test()	-	-
세 집단	평균	aov()	bartlett.test()	-

## 단일 집단 비율 검정

단일 집단의 비율이 어떤 특정한 값과 같은 지를 검정하는 방법이다.

비율을 바탕으로 `binom.test()` 함수를 이용하여 이항 분포의 양측 검정을 통해서 검정 통계량을 구한 후, 이를 이용하여 가설을 검정한다.

예시 :

2017년도의 고객 불만율과 2018년도 고객 응대 교육 수료 후에 불만율에 차이가 없다.

## 구현 흐름도

내용	Flow Chart
<p>분석할 데이터를 대상으로 결측치와 이상치를 제거하는 전처리 과정을 우선 실행한다</p> <p>기술 통계량으로 빈도 분석을 계산한다.</p> <p>이 결과를 <code>binom.test()</code> 함수의 인수로 사용하여 비율 차이에 대한 검정을 수행한다.</p> <p>비율 차이 검정 통계량을 바탕으로 귀무 가설의 기각 여부를 결정한다.</p>	<pre> graph TD     A[실습 파일 읽어 오기] --&gt; B[데이터 전처리]     B --&gt; C[기술 통계량(빈도 분석)]     C --&gt; D[binom.test()]     D --&gt; E[검정 통계량 분석]           </pre>

## `binom.test()` 함수

명목 척도의 비율을 바탕으로 `binom.test()` 함수를 이용하여 이항 분포의 양측 검정을 통해서 검정 통계량을 구한 후 이를 이용하여 가설을 검정한다.

즉, 특정 변수의 선택 항목이 2개 중 1개일 때, 선택 비율이 동일한지를 검정하고자 할 때 사용하는 분석 함수이다.

항목	설명
사용 형식	<code>binom.test( x, n, p=0.5, alternative, conf.level )</code>
x	성공의 수, 또는 성공과 실패 수를 각각 저장한 길이가 2 인 벡터
n	시행 횟수, x 의 길이가 2 이라면 무시된다.
p	성공 확률에 대한 가설이다.
alternative	대립 가설의 형태를 의미하는 데, 기본 값은 양측 검정이다.

	alternative = c("two.sided"(default), "less", "greater
conf.level	신뢰 수준(기본 값 : 0.95)

이항 분포 비율 검정 예시 :

150명의 고객에 대하여 만족(136), 불만족(14)인 데이터가 있다.  
136명의 만족 고객이 전체의 80% 이상의 만족율을 나타내는 지를 검정하세요.

```
binom.test(c(136, 14), p=0.8)
Exact binomial test
```

```
data: c(136, 14)
number of successes( 만족 ) = 136, number of trials( 전체 도수 ) = 150, p-value = 0.0006735
alternative hypothesis: true probability of success is not equal to 0.8
95 percent confidence interval: ( 신뢰 수준 )
0.8483615 0.9480298 ( 신뢰 구간 )
sample estimates:
probability of success ( 136/150의 연산 결과 )
0.9066667
```

풀이 하기  
유의 확률 p-value(0.0006735) < 0.05이므로 귀무 가설을 기각한다.  
즉, '작년과 올해의 불만율에 차이가 있다'라고 말할 수 있다.

신뢰 수준 95%에서 귀무 가설은 다음과 같다.

귀무 가설 : 2017년도의 고객 불만율과 2018년도 CS 교육 후에 불만율에 차이가 없다.

검정 대상	검정 방법	단측 비교 방법	유의 확률	귀무 가설	결론
만족율 비율 검정	양측 검정	-	0.0006735	기각	불만율에 차이가 있다.
만족율 비율 검정	단측 검정	만족율 > 80%	0.0003179	기각	만족율이 늘었다.
불만족율 비율 검정	양측 검정	-	0.0006735	기각	불만율에 차이가 있다.
불만족율 비율 검정	단측 검정	불만족율 > 20%	0.9999	채택	불만족이 20%보다 크지 않다.
불만족율 비율 검정	단측 검정	불만족율 < 20%	0.0003179	기각	불만족율이 줄었다.

## 단일 집단 평균 검정

단일 집단의 평균이 특정한 집단의 평균과 차이가 있는지를 검정하는 방법이다.  
 일표본 평균의 구간 추정 및 가설 검증에 사용하는 함수이다.

예시 :

국내에서 생산된 노트북과 A회사에서 생산된 노트북의 평균 사용 시간에 차이가 없다.

## 구현 흐름도

내용	Flow Chart
<p>분석할 데이터를 대상으로 전처리를 수행한다.                      평균 차이 검정을 위해서 기술 통계량으로 평균을 구한다.                      평균에 대한 차이 검정은 정규 분포 여부를 우선 판정해야 한다.</p> <p>정규 분포 여부는 shapiro.test() 분석 함수를 이용하면 된다.                      정규 분포라고 판정이 나면 T 검정을 수행하고, 반대이면 웰콕스(wilcox) 검정을 수행하면 된다.</p> <p>해당 조건에 맞는 평균 차이 검정을 실시하여 귀무 가설의 기각 여부를 결정하도록 한다.</p>	<pre> graph TD     A[실습 파일 읽어 오기] --&gt; B[데이터 전처리]     B --&gt; C[기술 통계량(평균)]     C --&gt; D{정규 분포}     D --&gt; E[shapiro.test()]     E --&gt; D     D --&gt; F[t.test()]     D --&gt; G[wilcox.test()]     F --&gt; H[검정 통계량 분석]     G --&gt; H                     </pre>

## shapiro.test() 함수

샤피로-윌크 검정이라고 하며, 어떠한 데이터의 분포가 정규 분포인지를 검정하는 방법이다.  
 95%의 신뢰 수준이라고 할 때,  $p\text{-value} > 0.05$ 이면 귀무 가설을 채택한다.

항목	설명
사용 형식	shapiro.test( x )
함수의 기본 귀무 가설	정규 분포를 따르고 있다.
x	숫자 형식의 vector이다.

## 정규 분포 시각화 관련 함수

정규 분포의 시각화는 다음 함수를 이용하면 확인 가능하다.

항목	설명
hist()	히스토그램을 그려 준다.
qqnorm()	주어진 데이터와 정규 분포를 비교한다.
qqline()	데이터와 분포를 비교하여 이론적으로 성립해야 하는 직선 관계를 그려 준다.

## t.test() 함수

stats 패키지에서 제공하는 함수이다.

단일 집단의 평균이 특정한 집단의 평균과 차이가 있는지를 검정하는 방법으로써, t 검정이라고 한다.

일표본 평균의 구간 추정 및 가설 검증에 사용하는 함수이다.

항목	설명
사용 형식	t.test(x, y = NULL, alternative, mu = 0, conf.level)
함수의 기본 귀무 가설	모 평균이 mu와 같다
x	일표본 t 검정의 경우에는 x에만, 이표본 t 검정의 경우에는 x, y에 모두 숫자 벡터를 지정한다.
y	이 표본일때의 숫자 벡터를 의미한다.
alternative	검정 방법, 대립 가설의 형태를 의미하는 데, 기본 값은 양측 검정이다. 'two.sided', 'greater', 'less'
mu	모집단의 평균을 의미한다.
conf.level	신뢰 구간( 기본 값 : 0.95 )
paired	False 이면 독립 이표본 검정이라는 의미이다.
var.equal	-

### t 검정 예시 :

```
t.test(x1, mu=5.2)
```

One Sample t-test

data: x1

t = 3.9461, df = 108, p-value = 7.083e-05 <== t 검정 통계량, 자유도, 유의 확률

alternative hypothesis: true mean is not equal to 5.2

95 percent confidence interval: <== 95% 신뢰 수준

5.377613 5.736148 <== 95%에서의 신뢰 구간(여기에 5.556881 가 들어 있다.)

sample estimates:

mean of x

5.556881 <== 평균에 대한 점 추정 값

## qt() 함수

stat() 패키지에서 제공하는 qt() 함수를 사용하면, 귀무 가설의 임계값을 확인할 수 있다.

항목	설명
사용 형식	유의 확률(p-value)와 자유도(df)의 값을 입력하면 임계 값을 구할 수 있다. qt( p-value, df ) = qt(7.083e-05, 108 ) = -3.946073
귀무 가설 채택 조건	"임계값 >= t 검정통계량"인 경우에는 귀무 가설을 채택한다. 위의 예시에서 3.946073 < 3.9461 이므로 귀무 가설을 기각한다.

## 단일 집단 T 검정 결과 정리 및 기술

논문이나 보고서에 단일 표본 평균 검정 결과를 제시하기 위해서는 다음과 같은 형식으로 일목 요연하게 기술하는 것이 좋다.

항목	설명
가설 설정	귀무 가설(H0) 국내에서 생산된 노트북과 A 회사에서 생산된 노트북의 평균 사용 시간에 차이가 없다.
연구 환경	국내에서 생산된 노트북 평균 사용 시간이 5.2 시간으로 파악된 상황에서 A 회사에서 생산된 노트북 평균 사용 시간과 차이가 있는 지를 검정하기 위해서 A 회사 노트북 150 대를 랜덤으로 선정하여 검정을 실시한다.
유의 수준	$\alpha = 0.05$
분석 방법	단일 표본 T 검정
검정 통계량	$t = 3.94606$ , $df = 108$
유의 확률	$p = 7.083e-05$
결과 해석	유의 수준 0.05 에서 귀무 가설이 기각되었다. 따라서, 국내에서 생산된 노트북과 A 회사에서 생산된 노트북의 평균 사용 시간에 차이를 보인다고 할 수 있다. 즉, 국내에서 생산된 노트북의 평균 사용 시간은 5.2 이며, A 회사에서 생산된 노트북의 평균 사용 시간은 5.556881 으로 국내 평균 사용 시간 보다 더 길다고 할 수 있다.

## 두 집단 비율 검정

두 집단의 비율이 같은지/다른지를 검증하는 것을 두 집단 비율 차이 분석이라고 한다.

분석할 데이터를 대상으로 결측치와 이상치를 제거하는 전처리 과정을 거친 후, 비교 대상의 두 집단을 분류하고, 이를 prop.test() 함수의 인수로 비율 차이 검정을 수행한다.

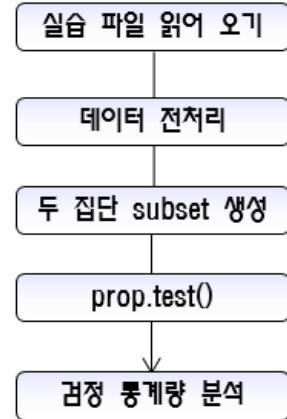
예시 :
교육 기관의 두 가지 교육 방법에 대한 교육생의 만족도에 차이가 없다.

## 구현 흐름도

내용	Flow Chart
----	------------



분석할 데이터를 대상으로 결측치와 이상치를 제거하는 전처리 과정을 우선 실행한다.  
비교 대상의 두 집단을 분류한다.  
이 결과를 prop.test() 함수의 인수로 사용하여 비율 차이에 대한 검정을 수행한다.  
비율 차이 검정 통계량을 바탕으로 귀무 가설의 기각 여부를 결정한다.



### prop.test() 함수

두 집단에 대한 비율에 대한 가설 검정을 수행한다.  
비율에 즉, 2개의 데이터 사이에 비율의 차이가 있는 지에 대한 검정을 수행한다.

항목	설명
사용 형식	prop.test(x=c(110,135), n=c(150,150), alternative, conf.level, correct=TRUE ) 두 교육의 중 빈도수가 150이고, A는 110회/B는 135인 비율 검정을 수행하시오.  prop.test( x=56, n=100, p=0.5, alternative='two.sided' )
함수의 기본 귀무 가설	'두 그룹의 비율이 같다' 또는 '비율이 p와 같다'이다.
x	성공 회수를 저장한 벡터 또는 성공과 실패 수를 저장한 1*2 또는 2*2 표(행렬)
n	전체 시행수를 의미한다.
p	성공 확률(비율)에 대한 가설이다.
alternative	대립 가설의 형태이고, 기본 값은 양측 검정이다. 'two.sided', 'greater', 'less'
conf.level	신뢰 수준(기본 값 : 0.95)

### prop.test() 함수에 따른 검정 결과 판단

항목	설명
p-value	p-value의 값이 95%의 신뢰 수준에서 0.05보다 크면 귀무 가설을 채택한다. 즉, "두 그룹의 비율이 같다"라고 판단한다.
검정 통계량	X-squared 검정 통계량(변수a) < 카이 제곱 검정 분포표의 값(변수b)이면 귀무 가설을 채택한다. 변수a : prop.test() 함수를 이용하여 구한 값(X-squared 항목 참조)을 의미한다. 변수b : 자유도와 유의 수준 값을 이용하여 카이 제곱 검정 분포표에서 값을 찾는다.

### prop.test() 함수 사용 예시 01

## 사용 예시

```
prop.test(42, 100)
#
#      1-sample proportions test with continuity correction
#
# data: 42 out of 100, null probability 0.5
# X-squared = 2.25, df = 1, p-value = 0.1336 <== 자유도 1
# alternative hypothesis: true p is not equal to 0.5
# 95 percent confidence interval: <== 95% 신뢰 수준
# 0.3233236 0.5228954 <== 95%에서의 신뢰 구간(여기에 0.5 가 들어 있다.)
# sample estimates:
#      p
# 0.42 <== 42/100
```

## prop.test() 함수 사용 예시 02

### 사용 예시

```
# c(34,37,39) : 방법 1, 방법 2, 방법 3 에 대한 만족도 수이다.
# c(50,50,50) : 세가지 교육 방법에 대한 변량의 길이다.

prop.test(c(34,37,39), c(50,50,50))
#      3-sample test for equality of proportions without continuity correction
#
# data: c(34, 37, 39) out of c(50, 50, 50)
# X-squared = 1.2955, df = 2, p-value = 0.5232 <== 자유도 2
# alternative hypothesis: two.sided <== 양측 검정
# sample estimates:
# prop 1 prop 2 prop 3
# 0.68 0.74 0.78 <== 34/50, 37/50, 39/50

유의 확률(p-value) = 0.5232 > 유의 수준(0.05)이므로, 신뢰 수준 95%에서 귀무 가설을 채택한다.
세 가지 교육 방법에 따른 집단 간 만족율에 차이가 없다. 라는 결론을 내릴 수 있다.
```

## 두 집단 평균 검정(독립 표본 T 검정)

독립 표본 T검정이란 서로 다른 두 집단의 비교하는 검정을 말한다.

## 구현 흐름도

내용

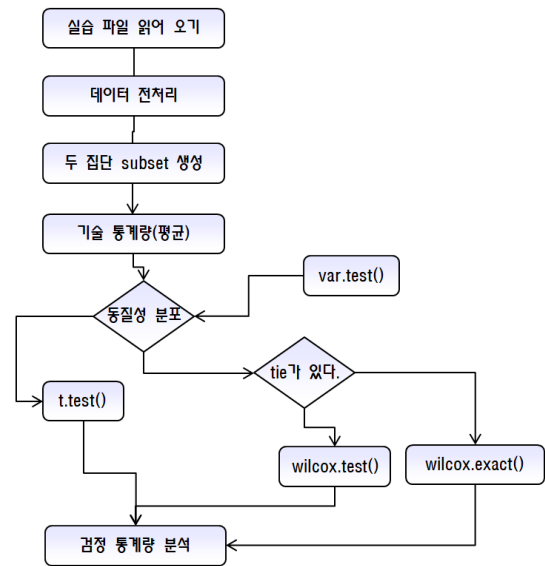
Flow Chart

분석할 데이터를 대상으로 결측치와 이상치를 제거하는 전처리 과정을 우선 실행한다.  
비교 대상의 두 집단을 분류한다.  
평균 차이 검정을 위해서 기술 통계량으로 평균을 구한다.

평균에 대한 차이 검정은 두 집단 간의 동질성 검증(정규 분포 검정)을 우선 수행해야 한다.  
동질성 분포 검정은 `var.test()` 분석 함수를 이용하면 된다.

판정 결과에 따라서 T 검정 또는 웰콕스(wilcox) 검정을 수행하면 된다.

두 검정 방법의 선택은 단일 표본 평균 검정과 동일하다.  
해당 조건에 맞는 평균 차이 검정을 실시하여 귀무 가설의 기각 여부를 결정하도록 한다.



## var.test() 함수

stats 패키지에서 제공하는 `var.test()` 함수를 이용하면 동질성 검사를 수행할 수 있다.

동질성 검정이란 **두 집단의 분포가 동일한 분포를 갖는 모집단에서 추출된 것인지를 검정하는** 방법이다.

즉, 모집단에서 추출된 표본을 대상으로 분산 동질성 검정을 수행해주는 함수이다.

항목	설명
사용 형식	<code>var.test( x=method01_score, y=method02_score )</code>
함수의 기본 귀무 가설	"두 집단간 분포의 모양이 동질적이다." 즉, 동일한 분포에서 나왔다.

## var.test() 함수 사용 예시

### 사용 예시

```
var.test( method01_score, method02_score )
```

F test to compare two variances

data: method01\_score and method02\_score

F = 1.2158, num df = 108, denom df = 117, p-value = 0.3002 <= 유의 확률

```
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval: <== 95% 신뢰 수준
0.8394729 1.7656728 <== 95%에서의 신뢰 구간
sample estimates:
ratio of variances
1.215768
```

검정 통계량은  $p\text{-value}(0.3002) > 0.05$  이므로 귀무 가설을 채택한다.  
 귀무 가설을 채택하므로 두 집단간의 분포 형태가 동일하다고 볼 수 있다.

### 독립 표본 t-검정 결과 정리 및 기술

논문이나 보고서에서 독립 표본 평균 검정 결과를 제시하기 위해서는 다음과 같은 형식으로 기술한다.

항목	설명
가설 설정	귀무 가설 : 교육 방법에 따른 두 집단 간 실기 시험의 평균에 차이가 없다.
	연구 가설 : 교육 방법에 따른 두 집단 간 실기 시험의 평균에 차이가 있다.
연구 환경	IT 교육 센터에서 PT를 이용한 프리젠테이션 교육 방법과 실시간 코딩 교육 방법을 적용하여 1개월 동안 교육받은 교육생 각 150명을 대상으로 실기 시험을 실시하였다. 두 집단간 실기 시험의 평균에 차이가 있는 지 검정한다.
유의 수준	$\alpha = 0.05$
분석 방법	독립 표본 T 검정
검정 통계량	$t = -2.0547, df = 218.19$
유의 확률	$p\text{-value} = 0.0411$
결과 해석	유의 수준 0.05에서 귀무 가설이 기각되었다. 따라서, 교육 방법에 따른 두 집단간 실기 시험의 평균에 차이가 있다. 라고 말할 수 있다. 단측 검정을 실시한 결과 첫 번째 교육 방법이 두 번째 교육 방법보다 크지 않은 것으로 나타났다. 즉, 실시간 코딩 교육 방법이 교육 효과가 더 높은 것으로 분석된다.

### 동질성 검정

분산 분석의 동질성 검정은 stats 패키지에서 제공하는 `bartlett.test()` 함수를 이용한다.

이 함수에 대한 귀무 가설은 "분포가 동질적이다."이다.

유의 확률( $p\text{-value}$ )이 유의 수준( $\alpha$ ) 보다 큰 경우 세 집단 간 분포의 모양이 동질하다고 할 수 있다.

항목	설명
----	----

사용 형식	bartlett.test(종속변수 ~ 독립변수, data=dataset) bartlett.test(score ~ method2, data=mydata2)
-------	--

### bartlett.test() 함수 사용 예시

사용 예시
<pre> bartlett.test(score ~ method2, data=mydata2) #      Bartlett test of homogeneity of variances # # data:  score by method2 # Bartlett's K-squared = 3.3157, df = 2, p-value = 0.1905 &lt;== 유의 확률 </pre>

### 세 집단 평균 검정(분산 분석)

분산 분석(ANOVA Analysis)은 세 집단 이상의 평균에 의한 차이를 검정하는 방법이다.

예를 들어 의학 연구 분야에서 개발된 3가지 치료제가 있다고 가정하자.

이 3가지 치료제의 효과에 차이가 있는지를 검정하는 경우가 분산 분석이다.

가설 검정을 위해 F 분포를 따르는 F 통계량을 검정 통계량으로 사용하기 때문에 F 검정이라고도 한다.

### 구현 흐름도

내용	Flow Chart
<p>분석할 데이터를 대상으로 결측치와 이상치를 제거하는 전처리 과정을 우선 실행한다</p> <p>비교 대상의 세 집단을 분류한다.</p> <p>평균 차이 검정을 위해서 기술 통계량으로 평균을 구한다.</p> <p>분산 분석에서 집단 간의 동질성 여부를 검정하기 위해서는 bartlett.test() 분석 함수를 이용하면 된다.</p> <p>집단 간의 분포가 동질한 경우 분산 분석을 수행하는 aov() 분석 함수를 사용하도록 하며, 그렇지 않은 경우에는 비모수 검정 방법인 kruskal.test() 분석 함수를 이용하여 분석을 수행한다.</p> <p>마지막으로 TukeyHSD() 함수를 이용하여 사후 검정을 수행하도록 한다.</p>	<pre> graph TD     A[실습 파일 읽어 오기] --&gt; B[데이터 전처리]     B --&gt; C[세 집단 subset 생성]     C --&gt; D[기술 통계량(평균)]     D --&gt; E{동질성 분포}     E --&gt; F[bartlett.test()]     E --&gt; G[aov()]     F --&gt; G     G --&gt; H[kruskal.test()]     H --&gt; I[TukeyHSD()] </pre>

### aov() 함수

세 집단의 평균 검정은 aov() 함수를 이용하면 된다.

만약 동질하지 않은 경우에는 kruskal.test() 함수를 이용하여 비모수 검정을 수행한다.

항목	설명
사용 형식	aov(종속변수 ~ 독립변수, data=df_data set)

	result = aov(score ~ method2, data=mydata2)
함수의 기본 귀무 가설	"세 집단의 평균에 차이가 없다."

## aov() 함수 결과 확인

사용 예시	
names(result)	
# aov()의 결과 값은 summary() 함수를 사용하여 유의 확률(p-value)을 확인할 수 있다.	
summary(result)	
	<pre> Df Sum Sq Mean Sq F value Pr(&gt;F) &lt;== F 검정 통계량 유의 확률 method2      2  99.37   49.68   43.58 9.39e-14 *** Residuals    85  96.90    1.14 --- Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.0</pre>
유의 확률이 9.39e-14 이므로 귀무 가설을 기각한다.	
<p>F 검정 통계량으로 가설 검정</p> <p>F 검정 통계량 &lt; 절대 값(Z)이면 귀무 가설을 채택한다.</p> <p>분산 분석에서 신뢰 수준 95%에서는 -1.96 ~ +1.96 의 범위가 귀무 가설의 채택역이다.</p> <p>따라서 F 검정 통계량이 채택역에 해당하지 않으면 귀무 가설을 기각할 수 있다.</p> <p>현재 F 검정 통계량 43.58 은 ±1.96 보다 크기 때문에 귀무 가설을 기각하고, 연구 가설이 채택된다.</p> <p>분산 분석에서 F 검정 통계량과 유의수준 <math>\alpha</math> 의 관계는 다음 표와 같다.</p>	

## 분산 분석에서 F 검정 통계량과 유의 수준 $\alpha$ 의 관계

F값(절대치)	유의 수준(양측 검정시)
F 값(절대치) >= 2.58	$\alpha$ = 0.01(의학/생명 분야)
F 값(절대치) >= 1.96	$\alpha$ = 0.05(사회 과학 분야)
F 값(절대치) >= 1.645	$\alpha$ = 0.1(기타 일반 분야)

## 사후 검정

사후 검정은 분산 분석의 결과에 대하여 구체적으로 어떻게 차이가 나는 지를 보여주는 부분이다.

사용 예시
<pre> TukeyHSD(result) # Tukey multiple comparisons of means</pre>

```
# 95% family-wise confidence level
#
# Fit: aov(formula = score ~ method2, data = mydata2)
#
# $method2
#           diff      lwr      upr    p adj <== 하단 설명 참조 요망
# 방법 2-방법 1  2.612903  1.9424342  3.2833723  0.0000000
# 방법 3-방법 1  1.422903  0.7705979  2.0752085  0.0000040
# 방법 3-방법 2 -1.190000 -1.8656509 -0.5143491  0.0001911

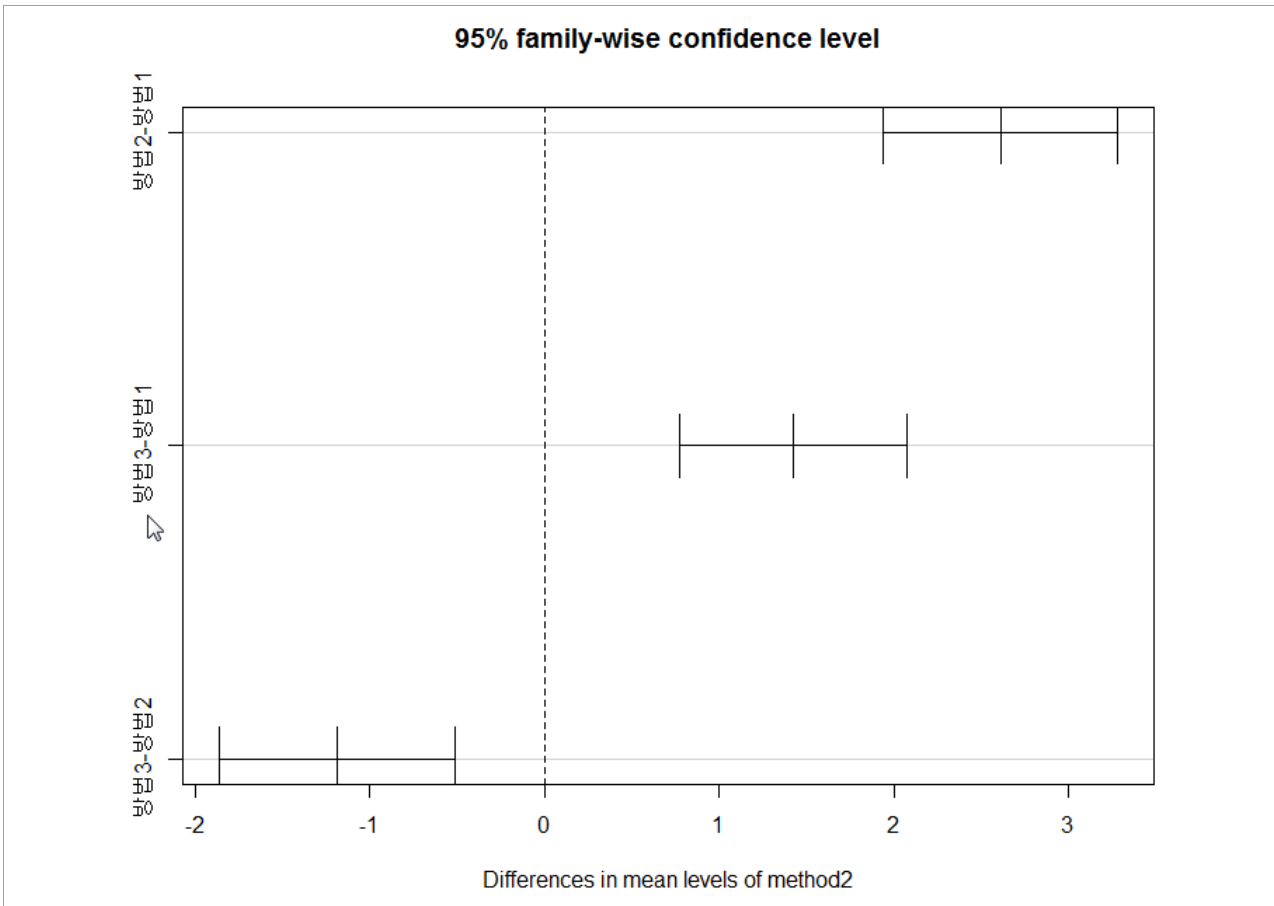
plot(TukeyHSD(result))
```

diff 는 두 집단의 평균의 차이를 말한다.

방법 2 와 방법 1 의 평균 차이가 가장 큰 것으로 나타났다.

lwr 은 신뢰 구간의 하한 값, upr 은 신뢰 구간의 상한 값이다.

p adj 은 다중 비교를 위하여 조절된 p-value 값이다.



## 분산 분석 검정 결과 정리 및 기술

가설 설정	연구가설(H1) : 교육 방법에 따른 세 집단 간 실기 시험의 평균에 차이가 있다.
	귀무가설(H0) : 교육 방법에 따른 세 집단 간 실기 시험의 평균에 차이가 없다.
연구 환경	세 가지 교육 방법을 적용하여 1 개월 동안 교육받은 교육생 각 50 명씩을 대상으로 실기 시험을 실시하였

	다. 세 집단간 실기 시험의 평균에 차이가 있는가 검정한다.
유의 수준	$\alpha = 0.05$
분석 방법	ANOVA 검정
검정 통계량	F = 43.58, Df =2, Sum Sq=99.37, Mean Sq = 49.68
유의 확률	P = 9.39e-14 ***
결과 해석	유의 수준 0.05 에서 귀무 가설이 기각되었다. 따라서 교육 방법에 따른 세 집단 간 실기 시험의 평균에 차이가 있는 것으로 나타났다. 또한 사후 검정을 위한 Tukey 분석을 실시한 결과 "방법 2-방법 1"의 평균 점수의 차이가 가장 높은 것으로 나타났다.

## 상관 관계 분석

상관 관계 분석이란 변수들간의 연관성을 분석하기 위하여 사용하는 방법으로써, **한 개의 변수가 다른 변수와 어느 정도의 관련성이 있는 지**를 개관할 수 있는 분석 기법이다.

"하나의 변수가 다른 변수와 관련성이 있는가?" 또는 "관련성이 있다면 어느 정도 있는 가?" 등등이다.



### 사용 예시

광고의 양과 브랜드의 인지도는 얼마나 연관성이 있는가?

광고비와 매출액은 얼마나 연관성이 있는가?

### 사용처

회귀 분석에서 변수들 간의 인과 관계를 분석하기 전에 선행 자료로 사용될 수 있다.

### 피어슨 상관 계수

상관 계수란 변수들 간의 관련성을 가늠하기 위한 척도이다.

두 변수간의 가장 높은 상관 관계의 상관 계수는 1이다.

두 변수간에 상관 관계가 전혀 없으면 상관 계수는 0이다.

피어슨 상관 계수 R	상관 관계의 정도
$\pm 0.9$ 이상	매우 높은 상관 관계
$\pm 0.9 \sim \pm 0.7$	높은 상관 관계
$\pm 0.7 \sim \pm 0.4$	다소 높은 상관 관계
$\pm 0.4 \sim \pm 0.2$	낮은 상관 관계
$\pm 0.2$ 미만	상관 관계 없음

### 상관 계수 보기

변수 간의 상관 계수는 stats 패키지에서 제공하는 cor() 함수를 이용하며 다음과 같은 형식이다.

### cor() 함수

상관 계수를 구해주는 함수로써, 반환 값은 상관 계수 행렬(matrix) 객체이다.

항목	설명
사용 형식	cor(x, y, use, method)
x	vector, matrix, data frame 등이 사용될 수 있다.
y	기본 값은 NULL 이다.
use	-
method	c('pearson', 'kendall', 'spearman') 기본 값은 pearson 이다. 보편적으로 피어슨 상관 계수를 이용한다.

### 상관 계수 시각화

corrgram() 함수를 사용하면 상관 계수와 상관 계수에 따른 색의 농도로 시각화 해준다.

즉, 상관 계수를 시각적으로 보여 주는 패키지이다.

항목	설명
----	----

사용 형식	<code>corrgram(x, lower.panel = panel, upper.panel = panel)</code>
x	작업을 수행할 데이터
lower.panel	아래쪽 판넬에 상관 계수를 추가한다. <code>lower.panel = panel.conf</code>
upper.panel	위쪽 판넬에 상관 계수를 추가한다. <code>upper.panel = panel.conf</code>

### corrplot 패키지

corrplot 패키지는 상관 행렬(correlation matrix)의 결과와 신뢰 구간(confidence interval)을 그래프로 그려 주는 패키지이다.

또한 행렬의 재정렬(reordering) 알고리즘을 포함하고 있으며 색의 선택, 텍스트 라벨링, 칼라 라벨 등을 포함하고 있다.

참조 사이트

<https://rpubs.com/cardiomoon/27080>

[https://rstudio-pubs-static.s3.amazonaws.com/27134\\_f8052fbae4fe4402824ebb9fe080d876.html](https://rstudio-pubs-static.s3.amazonaws.com/27134_f8052fbae4fe4402824ebb9fe080d876.html)

### corrplot() 함수

항목	설명
사용 형식	<code>corrplot(mcor, method="shade", shade.col=NA, tl.col="black", tl.srt=45)</code>
mcor	matrix 객체(상관 계수 행렬 객체)
method	원 : 디폴트, shade : 네모칸, ellipse : 양수(오른쪽), 음수(왼쪽)으로 뻗어있는 타원
shade.col	-
tl.col	글자 색
tl.srt	변수의 기울기
addCoef.col	내부 색상
addcolorlabel	-
order	순서를 조절할 수 있다.("AOE", "FPC", "hclust")

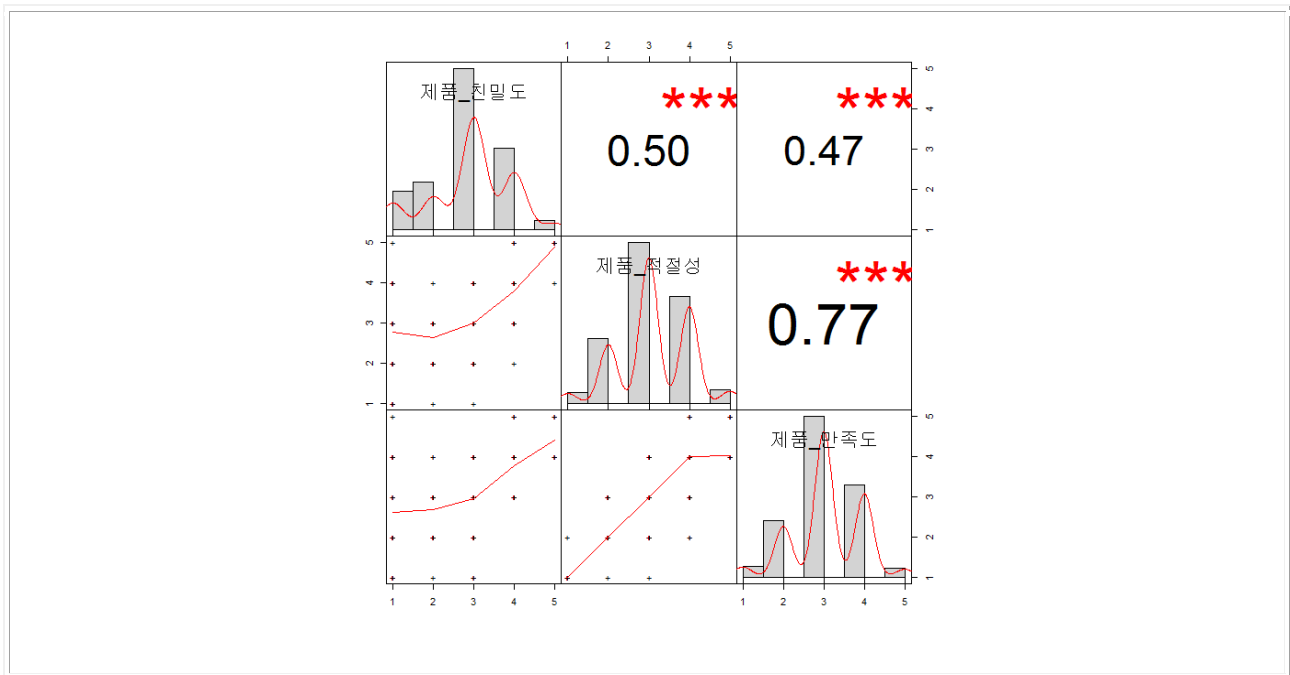
### 밀도 곡선, 상관성, 유의 확률(별표)

차트에 밀도 곡선, 상관성, 유의 확률(별표)를 추가하려면 다음과 같이 작성하면 된다.

#### 사용 예시

```
# 차트에 밀도 곡선, 상관성, 유의 확률(별표) 추가
install.packages('PerformanceAnalytics')
library(PerformanceAnalytics)

# 상관성, p값(*), 정규 분포(모수 검정 조건) 시각화
chart.Correlation( product, histogram=, pch="+")
savePlot('chart.Correlation 실행 결과.png', type="png")
```



### 상관 관계 분석 결과표

일반적으로 상관 관계 분석 결과를 논문이나 보고서에 제시할 경우에는 기본적인 기술 통계량과 피어슨 상관 계수를 함께 제시하는 것이 좋다.

분석 단위	평균(Mean)	표준 편차(Std. Deviation)	분석 단위간 상관 관계 (Inter-Analysis Correlations)		
			1	2	3
1.친밀도	2.928	0.9703446	1		
2.적절성	3.133	0.8596574	0.50***	1	
3.만족도	3.095	0.8287436	0.47***	0.77***	1

### 요인 분석

요인 분석은 다수의 변수들을 대상으로 변수들 간의 관계를 분석하여 **공통 차원으로** 축약하는 통계 기법이다.

데이터를 축소하는 변수의 정제 과정이다.

### 요인 분석 예시

5명의 고객에 대한 "대한 fast-food" 점의 평가 내용을 살펴 보자.

소비자	대기 시간	청결	종업원	음식 맛	음식 온도	음식 신선도
김유신	2	2	1	6	5	5
이순신	1	1	1	4	5	4
강감찬	2	2	2	5	5	5
이율곡	2	1	2	4	6	5
신사임당	1	3	1	6	5	5
분류	서비스 품질			음식 품질		

"대한 fast-food" 점은 대체적으로 "음식의 질은 좋으나 서비스는 별로 좋지 않다"라고 해석할 수 있다.

또 다른 예를 들어 보면, 학생들 100명을 대상으로 국어, 영어, 수학, 물리 등의 시험을 실시하여 성적을 구하였을 때 공통적으로 설명할 수 있는 공통 인자(변수)를 파악하는 것이다.

즉, 언어 능력(국어, 영어), 수리 능력(수학, 물리) 등으로 분리해 내는 것이다.

국어	영어	수학	물리
언어 능력		수리 능력	

### 요인 분석의 전제 조건

#### 요인 분석의 전제 조건 :

하위 요인으로 구성되는 데이터 셋이 준비되어 있어야 한다.

분석에 사용되는 척도는 등간 척도나 비율 척도이어야 한다.

표본의 크기는 최소 50 개 이상이 바람직하다.

상관 관계가 높은 것끼리 그룹화하는 것이므로, 변수들 간의 **상관 관계가 낮다면(보통  $\pm 3$ 이하) 요인 분석에 적합하지 않다.**

### 요인 분석의 목적

요인 분석의 일반적인 목적은 다음과 같다.

항목	설명
자료의 요약	변인(변수)을 몇 개의 공통된 변인으로 묶어 준다.
변인 구조 파악	변인들의 상호 관계를 파악하기 위함이다.(독립성 등등)
불필요한 변인 제거	중요도가 떨어지는 변수는 제거한다.
특정 도구 타당성 검증	변인들이 동일한 요인으로 묶이는 지 여부를 확인한다.

요인 분석의 종류

요인 분석의 종류에는 다음과 같은 항목들이 있다.

항목	설명
탐색적 요인 분석	요인 분석시 사전에 어떤 변수들끼리 묶어야 한다는 전제를 두지 않고 분석하는 방법이다.
확인적 요인 분석	사전에 묶여질 것으로 기대되는 항목끼리 묶여지는지를 조사하는 방법이다.

요인 분석의 결과에 대한 활용 방안

요인 분석의 결과에 대한 활용 방안은 다음과 같은 것들이 있다.

항목	설명
타당성 검증	측정 도구가 정확히 측정되었는지를 알아 보기 위해서 측정 변수들이 동일한 요인으로 묶이는지를 검증한다.
변수 축소	변수들의 상관 관계가 높은 것끼리 묶어서 변수를 정제시킨다.
변수 제거	변수의 중요도를 나타내는 요인 적재량이 0.4 미만이면 설명력이 부족한 요인으로 판단하여 제거한다.
설명(독립) 변수 활용	요인 분석에서 얻어 지는 결과를 이용하여 상관 분석이나, 회귀 분석의 설명 변수로 활용한다.

주요 용어

요인 분석을 위하여 사용되는 용어들을 간단히 정리해본다.

항목	설명
주성분	Principal component 분산(변동량)에 영향을 많이 주는 성분을 주요 성분이라고 한다.
변수(variable)	분석에 사용하고자 하는 내용이 있는 데이터를 의미한다. 변수 : s1~s6 를 저장하고 있는 subject 변수
factor(요인)	서로 상관 계수가 높은 것끼리 모아 변수의 집단으로 나눠 놓은 결과물을 의미한다.
요인 적재값(Loadings)	각 변수(s1~s6)와 요인(factor) 간의 상관 계수를 의미한다. 값이 +0.4 미만이면 중요도가 그리 높지 않는 변수로 이해하면 된다. 높은 적재 값은 해당 변수들이 이 요인으로 잘 설명할 수 있다는 의미이다.
요인 점수(scores)	관측치와 요인 간의 관계를 통하여 구해진 점수를 말한다.
SS Loadings	각 요인 적재 값들의 제곱의 총합을 말한다. 이 수치는 각 요인의 설명력을 보여 주는 것으로써 값이 크면 높은 설명력을 보여 주는 것이다.
Uniquenesses	유효성을 판단하여 제시한 값으로 통상 0.5 이하이면 유효한 것으로 본다.
정보 손실	1 - 누적 분산 비율
고유 값	변수 속에 담겨진 정보가 어떤 요인에 어느 정도 표현될 수 있는 가를 말해주는 비율이다. 먼저 추출된 요인의 고유 값은 다음에 추출되는 요인의 고유 값보다 크다.

요인 수를 결정하는 방법

요인수를 결정하는 방법은 크게 주성분 분석 방법과 상관 계수 행렬을 이용한 초기 고유 값을 이용하는 방법이 있다.

### 고유 값(eigen value)의 사용처 :

선형 연립 방정식  
특이 값 분해  
주성분 분석

### prcomp 함수

주성분 분석이란 변동량(분산)에 영향을 주는 주요 성분을 분석하는 방법이다.

prcomp 함수는 주성분 분석을 수행해주는 함수인데, 주성분 분석을 이용하여 어떤 요소가 변동량에 영향을 가장 많이 주는 가를 찾아 내는 방법이다.

요인 분석에서 사용될 요인의 개수를 결정하는 데 주로 사용한다.

항목	설명
사용 형식	<code>pc &lt;- prcomp( data )</code>
data	주성분 분석을 수행할 데이터 프레임

### factanal 함수

요인 분석(factor analysis) 함수의 줄임말로써 요인 분석에서 해석이 어려운 어느 한 요인을 높게 나타나도록 하기 위하여 요인 축을 회전하는 방법이 있다.

일반적으로, varimax 회전법을 많이 사용한다.

유의 확률(p-value)의 값이 0.05보다 적으면 요인 수가 부족하다는 의미이다.

항목	설명
사용 형식	<code>result &lt;- factanal(dataset, factors=2, rotation='varimax')</code>
dataset	요인 해석을 위한 데이터 셋을 지정한다.
factor	주성분 변수(요인 갯수)의 갯수를 지정한다.
rotation	요인 회전법 이름(varimax, promax, none)
scores	요인 점수를 계산하는 방법이다. <code>c('none', 'regression', 'Bartlett')</code> 예시 : regression(회귀 분석으로 요인 점수를 계산하는 방식)
na.action	결측치에 대한 처리

### factanal 출력 결과 예시

```
result <- factanal(subject, factors=3, rotation='varimax', scores='regression')
result
```

Call:

```
factanal(x = subject, factors = 3, scores = "regression", rotation = "varimax")
```

Uniquenesses: # 유효성을 판단하여 제시한 값으로 통상 0.5 이하이면 유효한 것으로 본다.

국어 수학 인문 물리 사회 영어

0.005 0.051 0.240 0.005 0.005 0.056

Loadings: # 변수와 요인 간의 상관 관계이다.

# 예를 들어 Factor1 과 인문 과목과 사회 과목이 상관 계수가 높다

	Factor1	Factor2	Factor3
--	---------	---------	---------

국어	-0.379		0.923
----	--------	--	-------

수학	0.236	0.931	0.166
----	-------	-------	-------

인문	0.771	0.297	-0.278
----	-------	-------	--------

물리	0.120	0.983	-0.118
----	-------	-------	--------

사회	0.900	0.301	-0.307
----	-------	-------	--------

영어	-0.710	0.140	0.649
----	--------	-------	-------

	Factor1	Factor2	Factor3
--	---------	---------	---------

SS loadings	2.122	2.031	1.486
-------------	-------	-------	-------

 # 각 요인 적재 값들의 제곱의 총합을 말한다.

Proportion Var	0.354	0.339	0.248
----------------	-------	-------	-------

Cumulative Var	0.354	0.692	0.940
----------------	-------	-------	-------

The degrees of freedom for the model is 0 and the fit was 0.7745

## 기계 학습(머신 러닝)

기계 학습 즉, 머신 러닝은 빅 데이터와 사물 인터넷(IoT) 시대에서 유용한 정보를 생성해주는 중요한 역할을 제공한다.

기계 학습은 정해진 특정 알고리즘을 통해서 데이터를 예측하는 인공지능의 일종이다.



되도록 사람의 개입을 적게 하고 컴퓨터가 데이터에 의한 학습을 통해 최적의 판단이나 예측을 가능하게 해주는 것을 말한다.

지도 학습과 비지도 학습(자율 학습)으로 분류가 된다.

지도 학습은 사전에 입력출력 정보를 제공하고, 해당 입력에 대한 출력 값이 나타나는 규칙을 발견하고, 이를 통해서 만들어진 모델(model)을 통해서 새로운 데이터를 추정 및 예측하는 학습 패턴을 의미한다.

비지도 학습은 최종적인 정보가 없는 상태에서 컴퓨터 스스로 공통점과 차이점 등의 패턴을 이용해서 규칙을 생성하고, 이를 통해서 분석 결과를 도출해내는 방식이다.

따라서 유사한 데이터를 그룹화해주는 군집화와 군집내의 특성을 나타내는 연관 분석 방법에 주로 이용된다.

## 회귀 분석

선형 회귀를 설명하기 전에 독립 변수와 종속 변수에 대한 개념을 간단히 살펴 보자.

독립 변수는 영향을 주는(미치는) 변수이고 종속 변수는 영향을 받는 변수이다.

예를 들어서, 직선의 방정식  $y = 2 * x + 1$ 에서  $x$ 는 독립 변수이고,  $y$ 는 종속 변수이다.

$x$ 의 값에 따라서  $y$ 의 값이 달라진다.

즉,  $x=2$ 이라고 가정하면  $y=2*2+1$  라는 연산식에 의하여 값이 5가 된다.

회귀 분석(Regression Analysis)이란 특정 변수(독립 변수)가 다른 변수(종속 변수)에 어떠한 영향을 미치는가를 분석하는 방법이다.

즉, **인과 관계**가 있는지 등을 분석하기 위한 방법으로 한 변수의 값을 가지고 다른 변수의 값을 예측해 주는 분석 방법이다.

## 선형 회귀 분석

선형 회귀는 독립 변수와 종속 변수간의 관계를 모델링하는 기법을 말한다.

이 때 독립 변수가 하나인 경우 단순 선형 회귀(Simple Linear Regression)라하고 독립 변수가 2개 이상인 경우 중선형 회귀(Multiple Linear Regression)이라고 한다.

선형 회귀는 다음과 같이 독립 변수  $X$ 와 종속 변수  $y$ 로 표현한다.

$$y = \beta_0 + \beta_1 * x + \varepsilon = wx + b$$

# 회귀 계수 =  $\beta_0$ (절편) +  $\beta_1$ ( $x$ 의 계수)

#  $\varepsilon$ 는 오차를 나타낸다.

## 관련 용어

회귀 분석을 함에 있어서 자주 사용되는 용어들을 정리해본다.

관련된 함수들은 대부분 `stats` 패키지에 들어 있다.

항목	설명
독립 변수	다른 변수에게 <b>영향을 주는 변수</b> 를 말한다.
종속 변수	다른 변수에게 <b>영향을 받는 변수</b> 를 말한다.

회귀 계수	절편과 기울기를 의미한다. ( $\beta_0$ (절편) + $\beta_1$ (x의 계수)) <code>coef( model )</code> 함수는 회귀 계수(절편과 기울기)를 구해 주는 함수이다.
회귀 방정식	회귀 계수를 이용하여 생성된 방정식을 말한다.
회귀 선	독립 변수와 종속 변수에 대한 분포를 나타내기 위한 가장 적합한 직선을 말한다.
적합된 값	각 독립 변수(x)에 대한 모델의 예측된 y 값을 적합된 값(fitted value)이라고 부른다. 예측치라고 이해하면 된다. 적합된 값은 <code>fitted(mydata)</code> 함수를 사용하면 구할 수 있다.
잔차(residuals)	모델로부터의 구한 예측 값과 실제 값 사이의 차이를 말한다. (H - y)를 의미한다. 잔차는 <code>residuals()</code> 함수를 이용하면 구할 수 있다.
잔차 제곱 합	선형 회귀에서는 오차의 제곱의 합이 최소가 되도록 회귀 계수를 정한다. stats 패키지에서 제공하는 <code>deviance()</code> 함수를 이용하면 잔차 제곱의 합을 구할 수 있다.
다중 공선성 (multicollinearity)	독립 변수들 간의 강한 상관 관계로 인하여 회귀 분석의 결과를 신뢰할 수 없게 되는 현상을 말한다. 강한 상관 관계를 갖는 독립 변수를 제거하여 해결한다. 다중 공선성 문제가 의심이 되는 경우에는 반드시 상관 계수를 구해보도록 한다.
분산 팽창 요인 (VIF)	분산 팽창 요인(Variance Inflation Factor)은 공차 한계의 역수로 표시한다. 공차 한계 : 한 독립 변수가 다른 독립 변수들에 의하여 설명이 되지 않는 부분을 의미한다. VIF는 10 이상이면 다중 공선성을 의심해봐야 한다.

## 단순 회귀 분석

제품의 적절성(독립 변수)이 제품의 만족도(종속 변수)에 영향을 주는 가에 대한 회귀 분석을 수행해본다.

항목	설명
귀무 가설	제품 적절성은 제품의 만족도(y)에 영향을 <b>미치지 않는다.</b>
연구 가설	제품 적절성은 제품의 만족도(y)에 영향을 <b>미친다.</b>

## 선형 회귀 모델 구하기

선형 회귀 모델은 stats 패키지에서 제공하는 `lm()` 함수를 사용하여 만들 수 있다.

반환 값은 선형 회귀 모델 클래스의 `lm` 인스턴스이다.

항목	설명
사용 형식	<code>lm(formula = y ~ x, data = df)</code>
formula	종속_변수 ~ 독립_변수 형태로 지정한 포물러를 지정한다.
data	포물러를 적용시키고자 하는 데이터를 지정한다.

## lm() 함수의 반환 결과

```
# 회귀 계수 = 절편 + 기울기
# Call:
# lm(formula = ydata ~ xdata, data = myframe)
#
# Coefficients:
# (Intercept)          xdata
#      0.7789 ← 절편      0.7393 ← 기울기
```

회귀 모델에 대하여 세부적인 내용을 확인하려면 summary() 함수를 이용하면 된다.

## summary를 이용한 확인 :

```
# 선형 회귀 분석 결과 보기
summary(result)
# Call: 항목 ①
# lm(formula = ydata ~ xdata, data = myframe)
# Residuals: 항목 ②
# Min      1Q      Median      3Q      Max
# -1.99669 -0.25741  0.00331  0.26404  1.26404
# Coefficients: 항목 ③
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept)  0.77886    0.12416   6.273 1.45e-09 ***
# xdata        0.73928    0.03823  19.340 < 2e-16 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
# 항목 ④
# Residual standard error: 0.5329 on 262 degrees of freedom
# Multiple R-squared:  0.5881, Adjusted R-squared:  0.5865 <== 조정 결정 계수
# 조정 결정 계수 : 결정 계수에 대하여 오차를 감안하여 조정된 결정 계수를 말한다
# F-statistic: 374 on 1 and 262 DF, p-value: < 2.2e-16

# R-squared: 0.5881의 값이 독립 변수와 종속 변수와의 상관 관계를 의미한다.
# 1에 가까울수록 변수의 모델이 잘 되었다는 의미이다.
# 유의 확률이 0.05 이상인 경우에는 회귀선이 모델에 부적합하다는 의미이다.
# 예시에서는 p-value(< 2.2e-16)이므로 적합하다고 볼 수 있다.
```

구분	특징
항목 ①	사용된 포물선에 대한 정보를 보여 주고 있다.
항목 ②	잔차에 대한 정보를 보여 준다.

항목 ③	<p><b>독립 변수</b>(설명 변수) <b>평가 영역</b>이다.</p> <p>Estimate 열은 절편과 계수의 추정치를 보여 준다.</p> <p>Pr(&gt; t ) 열은 p-value 값을 알려 준다.</p> <p>귀무 가설은 절편이 0 이다.</p> <p>p-value 값이 0.05 보다 크면 무의미하다.</p> <p>바로 뒤에 * 또는 ***로 표시된 문자열은 p-value 의 범위를 알려 준다.</p> <p>*, **, ***등이 있다면 유의미한 값이라고 보면 된다.</p>
항목 ④	<p>F 통계량을 보여 주는 곳이다.</p> <p>R-squared 를 결정 계수 또는 상관 계수라고 부른다.</p> <p><b>독립 변수와 종속 변수와의 상관 관계</b>를 의미해주는 변수이다.</p> <p>1 에 가까울수록 가장 <b>이상적인 값</b>이라고 이해하면 된다.</p>

### 단순 회귀 분석 결과 제시 방법

제품의 가격과 품질을 결정하는 제품 적절성은 제품 만족도에 정(正)의 영향을 미칠 것이라는 연구 가설을 검정한 결과 검정 통계량  $t = 19.340$ ,  $p = <2e-16$  으로 통계적 유의 수준 하에서 영향을 미치는 것으로 나타났다.

따라서, 연구 가설을 채택한다.

회귀 모형은 상관 계수  $R=0.588$  로 두 변수 간에 다소 높은 상관 관계를 나타내고 있다.

$R^2=.587$  로 제품 적절성 변수가 제품 만족도를 58.7% 설명하고 있다.

또한 회귀 모형의 적합성은  $F=374(p=<2.2e-16)$ 으로 회귀선이 모형에 적합하다고 볼 수 있다.

### 논문 보고서 제시 예제

단순 회귀 분석 결과를 논문/보고서에 제시한 예는 다음과 같다.

아래 예시는 제품 적절성에 따른 제품 만족도 영향 분석을 한 예시이다.

"제품 만족도"와 "분석 통계량"은 summary() 함수에서 구할 수 있다.

"회귀식"은 lm()함수에서 구할 수 있다.

종속 변수	독립 변수	표준 오차 (Std. Error)	베타	검정 통계량 (t value)	유의 확률 (p)
제품 만족도	상수	0.12416	-	6.273	1.45e-09 ***
	제품 적절성	0.03823	0.739	19.340	<2e-16 ***
분석 통계량	Model Summary : $R = 0.767$ , $R^2 = 0.5881$ ANOVA : $F = 374$ , $p \leq 2.2e-16$ ***				
회귀식	$Y(\text{제품 만족도}) = 0.77886 + 0.73928 * X(\text{제품 적절성})$				

### 다중 회귀 분석

여러 개의 독립 변수가 동시에 1개의 종속 변수에 미치는 영향을 분석할 때 사용하는 분석 방법이다.

#### 다중 회귀 분석 결과 보기

##### 다중 회귀 분석 결과

```
summary(result.lm)
# Call:
# lm(formula = y ~ x1 + x2, data = df)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -2.01076 -0.22961 -0.01076  0.20809  1.20809
#
# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept)  0.66731    0.13094   5.096 6.65e-07 ***
# x1           0.09593    0.03871   2.478  0.0138 *
# x2           0.68522    0.04369  15.684 < 2e-16 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 0.5278 on 261 degrees of freedom
# Multiple R-squared:  0.5975,    Adjusted R-squared:  0.5945
# F-statistic: 193.8 on 2 and 261 DF, p-value: < 2.2e-16
```

상관 계수 R은 0.5975로 독립 변수와 종속 변수 간에 다소 높은 상관 관계를 보이고 있다.

## 일반화 선형 모델(Generalized Linear Model)

회귀 분석은 종속 변수가 정규 분포화 되어 있는 연속형 변수이다.  
하지만 다음과 같은 경우에는 종속 변수가 정규 분포화되어 있고 볼 수 없다.

### 일반화 선형 모델이 필요한 경우

항목	설명
이항 변수	종속 변수가 범주형 변수인 경우로써 두 가지 경우에서 1가지를 고르는 경우를 말한다. 예시 : 0 또는 1, 합격/불합격, 사망/생존 등
다항 변수	상품의 품질이 poor/good/excellent 중에서 1개이다. 선호하는 당이 공화당/민주당/무소속 등인 경우에 해당한다.

일반화 선형 모형은 종속 변수가 정규 분포되어 있지 않는 경우를 포함하는 선형 모형의 확장 모형이다.

대표적으로 로지스틱 회귀(Logistic regression)와 포아송 회귀(Poisson regression) 등이 있다.

## 로지스틱 회귀 분석

로지스틱 회귀 분석(Logistic Regression Analysis)은 종속 변수와 독립 변수 간의 관계를 나타내어 예측 모델을 생성한다는 점에서는 선형 회귀 분석 방법과 동일하다.

하지만 독립 변수(x)에 의해서 **종속 변수(y)의 범주로 분류**한다는 측면은 분류 분석 방법으로 분류된다.

### 로지 스틱 회귀의 특징

항목	설명
분석 목적	종속 변수와 독립 변수 간의 관계를 통해서 예측 모델을 생성하는데 있다.
회귀와의 차이점	종속 변수는 반드시 범주형 변수이어야 한다. 이항 분류 : Yes/No 다항 분류 : iris의 Spices 갈럼
정규성	정규 분포 대신에 이항 분포를 따른다.
로지 변환	종속 변수의 출력 범위를 0 과 1 로 조정하는 과정을 의미한다. 혈액형 A 형인 경우 -> [1,0,0,0], B 형인 경우 -> [0,1,0,0]
활용 가능 분야	의료, 통신, 날씨 등 다양한 분야에서 활용 가능하다.

### 로지스틱 모델 구하기

stats 패키지에서 제공하는 glm() 함수를 사용하면 로지스틱 회귀 모델을 구해준다.

항목	설명
----	----

사용 형식	glm( y~ x, data = train, family = 'binomial' )
formula	종속_변수 ~ 독립_변수 형태로 지정한 포물러이다.
data	포물러를 적용시킬 데이터를 지정한다.
family	로지 스틱 분류시 'binomial'이라는 값을 사용하면 된다.

### 다중 공선성 문제 해결

다중 공선성 문제는 독립 변수 간의 강한 상관 관계로 인해서 회귀 분석의 결과를 신뢰할 수 없는 현상을 의미한다. 다중 공선성 문제가 의심이 되는 경우 상관 계수를 구하여 변수간의 강한 상관 관계를 명확히 분석하는 것이 좋다. 변수간의 강한 상관 관계를 갖는 독립 변수를 제거하여 해결할 수 있다.

#### 다중 공선성 문제 해결

```
model <- lm(formula=Sepal.Length ~ Sepal.Width + Petal.Length + Petal.Width , data=training)
model

vif(model)
# Sepal.Width Petal.Length Petal.Width
# 1.218701 14.422474 13.654679

# vif의 값이 10 이상이면 다중 공선성 문제 의심
sqrt(vif(model)) > 2
# Sepal.Width Petal.Length Petal.Width
# FALSE TRUE TRUE

# 다중 공선성 문제가 의심이 되는 경우 상관 계수를 구하여 변수간의 강한 상관 관계를 명확히 분석하는 것이 좋다.
cor(iris[, -5]) # Species(종) 제외
# Sepal.Length Sepal.Width Petal.Length Petal.Width
# Sepal.Length 1.0000000 -0.1175698 0.8717538 0.8179411
# Sepal.Width -0.1175698 1.0000000 -0.4284401 -0.3661259
# Petal.Length 0.8717538 -0.4284401 1.0000000 0.9628654 <-- 높은 상관 관계
# Petal.Width 0.8179411 -0.3661259 0.9628654 1.0000000
```

## 모델 성능 평가

### 모델 예측하기

검정 데이터를 이용하여 회귀 모델의 예측치를 생성하려면 `predict()` 함수를 사용한다.( stats 패키지 )

구분	특징
사용 형식	<code>predict(model, newdata)</code>
model	회귀 모델(회귀 분석 결과가 저장된 객체)를 의미한다.
newdata	독립 변수(x)가 존재하는 검정 데이터 셋을 지정한다.
interval	부가적인 옵션이다. "confidence"라고 입력하면, 해당 데이터의 상하한 신뢰 구간까지 보여 준다. "prediction"라고 입력하면 오차까지 고려하겠다는 의미이다.
type	로지스틱 회귀 모델이면 'response'을 값을 사용하면 된다.

### 모델 평가하기

**선형 회귀 모델**은 회귀 방정식에 의해서 숫자로 예측을 하기 때문에 모델 평가는 일반적으로 **상관 계수**를 이용한다.  
즉, `cor(예측치, 실제_정답)` 명령어를 사용하면 된다.

반면 **분류 모델**은 y 변수를 범주로 예측하므로 일반적으로 **혼돈 매트릭스**로 판단한다.



---

### ROC 곡선

ROC(Receiver Operating Characteristic) Curve는 **이진 분류의 진단 능력을 보여주는 곡선**이다.  
ROC Curve는 x축, y축 모두  $[0,1]$ 의 범위의 값을 가지고,  $(0,0)$  에서  $(1,1)$ 을 잇는 곡선이다.

FPR(False Positive Rate)은 특이도라고 하는 데, 0(False)인 케이스에 대해 1(True)로 잘못 예측한 비율을 말한다. 예를 들어서 암 환자가 아님에도 불구하고 암이라고 진단 받는 경우이다.

TPR(True Positive Rate)은 민감도라고 하는 데, 1(True)인 케이스에 대해 1로 잘 예측한 비율을 말한다. 예를 들어서 암 환자를 진찰해서 암이라고 진단을 하는 경우이다.

ROC Curve는 x축에 FP Rate(FPR)를, y축에 TP Rate(TPR)를 표시한다.  
ROC 곡선에서 왼쪽 상단의 계단 모양의 빈 공간이 분류 정확도에서 오분류(missing)를 의미한다.

### ROC 곡선 그리는 절차

```
pr <- prediction( 예측_값, 정답_label )  
prf <- performance(pr, measure = "tpr", x.measure = "fpr")  
plot(prf, main='ROC Curve 그래프')
```

ROC 커브는 그 면적이 1에 가까울수록 (즉 왼쪽위 꼭지점에 다가갈수록) 좋은 성능이다. 그리고 이 면적은 항상 0.5~1의 범위를 갖는다.(0.5면 성능이 전혀 없음. 1이면 최고의 성능)

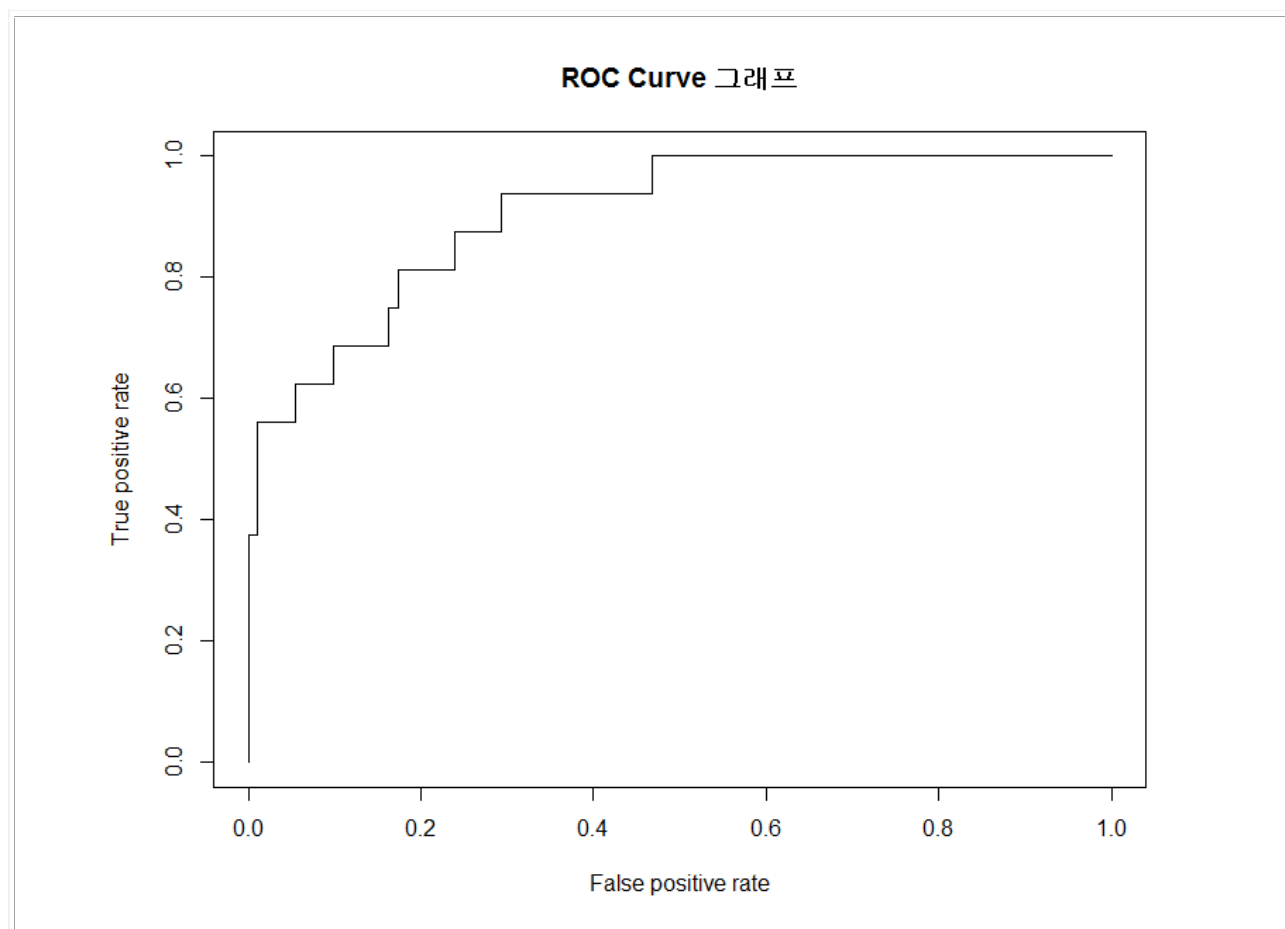
TPR과 FPR은 서로 비례하는 관계에 있다. 암환자를 진단할 때, 성급한 의사는 아주 조금의 징후만 보여도 암인 것 같다고 할 것이다. 이 경우 TPR은 1에 가까워질 것이다. 그리고 동시에 FPR도 1에 가까워진다. (정상인 사람도 전부 암이라고 하니 까)

반대로 실력이 없는 의사라서 암환자를 알아내지 못한다면, 모든 환자에 대해 암이 아니라고 할 것이다. 이 경우 TPR은 매우 낮아져 0에 가까워 질 것이고, 마찬가지로 FPR 또한 0에 가까워질 것이다.(암환자라고 판단 자체를 안하므로, 암환자라고 잘못 진단 하는 경우가 없어짐)

그런데 좋은 성능에 대한 지표인 TPR을 높이려다보면, 나쁜 성능에 대한 지표인 FPR도 같이 높아져버린다. 따라서 어떤 의사의 실력을 판단하기 위해서는 특정 기준(언제 암이라고 예측 할 지)을 연속적으로 바꾸어 가면서 TPR과 FPR을 측정을 해야한다. 그리고 이것을 한눈에 볼 수 있게 시각화 한 것이 바로 ROC 커브이다.

이 ROC커브는 두가지 장점이 있다. 먼저 그 커브의 면적을 재어 다양한 기준에서의 TRP과 FPR을 복합적으로 평가할 수 있다는 점이고, 또 한가지는 실제로 암을 판단할 때, 어디를 기준으로 잡을 지 결정하는 데 도움이 될 수 있다.

단순히  $TPR + (1-FPR)$ 이 최대가 되는 지점을 잡아도 되지 않을까 생각할 수 있지만, 실제로는 병에 따라서 어느 쪽에 좀 더 강조를 둘 것인가가 매우 중요할 수 있다. 예를 들어 걸릴 확률은 매우 낮지만 치사율이 극히 높은 병은 일단 환자라고 의심할 수록 좋기 때문에 FPR이 높더라도 괜찮을 수 있다. 반대로 걸릴 확률은 높지만 위험성이 매우 낮은 병은 FPR이 좀 낮은 기준을 선택하는 것이 괜찮을 것이다.



AUC = AUROC (the Area Under a ROC Curve) : ROC 커브의 밑면적을 구한 값이 바로 AUC. 이 값이 1에 가까울 수록 성능이 좋다.

AUC 해석 : 1로 예측하는 기준을 쉽게 잡으면 민감도는 높아진다. 그대신 너무나 쉽게 1이라고 판단하므로 따라서 특이도가 낮아진다. 그런데 이 두 값이 모두 1에 가까워 질 수록 좋은 성능을 의미한다. 만약 AUC = 0.5라면, 특이도가 감소하는 딱 그만큼만 민감도가 증가하는 것으로, 즉 어떤 기준으로 잡아도 민감도와 특이도를 동시에 높일 수 있는 지점이 없다는 것이다.

AUC가 0.5라면, 특이도가 1일때 민감도는 0, 특이도가 0일때 민감도는 1이되는 비율이 정확하게 trade off관계로, 두값의 합이 항상 1이다.

그러므로 AUC값은 전체적인 민감도와 특이도의 상관 관계를 보여줄 수 있어 매우 편리한 성능 척도에 기준이다.

출처: <https://newsight.tistory.com/53> [New Sight]

# 과적합(overfitting) 피하는 방법

# 테스트 결과 특정한 데이터 셋에만 잘 들어 맞는 현상

# K 겹 교차 검증 : 전체 집합을 K 등분 시켜서 각각의 그룹들을 testing 데이터로 사용하는 알고리즘

# min-max 알고리즘 :

# 특정 컬럼이 상대적으로 나머지 컬럼에 비하여 값이 너무 크거나 작은 경우

#  $\text{min-max} = (x - \text{min}) / (\text{max} - \text{min})$

# training 데이터와 testing 데이터는 반드시 분리하라.

## 분류 분석

카페 문서 : 2290

분류 분석(Classification Analysis)은 다수의 변수를 갖는 데이터 셋을 대상으로 특정 변수 값을 조건으로 지정하여 데이터를 분류하여 **트리 형태의 모델을 생성**하는 분석 방법이다.

분류 분석은 학습 데이터(training data)를 이용하여 분류 모델을 만든 다음에 이를 이용하여 새로운 데이터에 대하여 분류 값을 예측한다.

R에서 제공하는 분석 방법은 의사 결정 트리, 랜덤 포레스트, 인공 신경망 등이 있다.

의사 결정 트리(Decision Tree) 방식과 랜덤 포레스트(Random Forest) 방식 기법으로 데이터를 분류하는 방법에 대해서 알아 보도록 한다.

### 분류 분석의 특징

분류 분석의 일반적인 특징은 다음과 같다.

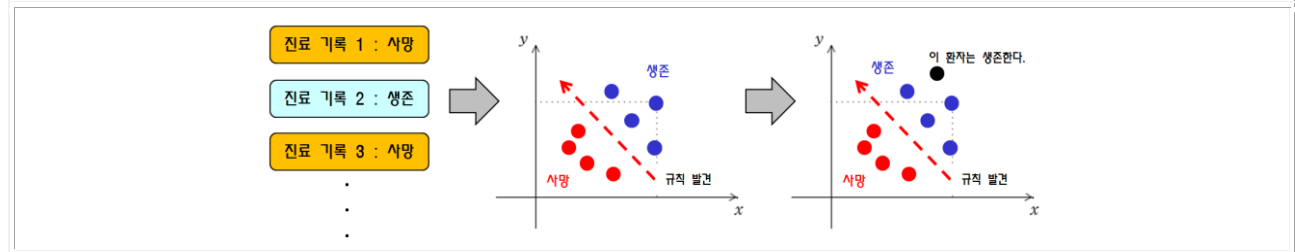
항목	설명
y 변수 존재	설명 변수(x 변수)와 반응 변수(y 변수)가 존재한다.
의사 결정 트리	분류 예측 모델에 의해서 의사 결정 형태로 데이터가 분류 된다. 분류 결과를 시각화할 수 있다.
비모수 검증	선형성, 정규성, 등분산성 가정이 필요 없다.
주론 기능	유의 수준 판단 기준이 없다.(주론 기능이 없다)
활용 분야	이탈 고객과 지속 고객의 분류 신용 상태의 좋고 나쁨을 분류 번호 이동 고객과 지속 고객 분류

### 분류 분석의 활용 예시

고객 분류하기, 기업의 부도 예측, 주가 예측, 환율 예측, 경제 전망 등등

기존 고객들의 여러 가지 정보를 이용하여 신용 상태를 파악한 다음, 새로운 고객에 대하여 향후 신용 상태 예측하기  
과거 환자들의 중앙 검사를 토대로, 이를 통해서 새로운 환자에 대한 암을 진단하는데 이용하기

즉, 고객을 분류할 수 있는 변수들에 대한 규칙/특성들을 찾아내어, 미래 잠재 고객의 행동이나 반응을 예측하거나 유도하는데 활용된다.



## 의사 결정 트리

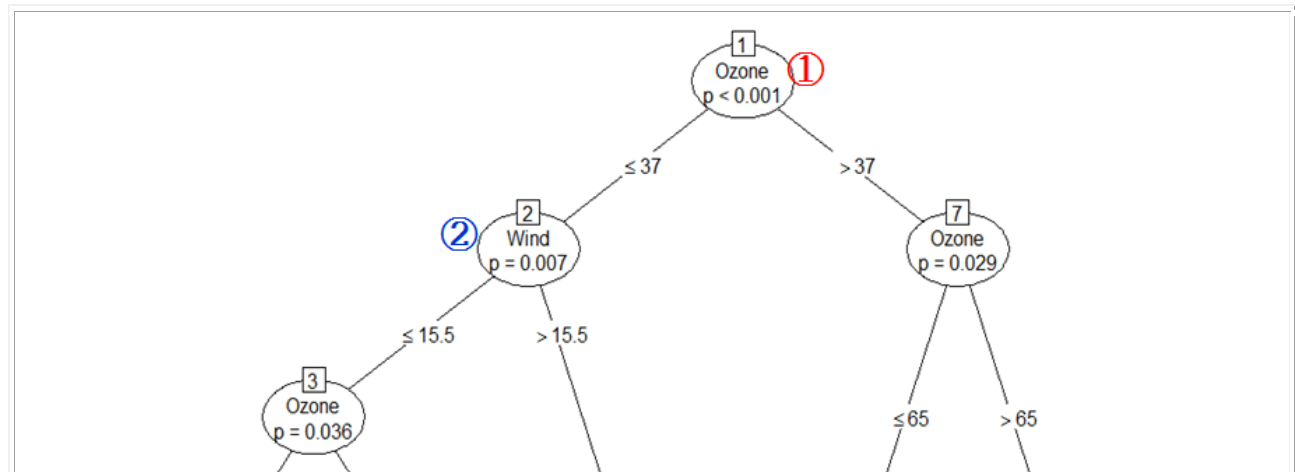
의사 결정 트리(Decision Tree)는 여러 가지 규칙을 순차적으로 적용시키면서 독립 변수 공간을 지속적으로 분할해 나가는 분류 모형이다.

### 결정 트리 이해하기

다음 그림은 온도에 가장 많은 영향을 미치는 요소들에 대한 의사 결정 트리이다

그림에서 보듯이 나무 구조 형태로 분류 결과를 도출해내는 방식이다.

분류시 **독립(입력) 변수 중에서 가장 영향력이 있는 변수를 기준으로 이진 분류**해 나가면서 크리스마스 트리처럼 나무(Tree) 구조 형태로 데이터를 분류해 나가는 방식이다.



결정 노드(node)는 속성에 따라서 가지(branches)들로 나뉘진다.

Root 노드에서 시작하여 각 속성에 따라 다양한 결정으로 범주를 예측하는 잎 노드까지 내려 간다.

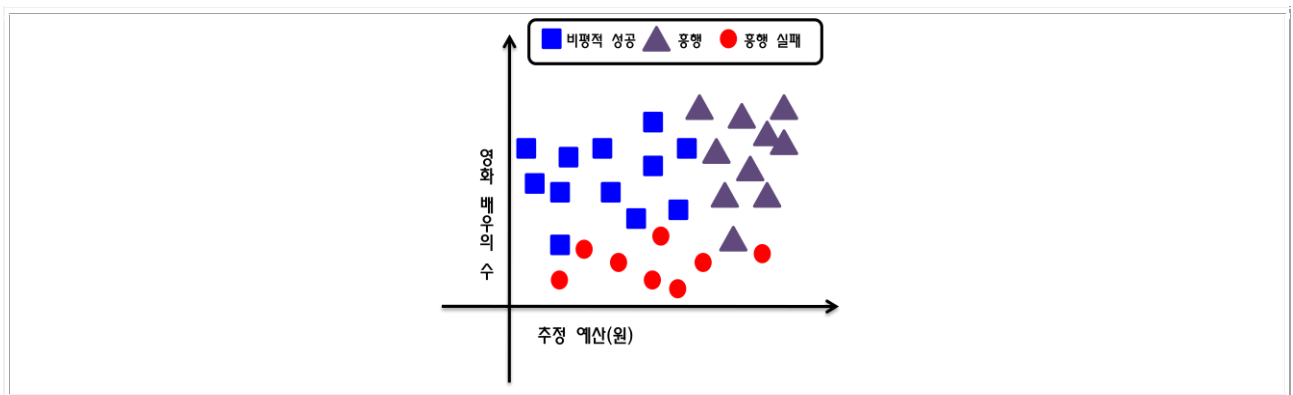
Tree 는 결정들의 조합 결과인 잎 노드(leaf node)로 끝난다.

분류(classification)와 회귀 분석(regression)에 모두 사용될 수 있기 때문에 CART(Classification And Regression Tree)라고도 한다.

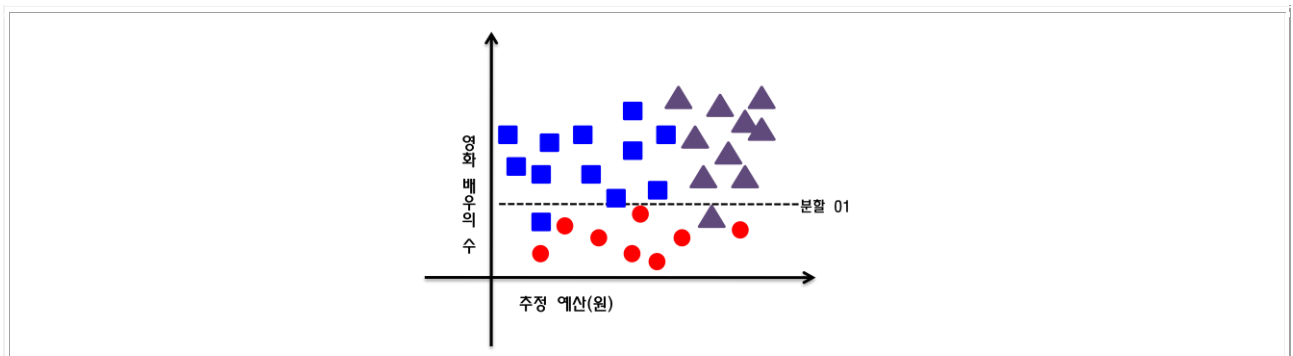
참조 문서 : [결정 트리 학습법\(위키 백과\)](#)

### 트리 생성 예시

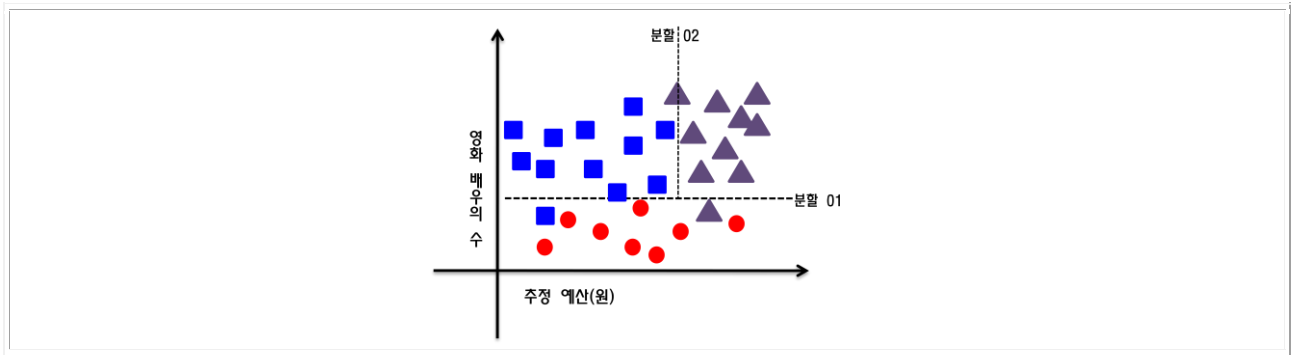
모든 영화에 대하여 3 가지 범주(비평적 성공, 흥행, 흥행 실패)로 분류하는 결정 트리 알고리즘을 개발한다고 가정하자. 영화의 제안된 예산과 영화 배우의 수에 있어서의 산포도는 다음과 같다.



이 데이터로 단순한 결정 트리를 만들기 위하여 '영화 배우의 수' 속성으로 나누면 다음과 같은 그림이 된다. 영화 배우가 많은 영화와 그렇지 않은 영화로 분류가 된다.



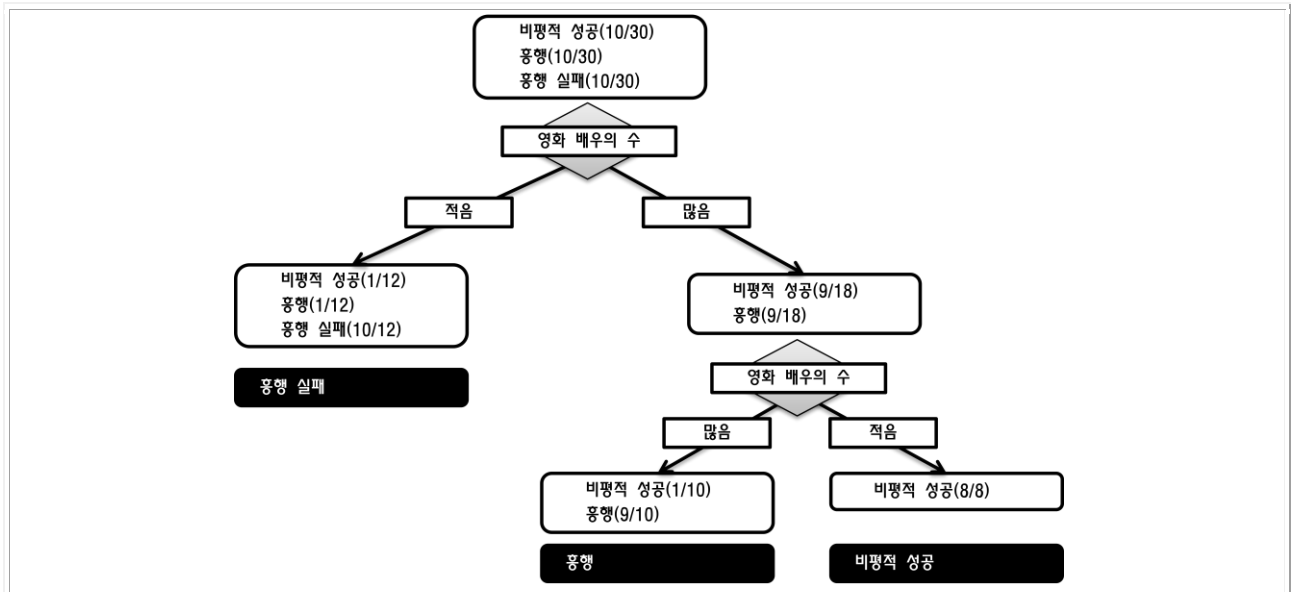
다음으로 영화 배우의 수가 많은 부분은 '고예산 영화'와 '저예산 영화'로 다시 분류할 수 있다.



#### 각 그룹에 대한 설명

항목	설명
왼쪽 위 그룹	비평이 성공적인 영화이며, 많은 영화 배우들과 상대적으로 적은 예산이 들었다.
오른쪽 위 그룹	많은 예산이 들어 갔으며, 많은 영화 배우가 나오는 흥행된 영화이다.
마지막 그룹	출연한 영화 배우의 수가 적고, 예산과 상관 없이 망해버린 영화이다.

미래에 성공할 영화를 예측하는 모델을 다음 다이어그램처럼 표현할 수 있다.



#### 결정 트리 관련 알고리즘

결정 Tree 를 구성하는 알고리즘에는 주로 하향식 기법이 사용되며, 각 진행 단계에서는 주어진 데이터 집합을 가장 적합한 기준으로 분할하는 변수 값이 선택된다.

지니 불순도, 정보 획득량, 분산 감소 등이 있다.

#### 엔트로피(Entropy)

Entropy 는 열역학에서 말하는 의미와 비슷하게 **정보의 무질서도 혹은 복잡도나 불확실성의 정도**를 나타낸다.

주어진 데이터 집합에 서로 다른 종류(클래스)들이 많이 섞여 있으면 엔트로피가 높고, 같은 종류(클래스)들이 많이 있으면 엔트로피가 낮는데, 이것을 이용하여 **범주 데이터가 얼마나 섞여 있는 지를 나타내는 지표**가 된다.

엔트로피의 값은 최소 0 이고, 최대 값은 1 이다.

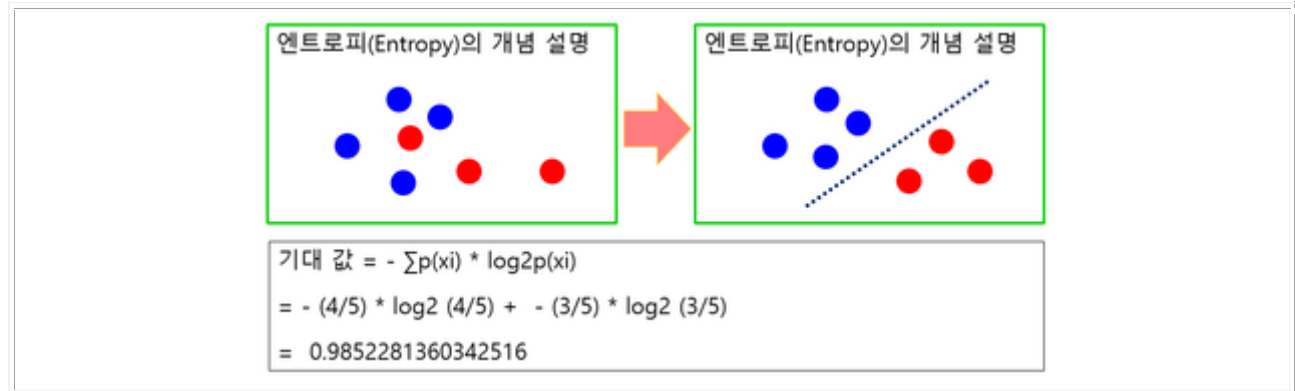
바구니에 색상이 다른 두 종류의 공이 있다고 가정하자.

이 때 정보 엔트로피(정보의 기대 값)는 다음과 같이 구할 수 있다.

$$\text{기대 값} = -\sum p(x_i) \cdot \log_2 p(x_i)$$

만약 바구니에 1 색상만 들어 있다면  $\log_2(1) = 0$  이므로 엔트로피는 0 이다.

즉, 데이터가 무질서하지 않고 균일하다는 의미이다.



2 가지 이상의 공이 서로 섞여 있다면, 2 가지 속성으로 서로 완전 분리를 해야 한다는 의미이다.

이 원리가 **의사 결정 트리**에서 어떤 변수를 어떤 속성 값으로 나누는 판단 기준이 된다는 것이다.

의사 결정 트리는 결국 엔트로피가 높은 상태에서 낮은 상태가 되도록 데이터를 특정 조건을 찾아 나무 모양으로 구분해 나간다.는 알고리즘이다.

### 의사 결정 트리 예시

그림에서 동그라미 부분에 해당하는 부분을 노드(Node)라고 한다.

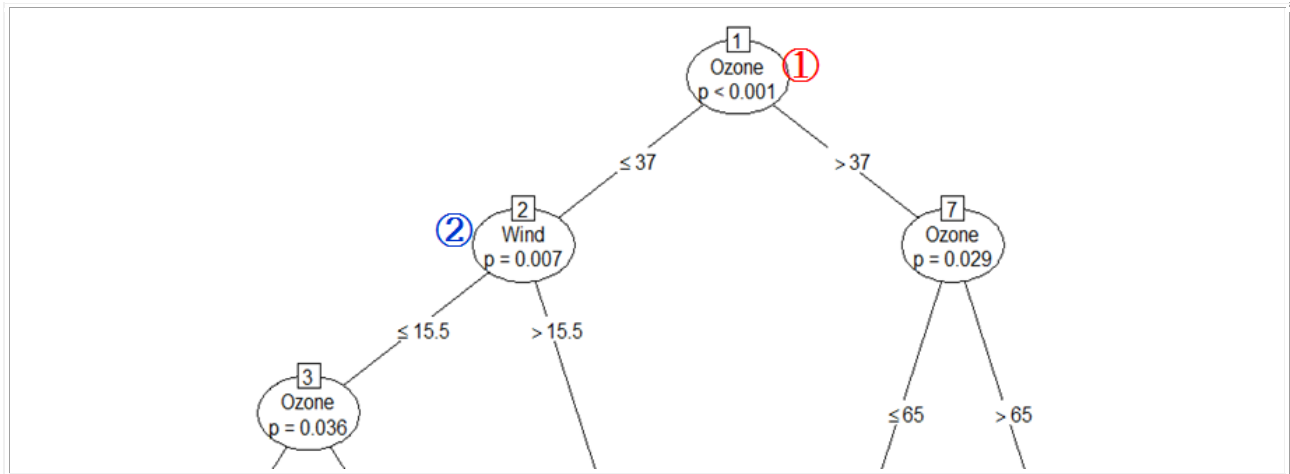
노드는 데이터 셋에 있는 변수의 이름(엑셀의 컬럼)에 해당하고, 이 변수는 반드시 범주형이어야 한다.

아래로 이어지는 선은 nominal(범주형) 변수가 가질 수 있는 속성의 종류만큼 분기가 된다.

제어문의 if~else 형식으로 조건문을 만들게 된다.

항목	설명
root node	트리 구조의 가장 상단에 있는 노드를 말한다.
label	마지막 하단에 있는 네모 상자를 의미하며, 이것은 우리가 원하는 solution 이다. 이것을 leaf(잎사귀)라고 부른다.





비교적 모델 생성이 쉽고, 단순하지만 명료한 결과를 제공하기 때문에 현업에서 가장 많이 사용하는 지도 학습 모델이다. party 패키지와 rpart 패키지를 이용하여 의사 결정 트리 방식으로 분류 모델을 사용하는 방법을 숙지해보도록 한다.

## party 패키지

party 패키지에서 제공되는 ctree() 함수를 이용하면 특정 변수 값을 기준으로 분류 분석된 결과를 의사 결정 트리 형태로 제공하는 함수이다.

반환되는 타입은 "BinaryTree"이다.

항목	설명
사용 형식	ctree(formula, data)
formula	의사 결정 형태로 분석할 포물러를 지정한다. 포물러 사용 예시 $\text{myformula} = \text{종속변수} \sim \text{독립변수1} + \text{독립변수2}$ $\text{result} = \text{ctree}(\text{formula}=\text{myformula}, \text{data}=\text{데이터셋})$
data	의사 결정 트리를 수행할 데이터 셋을 의미한다.

## ctree() 함수 반환 결과 예시

ctree() 함수의 반환 결과는 BinaryTree 구조이다.

반환된 결과에 대한 세부 내용은 다음과 같다.

반환 결과 예시
<pre> # Conditional inference tree with 5 terminal nodes # Response: Temp # Inputs: Solar.R, Wind, Ozone # Number of observations: 153                     </pre>

```
# 1) Ozone <= 37; criterion = 1, statistic = 56.086
#      2) Wind <= 15.5; criterion = 0.993, statistic = 9.387
#      3) Ozone <= 19; criterion = 0.964, statistic = 6.299
#      4)* weights = 29
#      3) Ozone > 19
#      5)* weights = 69
#      2) Wind > 15.5
#      6)* weights = 7
# 1) Ozone > 37
#      7) Ozone <= 65; criterion = 0.971, statistic = 6.691
#      8)* weights = 22
#      7) Ozone > 65
#      9)* weights = 26
```

7) Ozone <= 65; criterion = 0.971, statistic = 6.691  
 ① ② ③ ④

## 번호 별 상세 설명

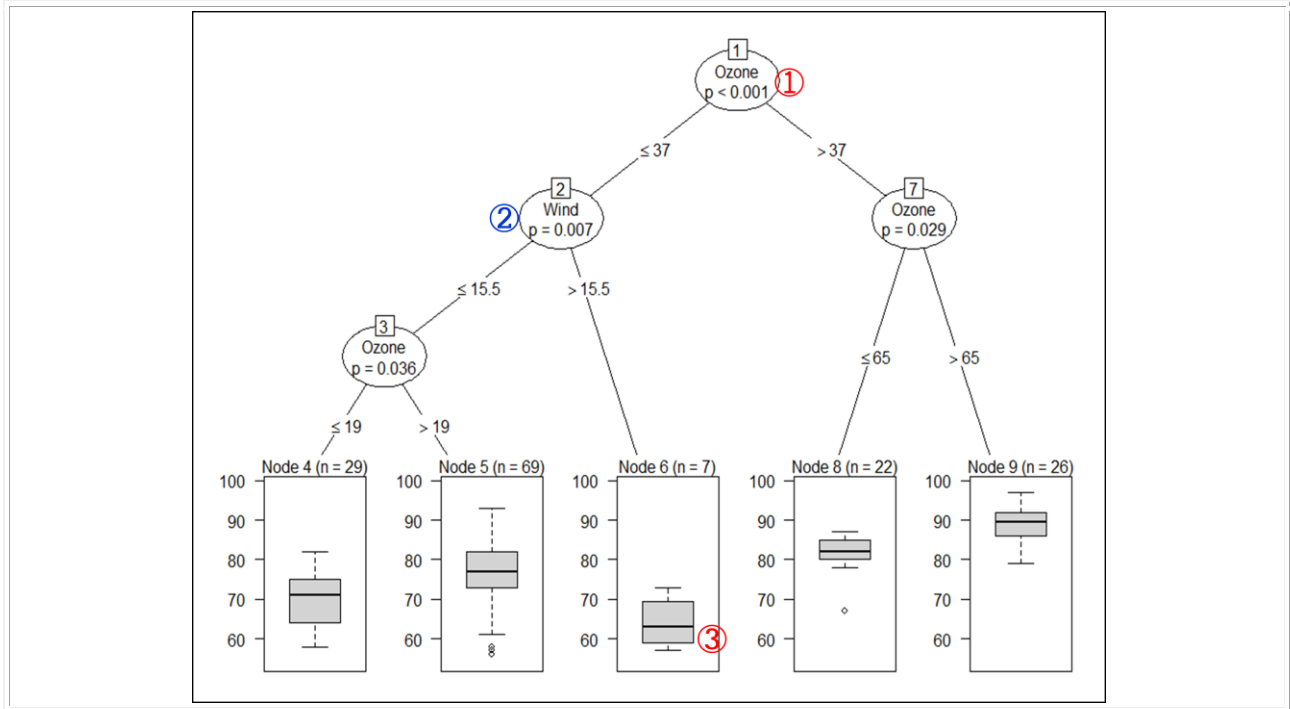
그림에 새겨진 숫자들에 대한 세부 설명이다.

번호	설명
①	종속 변수(Temp)에 대하여 독립 변수(태양광, 바람, 오존 수치)가 영향을 미치는 중요 변수의 척도이다. 수치가 작을 수록 영향을 미치는 정도가 크다. 온도에 가장 영향을 끼치는 변수는 오존 수치이다. 두 번째는 바람이다. 순서는 분기되는 순서를 말한다.
②	의사 결정 트리의 노드 이름이다. * 기호는 해당 노드의 마지막 노드(leaf node)를 의미한다. * 기호가 없으면 임계 값이 조건식으로 온다.
③	노드의 분기 기준(criterion)이 되는 수치이다. criterion = 0.971 이라면 p-value = (1 - 0.971)= 0.029 이다.
④	종속 변수의 통계량으로써, 수치 값이 크면 영향력이 크다.는 의미이다.
특이 사항	마지막 노드이거나 또 다른 분기 기준이 있는 경우에는 세 번째, 네 번째 수치는 표시되지 않는다. 마지막 노드의 weights 는 도수를 의미한다.

## 분류 분석 결과 보기

ctree() 함수의 반환 결과에 대하여 plot() 함수를 사용하면 분류 분석 결과를 저장하고 있는 변수에 대하여 차트를 다음과 같이 그릴 수 있다.

plot() 함수 사용시 간단히 보려면 type="simple" 옵션을 사용하면 된다.



### 분류 분석 결과 세부 설명

분류 분석 결과 그림에 새겨진 숫자들에 대한 세부 설명이다.

번호	설명
①	온도에 가장 영향을 주는 변수는 오존 수치(Ozone)이다. 오존량이 감소하면 대체적으로 온도가 감소하는 경향을 보인다.
②	두 번째로 영향을 주는 변수는 바람(Wind)이다.
③	오존량이 37 이하이면서 바람의 양이 15.5 이상이면 평균 온도가 63 정도이다.
④	바람의 양이 15.5 이하인 경우에는 평균 온도가 70 이상으로 나타난다.
⑤	태양광(Solar.R)은 다른 설명 변수에 비해서 온도에 크게 영향을 미치지 않는 것으로 분석이 된다.

### rpart 패키지

rpart 패키지에서 제공되는 rpart() 함수는 재귀 분할(recursive partitioning)이라는 의미를 갖는다.

기존 ctree() 함수에 비해서 2수준 요인으로 분산 분석을 실행한 결과를 트리 형태로 제공하여 모형을 단순화해준다.

Gini Index가 작아지는 방향으로 움직이며, Gini Index를 가장 많이 감소 시키는 변수가 가장 큰 영향을 미치는 변수가 된다.

항목	설명
----	----

사용 형식	rpart(formula, data, cp)
formula	의사 결정 형태로 분석할 포물러를 지정한다.
data	의사 결정 트리를 수행할 데이터 셋을 의미한다.
cp	cp 속성 값을 높이면 가지 수가 적어지고, 낮추면 가지 수가 많아진다.(기본 값 : 0.01)

## K겹 교차 검증

테스트를 위한 데이터의 수가 충분하지 않는 경우 모든 데이터를 테스트 데이터 셋으로 만드는 방법이 있다.

K겹 교차 검증이란, 데이터 셋을 여러 개로 나누어 하나씩 테스트 데이터 셋으로 사용하고 나머지를 모두 학습용 데이터 셋으로 사용하는 방법이다.

이러한 단순한 방법을 단일 잔류(leave-one-out)라고 한다.



## cvTools를 사용한 교차 검증

n개의 관찰치를 K겹 교차 검증의 R회 반복으로 분할한다.

항목	설명
사용 형식	cross <- cvFolds(n=nrow(iris), K=3, R=2, type)
n	관찰치의 수 또는 데이터의 크기
K	K 겹 교차 검증
R	R 회 반복
type	type('random', 'consecutive', 'interleaved')

## K겹 교차 검증 회전수와 데이터 셋 생성

R	K	검정(Test) 데이터	훈련(Train) 데이터
---	---	--------------	---------------

1	1	subsets[1, 1]	subsets[2, 1]	subsets[3, 1]
1	2	subsets[2, 1]	subsets[1, 1]	subsets[3, 1]
1	3	subsets[3, 1]	subsets[1, 1]	subsets[2, 1]
1	1	subsets[1, 2]	subsets[2, 2]	subsets[3, 2]
1	2	subsets[2, 2]	subsets[1, 2]	subsets[3, 2]
1	3	subsets[3, 2]	subsets[1, 2]	subsets[2, 2]

---

## 비지도 학습

비지도 학습(unSupervised Learning)은 입력 데이터에 의한 학습을 통하여 사전 지식이 없는 상태에서 컴퓨터 스스로 공통 점과 차이점 등의 패턴을 찾아서 규칙(rule)을 생성하고, 이를 통해서 결과를 도출해내는 방식이다.

---

## 군집 분석

데이터를 몇 개의 그룹으로 나누어 그룹 간의 비교 분석을 통하여 전체의 구조에 대한 이해를 돕고자 하는 탐색 기법이다. 군집 분석은 계층적 군집 분석(탐색적)과 비계층적 군집 분석(확인적)으로 나누어 진다.

## 군집 분석의 특징

군집 분석은 다음과 같은 특징을 가지고 있다.

### 군집 분석의 특징

전체적인 데이터 구조를 파악하는 데 이용된다.  
유사한 특성을 가진 항목들을 합쳐 가면서 유사한 특성을 군집을 찾아낸다.  
분류하려고 하는 항목에 대한 사전 지식이 없는 상태에서 이루어지므로 비지도 학습이다.  
유사도에 근거하여 군집들을 분석하는 데, 유사성은 유클리디언 거리를 사용한다.  
분석 결과에 대한 가설 검정이 없다.  
규칙을 기반으로 계층적인 트리 구조를 생성한다.

## 참조 사이트

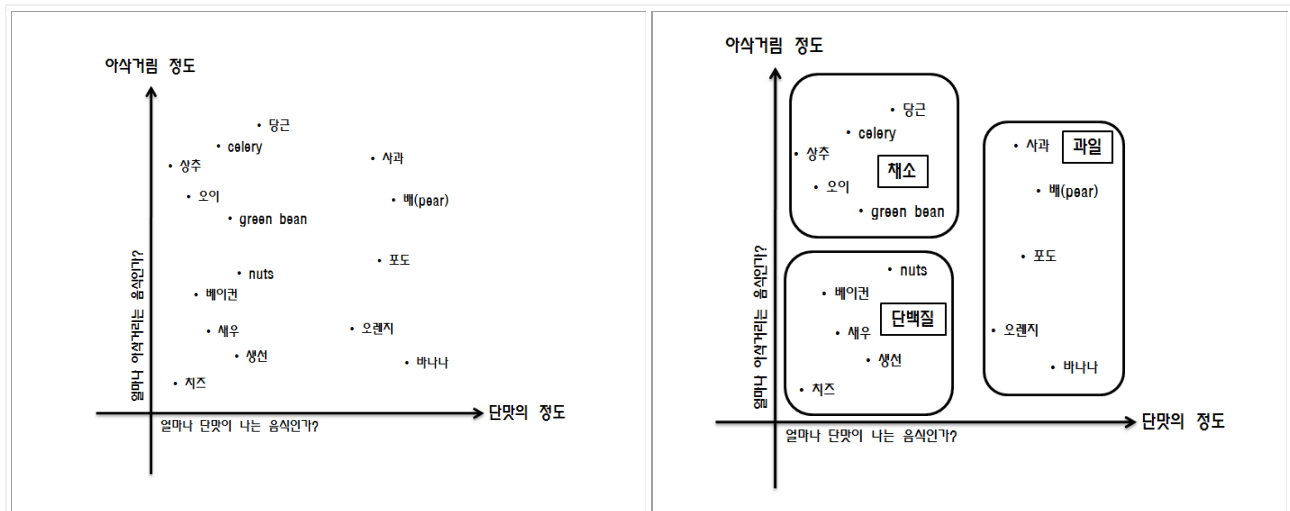
### 참조 사이트

[https://rstudio-pubs-static.s3.amazonaws.com/249084\\_09c0daf4ceb24212a81ceddca97ba1ea.html#-r-](https://rstudio-pubs-static.s3.amazonaws.com/249084_09c0daf4ceb24212a81ceddca97ba1ea.html#-r-)  
<http://blog.naver.com/liberty264/221017194817>

## 유사도(Similarity)

특정 데이터들을 분류하려면 분류를 위한 판단 근거가 필요하다.  
다음과 같이 두 가지 속성의 값을 이용하여 분류하는 예를 살펴 보자.  
참고하고자 하는 두 가지 속성은 "단맛"과 "아삭거림"이다.  
x축은 단맛의 정도를, y축은 아삭거림의 정도를 나타내는 좌표계로 그려 보면 다음과 같다.

그림에서 "바나나"와 "오렌지"의 거리는 가깝지만, "상추"와 "바나나"의 거리는 비교적 멀다.  
즉, 두 객체간의 "물리적 거리가 가까운 항목들은 동일 집단으로 묶을 수 있다"라는 의미이다.



이때 필요한 것이 거리를 구하는 함수의 개념이 필요하다.  
일반적으로 유클리디언 거리 공식이 많이 사용된다.

### 유사도 측정 방법

특성을 기반으로 유사도를 측정하는 방법은 여러 가지가 존재한다.  
주로 특성 값을 벡터 공간에 맵핑한 후, 벡터간의 거리를 기반으로 계산하는 방법이 많이 사용된다.

유클리디언 거리, 코사인 거리, 자카드 거리, 맨해튼 거리 등이 있다.

#### 유클리디언 유사도(Euclidean similarity )

유클리디언(Euclidean distance) 거리는 두 점 사이의 거리를 계산하는 방법이다.  
관측 대상 p 와 q 의 대응하는 변량 값의 차가 작으면, 두 개의 관측 대상은 유사하다고 정의하는 방식이다.

유클리디언 거리는 **두 점 사이의 거리를 계산 할 때 흔히 쓰는 방법**으로 두 문서 벡터의 상대적인 거리 차를 측정 하는 방법이다.

유클리디언 **유사도는 거리 값이 작을 수록 두 문서가 유사함**을 나타낸다.

#### 유클리디언 거리 계산식

임의의 점  $p = (p_1, p_2, \dots, p_n)$ 와  $q = (q_1, q_2, \dots, q_n)$ 의 유클리디언 거리는 다음과 같이 구한다.

$$\text{euclidean distance} = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$$

즉, 유클리디언 거리 계산식은 p와 q의 대응하는 변량 값의 차가 작으면, 두 관측 대상은 유사하다고 정의하는 식이다.  
변량의 차가 작다는 것은 "**거리가 가깝다**"라는 의미이다.

#### 코사인 유사도

특성 값을 나타내는 벡터 a와 b가 있을 때, 두 벡터 간 각도의 코사인 값을 이용하여 벡터 간의 유사도를 측정하는 방법이다.

코사인은 단위가 1인 벡터 x와 y 사이의 각도(코사인)값에 해당된다.

**벡터 x와 y가 유사한 경우 각도가 작아지고**, 즉 코사인 값이 1에 가까워지고 유사하지 않는 경우 각도가 커진다.

즉 **a와 b의 각도인  $\theta$ 가 최소(0에 가까울수록)일수록 값이 유사하다고 판단**하는 방식이다.

그러면 벡터 a와 b만을 가지고, 어떻게 각도  $\theta$ 를 구할 수 있는가?

벡터의 내적을 사용하면 이  $\theta$ 가 크고 작음을 알아낼 수 있는데 기본 원리는 다음과 같다.

$$\cos(\theta) = a \cdot b / |a| \cdot |b|$$

$a \cdot b$ 은 벡터 a행렬의 각 항의 값  $a_i$ 과 b행렬의 각 항의 값  $a_i$ 를 순차적으로 곱하여 더하면 된다.

$$a \cdot b = (a_1 \cdot b_1 + a_2 \cdot b_2 \cdots + a_n \cdot b_n)$$

$|a|$ 는  $a$  벡터의 길이를 의미하는 데,  $|a| = \sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}$ 의 공식으로 구하면 된다.  
 $|b|$ 는  $|a|$ 와 동일한 개념으로 이해하면 된다.

### 코사인 유사도

벡터  $a$ 와  $b$ 의 코사인 유사도는 다음과 같이 구할 수 있다.

$$\text{cosine similarity} = \cos(\theta) = \frac{a \cdot b}{|a||b|} = \frac{\sum_{i=1}^n a_i \times b_i}{\sqrt{\sum_{i=1}^n (a_i)^2} \times \sqrt{\sum_{i=1}^n (b_i)^2}}$$

### 유클리디언 거리 생성 함수

matrix 객체에 대하여 `dist()` 함수를 사용하면 유클리디언 거리를 생성할 수 있다. (stats 패키지)  
matrix 객체 내의 값이 서로 가까울수록 적은 값으로 나타난다.

항목	설명
사용 형식	<code>dist(x, method, diag)</code>
x	행렬, dataframe
method	거리를 측정하는 방법이다. 'euclidean', 'maximun', 'manhattan', 'canberra' 등등
diag	대각 행렬 출력 여부

### 클러스터링 방법

군집 분석을 위한 클러스터링 방법에는 다음과 같은 방법이 있다.  
표에 `method`는 `hclust` 함수의 `method` 옵션에 사용되는 값을 의미한다.

항목	설명	method
최단 연결법	군집 A와 군집 B에 속하는 데이터 중에서 가장 가까운 데이터들을 군집으로 묶는 방법이다.	single
최장 연결법	군집 A와 군집 B에 속하는 데이터 중에서 가장 먼 데이터들을 군집으로 묶는 방법이다.	complete
와드 연결법	군집을 묶을 때 생기는 새로운 군집의 오차 제곱합을 이용한 방식이다.	ward.D2
평균 연결법	군집 A의 요소와 군집 B의 요소 각각의 거리를 모두 구하여, 이것의 평균 값을 이용하는 방식이다.	average



최단 연결법 클러스터링 예시

실습 파일 : 클러스터링 예시.R

단계 01

데이터 중에서 A와 B와의 거리가 가장 가까워서 둘을 하나의 군집으로 처리한다.

데이터	(x1, x2)	유클리드 제곱 거리	A	B	C	D	E
A	(2, 5)	A	0				
B	(2, 4)	B	1	0			
C	(4, 5)	C	4	5	0		
D	(5, 6)	D	10	13	2	0	
E	(4, 1)	E	20	13	16	26	0

단계 02

데이터 중에서 C와 D와의 거리가 가장 가까워서 둘을 하나의 군집으로 처리한다.

데이터	(x1, x2)	유클리드 제곱 거리	(A, B)	C	D	E	
A	(2, 5)	(A, B)	0				
B	(2, 4)	C	4	0			
C	(4, 5)	D	10	2	0		
D	(5, 6)	E	13	16	26	0	
E	(4, 1)						

단계 03

데이터 중에서 C와 D와의 거리가 가장 가까워서 둘을 하나의 군집으로 처리한다.

데이터	(x1, x2)	유클리드 제곱 거리	(A, B)	(C, D)	E		
A	(2, 5)	(A, B)	0				
B	(2, 4)	(C, D)	4	0			
C	(4, 5)	E	13	16	0		
D	(5, 6)						
E	(4, 1)						

### 단계 04

동일한 방식으로 계속 반복한다.

#### 클러스터 생성 함수

클러스터를 만들기 위해서는 hclust() 함수를 사용하면 된다.

계층적 군집 분석을 위하여 클러스터링을 수행하는 함수이다.

항목	설명
사용 형식	<code>hclust(dist, method='complete', member)</code>
dist	거리 매트릭스를 의미한다. <code>dist()</code> 함수에 의하여 구해진 객체를 사용하면 된다.
method	clustering 방법이다. ward(워드), single(최단 거리법), complete(최장), average(평균), median(중심), centroid(중심)
member=NULL	군집 수

#### 동일 군집에 테두리 입히기 함수

덴드로그램에 다른 군집과 분류하기 위하여 사각형을 그려 준다. (stats 패키지)

즉, 군집 별로 묶음 처리할 때 테두리를 입힐 수 있다.

항목	설명
사용 형식	<code>rect.hclust(hc, k=3, border='red')</code>
hc	hclust 객체이다.
k	클러스터의 갯수이다.
border	테두리의 색상을 결정해주는 Vector 예시 ) <code>border = c('red', 'blue', 'yellow')</code> # 3 개의 그룹에 각각 다른 색상을 지정한다.

#### 군집수 자르기

계층형 군집 분석 결과에서 분석자가 원하는 군집의 개수 만큼 잘라서 인위적으로 군집을 만들 수 있다. (stats 패키지)

`cuttree()` 함수는 군집 수만큼 그룹을 지어 잘라 내어 주는 함수이다.

항목	설명
사용 형식	<code>cuttree(hc, k=3)</code>
hc	hclust 객체(군집 분석 결과)이다.
k	군집의 수이다.

## 계층적 군집 분석

개별 대상 간의 거리에 의하여 **가장 가까이에 있는 대상들로부터 결합해 나가는 방식**이다.

**나무 모양의 계층 구조를 상향식(bottom-up)**으로 만들어 가면서 덴드로그램을 그려 나가는 방식이다.

항목	설명
장점	군집이 형성되는 과정을 파악할 수 있다.
단점	자료의 크기가 큰 경우 분석하기가 어렵다.

## 비계층적 군집 분석

군집의 수가 정해진 상태에서 군집의 중심에서 가장 가까운 개체를 하나씩 포함해 나가는 방법이다.

가장 대표적인 방법으로 k-means clustering이 있다.

k-means clustering는 군집의 초기 값을 지정해주면, 초기 값에서 가장 가까운 거리에 있는 대상을 하나씩 더해 가는 방식으로 군집화를 수행한다.

따라서, 계층적 군집 분석을 통하여 대략적인 군집의 수를 파악하고, 이를 초기 군집 수로 설정하여 비계층적 군집 분석으로 수행하는 것이 효과적이다.

항목	설명
장점	대량의 자료를 빠르고, 쉽게 분류할 수 있다
단점	군집의 수를 미리 알고 있어야 한다.

## 군집화를 위한 K 평균 군집 알고리즘

데이터 과학 주제와 관련된 학회를 조직한다고 가정하자.

3가지 전문 연구 분야(컴퓨터 과학 전문가, 머신 러닝 전문가, 수학 전문가)로 각각의 그룹을 만들려고 한다.

학자들이 발표한 논문 내용을 검토하여 각 학자의 전문 연구 분야를 대략적으로 유추할 수 있다.

학자들에 대한 수집된 데이터를 이용하여 산포도를 다음과 같이 만들어 본다.

K 평균 알고리즘은 군집의 중앙으로, K개의 점을 선택하여 시작한다.

이 점들은 훈련 데이터에서 무작위로 선택하거나, 임의의 점을 선택하기도 한다.

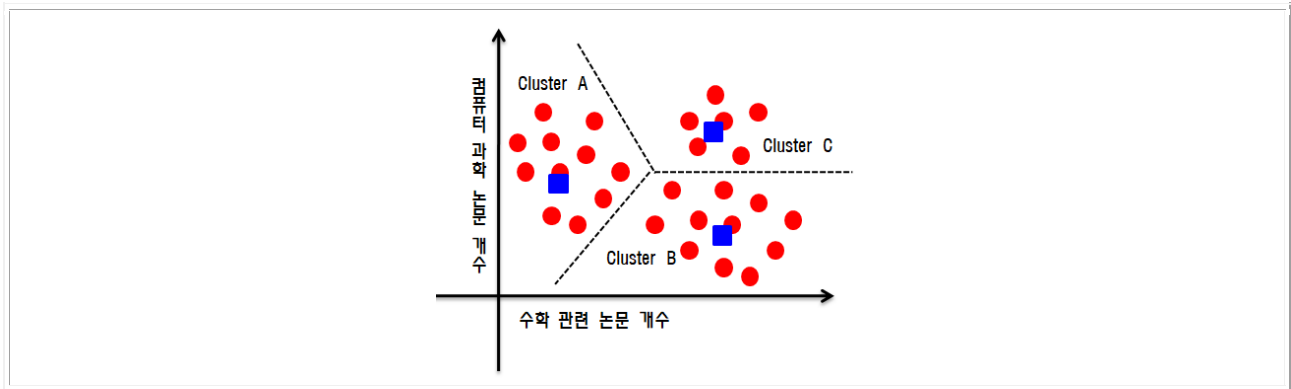
그림에서 보듯이 3개의 군집 중앙은 A, B, C로 레벨된 3개의 영역으로 예제를 나누고 있다.

초기 군집 중앙을 선택한 후 다른 항목들과 거리 함수를 이용하여 거리 측정을 한 다음 가장 가까운 군집에 따라서 클러스터링을 수행한다.

보로다이 다이어그램은 다른 군집의 중앙보다 가까운 영역을 나타낸다.

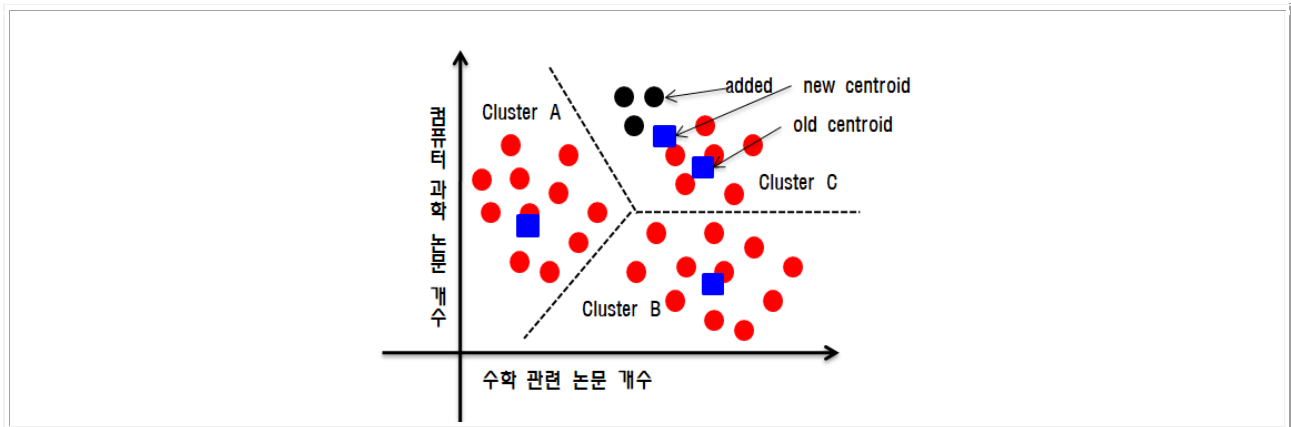
K 평균 알고리즘은 현재 군집에 속해 있는 점들의 평균 값인 중앙점(Centroid)으로 초기 중앙을 옮긴다.

그림에서 파란 색상의 사각형이다.



군집에 새로운 데이터가 추가된다고 가정하자.(added)

그러면 군집의 이전의 중앙점(old centroid)을 다시 재계산하여 새로운 중앙점(new centroid)으로 다시 이동한다.



### 적절한 군집 수 정하기

이상적으로 군집에 대한 사전 지식이 충분하다면, K의 값을 쉽게 정할 수 있다.

예를 들어서, 영화를 분류하고자 할 때 장르의 개수를 안다면 장르의 수 만큼 군집을 나누면 된다.

이전 예시에서 데이터 과학회 문제에서 K는 초대할 연구의 학문적 분야 개수를 반영했다.

전혀 사전 지식이 없다면 데이터 셋의 총 개수  $n$ 의 절반인  $n/2$ 의 제곱근으로 설정하도록 하는 경험 법칙(rule of thumb)을 제안하기도 한다.

적당한 K개를 구하는 통계적 기법이 있다.

엘보우 기법은 여러 개의 K 값에 대하여 동질성과 이질성을 측정하는 방법이다.

일반적으로 동질성은 K의 값이 커질수록 값이 증가하는 반면, 이질성은 그 반대이다.

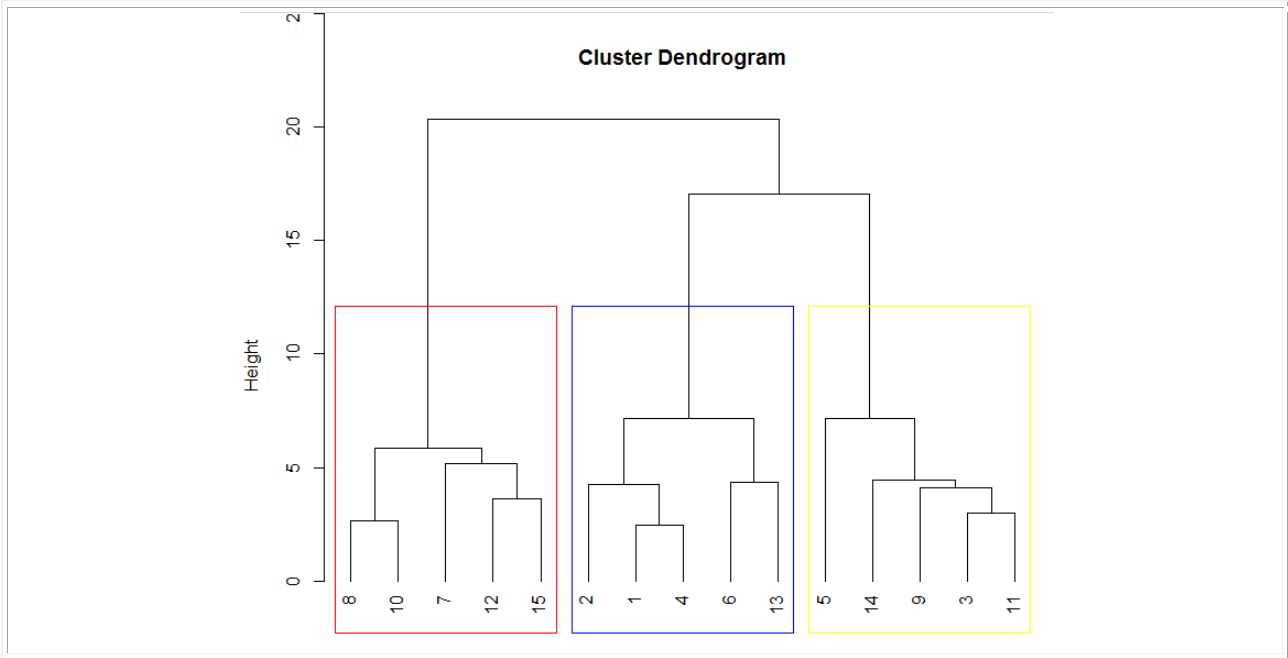
이를 둘다 만족시키는 적정 K 값을 찾으면 된다.

### 군집 분류 함수

군집 분류를 위하여 사용하는 함수는 stats 패키지의 kmeans() 함수를 사용하면 된다.

항목	설명
----	----

사용 형식	kmeans(x=data, centers=3)
x	군집 분류를 위한 데이터
centers	클러스터링할 개수



각 군집별 특징 요약

다음과 같이 각 군집별로 특징들을 간략히 요약해 본다.

구분	제1그룹	제2그룹	제3그룹
요약 통계량	종합 점수 평균 : 71.6 인성 평균 : 9.4	종합 점수 평균 : 75.6 인성 평균 : 14.8	종합 점수 평균 : 62.8 인성 평균 : 11
자격증 유무	자격증 없음	자격증 있음	자격증 있음
군집 특징	종합 점수가 평균 71 점 이상이고, 인성 점수가 10 점 미만으로 모두 불합격 대상자의 군집이다.	종합 점수가 평균 75 점 이상이고, 인성 점수가 10 점 이상으로 모두 합격 대상자의 군집이다.	종합 점수가 평균 70 점 미만이고, 인성 점수가 11 점으로 모두 불합격 대상자의 군집이다.

## 비지도 학습

비지도 학습(unsupervised Learning)은 입력 데이터에 의한 학습을 통하여 사전 지식이 없는 상태에서 컴퓨터 스스로 공통 점과 차이점 등의 패턴을 찾아서 규칙(rule)을 생성하고, 이를 통해서 결과를 도출해내는 방식이다.

## 연관 분석

마트에 가면 맥주 옆에 항상 땅콩이 같이 진열 되어 있다.

인터넷에서 책을 구매하고자 할 때 연관된 도서를 추천하는 경우를 본적이 있을 것이다.

언급한 예시에서와 같이 연관 분석(Association Analysis)은 **항목들간의 관련성을 파악**하여 둘 이상의 항목들로 구성된 **연관성 규칙을 도출하는 탐색적인 분석 방법**이다.

군집 분석에 의해서 생성된 군집(cluster)의 특성을 분석하는 **장바구니 분석**으로 알려져 있다.

주로 마케팅에서 고객의 장바구니에 들어있는 품목 간의 관계를 분석하여 마케팅에 사용할 수 있다.

장바구니 분석 결과는 효과적인 매장 진열, 패키지 상품의 개발, 교차 판매 전략 수립 등에 이용된다.

### 활용 분야 :

고객 대상 상품 추천 및 상품 정보 발송 : A 고객에 대한 B 상품 쿠폰 발송

상점대 상품 진열 및 쇼윈도의 상품 디스플레이

텔레 마케팅을 통해서 패키지 상품 판매 기획 및 홍보

### 활용 예시 :

고객들은 어떤 상품들을 동시에 구매하는가?

맥주를 구매한 고객은 주로 어떤 상품을 함께 구매하는가?

인터넷에서 서적을 구매할 때, 이 독자가 많이 구매한 관련 서적을 소개하기

## 연관 규칙의 평가 척도

연관성을 비교할 수 있는 규칙으로써, 지지도(support), 신뢰도(confidence), 향상도(lift) 등을 평가 척도로 사용한다. 높은 지지도와 신뢰도를 갖는 것은 강한 규칙(strong rule)이라고 한다.

항목	설명
지지도(support)	전체 자료에서 관련 품목의 거래 확률을 의미한다. <b>값이 작다는 것은 다른 항목에 자주 발생하지 않는다는 것을 의미한다.</b> $\text{support} = (\text{상품 A 와 상품 B 를 포함한 거래수}) / (\text{전체 거래수})$
신뢰도(confidence)	A 가 구매될 때 B 가 구매될 확률(조건부 확률) $\text{confidence} = (\text{A 와 B 를 포함한 거래수}) / (\text{A 를 포함한 거래수})$
향상도(Lift)	상품 간의 독립성과 상관성을 나타내는 척도 $\text{Lift} = (\text{신뢰도}) / (\text{B 가 포함될 거래율})$

## 평가 척도의 수치적 의미

항목	설명
높은 지지도	해당 조합의 거래 건수가 다른 것에 비하여 상대적으로 많다는 의미이다.
높은 신뢰도	A 상품 구매시 B 상품을 구매하는 거래수가 많다는 의미이다.
향상도	$= 1$ A와 B는 서로 독립적이다. $> 1$ 맥주는 보통 치킨과 동시에 구입한다. $< 1$ 성경책 구입시 불경책을 사지는 않는다.

다음과 같은 상품 거래 정보가 있다고 가정하자.

참고로 "상품 거래 정보"를 트랜잭션(transaction)이라고 부른다.

이 문제에 대하여 지지도와 신뢰도 향상도를 구해보세요.

### 상품 거래 정보

```
# t1 : 라면, 맥주, 우유
# t2 : 라면, 고기, 우유
# t3 : 라면, 과일, 고기
# t4 : 고기, 맥주, 우유
# t5 : 라면, 고기, 우유
# t6 : 과일, 우유
```

이 항목에 대한 연관 규칙의 평가 척도 결과는 다음과 같다.

A -> B	지지도(support)	신뢰도(confidence)	향상도(Lift)
맥주 -> 고기	① $1/6=0.166$	② $1/2=0.5$	③ $0.5/0.66(4/6)=0.75$
라면, 맥주 -> 우유	④ $1/6=0.166$	⑤ $1/1=1$	⑥ $1/0.83(5/6)=1.2$

위 예시에 대한 규칙을 풀어 보면 다음과 같다.

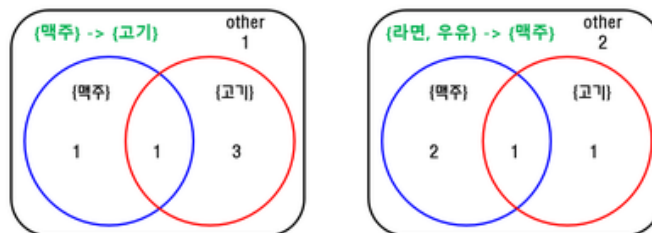
\* 상품 거래 정보

- # t1 : 라면, 맥주, 우유
- # t2 : 라면, 고기, 우유
- # t3 : 라면, 과일, 고기
- # t4 : 고기, 맥주, 우유
- # t5 : 라면, 고기, 우유
- # t6 : 과일, 우유

$$\begin{aligned}
 ① \quad & \frac{\text{맥주 \& 고기 동시 구입}}{\text{전체 거래수}} = \frac{1\text{건}(t4)}{6\text{건}} = \frac{1}{6} \approx 0.1667 \\
 ② \quad & \frac{\text{맥주 \& 고기 동시 구입}}{\text{맥주 구입 건수}} = \frac{1\text{건}(t4)}{2\text{건}(t1, t4)} = \frac{1}{2} = 0.5 \\
 ③ \quad & \frac{\text{신뢰도}}{\text{고기 포함한 거래율}} = \frac{1/2}{4/6} = \frac{6}{8} = 0.75 \\
 ④ \quad & \frac{\text{라\&맥\&우 동시 구입}}{\text{전체 거래수}} = \frac{1\text{건}(t1)}{6\text{건}} = \frac{1}{6} \approx 0.1667 \\
 ⑤ \quad & \frac{\text{라\&맥\&우 동시 구입}}{\text{라\&맥 동시 구입}} = \frac{1\text{건}(t1)}{1\text{건}(t1)} = \frac{1}{1} = 1 \\
 ⑥ \quad & \frac{\text{신뢰도}}{\text{우유 포함한 거래율}} = \frac{1}{5/6} = \frac{6}{5} = 1.2
 \end{aligned}$$

이것은 집합의 의미로 생각하면 이해가 빠르다.

항목	설명
지지도(support)	전체 건수 중에서 A와 B가 모두 포함되어 있는 건수의 비를 말한다. 개념적으로 보면 <b>동시에 존재하는 확률</b> 이다.
신뢰도(confidence)	항목 A를 포함하는 건수 중에서 A와 B가 모두를 포함하고 있는 건수의 비를 말한다. <b>조건부 확률</b> 로써, A사건이 발생 되었다고 가정할 경우에 B의 사건이 발생할 확률이다.



## arules 패키지

어느 상점에 30개의 품목을 판매한다고 하자.

그러면, 발생 가능한 품목의 수는  $2^{30} = 1,073,741,824$ (10억개 이상)이다.

이것은 현실적으로 불가능해 보이는 작업이다.

다행스럽게도 R에서는 arules(Association Rule)라는 패키지를 이용하면 좀 더 수월하게 해결할 수 있다.

arules는 최소 지지도 가지 치기 알고리즘인 Apriori(아프리오리) 알고리즘을 이용하여 연관 규칙을 분석한다.

연관 규칙 검색 공간을 줄이고자 하는 기준선으로, 이전에 믿고 있었던 내용을 다시 사용한다.

예를 들어서, {병문안 카드}와 {꽃}의 빈도가 각각 많았다면 {병문안 카드, 꽃}의 빈도도 역시 많을 것이다. 라는 것이다.



Adult, Groceries 데이터 셋을 제공한다.

또한 as(), labels(), crossTable() 등의 연관 분석에 필요한 여러 가지 함수들을 제공한다.

### apriori 의 원칙 :

빈번한 itemSet 의 부분 집합 역시 빈번해야 한다.

### apriori() 함수

트랜잭션 객체를 대상으로 **연관 규칙을 발견해 주는 함수**이다.

반환 타입은 rules 객체이다.

항목	설명
사용 형식	apriori( data, parameter=list(supp, conf), control = NULL )
data	transaction data 또는 matrix, data frame 등등
controll	통제 계수(control parameter)
parameter	기본 값은 다음과 같다. supp : 규칙을 생성하기 위한 최소한의 지지도(기본 값 : 0.1) conf : 규칙을 생성하기 위한 최소한의 신뢰도(기본 값 : 0.8) maxlen(10) : 규칙에 포함되는 최대 길이(lhs 와 rhs 의 길이를 합친 길이) minlen(1) : 규칙에 포함되는 최소 길이 smax(1)

### image() 함수

image() 함수는 바둑판처럼 격자 형식으로 데이터를 시각적으로 보여준다.

행은 트랜잭션, 열은 항목을 보여 준다.

항목	설명
사용 형식	image(tran)

### inspect() 함수

트랜잭션 객체에 대한 **연관 규칙의 내용을 확인**하기 위한 함수이다.

LHS : 규칙을 구성하는 왼쪽을 의미한다.

RHS : 규칙을 구성하는 오른쪽을 의미한다.

항목	설명
사용 형식	inspect(rules객체) 또는 inspect(tran객체)
rules객체	apriori() 함수가 반환하는 객체를 사용하면 된다.

---

### itemFrequency() 함수

제품이 포함된 거래의 비율을 보여 준다.

항목	설명
사용 형식	itemFrequency(x = groceries[, 1:3])
x	transactions 객체(회소 행렬이 들어 있는)

---

### sort() 함수

rules 객체에 대하여 정렬을 수행해준다.

항목	설명
사용 형식	sort(rules, decreasing=T, by='confidence')
decreasing	정렬 방식을 지정한다( 기본값 : TRUE )
by	정렬을 수행할 항목을 지정한다.( 기본 값은 "support"이다. )

---

### subset() 함수

rules 객체 중에서 특정 조건을 만족하는 항목에 대한 부분 집합을 추출한다.

항목	설명
사용 형식	wmilk <- subset(rules, rhs %in% 'whole milk')
rules	rules 객체이다.
연산자	%in% 'a' : a인 항목을 추출한다. %in% c('a','b') : a또는 b인 항목을 추출한다. %pin% 'a' : 항목에 문자 a를 포함하고 있는 항목을 추출한다.

---

### summary() 함수

item의 수와 트랜잭션의 수를 요약 통계량으로 보여 준다.

항목	설명
사용 형식	summary( tran객체 )

## summary 함수 예시

트랜잭션에 대한 통계량 정보를 보여 주는 함수이다.

```
transactions as itemMatrix in sparse format with
① 9835 rows (elements/itemsets/transactions) and
169 columns (items) and a density of ②0.02609146

most frequent items:
  whole milk other vegetables rolls/buns soda yogurt (Other)
  ③ 2513      1903          1809      1715    1372    34055

element (itemset/transaction) length distribution:
sizes ④
  1  2  3  4  5  6  7  8  9 10 ... 22 23 24 26 27 28 29 32
2159 1643 1299 1005 855 645 545 438 350 ... 11 4 6 1 1 1 3 1

  Min. 1st Qu. Median Mean 3rd Qu. Max.
  1.000 2.000 3.000 4.409 6.000 32.000
```

- ① 데이터가 9835행 169열을 가지고 있다.  
 ② a density of 0.02609146는 매트릭스에서 0이 아닌 칸의 비율을 말한다.  
 따라서, 영업 시간 중 거래건수 = 9835(행) \* 169(열) \* 0.02609146 = 43,367 개이다.  
 거래 건수당 평균 구매 개수는 43,367 / 9835 = 4.409456024개이다.
- ③ 가장 빈도가 많은 품목은 전지 우유(whole milk)로써 2513건이다.  
 whole milk의 거래 비율은 2513 / 9835 = 0.255516014 = 25.55%이다.  
 즉, 거의 4회당 1번은 판매가 된다고 볼 수 있다.
- ④ 물건 1개만 팔린 건수가 2159건이다.  
 물건 32 개가 팔린 적이 단 1 건이 있었다.

## transactions 함수

트랜잭션 객체를 생성해주는 함수이다.

거래한 데이터에 대하여 희소 매트릭스를 생성해준다.

항목	설명
사용 형식	read.transactions( file, format=c('basket', 'single'), sep=NULL, cols=NULL, rm.duplicate=FALSE, encoding='unknown' )
file	분석 수행을 하기 위한 파일
foramt	트랜잭션 데이터 셋의 형식 single : 트랜잭션 구분자(transactionID)에 의해서 상품(item)이 대응된 경우 basket : 여러 개의 상품(item)으로 구성된 경우
sep	각 상품을 구분하는 구분자

cols	single 인 경우 원의 컬럼 수를 지정(basket 은 생략)
rm.duplicate	중복된 트랜잭션의 상품(item)들을 제거한다.
encoding	데이터 셋의 인코딩 방식 지정

### 출력 결과 예시

다음 예시는 해당 거래 건수가 6건이고, 총 5개의 거래 품목이 있다는 의미이다.  
매트릭스의 각 칸은 구매한 제품이면 숫자 1을, 그렇지 않으면 0의 값으로 채워진다.

#### 결과 예시

transactions in sparse format with  
6 transactions (rows) and  
5 items (columns)

## arulesViz 패키지

연관성 규칙에 대한 데이터를 시각화하기 위한 패키지이다.

### plot 함수

연관 규칙(association rules)과 itemsets을 시각화하기 위한 함수이다.

항목	설명
사용 형식	plot(rules, method='graph', control=list(type='items'))
rules	rules 객체
method	그리는 방법을 지정하는 데, 하단의 세부 설명을 참조하도록 한다.
control	control = list(type = "itemsets")  control = list(type = "items")  # 마지막으로 method 인자를 "paracoord"로 하면 각 물품 간의 연관관계를 병렬적으로 확인할 수 있다. 가로축의 숫자는 조건(LHS)의 물품 수를 의미한다. plot(dvd.rules, method = "paracoord", control = list(reorder = TRUE))

### plot() 함수의 method 옵션

method 옵션	설명
graph	graph : 큰 원과 작은 원을 이용하여 그림을 그려 준다.

	<p>물품들 간의 연관성을 그래프로 보여준다.</p> <p>화살표의 두께는 <b>지지도</b>를 화살표의 색상의 진하기는 <b>향상도(lift)</b>를 나타낸다.</p>
grouped	<p>좌측을 연관 규칙의 조건(LHS)과 결과(RHS)를 우측으로 하여 그래프를 보여 준다.</p> <p>원의 크기는 각 규칙의 지지도(support)를 의미하고, 색상의 진하기는 향상도(lift)를 의미한다.</p> <p>조건(LHS) 이름 앞의 숫자는 그 조건으로 되어 있는 연관 규칙의 수를 의미한다.</p> <p>조건(LHS)에 "+"와 함께 표시된 숫자는 표시가 생략된 물품 수를 의미한다.</p>
scatterplot	<p>지지도와 신뢰도를 산점도로 보여 준다.</p> <p>x 축은 지지도(support), y 축은 신뢰도(confidence), 색상은 향상도(Lift)를 의미한다.</p>

## SVM(Support Vector Machine)

카페 문서 : 1203

SVM은 서로 다른 분류에 속한 데이터 간에 간격이 최대가 되는 선(또는 평면)을 찾아 이를 기준으로 데이터를 분류하는 모델이다.

즉, 선을 구성하는 매개 변수를 조정하여 요소들을 구분 짓는 선을 찾고, 이것을 기반으로 pattern을 인식하는 기법이다. 식별 선이나 평면에서 패턴들과의 거리(마진)를 최대로 만들어 내는 것이 목표이다.

### 사용처

유전자 데이터의 분류(classification)  
 텍스트 분류(주제별 분류, 언어의 식별)  
 회귀 분석(regression)

기본적으로 지도 학습 알고리즘이며, HyperPlane(조평면)을 이용하여 카테고리들을 나눈다.

SVM을 사용하게 되면 알 수 없는 패턴도 제대로 분류할 확률이 굉장히 높다.(일반화 능력이라고 부른다.)

## 제한/참고 사항

SVM 학습기가 **처리하는 모든 속성은 수치 값**이어야 한다.

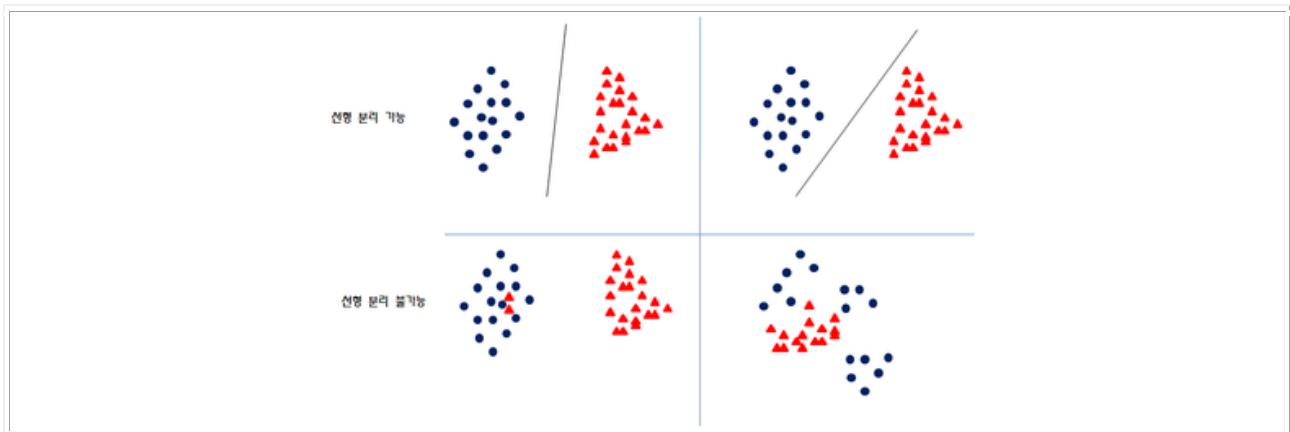
모든 속성들의 수치 값의 분포가 너무 크지 않아야 하고, 이런 경우에는 반드시 정규화나 표준화를 수행할 필요가 있다.

## 선형 분리의 가능/불가능

선형 분리란 임의의 공간에 직선을 그어서 서로 다른 분류를 분리시키는 것을 말한다.

예를 들어 다음 그림에서 상단의 두 가지 예시는 선형 분리가 가능한 경우이다.

하단의 두 가지 예시는 선형 분리가 불가능한 예시이다.



## 가장 최적화된 분리선은?

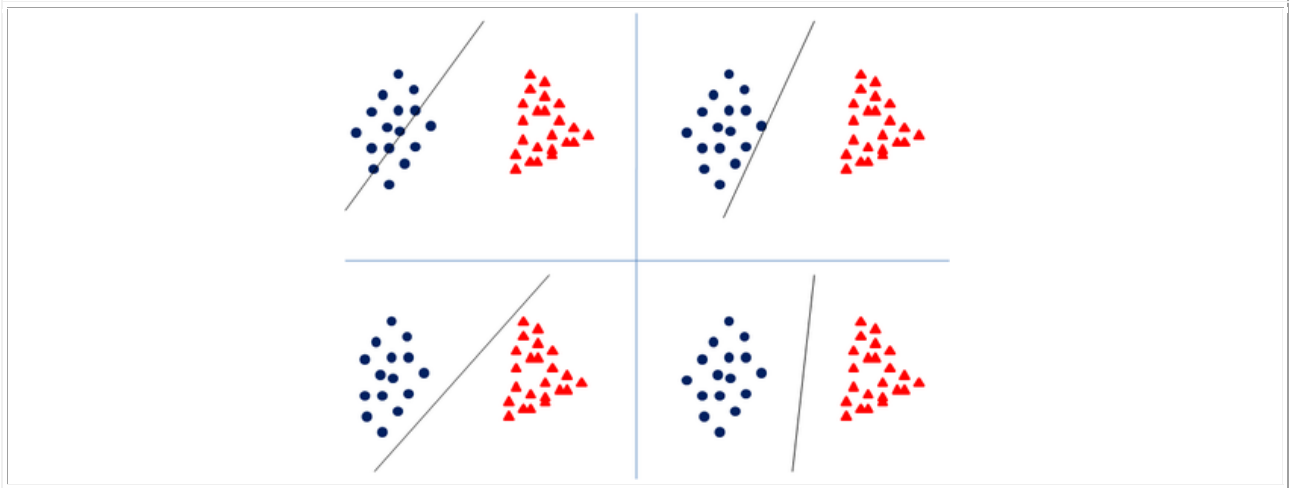
다음 4개의 그림에서 **가장 최적으로 분류를 한 항목은** 어느 것인가?

왼쪽 상단을 제외한 3개의 그림 모두 선형 분리를 하고 있다.

하지만, 가장 잘 분리한 것은 우측 하단의 그림이다.

왜냐하면, SVM은 최고의 마진을 가져 가는 방향으로 분류를 수행(**최대 마진화 방침**)해야 하기 때문이다.

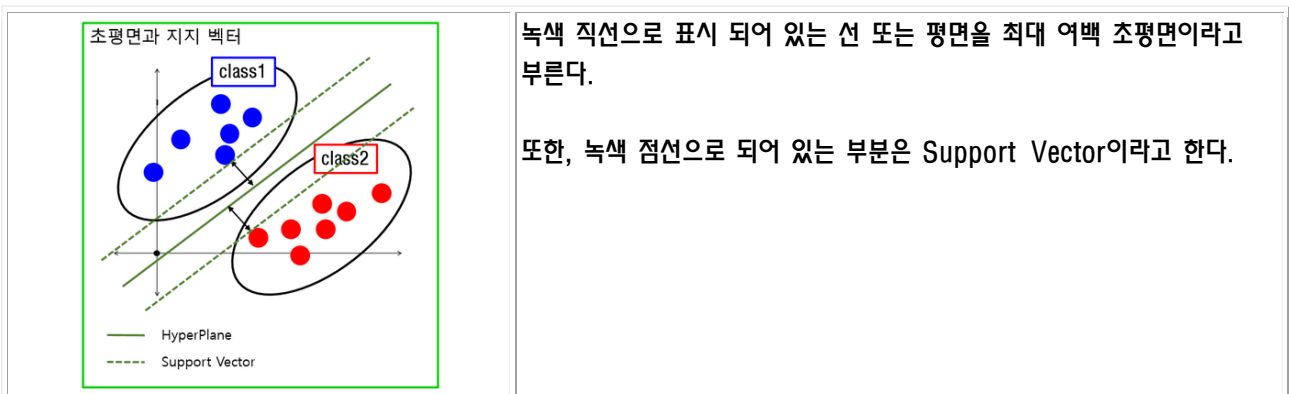
그래야만, 차후에 새로운 데이터가 들어 오더라도 **잘 분류될 가능성이 커지기 때문**이다.



### 초평면과 지지 벡터

그림에서 두 분류(파란 색과 빨간 색)를 할 수 있는 직선은 엄청 많다.

하지만, 두 개의 분류에서 가장 멀리 떨어진 구분선을 "최적 선형 구분자"라고 할 수 있다.



### 참조 문서

[https://en.wikipedia.org/wiki/Support\\_vector\\_machine](https://en.wikipedia.org/wiki/Support_vector_machine)

<http://operatingsystems.tistory.com/128>

### 용어 정리

SVM과 관련된 용어를 정리해본다.

항목	설명
최대 여백 초평면	Maximum Margin Hyperplane 한쪽면으로 동일한 데이터가 놓이게 하는 선형 경계면을 말한다.

	각 분류에 속하는 데이터로부터 같은 간격으로, 그리고 최대로 멀리 떨어진 선(2차원) 또는 평면(3차원)을 말한다. 이 평면(또는 직선)이 분류를 나누는 기준이 된다.
서포트 벡터	최대 여백 초평면과 가장 가까운 각 분류에 속한 점들을 말한다. margin을 구하는데 supporting을 하기 때문에 support vector라고 부른다.
Kernel Trick	低차원에서는 선형적으로 구분이 안되지만, mapping(사상)을 통하여 高차원적으로 변형시키면 선형적으로 구분이 가능하도록 하는 기법
Kernel Function	Kernel Trick에 사용되는 함수를 말한다. 커널 함수의 예는 다항(Polynomial) 커널과 가우시안(Gaussian) 커널이었다.
Convex hull	각 그룹의 데이터 점들의 가장 외곽 경계를 말한다. 두 컨벡스 홀의 거리가 가장 짧은 지점을 수직 이등분한다.

## 커널 트릭

비선형적인 데이터를 구분하고자 할 때 특성을 추가하면 선형적으로 구분할 수 있다.

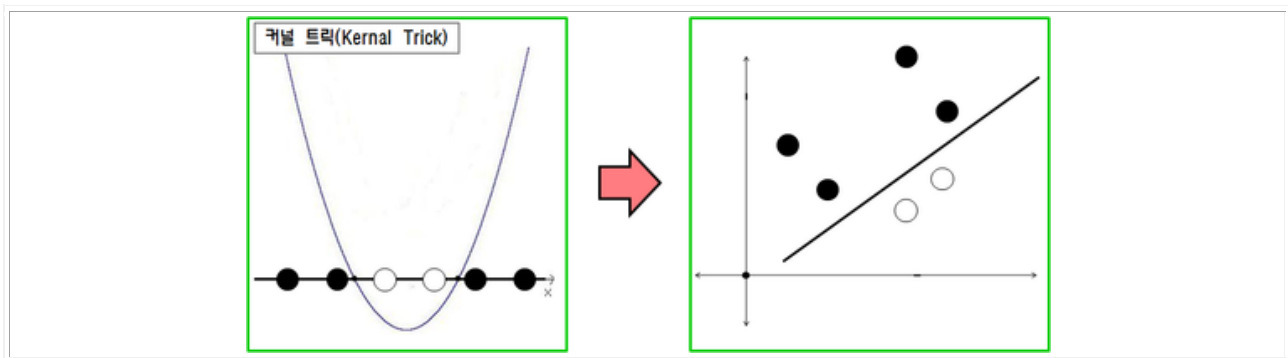
하지만 낮은 차수의 다항식은 매우 복잡한 데이터 셋을 잘 표현하지 못하고, 높은 차수의 다항식은 굉장히 많은 특성을 추가하므로 모델을 느리게 만든다.

특성을 추가하지 않고, 마치 특성을 추가한 듯한 효과를 낼수 있는 기법이 커널 트릭(kernel trick)이다.

선형적으로 구분이 되지 않는 데이터를 분류하기 위해서는 커널 트릭을 사용한다.

주어진 데이터를 고차원으로 옮긴 뒤 변환한 차원에서 서포트 벡터 머신을 이용하여 초평면을 찾는 것이다.

다음 그림과 같이 1 차원의 데이터를 2 차원으로 옮기면 최적의 직선을 찾을 수 있다.



## 커널 트릭의 사상(mapping)

저차원에서는 선형적으로 구분이 불가능하지만, mapping(사상)이라는 개념을 통하여 고차원으로 변형 시키면 선형적으로 구분이 가능하도록 하는 기법이다.

이때 사용되는 함수를 커널 함수라고 부른다.

커널 함수의 대표적인 예시는 다항 커널과 가우시안 커널 방식이 있다.



### R에서 사용하는 SVM 패키지

R로 SVM 모델을 적합화하려 할 때 선택할 수 있는 몇 개의 패키지가 있다.

패키지	설명
e12071	비엔나 기술 대학 통계학과 효율적인 SVM 구현체로 잘 알려진 LIBSVM을 R에서 사용할 수 있도록 한 패키지이다. C++로 작성된 오픈 소스 SVM 프로그램이다.  참조 : <a href="https://www.csie.ntu.edu.tw/~cjlin/libsvm/">https://www.csie.ntu.edu.tw/~cjlin/libsvm/</a>
klaR	도르트문트 기술 대학의 SVMlight 알고리즘  참조 : <a href="http://svmlight.joachims.org/">http://svmlight.joachims.org/</a>
kernlab	커널 기반의 기계 학습 알고리즘을 R에서 구현한 것이다. C++ 코드의 수정 없이 손쉽게 기능을 확장할 수 있다.  SVM 모델의 여러 가지 자동화된 기법을 사용해 훈련할 수 있는 caret 패키지와 같이 사용할 수 있다.  참조 : <a href="http://www.jstatsoft.org/v11/i09">http://www.jstatsoft.org/v11/i09</a>

### 서포트 벡터 머신 문법

서포트 벡터 머신 알고리즘을 수행하려면 kernlab 패키지의 ksvm() 함수를 사용하면 된다.

항목	설명
사용 형식	letter_classifier <- ksvm( target ~ predictors, data, kernel, C, scaled = TRUE, kpar = "automatic" )

target	모델링하고자 하는 데이터, 즉 종속 변수를 의미한다.
predictors	예측에 사용하고자 하는 독립 변수의 모음의 지정한다.
data	학습을 수행하기 위한 훈련용 데이터 셋을 의미한다.
kernel	학습과 예측에 사용할 커널 함수를 명시한다. vanilladot(Linear kernel) : 커널 트릭 없이 단순히 벡터의 내적만을 이용하는 방식이다. rbfdot(Radial Basis kernel "Gaussian") : 가우시안 커널이라고 한다.(기본 값이다.) polydot(Polynomial kernel) : 다항 커널을 사용한다. tanhdot(Hyperbolic tangent kernel) 등과 같은 알고리즘 등이 있다.
C	cost(C로 지정)는 과적합을 막기 위한 정도를 지정하는 파라미터이다. 적절한 C를 구하기 위해서는 tune() 라는 함수를 사용하여 조정을 하면 된다.
scaled	데이터를 정규화할지의 여부를 결정한다. 기본 값인 TRUE는 평균 0, 분산 1이 되도록 데이터를 변경해준다. 이것은 표준화 또는 z-score 정규화라고 부른다.
kpar	커널 파라미터를 list 형식으로 작성할 수 있다. 다음 예시는 degree를 사용하여 3차 다항 커널을 사용하는 예시이다.

예측하기 위하여 사용하는 함수이다.

항목	설명
사용 형식	letter_predictions <- predict( model, testing, type)
model	ksvm() 함수를 사용하여 훈련된 모델을 의미한다.
testing	예측을 수행할 검증용(testing) 데이터 셋을 의미한다.
type	예측이 'response'(예측된 범주)인지, 'probabilities'(예측된 확률)인지 명시하도록 한다.

## 서포트 벡터 머신 문법

### svm() 함수

e1071 패키지의 svm() 함수는 서포트 벡터 머신 알고리즘을 수행해준다.

항목	설명
----	----

사용 형식	svm( formula, data=NULL, scale = TRUE, type=NULL, kernel='radial', gamma, cost=1)
formula	모델을 위한 포물러
data	데이터
scale	정규화를 수행할것지의 여부
type	분류 회귀 등의 모델 중에서 만들 모델을 지정한다. 다음과 같은 항목들이 존재한다. C-classification, nu-classification, one-classification (for novelty detection), eps-regression, nu-regression
kernel	커널 함수를 명시한다.
gamma	커널 파라미터 gamma
cost	커널 파라미터 cost

### tune() 함수

그리드 탐색을 사용한 파라미터 튜닝을 수행한다.

항목	설명
사용 형식	tune( method, train.x , train.y, data )
method	최적화할 함수를 지정한다.
train.x	포물러 또는 독립 변수의 행렬을 지정한다.
train.y	예측할 분류, 만일 train.x가 포물러이면 무시된다.
data	포물러를 적용하고자 하는 데이터 셋이다.
gamma	최적화하고자 하는 gamma 값
cost	최적화하고자 하는 cost 값을 의미한다. cost는 과적합을 막는 정도를 지정하는 파라미터이다.

### attributes() 함수의 반환 값

\$names

```
[1] "best.parameters" "best.performance" "method"          "nparcomb"        "train.ind"
[6] "sampling"        "performances"     "best.model"
```

\$class

```
[1] "tune"
```

## KNN

KNN(K-Nearest Neighbor) 알고리즘은 범주를 모르는 어떠한 데이터에 대하여 분류 되어 있는 가장 유사한 예제의 범주로 지정해주는 알고리즘이다.

입력 데이터와 유사한 K개의 데이터를 구하고, 그 K개 데이터의 분류 중 가장 빈도가 높은 클래스를 입력 데이터의 분류로

결정하는 알고리즘이다.

참조 문서 : <https://kkokkilkon.tistory.com/14>

### 사용 예시

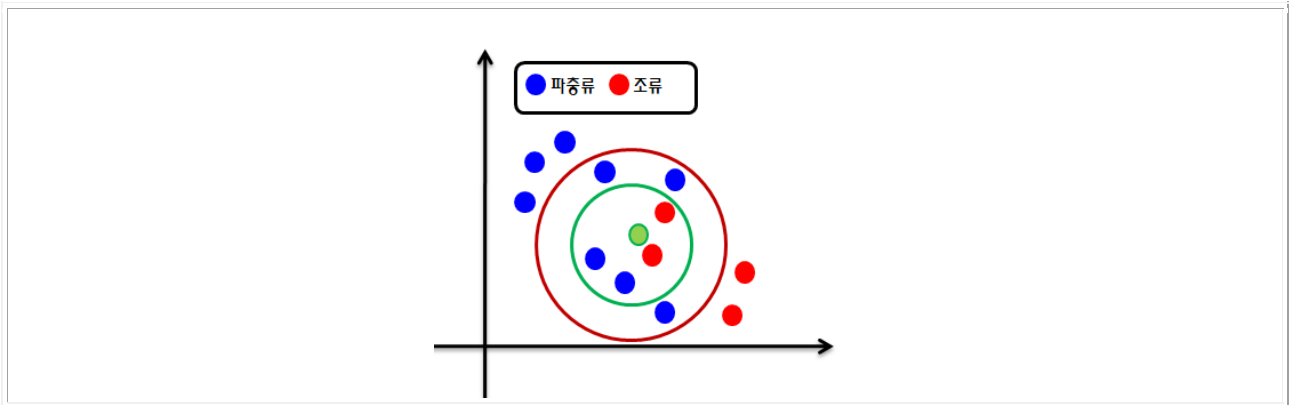
개인별 추천 영화 예측하기

유전자 데이터의 패턴 식별하기

얼굴과 글자를 인식하는 컴퓨터 비전 application

### KNN 이해하기

다음 그림에서 파란색은 파종류로 분류된 데이터이고, 빨간색은 조류로 분류된 데이터이다.



녹색으로 되어 있는 새로운 데이터가 파종류인지 조류 인지 분류를 하고자 한다.

이를 분류할 수 있는 방법 중의 하나는 가장 가까이 있는 항목으로 분류하는 방법이다.

이러한 기법을 **Nearest Neighbor**이라고 한다.

예시에서는 빨간색이 가장 가까우므로 조류라고 볼 수 있다.

하지만, 실제 데이터 분포를 보면 파란색이 훨씬 더 많고 범위를 조금만 더 넓혀 보면 파란색이 더 많다.

녹색 원으로 되어 범위까지 확대해 보면 파란색과 빨간색의 개수가 동일하다.

좀더 확장하여 갈색 원까지 범위를 넓혀 보면 파란색이 더 많다.

이와 같이 주어진 개수(K 개) 만큼 가까운 멤버들과 비교하여 판단하는 방법을 KNN(K-Nearest Neighbor) 알고리즘이라고 한다.

추가적으로 고려해야 할 사항으로 가까운 거리에 있는 항목은 가중치를 크게, 멀리 있는 항목은 가중치를 적게 적용하는 방식도 있다.

이 방식은 수정된 KNN이라고 부른다.

### K 값의 선택

실제로 K의 값은 학습해야 할 데이터의 개수에 달려 있다.

보통 3~10 사이에서 결정한다.

일반적인 방법으로 훈련 데이터 개수의 제곱근으로 설정하기도 한다.

이러한 방법이 항상 최적의 결과를 만들지 못할 수도 있다.

다양한 테스트 데이터 셋에 대하여 일부 K 값을 테스트하여 최적의 분류 성능을 내는 K 값을 선택하는 것이다.

### KNN을 사용하기 위한 데이터 준비

사전에 고려해야 할 사항은 각 속성들의 값의 분포를 상대적으로 균등하게 하기 위하여 최대-최소 정규화(Min-Max Normalization) 또는 Z 점수 표준화(Z-Score Standardization)를 수행하는 것이 좋다.

### knn 함수

class 패키지에 들어 있는 knn() 함수는 데이터에 들어 있는 각 인스턴스에 대하여 유클리드 거리를 사용하여 k 근접 이웃을 찾는다.

k 개의 인스턴스에 다수결의 '투표'로 분류를 한다.

만약 동수 투표인 경우에는 임의로 선택을 하게 되는데, 일반적으로 홀수 정수 값을 사용하여 동수가 나오지 않도록 한다.

항목	설명
사용 형식	knn(train = wbcd_train, test = wbcd_test, cl = wbcd_train_labels, k=21)
train	훈련용 데이터 프레임
test	테스트용 데이터 프레임
cl	훈련용 데이터의 각 행에 대한 정답(label)을 가지고 있는 벡터
k	최근접 이웃의 수를 명시하는 정수 값이다. 일반적으로 권장하는 수는 training 데이터 셋에 루트를 씌운 값에 홀수의 정수 값을 사용하기를 권장한다. 권장하는 방법이지, 이것이 가장 좋은 효율이 나온다고 장담은 못한다. 사용 예시 k_size <- floor(sqrt(training_row)) k_size <- ifelse(k_size %% 2 == 0, k_size + 1, k_size)

### 파이썬 프로그래밍

scikit learn에서는 KNeighborsClassifier로 구현이 된다.

## 나이브 베이즈

18세기의 수학자 토마스 베이즈가 제안한 기법이다.

속성을 사용하여 분류를 하고자 할 때 확률의 원리를 활용하는 알고리즘이다.  
미래의 사건 확률을 추정하기 위하여 과거의 데이터를 활용한다.  
조건부 확률이라는 개념과 베이즈 정리를 이용하여 만든 알고리즘이라고 이해하면 된다.  
카페 실습(2205), 이론(2291, 2242)

### 나이브 베이즈의 사용처

#### 사용처

스팸 메일 분류기  
저자 식별이나 문서 분류  
질병에 대한 진차(관찰된 증상)

### 베이즈 기법의 기본적인 개념

베이즈 확률 이론은 유사한 증거를 기반으로 한 사건의 유사성을 추정하는 개념에 근거를 두고 있다.  
사건(event)이란 확률을 계산하기 위한 결과의 집합을 의미한다.  
예를 들어, 확장하거나 비가 올 날씨, 동전의 앞면과 뒷면 등이다.

### 확률(Probability)

사건의 확률은 사건이 일어난 시도의 숫자를 중 사건의 수로 나눈 수를 의미한다.

용어	설명
확률	사건이 일어난 도수를 총사건의 수로 나눈 값을 말한다. 모든 확률의 총합은 1 이다.  0%라 함은 절대로 일어나지 않는 확률을 의미한다. 100%라 함은 이 사건은 반드시 일어난다는 의미이다.
베이즈 확률 이론	유사한 증거를 기반으로 한 사건성의 유사성을 추정한다.
독립 사건	두 개의 사건이 <b>전혀 연관성이 없는 사건</b> 을 말한다. 동전 던지기와 내일 날씨와는 전혀 상관성이 없다. $P(A \cap B) = P(A) * P(B)$
종속 사건	두 개의 사건이 <b>서로 연관성이 있는 사건</b> 을 말한다. 구름이 낀 걸 보니 비가 올 것 같다. Vigra 라는 단어가 있는 것을 보니 이 메일은 스팸일 확률이 높다.

### 확률의 표기법

다음과 같은 도서 대출 현황표가 있다고 가정하자.

예시 : 도서 대출 현황

전체 고객 수 : 100

Tensorflow 책을 구매한 고객 수 : 50

Database 책을 구매한 고객 수 : 20

두 권 모두를 구매한 고객 수 : 10

이에 대한 확률 표기법은 다음과 같은 항목들이 있다.

$P(B|A)$ 는 사건 A 가 발생했다는 전제 하에서의 사건 B 의 확률을 의미한다.

확률 표현	수식 표현	설명
$P(A)$	50/100	Tensorflow 책을 구매한 확률
$P(B)$	20/100	Database 책을 구매한 확률
$P(B A)$	10/50	Tensorflow 책 구매자 중에서 Database 책도 같이 구매한 확률
$P(A B)$	10/20	Database 책 구매자 중에서 Tensorflow 책도 같이 구매한 확률
$P(\sim B)$	$(100-20)/100$	Database 책을 구매하지 않을 확률

### 스팸 필터

다음과 같이 전체 메일 중에서 스팸 메일을 걸러 주는 Spam-Filter 기가 있다고 가정하자.

또한 전체 메일에서 스팸 메일은 30%라고 가정하자.

사건 A 를 스팸 메일이라고 한다면 표기법은 다음과 같다.

표기법	설명
$P(A) = 0.3$	스팸 메일의 확률은 30%이다.
$P(\sim A) = 0.7$	스팸 메일이 아닐 확률은 80%이다.

이메일에는 '비아그라'라는 단어가 들어 있을 수 있다.

전체에서 스팸 메일은 30%를 차지하고, '비아그라'라는 단어가 들어 있는 메일은 20%이다.

스팸 메일이면서, '비아그라'라는 단어가 들어 있는 메일은 전체의 10%이다.

### 현황

전체 메일 수 : 100

스팸 메일 수 : 30

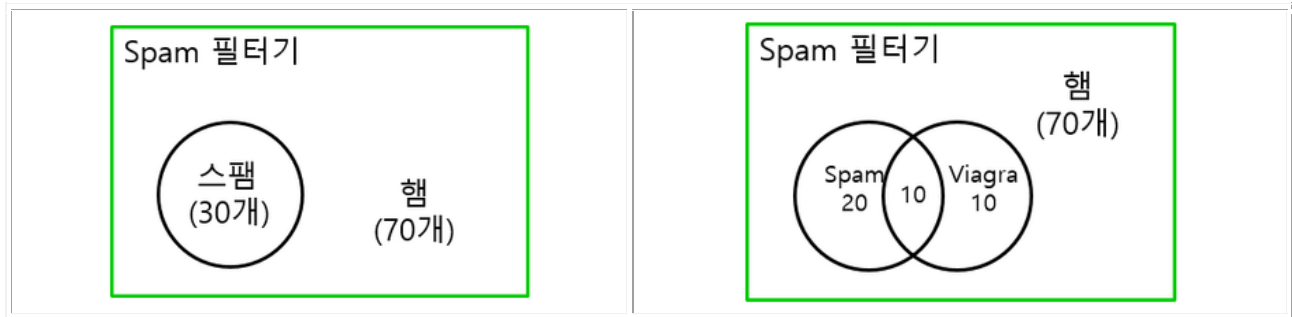
Viagra라는 단어가 들어 있는 메일 수 : 20

스팸 메일 중에서 Viagra라는 단어가 들어 있는 메일 수 : 10

### 관련 그림

'비아그라'라는 단어가 있다고 해서, 모든 메일이 'Spam'은 아니다.

스팸 메일이 무조건 'Vigra'라는 단어가 포함하고 있지 않다.



### 확률 표기

모든 사건들에 대한 확률은 다음과 같다.

확률 표현	수식 표현	설명
P(S)	30/100	스팸 메일의 확률
P(V)	20/100	Viagra 라는 단어가 들어 있는 메일의 확률
P(VIS)	10/30	스팸 메일에서 Viagra 라는 단어가 들어 있는 메일의 확률
P(SIV)	10/20	Viagra 라는 단어가 들어 있는 메일 중에서 스팸 메일의 확률

### 조건부 확률

조건부 확률이란 어떠한 사건 A가 일어났다는 전제하에서 다른 사건 B가 발생할 확률을 의미한다.

수식으로는  $P(B|A)$ 라고 표현한다.

확률의 기본 정리인 곱셈 정리를 응용한 것이다.

#### 조건부 확률 공식

$P(A | B)$ 는 사건 B가 일어났다는 전제하에서, 사건 A가 일어날 확률을 의미한다.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} = \frac{P(A \cap B)}{P(B)}$$

주사위를 두 번 던져서 "첫번째가 3", "두번째가 짝수"가 될 확률은 다음과 같다.

3은 확률이 1/6이고, 짝수는 3/6=1/2이므로, 곱셈 법칙에 의하여  $1/6 * 1/2 = 1/12$ 이다.

만약 두 사건이 독립 사건이라면 다음과 같은 공식이 성립한다.

$$P(A \cap B) = P(A) * P(B)$$

$P(A \cap B)$ 는 두 개의 사건이 동시에 일어날 확률을 의미한다.

### 베이즈 이론(정리)

베이즈 정리는 "조건부 확률"과 관련된 이론으로, 토머스 베이즈에 의하여 정립된 이론이다.

예시 : 서적 구매량에 대한 베이즈 정리



Tensorflow 책을 구매했다는 전제하에서 Database 책까지 구매한 확률은 다음과 같다.

$$P(A) * P(B|A) = 50/100 * 10/50 = 10/100 = 1/10$$

Database 책을 구매했다는 전제하에서 Tensorflow 책까지 구매한 확률은 다음과 같다.

$$P(B) * P(A|B) = 20/100 * 10/20 = 10/100 = 1/10$$

위 두 가지의 결과는 다음과 같이 동일하다.

$$P(A) * P(B|A) = P(B) * P(A|B)$$

### 예시 : 스팸 메일에 대한 베이즈 정리

스팸 메일 중에서 Viagra 라는 단어가 들어 있는 메일의 확률은 다음과 같다.

$$P(S) * P(VIS) = 30/100 * 10/30 = 10/100 = 1/10$$

Viagra 라는 단어가 들어 있는 메일중에서 스팸 메일의 확률은 다음과 같다.

$$P(V) * P(SIV) = 20/100 * 10/20 = 10/100 = 1/10$$

위 두 가지의 결과는 다음과 같이 동일하다.

$$P(S) * P(VIS) = P(V) * P(SIV)$$

### 스팸 메일 분류기 베이즈 이론

베이즈 이론이 어떻게 동작하는지 이해하기 위하여 수신 받은 이메일이 스팸일 확률을 구한다고 가정하자.

다음 용어들을 우선 정리해보자.

항목	설명
사전(Prior) 확률	이미 알고 있는 이전 사건들에 대한 확률. 이전 메일이 Spam 메일인 확률을 의미한다.
우도(Likelihood)	이미 알고 있는 사건들이 발생했다는 전제 하에서, 다른 사건이 발생할 확률 Vigra라는 단어가 Spam 메일에 사용 되었을 확률을 의미한다.
주변(Marginal) 우도	모든 메일에 Vigra가 나타난 확률을 의미한다.
사후(Posterior) 확률	사전 확률과 우도 확률을 통해서 알게 되는 조건부 확률 베이즈 이론을 적용하여 이 메일이 Spam 메일일 확률을 의미한다.

다음 공식은 스팸 메일 분류기에 대한 그림이다.

$$P(\text{Spam} \mid \text{Vigra}) = \frac{P(\text{Vigra} \mid \text{Spam}) * P(\text{Spam})}{P(\text{Vigra})}$$

사후(Posterior) 확률 ←  $P(\text{Spam} \mid \text{Vigra})$

우도(Likelihood) ←  $P(\text{Vigra} \mid \text{Spam})$

사전(Prior) 확률 ←  $P(\text{Spam})$

주변(Marginal) 우도 ←  $P(\text{Vigra})$

위의 용어에서 사후 확률이 50%보다 크면 이 메일은 Spam 메일로 보는 것이다.

$$\text{사후 확률} \geq 0.5 \quad \text{Spam 메일}$$

$$\text{사후 확률} < 0.5 \quad \text{일반 메일}$$

### 빈도 표와 우도 표

베이즈 이론의 요소를 계산하기 위하여 Spam과 Ham 메일에서 Vigra가 나타나는 회수를 빈도 표(Frequency Table)로 작성할 수 있다.

이원 교차표 형식으로 작성하는데, 행에는 범주 변수(Spam과 Ham)로 명시한다.

열에는 속성에 대한 등급(Vigra에 대한 yes/no)을 표시한다.

	Vigra			
빈도(Frequency)	yes	no	sumtotal	
Spam	10	20	30	
Ham	10	60	70	
sumtotal	20	80	100	

빈도 표는 다음과 같이 우도 표(Likelihood Table)를 구성하는 데 사용한다.

	Vigra			
우도(Likelihood)	yes	no	sumtotal	
Spam	10/30	20/30	30	
Ham	10/70	60/70	70	
sumtotal	20/100	80/100	100	

우도 표를 살펴 보면  $P(\text{Vigra} \mid \text{Spam}) = 10/30$ 이며, Vigra가 단어가 포함되어 있는 Spam 메일이 1/3임을 나타내고

있다.

$P(\text{Spam}/\text{Vigra})$  사후 확률을 계산해 보면,  $P(\text{Vigra}/\text{Spam}) * P(\text{Spam}) / P(\text{Vigra}) = (10/30) * (30/100) / (20/100) = 1/2 = 0.5$ 이다.

그러므로, 메일에 Vigra라는 단어가 포함이 되었을 경우 Spam 메일일 확률이 50%라는 것이다.

데이터의 개수가 많아지면 이 메일이 Spam일 확률은 높아 질 것이다.

### 라플라스 추정기

발생할 확률 값을 0이 되지 않도록 각 빈도의 값에 적은 수를 추가 하는 개념이다.

일반적으로 적어도 한 번 각 범주에 속한 것처럼 라플라스 추정기는 1로 설정한다.

### 패키지 설치

나이프 베이즈와 관련된 패키지는 'e1071'이다.

다음과 같이 설치하도록 한다.

참고로 klaR 패키지의 NaiveBayes()도 존재한다.

```
install.packages('e1071')  
library(e1071) # 나이브 베이즈 구현 패키지
```

### naiveBayes 함수

나이브 베이즈 분류기를 만들어 준다.

항목	설명
사용 형식	<code>someObj &lt;- naiveBayes(training, class, laplace=0)</code> 반환되는 객체(someObj)는 "naiveBayes" 타입이다.
training	훈련 데이터(training)가 포함되어 있는 dataframe 또는 matrix
class	훈련 데이터의 각 행에 대한 범주 정보를 담고 있는 팩터 벡터이다. 즉, 정답을 가지고 있는 label 이다.
laplace	프랑스 수학자 피에르 시몬 라플라스가 제안한 개념이다. 각 범주의 발생 확률이 0 이 되지 않도록 빈도 표의 각 값에 작은 수를 추가하는 기법이다. 라플라스 추정기를 제어하는 숫자이다. 일반적으로 적어도 1 번 정도는 각 범주에 속한 것처럼 값을 1 로 설정한다.

### predict 함수

나이브 베이즈 분류기에 대한 예측을 수행해주는 함수이다.

항목	설명
사용 형식	<code>pred &lt;- predict( navObj, testing[, type])</code>
navObj	naiveBayes 함수를 이용하여 구한 "naiveBayes" 타입의 객체를 의미한다.
testing	검정용 데이터(testing)가 포함되어 있는 dataframe 또는 matrix
type	예측이 범주 값(class)이든지 원시 예측 확률(raw)인지 명시하는 옵션이다.

## 인공 신경망 개요

인공 신경망(Artificial Neural Network)은 뉴런이라고 하는 두뇌 신경들이 상호 작용하여 경험과 학습을 통해서 패턴을 발견하고, 이를 통해서 특정 사건을 일반화 하거나 데이터를 분류하는데 이용되는 기계 학습 방법이다.

구글의 알파고(딥 러닝) 역시 인공 신경망의 이론을 바탕으로 하고 있다.

컴퓨터 스스로 인지, 추론, 판단을 하여 사물을 인식하거나 특정 상황의 미래를 예측하는데 이용될 수 있는 기계 학습 방법이다.

### 인공 신경망의 적용 분야

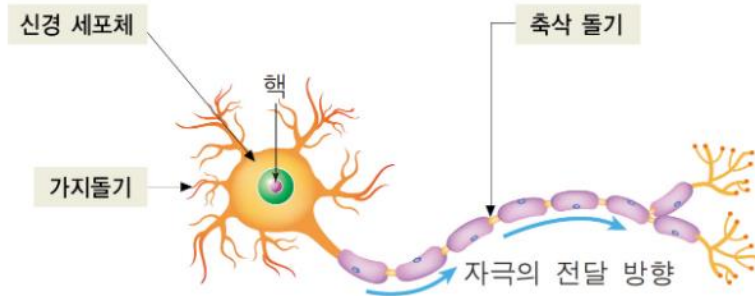
이미지등에 대한 인식 기능(Convolution Neural Net)  
문자, 음성, 증권 시장 예측하기 기능(Recurrent Neural Net)  
날씨 예보하기 등

## 생물학적 신경망의 구조

인간의 생물학적인 신경망의 구조는 다음과 같다.

먼저 수상 돌기가 외부 신호를 입력 받고, 시냅스에 의해서 신호의 세기를 결정한 후 이를 세포핵으로 전달하면 입력 신호와 세기를 토대로 신경 자극을 판정하여 축삭 돌기를 통해서 다른 신경으로 전달해준다.

뉴런의 구조 : 신경계의 구조적·기능적 기본 단위



신경 세포체	핵과 세포질이 모여 있는 뉴런의 본체로, 생명 활동이 일어남
가지돌기	다른 뉴런이나 감각 기관으로부터 자극을 받아들임
축삭 돌기	다른 뉴런이나 반응 기관으로 자극을 전달함

뉴런 내에서 자극의 전달 경로 : 가지돌기 → 신경 세포체 → 축삭 돌기

참조 사이트 : <http://study.zum.com/book/11779>

## 뉴런 용어 설명

뉴런의 각 용어에 대한 간략한 설명을 작성해 보았다.

항목	설명
----	----

수상 돌기(Dendrites)	외부 세계에서 신경 자극을 받아들이는 역할을 한다.
시냅스(Synapse)	신경과 신경의 연결 고리(뉴런 간의 교신)이다. 신경과 신경 간의 신호 전달 기능으로 <b>전달할 신호의 세기(Weight)</b> 를 결정한다.
세포핵(Soma)	여러 신경으로부터 전달되는 <b>신경 자극에 대한 판정</b> 과 다른 신경으로 <b>신호 전달 여부를 결정</b> 한다.
축삭 돌기(Axon)	전류와 비슷한 형태로 다른 신경으로 신호를 전달하는 기능을 수행한다.

뉴런을 인공 신경망과 비교해보기

생물학적 신경망을 컴퓨터로 처리할 수 있는 인공 신경망 구조와 비교를 해보면 다음과 같다.

생물학적 신경망	인공 신경망
수상 돌기	입력 신호(독립 변수 $x$ )에 해당한다.
시냅스	입력 신호에 가중치( <b>신호의 세기</b> )를 적용해주는 역할을 한다.
세포핵	입력 신호와 가중치를 이용하여 망(network)의 총합을 계산해준다. 활성화 함수를 이용하여 망의 총합을 출력 신호( $y$ )에 보내는 역할을 한다.

가중치 적용에 대한 이해

어떤 고객이 스마트 폰을 구매한다고 가정하자.

3 가지의 외부 조건에 따라서 스마트 폰 구매를 결정하기 위한 퍼셉트론이다.

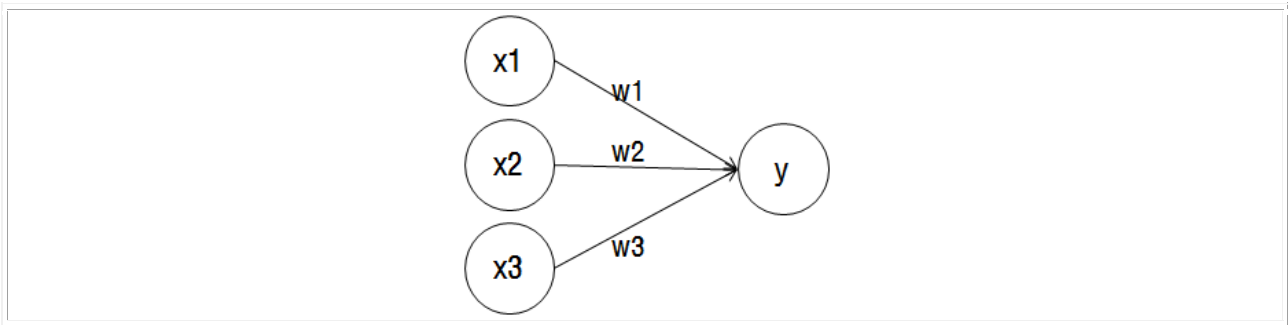
퍼셉트론이란 하나의 **신경 세포를 인공적으로 모델링한 개념**이다.

다음 토폴로지를 살펴 보도록 하자

토폴로지란 노드들과 이에 연결된 회선들을 포함한 네트워크의 배열이나 구성을 개념적인 그림으로 표현한 것이다.

동그란 원으로 된 항목은 Node 라고 한다.

세 개의 입력 노드에 가중치( $w_1$ ,  $w_2$ ,  $w_3$ )라고 하는 값을 곱하여 연산을 수행한 다음 출력 노드에 넘겨지는 그림이다.



입력 신호(독립 변수)에 대한 설명

입력이 되는 3 가지 변수에 대한 개략적인 설명이다.

변수	설명
$x_1$	이번 달의 수입이 충분한가?를 나타내는 척도이다.

x2	최신 기능을 가지고 있는가?를 나타내는 척도이다.
x3	기존의 스마트 폰에 문제가 있는가?를 나타내는 척도이다.
y	구매 함(1)/구매 안 함(0)

### 가중치의 필요성

예를 들어서 "나부자"씨는 돈이 많기(?) 때문에 x1의 값이 크게 중요하지 않다.  
또한, 스마트폰을 떨어뜨려서 액정이 깨진 "안보임"씨는 x3의 척도가 매우 중요한 항목이다.  
Earlyl Adapter인 "강감찬"씨는 x2의 척도가 가장 중요하다.

따라서, "단순히 몇 개의 조건을 만족한다"라는 것으로 판단을 내리면 안된다.  
이럴 경우 필요한 개념이 **가중치(weight)**이다.

즉, 고장인 사람은 x3의 값을 조금 키워 주는 w3의 값이 나머지에 비해서 상대적으로 값이 커야 한다.  
얼리 어댑터인 사람에게는 w2의 값이 매우 중요하다.

변수	w1	w2	w3	설명
부자	4	4	4	부자는 w의 값들에 크게 종속되지 않는다.
스마트 폰 고장	3	2	8	w3의 값이 다른 것에 비하여 크다.
얼리 어댑터	3	8	2	w2의 값이 다른 것에 비하여 크다.

이를 기반으로 코딩하면( b : 선택의 기준이 되는 값 )

#### 인공 신경망

```
if (x1*w1) + (x2*w2) + (x3*w3) > b :
    구매함 ;
else :
    구매안함 ;
```

이러한 형태에서 가중치와(weight)와 역치(bias)를 변경하면 의사 결정을 상황에 맞게 내릴 수 있을 것이다.

### 퍼셉트론과 신경망

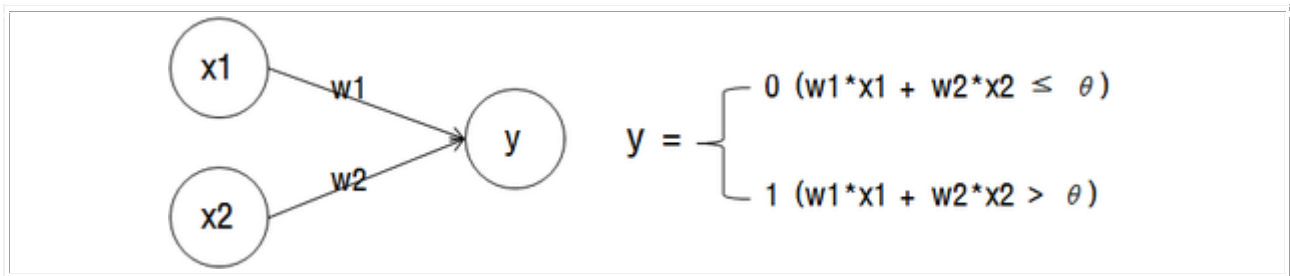
고객의 스마트 폰 구매에 대하여 가중치의 개념에 대하여 살펴 보았고, 이를 이용하여 토폴로지를 살펴 보았다.  
사람의 **신경 세포를 인공적으로 모델링한 개념**을 퍼셉트론이라고 한다.  
**다수의 신호를 입력 받아서 하나의 신호를 출력**해주는 알고리즘이다.  
이 퍼셉트론이 **신경망의 기원이 되는 알고리즘**이다.

퍼셉트론은 2 가지 신호를 가진다.

- 비활성화(숫자 0) : 신호가 흐르지 않는다.
- 활성화(숫자 1) : 신호가 흘러 간다.

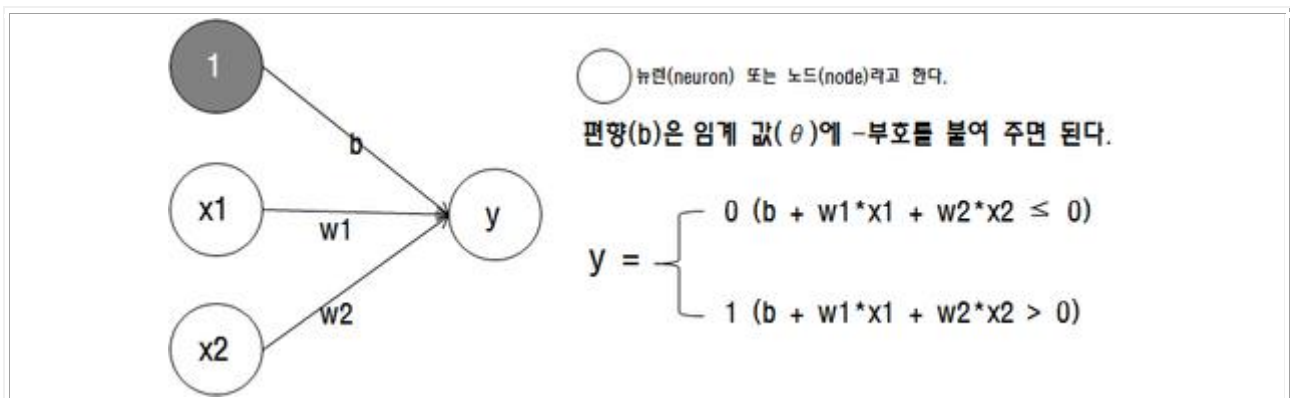
### 퍼셉트론

- 복수의 입력 신호에 각각의 고유한 가중치를 곱한다.
- 이것을 모두 더하여 1 개의 결과를 만들어 낸다.
- 이 결과 값이 정해진 한계(임계 값으로  $\theta$  으로 표현한다.)를 넘어 서면 1 을 출력해준다.



### 퍼셉트론 관련 용어들

신경망의 기초 개념이라고 하는 퍼셉트론과 관련된 용어들은 다음과 같은 것들이 있다.



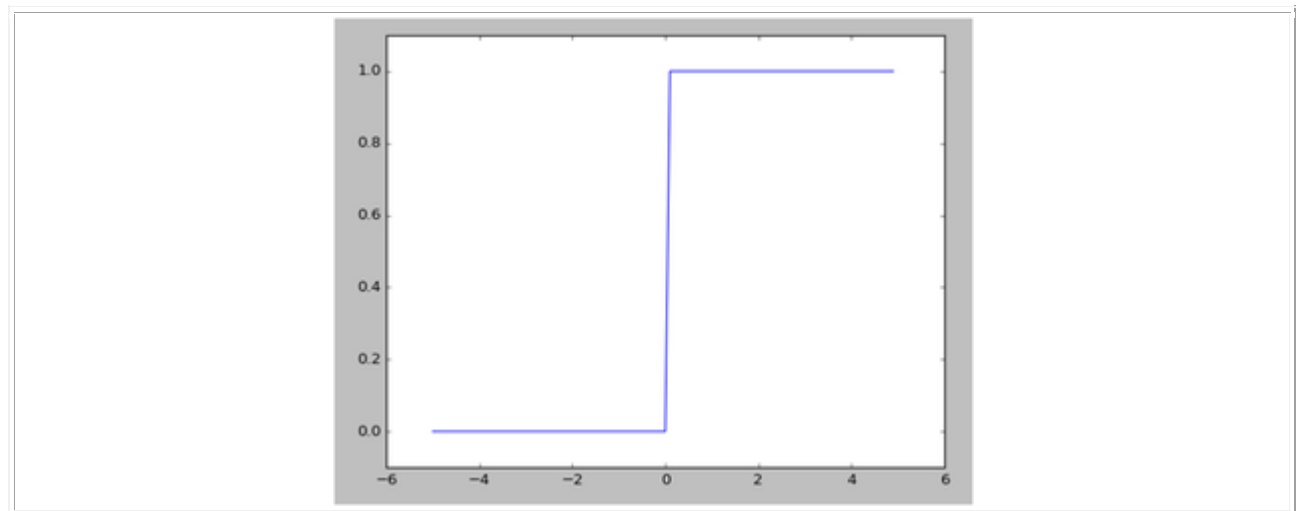
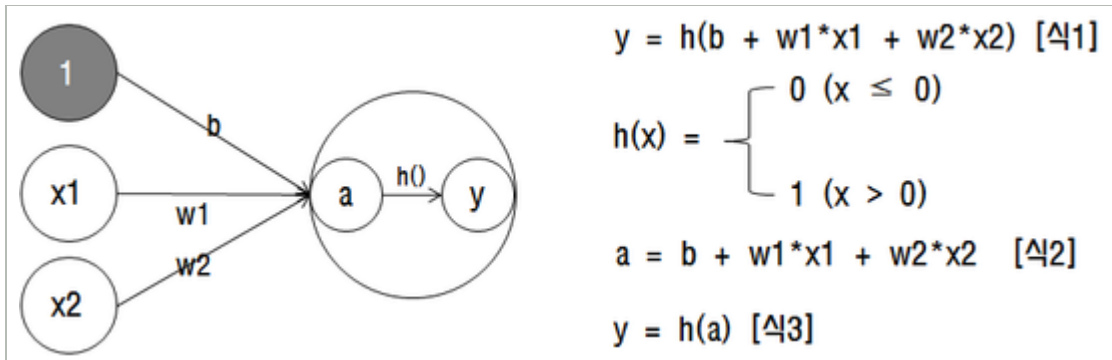
변수	설명
$x1, x2$	입력하는 신호이며, 독립 변수(설명 변수)라고 한다.
$w1, w2$	<b>가중치(weight)</b> 라고 한다. 입력 신호가 결과에 영향력을 조절하는 변수이다. $w1$ 의 값이 크다는 것은 $x1$ 의 입력 데이터가 중요하다는 의미이다.
$b$	<b>편향(bias)</b> : 입력이 1 이고, 가중치가 $b$ 라고 이해하면 좋을 듯하다. 뉴런이 얼마나 쉽게 활성화 되느냐를 제어해주는 <b>상수(constants)</b> 값이다.

### 활성화 함수(Activation Function)

활성화 함수란, 인공 뉴런이 정보를 처리하거나 망 전체로 전달하는 내부적인 작동을 의미한다  
일반적으로 입력 신호의 총합을 출력 신호로 변화시켜 주는 함수를 말한다.

퍼셉트론은 **계단 함수**를 활성화 함수로 사용한다.

계단 함수 이외의 다른 함수로 변경하는 것이 **신경망의 세계로 나아가는 열쇠**이다.



### 다른 활성화 함수

#### 시그모이드(Sigmoid) Function

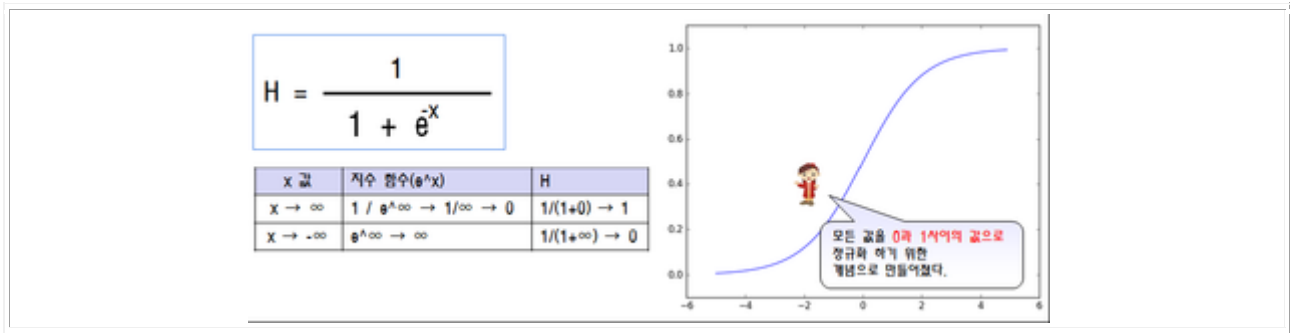
로지스틱 회귀 모델은 이항 분류 모델이다.

sigmoid(S자 모양의 곡선) function 함수를 사용하여 모든 데이터를 0 과 1 사이의 값으로 분류할 수 있다.

절반\_cutoff (0.5)을 기준으로 입력에 따른 정답을 2 가지(0, 1)로 구분한다.

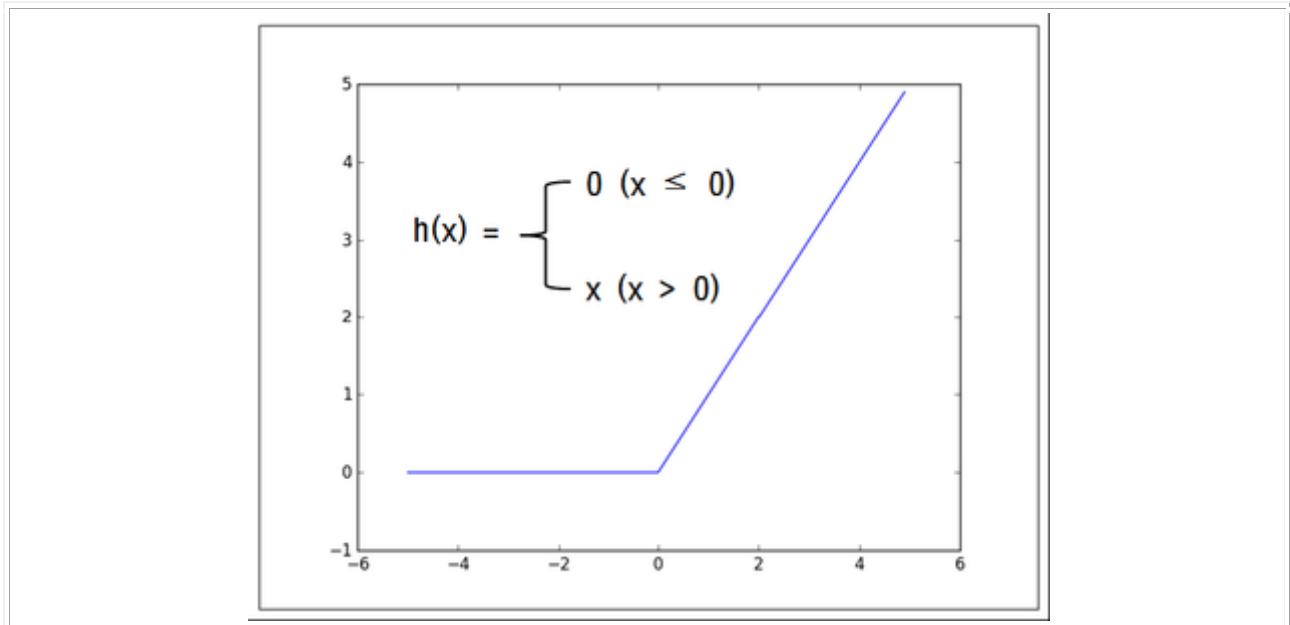
로지스틱 회귀라고 부르면 이 함수(sigmoid)를 사용하는 것이 훨씬 더 성능이 좋다. 전달





### 릴루(ReLU) Function

ReLU(Rectified Linear Unit) 함수는 입력이 0 을 넘으면 그대로 출력하고, 0 이하이면 0 을 출력해주는 함수이다. 정류(전기 용어)에서 나온 용어로  $\pm$ 가 반복이 되는 교류에서 마이너스(-) 흐름을 차단하는 회로이다. ReLU 함수를 수식으로 나타내면 다음과 같다.



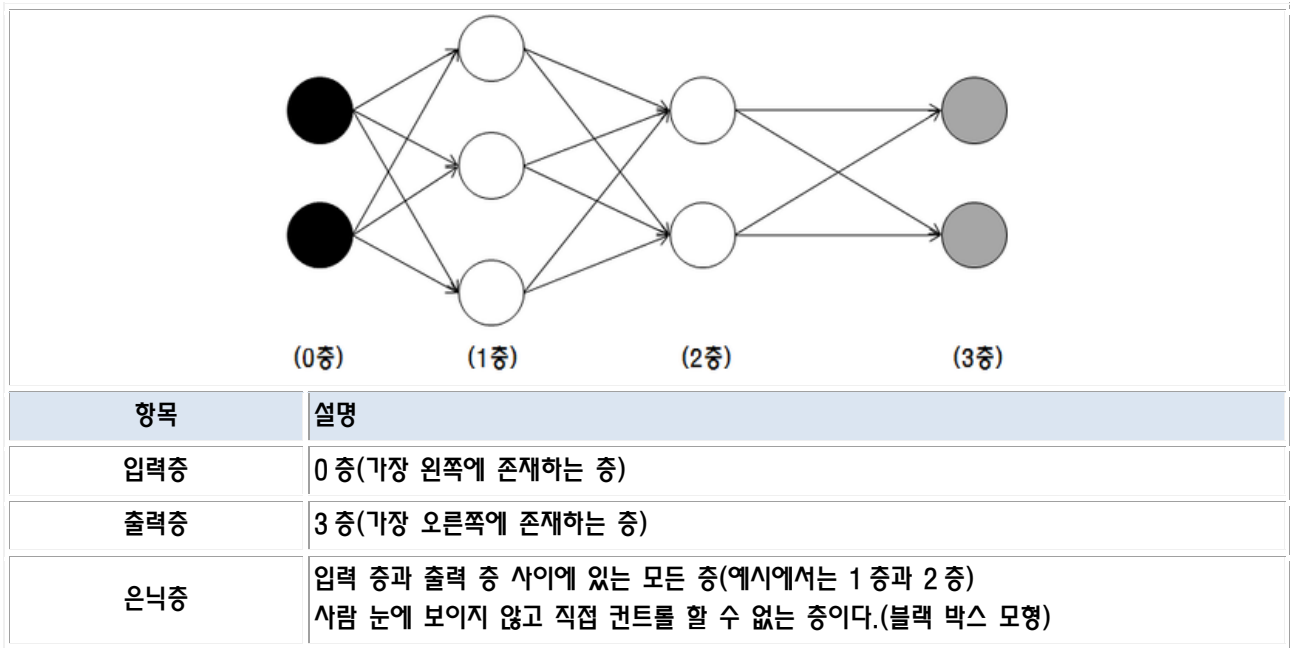
### 신경망(Neural Network)

사람 뇌의 신경망처럼 여러 신경들을 하나의 망(network) 형태로 나타내기 위해서 여러 개의 계층으로 다중화 하여 만들어진다.

즉, 신경망(Neural Net)은 여러 뉴런이 연결되어 있는 구조를 가지고 있는 망형(network type) 구조이다.

일반적으로 신경망을 3 개 이상 중첩하게 되면 **깊은 신경망(DeepNN)**이라고 부르는 데, 이를 활용한 기계 학습을 **딥러닝**이라고 부른다.

다음 그림은 입력층, 은닉층, 출력층으로 구성된 퍼셉트론의 모형이다.



이러한 모형에서 입력 데이터와 출력 데이터는 분석자가 직접 지정해준다.

따라서 인공 신경망 역시 **지도 학습**의 범주에 해당된다.

또 한편 인공 신경망은 은닉층에서의 연산 과정이 공개되지 않기 때문에 이러한 측면에서 블랙 박스 모형으로 분류되기도 한다.

### 신경망 구성하기

학습하는 신경망의 능력은 토폴로지나 서로 연결된 뉴런의 구조와 패턴에 따라 다를 수 있다.

토폴로지란 노드들과 이에 연결된 회선들을 포함한 네트워크의 배열이나 구성을 개념적인 그림으로 표현한 것이다.

#### 망(network) 구조를 구분하는 3 가지 특징

망의 구조는 수없이 많지만, 일반적으로 다음과 같이 3 가지로 특징 지을 수 있다.

- (1) layer 의 개수
- (2) 망에서 정보가 뒷단으로 진화 가능한가?
- (3) 각 층(layer)에 있는 노드의 갯수

##### (1) layer 의 개수

입력 노드는 입력된 데이터를 직접 가공하지 않는 신호를 받는다.

입력 노드는 받은 데이터들을 처리하고, 망은 연결 가중치들을 가지고 있다.

입력과 출력 노드만 가지고 있는 망을 단층망(single network layer)이라고 부른다.

좀 더 복잡한 망을 만들기 위해서는 새로운 layer(층)을 추가하여 다층망(multi network layer)을 만든다.

##### (2) 망에서 정보가 뒷단으로 진화 가능한가?

전방향(feed forward)

입력 신호가 한쪽 방향으로 꼭 진행하여 출력 신호까지 도달하는 망(network)

##### 재귀망/피드백망

순환(loop)을 통하여 양쪽 방향으로 진행한다.  
증권 시장 예측, 음성에 대한 이해, 날씨 정보 등등  
머신 러닝에서 RNN 이 여기에 해당한다.

### (3) 각 층(layer)에 있는 노드의 갯수

layer(층)	설명
입력 층	입력 데이터의 속성의 갯수가 노드의 수가 된다. 예를 들어서 iris 데이터 셋의 경우에는 노드의 수가 4 가 된다.
출력 층	결과의 분류의 갯수이다. 선형 회귀는 1, 로지스틱 회귀는 2, 소프트 맥스는 n(정수)이다. iris 데이터 셋의 경우에는 소프트 맥스(n=3)이다.
은닉 층	특별한 규칙은 없다. 입력 노드의 수, 훈련 데이터의 갯수, 노이즈한 데이터의 양, 기타 인자와 학습 태스크 등의 복잡성에 따라 달라질 수 있다.

### 뉴런의 갯수가 많다

과적합의 위험이 따른다.  
고비용이다.  
속도의 느림 현상이 발생할 수 있다.

즉, 뉴런의 수가 많아진다고 해서 무조건 좋아지는 것은 아니다.

## 인공 신경망 모델 패키지

### nnet 패키지 이용

nnet 패키지에서 제공되는 nnet() 함수는 1 개의 은닉층을 갖게 하는데 최적화된 함수로 사용 형식은 다음과 같다.  
인공 신경망 모델을 생성해준다.  
반환된 결과는 "nnet.formula", "nnet" 객체이다.

항목	설명
사용 형식	nnet(formula, data, weights, size)
formula	y ~ x 형식으로 반응(종속) 변수와 설명(독립) 변수 수식
data	모델 생성에 사용될 데이터 셋을 지정한다.
weights	각 case에 적용할 가중치(기본값 : 1)를 지정한다.
size	은닉층(hidden layer)의 수를 지정한다.
softmax	다중 분류를 위해서는 True를 사용하면 된다.

nnet의 결과물 객체에 entropy fitting 가 나오면 로지스틱 회귀이다.

### nnet() 함수의 결과물

반환된 결과는 "nnet.formula", "nnet" 객체이다.

이 객체와 관련된 함수들은 다음과 같은 목록들이 있다.

항목	설명

### class.ind 함수

Generates a class indicator function from a given factor.

nnet 패키지의 다음 클래스를 종속 변수(y 변수)에 대한 one-hot encoding 을 수행해주는 함수이다.

항목	설명
사용 예시	<pre>species.ind &lt;- class.ind(iris\$Species) # one-hot encoding iris &lt;- cbind(iris, species.ind) head(iris, 3)</pre>

### 신경망 시각화 라이브러리

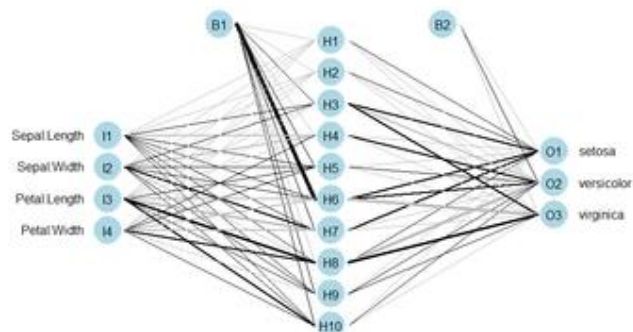
신경망에 대한 시각화를 사용하려면 다음과 같이 코딩하면 된다.

library("devtools") # 시각화 R 코드 함수 다운로드

source\_url('https://gist.githubusercontent.com/fawda123/7471137/raw/466c1474d0a505ff044412703516c34f1a4684a5/nnet\_plot\_update.r')

# 매개 변수 iris\_nn은 nnet 함수를 사용하여 구해진 nn 객체를 의미한다.

plot.nnet( iris\_nn )



## neuralnet 패키지 이용하기

nnet 패키지 보다 최근에 공포된 인공 신경망 모델 생성을 위한 패키지로써 나온 neuralnet 패키지는 역전파(Backpropagation) 알고리즘을 적용할 수 있다.

또한 가중치 망을 시각화해주는 기능도 제공한다.

주의 사항으로 출력 변수(y)는 'yes', 'no' 형태의 문자열 아닌 1과 0의 수치형 이어야 한다.

## neuralnet 함수

인공 신경망 모델을 생성해준다.

항목	설명
사용 형식	neuralnet(formula, data, hidden = 1, threshold = 0.01, stepmax = 1e+05, rep = 1, startweights = NULL, learningrate=NULL, algorithm = "rprop+")
formula	y ~ x 형식으로 반응(종속) 변수와 설명(독립) 변수 수식
data	모델 생성에 사용될 데이터 셋을 지정한다.
hidden	은닉층(hidden layer)의 노드 수를 지정한다. 만약 hidden = c(5,3)라고 하면 1번째 히든 레이어는 5개의 노드를, 2번째 히든 레이어는 노드가 3개가 된다.
threshold	경계값 지정
stepmax	인공 신경망 학습을 위한 최대 스텝을 지정한다.
rep	인공 신경망의 학습을 위한 반복 수를 지정한다.
startweights	랜덤으로 초기화된 가중치를 직접 지정
learningrate	backpropagation 알고리즘에서 사용될 학습 비율을 지정한다.(학습율이라고 한다.)
algorithm	backpropagation(역전파)과 같은 알고리즘 적용을 위한 속성
backprop	역전파를 통해서 가중치와 경계값을 조정하여 오차(E)를 줄이기 위해서 사용되는 속성이다.

## compute 함수

인공 신경망 모델의 테스트 데이터에 대한 예측치를 구해주는 함수이다.

항목	설명
사용 형식	compute(x, covariate, rep = 1)
x	neuralnet() 함수를 이용하여 구한 nn(neuralnet) 객체를 지정해준다.
covariate	예측을 수행할 점검용 데이터 셋을 의미한다.
rep	인공 신경망의 학습을 위한 반복 수를 지정한다.

### 반환 값 리스트

`neurons`

망에서 각 층에 있는 뉴런 정보를 가지고 있다.

a list of the neurons' output for each layer of the neural network.

`net.result`

분류에 대한 모델의 예측 값을 가지고 있다.

a matrix containing the overall result of the neural network.