사회조사
사례연구
12주차

- 훈련, 검증, 테스트 데이터
- Logistic 회귀분석
- Decision Tree
- R 실습

1. 훈련, 검증, 테스트 데이터

통계모형 활용 목적이 예측인 경우, 지도학습(Supervised Learning)을 주로 살펴봄

Predictive Modeling : 주어진 Data에 근거하여 Model을 만들고 이 Model을 이용하여 새로운 case들에 대한 예측을 하는 작업

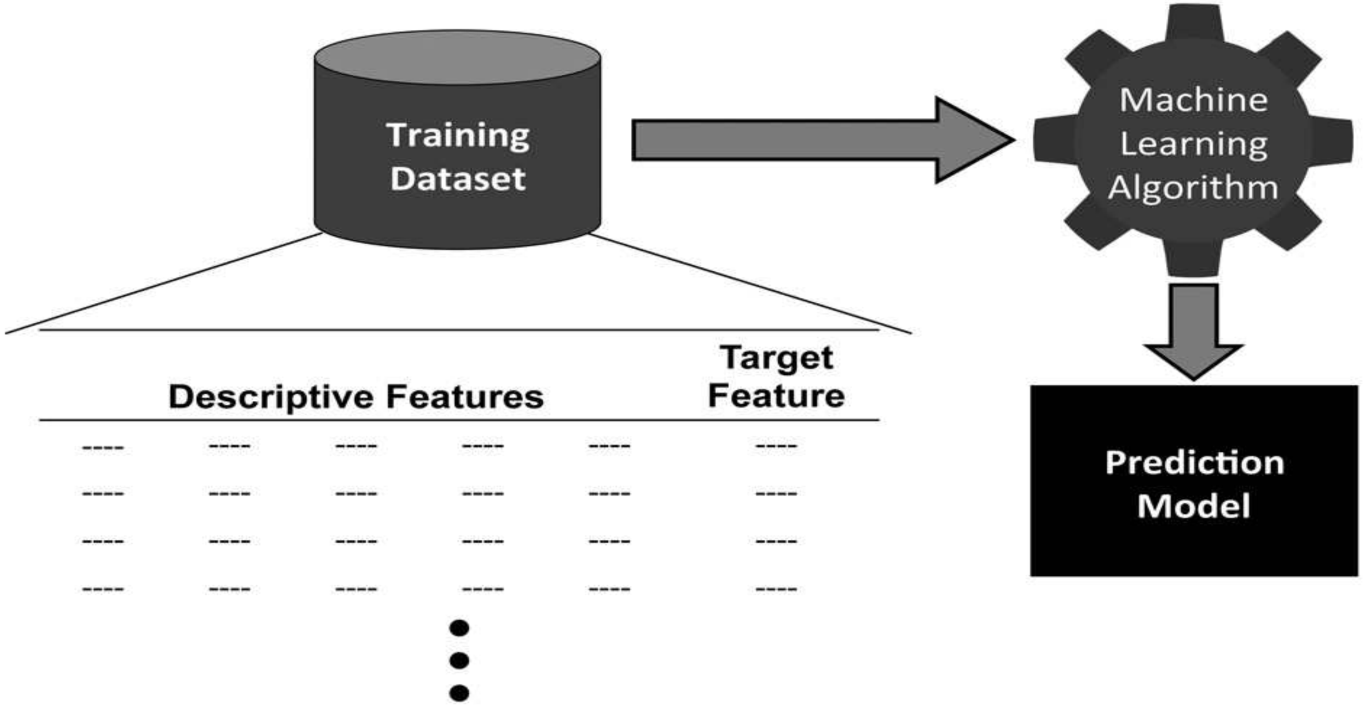
예) 분류 (Classification)
값 예측 (Value Prediction)

그동안 축적된 모형 구축을 위한 데이터셋(Training data)은 설명변수(Descriptive features)와 반응변수(Target feature)로 구성되어 있으며 이 데이터셋으로 기계학습 알고리즘을 활용하여 예측모형을 만듬

이렇게 만들어진 예측모형을 새로운 데이터셋(Test data)에 적용하여 새로운 데이터에 예측값을 추정할 수 있음

어떤 기법이 정확도가 높은지에 대한 부분은 모형 평가에서 다룰 예정임

학습 모델



자료: Kelleher, J. D., Mac Namee, B. D'Arcy, A. (2015). Fundamental of machine learing for predictive data analytics−algorithms, worked examples and case studies. London: The MIT Press.

Training set 모형 구축	Validation set 모형 선택	Test set 모형 평가

Training dataset : 모형의 적합과 모수의 추정에 사용

Validation dataset : 파라미터 튜닝과 변수 선택, 모형 선택에 사용됨. 검증 데이터셋의 오류확률은 테스트 오류확률을 추정하는 데 사용하지만, 일반적으로 테스트 오류확률보다 적음 => 검증 데이터셋이 모형의 튜닝에 사용되므로 과적합이 발생

Test dataset : 모형 적합과 모형 선택이 끝난 수 최종 모형의 오류확률(error rate)를 측정, 추정하기 위해 사용됨. 테스트 데이터는 모형의 선택과 튜닝에 사용하면 안됨!!!

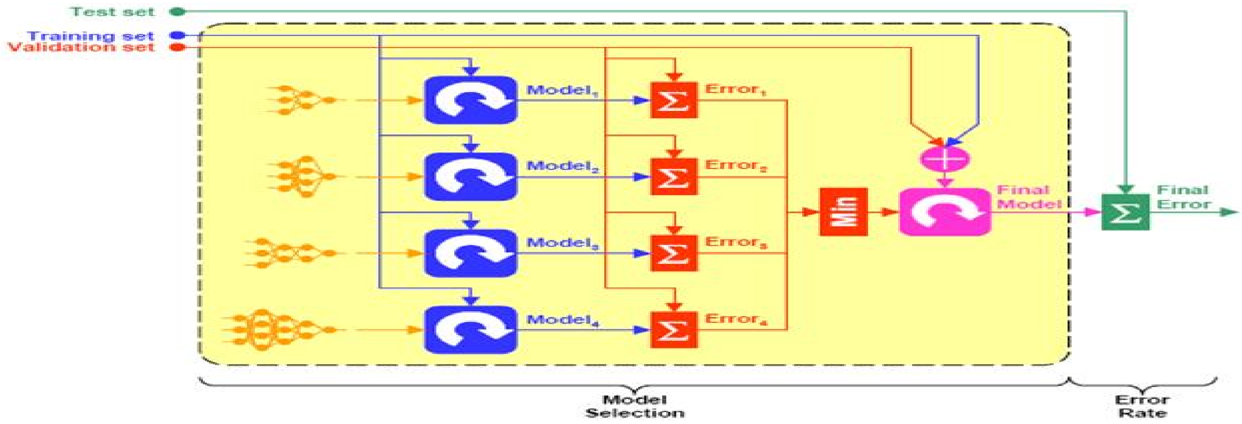
모형 선택과 평가를 위한 방법 중 하나는

1. 데이터를 랜덤하게(50:25:25 또는 60:20:20)의 비율로 훈련/검증/테스트 데이터셋으로 나눔

2. 훈련 데이터를 이용하여 모형을 적합

3. 검증 데이터셋을 이용하여 적합한 모형 중 최종 모형을 선택

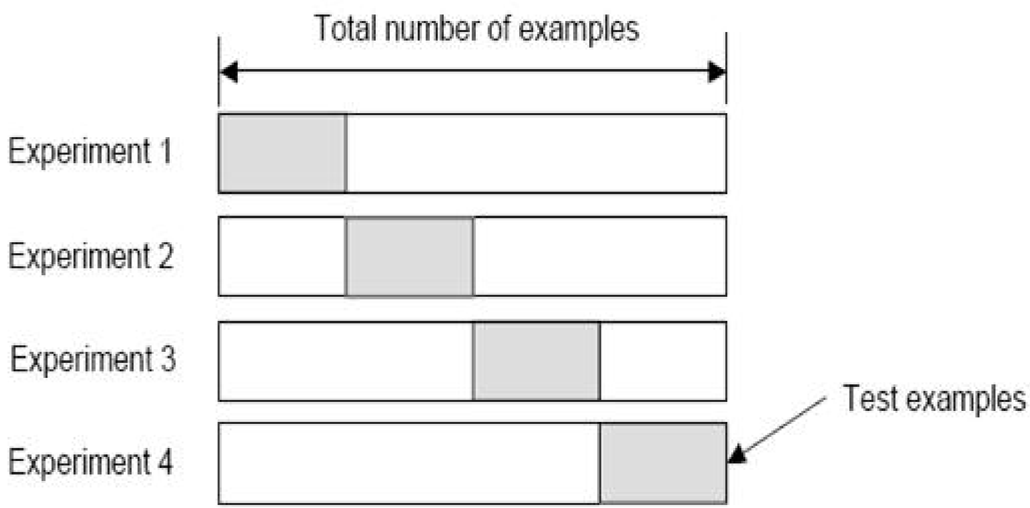
4. 테스트 데이터셋을 사용해 최종 모형의 성능을 측정하는 것임



K-Fold cross validation

훈련-검증 데이터셋을 처음부터 나누지 않고, 교차검증을 하기도 함

K-fold 교차검증은 데이터를 K개의 그룹으로 나눈 후 각각의 그룹을 차례로 검증 세트로 사용하여 정확도 지표를 계산 한 후, K개의 정확도 지표의 평균으로 최종 오차를 추정



분류/예측문제 접근 방법

1. Training set으로 다양한 모형 적합
2. Validataion set으로 모형을 평가, 비교, 최종 모형 선택
3. Test set으로 최종 모형의 일반화 능력을 계산

Data 분석 Step

1. 데이터의 구조 파악, y변수의 인코딩, x 변수의 변수형 분석

2. 데이터를 랜덤하게 훈련셋, 검증셋, 테스트셋으로 나눔, 보통 60-20-20 사용

3. 시각화와 간단한 통계로 y변수와 x변수 간의 관계 파악
   => x변수와 y변수의 상관관계는? 이상치는 있는지? 변환이 필요한 x변수는 없는지?

4. 시각화와 간단한 통계로 x변수들 간의 관계 파악
=> 상관관계가 아주 높은 것은 없는지? 비선형적인 관계는 없는지? 이상치는 없는지?

5. 다양한 분석 모형 적합
=> Logistic regression, Lasso, Elastic Net, Boosting 등

6. 각 모형에서 살펴볼 내용
★ 변수의 유의성: 모형이 적절한지? 기대한 변수가 중요한 변수로 선정되었는지?
★ 적절한 시각화 : 로지스틱 분석, 트리 모형 등 모형마다 도움이 되는 시각화 살펴보기
★ 모형의 정확도 : 교차검증을 이용하여 검증셋에서 계산하여야 함
=> 모형의 정확도 개념을 테스트셋에서 적용하기도 함

7. 검증셋을 사용하여 최종 모형을 선택
=> 다양한 모형을 검증셋을 사용해 평가하고, 가장 예측 성능이 좋은 모형을 최종 모형으로 선택

8. 테스트셋을 이용하여 최종 모형의 일반화 능력을 살펴봄

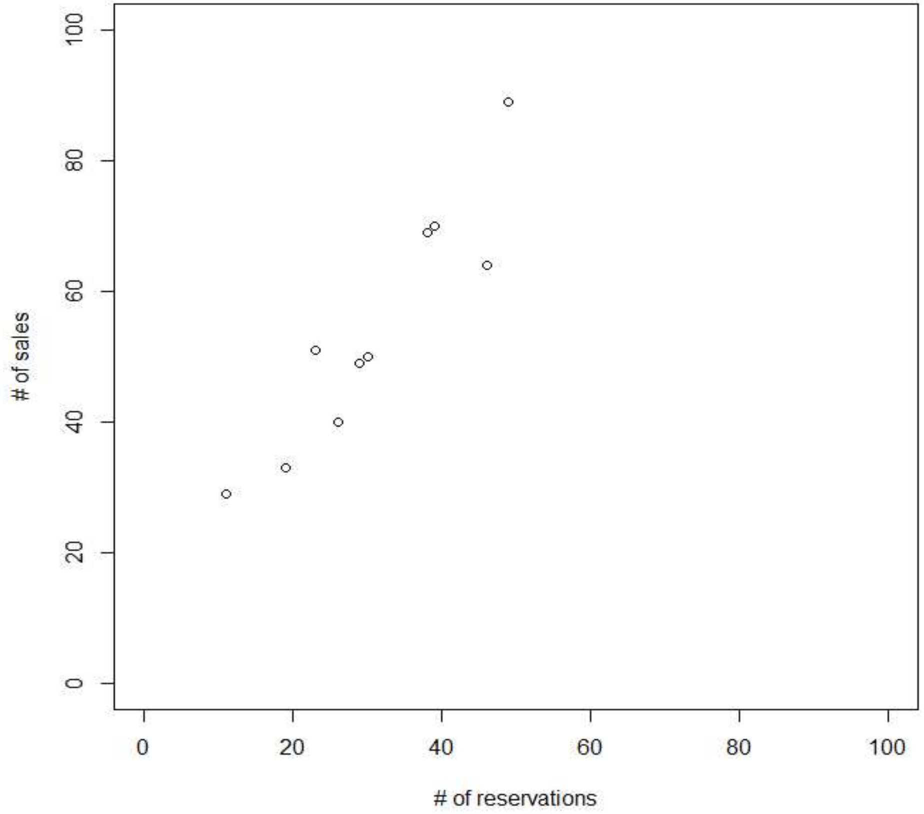
2. Logistic Regression

단순선형 회귀모형(Simple linear regression model)

(예제) 에어컨 판매대수 예측

- **목적**: 에어컨 예약대수를 이용하여 내년에 판매되는 에어컨 대수를 예측
- **자료**: 지난 10년간의 에어컨 예약대수와 판매대수
- **입력변수**: 에어컨 예약대수
- **출력변수**: 에어컨 판매대수

(예제) 자료 산점도



(1) 단순선형 회귀모형

● 모형

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \ldots, n.$$

예제자료에서

- x_i: 에어컨 예약대수
- y_i: 에어컨 판매대수
- ε_i: 오차항(평균이 0, 분산이 σ^2)

● **선형회귀모형**: 입력변수와 출력변수의 관계가 선형방정식

- 비고: 비선형회귀모형 (예: $y = \sin(2x) + \varepsilon$)

● **단순(simple) 선형회귀모형**: 입력변수가 하나인 선형회귀모형

● **다중(multiple) 선형회귀모형**: 입력변수가 2개 이상인 선형회귀모형

(2) 모수의 추정: 최소제곱법

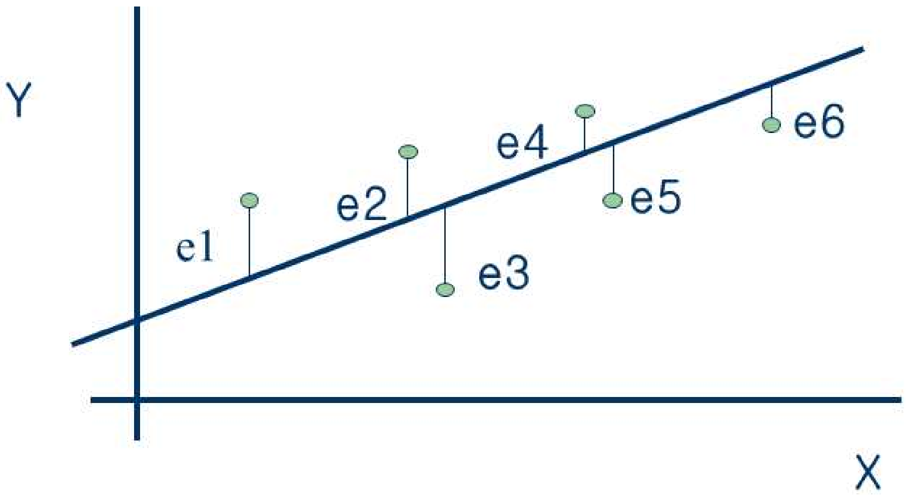
- 모수의 추정
  - **회귀모수(regression parameter)**: (α, β)
  - **자료**: $(y_1, x_1), \ldots, (y_n, x_n)$
  - **최소제곱추정법**: 잔차의 제곱의 합을 최소로 되게 하는 직선(α, β)을 구하는 것

최소제곱 추정치: $(\hat{\alpha}, \hat{\beta}) = \mathrm{arg}$삭‑櫶̂

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})\sum_{i=1}^{n}(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$$

● 최소제곱법 아이디어



$e_1^2 + \cdots + e_6^2$ 을 최소로 하는 직선을 찾는다.

(3) 모형의 해석 및 예측

- **회귀분석의 목적 (2가지)**
  - 모형의 해석(설명): estimation
  - 새로운 자료를 예측: prediction

- **예측**: 주어진 새로운 입력변수 x에 대하여 출력변수 y를 $\hat{y}$로 예측

$$\hat{y} = \hat{\alpha} + \hat{\beta} x$$

- **모형의 해석**
  - 회귀계수 $\hat{\beta}$가 가장 중요
  - x가 한 단위 증가할 때의 y의 증가량
  - $\hat{\beta}$가 양수이면 x가 증가하면서 y가 증가
  - $\hat{\beta}$가 음수이면 x가 증가할 때 y는 감소

(예제) 결과

- **예제: 에어컨 판매대수**

  - **최소제곱법으로 추정된 모형식**

$$y = 9.74 + 1.44x$$

  - **예측**: 올해 에어컨 예약대수가 45이면 내년 에어컨 판매대수는 $9.74 + 1.44 \times 45 = 74.54$로 예측할 수 있다.

  - **해석**: 에어컨 예약대수가 1단위 증가하면 에어컨 판매대수는 1.44단위 증가한다.

(4) 회귀계수의 검정

- 회귀계수 β가 0이면, 입력변수 x와 출력변수 y사이에 아무런 관계가 없게 된다.
  - 즉, 회귀계수 β가 0이면 적합된 추정식은 아무 의미가 없게 된다.

- 적합된 추정식이 의미가 있는지(자료를 잘 설명하는지)를 검정하는 것은 회귀계수 β가 0인지를 검정하는 것과 같다.

$$H_0 : \beta = 0$$

- **검정통계량**: $t = \dfrac{\hat{\beta}}{se(\hat{\beta})} \overset{H_0}{\sim} t(n-2)$

- **검정방법**: $|t|$가 크면 $\hat{\beta}$가 0이라는 가설(귀무가설)을 기각
  - 즉, 추정된 회귀식이 유의하다(의미있다)라고 결론

(예제) 결과

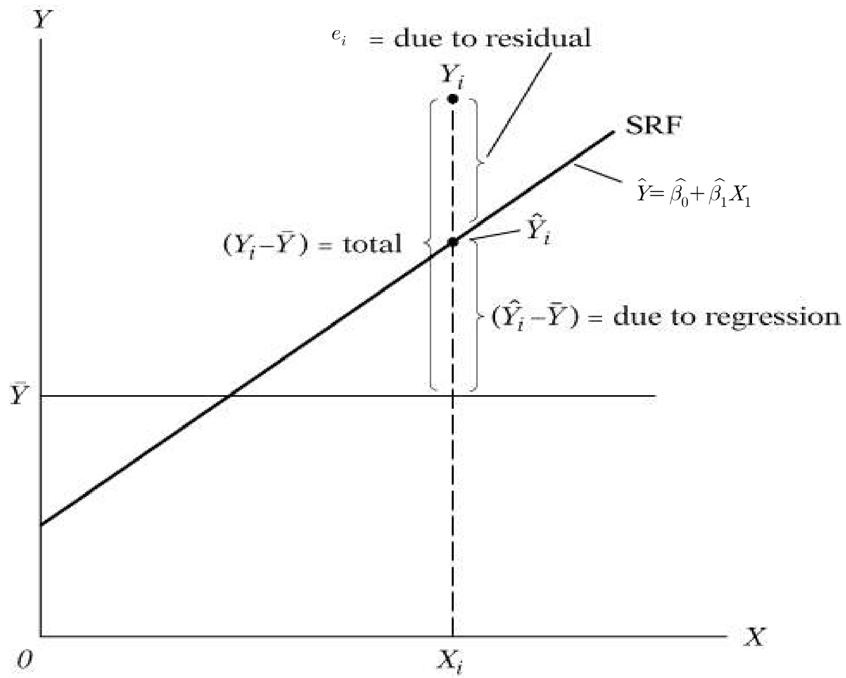
```
> # LSE by lm function
>   g0 = lm(sale~reserv, data=Data)
>   print(g0)

Call:
lm(formula = sale ~ reserv, data = Data)

Coefficients:
(Intercept)       reserv
      9.736        1.441

>   summary(g0)

Call:
lm(formula = sale ~ reserv, data = Data)

Residuals:
     Min       1Q   Median       3Q      Max
-12.0115  -3.8229   0.4485   4.4044   8.6662

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.7362     6.6206   1.471     0.18
reserv        1.4408     0.2004   7.188 9.35e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.227 on 8 degrees of freedom
Multiple R-squared:  0.8659,    Adjusted R-squared:  0.8492
F-statistic: 51.67 on 1 and 8 DF,  p-value: 9.354e-05
```

● **제곱합의 분할**

$$\sum(y_i - \bar{y})^2 = \sum(\hat{y_i} - \bar{y})^2 + \sum(y_i - \hat{y_i})^2$$

$$SST \quad = \quad SSR \quad + \quad SSE$$

$$n-1 \quad = \quad 1 \quad + \quad n-2$$



- 제곱합의 해석

  - 전체제곱합은 회귀모형을 사용하지 않았을 때의 자료의 편차(variation)

  - 잔차제곱합은 회귀모형을 사용하였을 때의 자료의 편차(회귀모형에서의 오차항의 추정치)

  - 따라서, 회귀제곱합은 전체제곱합 중 회귀직선을 사용하여서 설명된 편차로 생각할 수 있음

  - 회귀제곱합이 크면, 회귀직선이 자료를 잘 설명하는 것이고,

  - 반대로, 회귀제곱합이 작으면 회귀직선이 자료를 설명하는 양이 크지 않은 것을 의미함

- 결정계수(R^2)

- 정의: $R^2 = \frac{SSR}{SST}$

- **성질**: $0 \le R^2 \le 1$

- **의미**: R^2이 1에 가까울수록 적합한 회귀모형이 자료를 잘 설명

- **예제에서**, $R^2 = 0.8659$로 자료의 전체변동 중 회귀모형이 86.59% 설명

(5) 선형모형에서 사용되는 가정들

- 선형회귀모형에서 사용되는 가정들
  - **선형성**: 입력변수와 출력변수간의 관계가 선형적
  - **등분산성**: 오차의 분산이 입력변수와 무관하게 일정
  - **독립성**: 오차들이 서로 독립
  - **정규성**: 오차의 분포가 정규분포

- 선형모형을 자료에 적합하기 전에 위의 3가지 가정이 만족되는지를 확인하여야 한다.
  - 잔차분석을 통하여 위 가정들을 만족하는지 체크

- 특히, 선형성은 위의 가정 중 가장 중요한 가정으로 이 가정이 맞지 않은 경우에는 선형회귀 모형은 아주 나쁜 결과를 제공

로지스틱 회귀모형(Logistic regression model)

- **회귀 분석: 반응변수(Y)와 설명변수(X)간의 관계를 파악하는 분석**
  - 반응변수가 연속형: 선형 회귀 모형
  - 반응변수가 이산형(명목형/순서형): 로짓 모형(로지스틱 회귀 모형)

  ※ 설명변수의 형식은 상관없음(연속/이산/명목)

- **로지스틱회귀**: 반응값이 ***범주형***일 때, 주로 쓰이는 분석방법
  - 이항 반응변수: 성공/실패, 생존/사망, (대출)승인/거절, (시험)합격/불합격
  - 다항 반응변수: 순서형(누적 로짓모형), 명목형(일반화 로짓모형)
  - 관심사항
    - 설명변수들이 (이항/다항)반응 결과에 어떻게 영향을 미치는지에 대해 관심(estimation)
    - 새로운 개체의 반응결과를 예측(prediction) 및 분류(classification)

● 자료구조

- 반응변수 $Y_i \in \{0,1\}, \quad i=1,2,\cdots,n.$
- 설명변수 $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \cdots, X_{ip})^T, \quad i=1,2,\cdots,n.$

● 반응변수 y의 평균

$$E(Y) = 1 \times P(Y=1) + 0 \times P(Y=0) = P(Y=1) \text{ (성공확률)}$$

● 로짓 회귀모형은 반응변수(Y)의 성공확률이 X에 영향을 받는다고 가정

● 반응변수는 이항자료이므로, 성공확률이 p_x를 가진 베이누이 분포를 따른다고 가정

$$P(Y=y|X=x) = p_x^y (1-p_x)^{1-y}, \quad y=0,1$$

● 선형 확률모형(linear probability model)

$$p_x = \alpha + \beta x$$

- 확률은 0과 1사이 값,

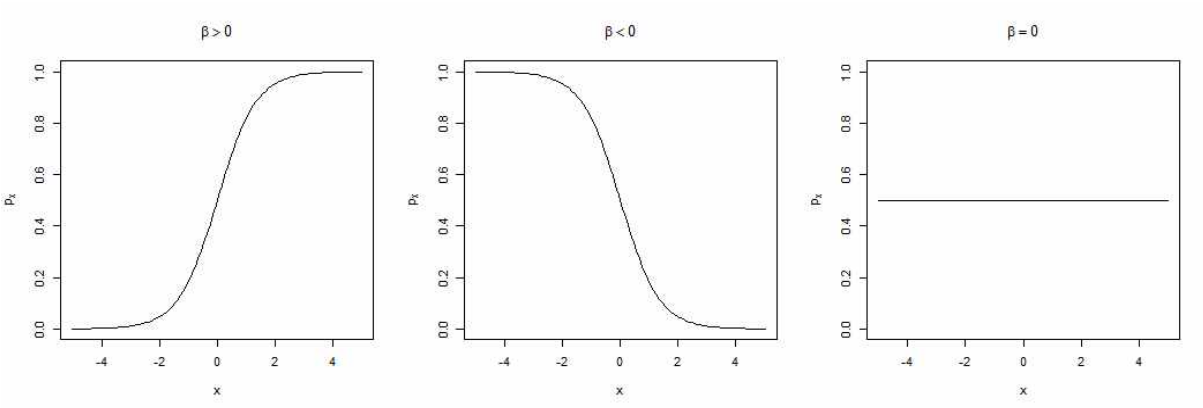
- 그러나, 선형 확률모형은 실수 전체 값을 가질 수 있음(1보다 크거나 음수)

● 로지스틱 회귀모형(logistic regression model)

$$p_x = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \qquad \text{or} \qquad P(Y=1|X=x) = \frac{\exp(x^T \boldsymbol{\beta})}{1 + \exp(x^T \boldsymbol{\beta})},\ (\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \ \ldots, \beta_p)^T)$$

$$\text{i.e. } \log\left(\frac{p(Y=1|X=x)}{p(Y=0|X=x)}\right) = x^T \boldsymbol{\beta}$$

- 위 모형은 확률의 조건을 모두 만족함
- s자 형태를 가진 곡선



- 곡선의 변화율은 $|\beta|$가 커짐에 따라 커짐

선형회귀모형에서는 x가 주어졌을 때 Y의 조건부 평균이지만, 로지스틱 회귀는 조건부확률을 연결함수(link function) p를 통해 모형화하는 것임

연결함수의 형태에 따라 $p(x)=\exp(x)/(1+\exp(x))$이면 로지스틱 모형, $p(x)=\exp(-\exp(x))$는 검벨(Gumbel) 모형, $p(x)$가 표준정규분포의 분포함수는 프로빗(probit) 모형이라고 하며 계산의 편리성으로 로지스틱 모형이 많이 사용됨

■ **성공확률에 대한 오즈(odds)**

$$\frac{p_x}{1-p_x}=\exp(\alpha+\beta x)$$

- 설명변수가 한 단위 증가하면, 오즈는 $\exp(\beta)$만큼 증가

$$\frac{P(Y=1|x+1)/P(Y=1|x)}{P(Y=0|x+1)/P(Y=0|x)}=\exp(\beta_1)$$

을 오즈비(odds ratio)라고 함

오즈비는 x가 한 단위 증가할 때 $y=1$일 확률과 $y=0$일 확률 비의 증가율을 의미

예를 들어, x는 소득이고 y는 어떤 제품에 대한 구입 여부(1=구입, 0=미구입)라 할 때, $\hat{\beta}_1 = 2$이면 소득이 한 단위 증가하면 물품을 구매하지 않을 확률에 대한 구매할 확률의 오즈비가 $\exp(2) = 7.38$배 증가하는 것으로 해석할 수 있음

로지스틱 회귀에 대한 우도함수(likelihood function)는

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} p(\beta_0 + \beta_1 x_i)^{y_i} (1 - p(\beta_0 + \beta_1 x_i))^{1-y_i},$$

$$where \;\; p(x) = \exp(x)/(1 + \exp(x))$$

임

로그 우도함수는 계수에 대한 비선형 함수이기 때문에 최대우도(maximum likelihood)추정치 $(\hat{\beta}_0, \hat{\beta}_1)$은 수치적 방법(numerical method)을 사용하여 구할 수 있음

설명변수 x가 y를 설명하는 데 유의한지에 대한 유의성 검정은 우도비 검정 통계량은

$$\chi^2 = -2\left(\max_{\beta_0} l(\beta_0, 0) - l(\widehat{\beta_0}, \widehat{\beta_1})\right) \text{ 임}$$

χ^2는 근사적으로 자유도가 1인 카이제곱 분포를 따르며 그 값이 크면 β_0가 0이 아니라고 결론을 내림

입력변수가 범주형인 경우 선형회귀와 같은 방법으로 가변수를 생성해서 분석할 수 있으며, 변수선택 역시 선형모형과 동일하게 적용할 수 있다. 선형회귀모형에서는 선택기준을 오차제곱합으로 사용하는 반면, 로지스틱 회귀에서는 로그 우도함수값을 사용한다. 로지스틱 회귀에서 변수선택을 전진선택법으로 하는 경우 각 단계마다 로그 우도함수값의 증가량을 구하고, 그 증가량이 가장 큰 변수부터 추가한다. 최종모형은 AIC 나 BIC 등의 선택기준을 최소화하는 모형으로 선택한다.

로지스틱 회귀는 주어진 설명변수 x에 대해 반응변수 Y가 1이 될 확률 $P(Y=1|x)$를 추정하는데, 0과 1 사이의 합리적인 기준값 α를 절단값으로 선택하여 $\hat{P}(Y=1|X=x) > \alpha$ 이면 자료를 $Y=1$인 클래스로 분류하고 $\hat{P}(Y=1|X=x) < \alpha$ 이면 자료를 $Y=0$인 클래스로 분류할 수 있다.

절단값 α를 결정할 때, 고려해야 하는 사항 첫 번째는 사전정보 고려이다.
사전정보에서 $y=1$인 자료가 상대적으로 많다면 절단값을 0.5보다 작은 값으로 고려할 수 있다.

두 번째로 적절한 손실함수를 고려해야 한다.
$y=1$인 자료를 잘못 분류하는 손실이 $y=0$인 자료를 잘못 분류하는 손실에 비해 손실 정도가 심각하게 크다고 판단하는 경우 절단값 α를 작게 잡을 수 있다.

그 밖에도 전문가 의견이나 민감도, 특이도 등을 고려하여 α값을 결정할 수 있다.

3. Decision Tree

● 의사결정나무 개요

- 주어진 입력값에 대한 출력값을 예측하는 것으로서 그 결과를 나무형태의 그래프로 표현

- 지도학습 기법으로 각 변수의 영역을 반복적으로 분할하여 전체 영역에서의 규칙을 생성

- 예측력은 다른 지도학습 기법들에 비해 떨어지나 해석력이 좋음

- 규칙은 if-then 형식으로 이해가 쉬움

● 예측력과 해석력

- 예측력만이 중요한 경우

  (예제) 홍보책자 발송회사가 기대집단의 사람들이 가장 많은 반응을 보일 고객 유치방안을 위한 예측

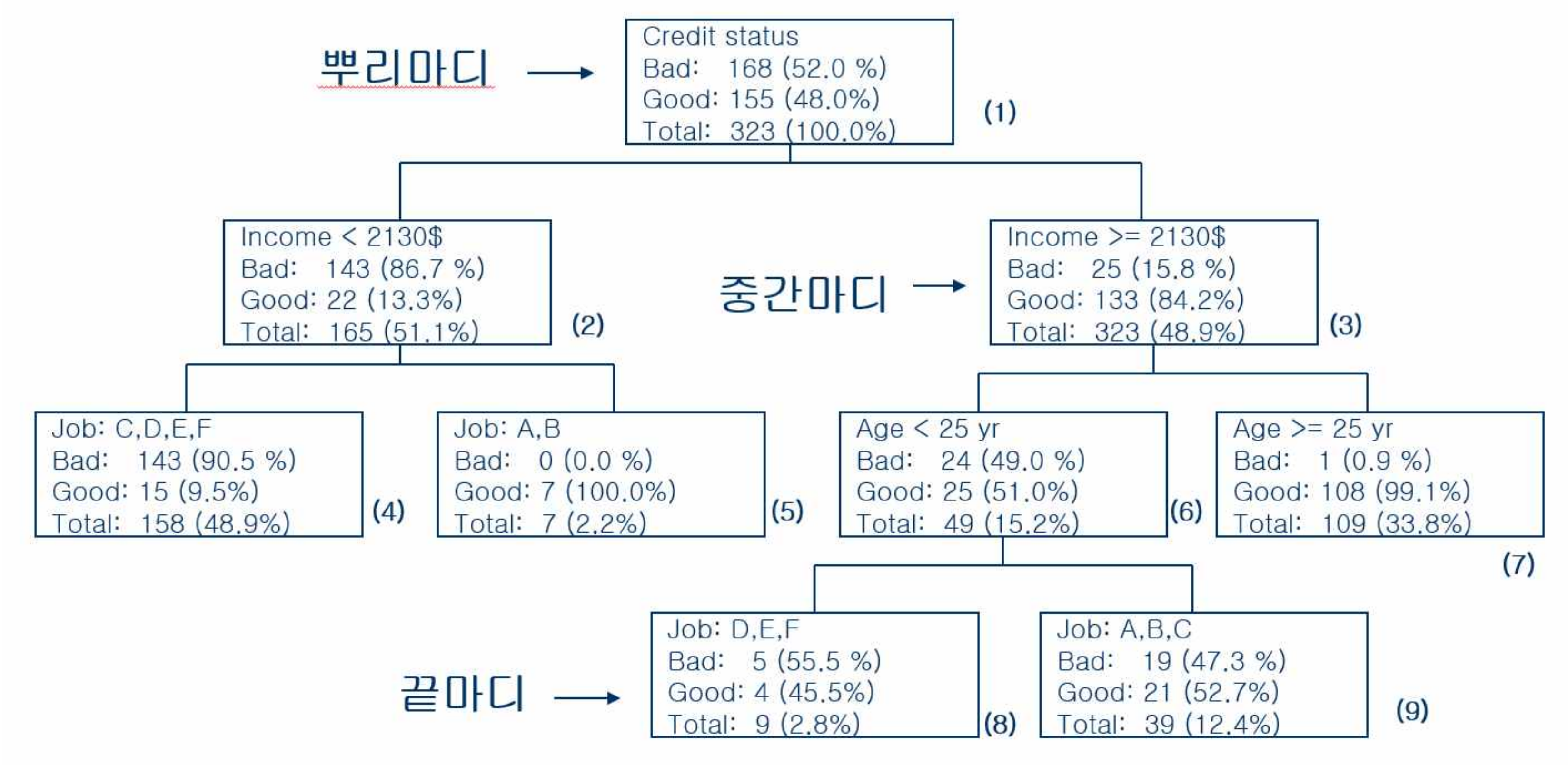
- 해석력이 중요한 경우

  : 많은 분야에서는 결정을 내리게 된데 대한 이유를 설명하는 능력이 중요함 (해석력)

  (예제) 은행의 대출심사 결과 부적격 판정이 나온 경우 고객에게 부적격 이유를 설명하여야 함

- 의사결정나무는 좋은 해석력을 갖음

● 의사결정나무 예시


뿌리마디
Credit status
Bad: 168 (52.0 %)
Good: 155 (48.0%)
Total: 323 (100.0%)
(1)
Income < 2130$
Bad: 143 (86.7 %)
Good: 22 (13.3%)
Total: 165 (51.1%)
(2)
중간마디
Income >= 2130$
Bad: 25 (15.8 %)
Good: 133 (84.2%)
Total: 323 (48.9%)
(3)
Job: C,D,E,F
Bad: 143 (90.5 %)
Good: 15 (9.5%)
Total: 158 (48.9%)
(4)
Job: A,B
Bad: 0 (0.0 %)
Good: 7 (100.0%)
Total: 7 (2.2%)
(5)
Age < 25 yr
Bad: 24 (49.0 %)
Good: 25 (51.0%)
Total: 49 (15.2%)
(6)
Age >= 25 yr
Bad: 1 (0.9 %)
Good: 108 (99.1%)
Total: 109 (33.8%)
(7)
끝마디
Job: D,E,F
Bad: 5 (55.5 %)
Good: 4 (45.5%)
Total: 9 (2.8%)
(8)
Job: A,B,C
Bad: 19 (47.3 %)
Good: 21 (52.7%)
Total: 39 (12.4%)
(9)

● 의사결정나무의 구성요소

- 뿌리마디(root node) : 시작되는 마디로 전체 자료 포함
- 자식마디(child node) : 하나의 마디로부터 분리되어 나간 2개 이상의 마디들
- 부모마디(parent node) : 주어진 마디의 상위마디
- 끝마디(terminal node) : 자식마디가 없는 마디
- 중간마디(internal node) : 부모마디와 자식마디가 모두 있는 마디
- 가지(branch) : 뿌리마디로부터 끝마디까지 연결된 마디들
- 깊이(depth) : 뿌리마디부터 끝마디까지의 중간마디들의 수

의사결정나무 형성

● 의사결정나무 구축을 위한 질문들

■ 질문

- 뿌리마디의 질문이 왜 소득인가?
- 4번, 5번, 7번 마디들은 끝마디인 반면 6번 마디는 왜 중간마디인가?
- 7번 마디에 속하는 자료는 신용상태를 어떻게 결정하여야 하는가?

■ 즉, 의사결정나무의 생성요소는 다음과 같음

- 분할 기준 (splitting rule)의 선택
- 분할을 계속할 것인지 그만 둘 것 인지를 결정 (stopping rule and pruning rule)
- 각 끝마디에 예측값의 할당

● **의사결정나무의 형성과정**

- 성장(growing) : 최적의 분리규칙을 찾아서 나무를 성장시키는 과정
- 가지치기(pruning) : 불필요한(오차 크게 할 위험, 부적절한 규칙) 가지 제거
- 타당성 평가 : 이익도표 혹은 시험자료로 평가
- 해석 및 예측 : 모형을 해석하고 예측모형을 설정한 후 예측에 적용

● **출력변수가 연속형인지 범주형인지에 따라 회귀나무와 분류나무로 구분**

4. R 실습 : Data 분석

UCI 머신러닝 예제 데이터 아카이브

Data : Adult
(https://archive.ics.uci.edu/ml/datasets/Adult)
(https://goo.gl/yV0qq)

목적: 설명변수에 근거해서 연소득(wage)이 $50K가 넘는지 예측

UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

About Citation Policy Donate a Data Set Contact

Search

Repository Web

View ALL Data Sets

Adult Data Set

Download: Data Folder, Data Set Description

Abstract: Predict whether income exceeds $50K/yr based on census data. Also known as "Census Income" dataset.

Data Set Characteristics:	Multivariate	**Number of Instances:**	48842	**Area:**	Social
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	14	**Date Donated**	1996-05-01
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	1060616

Source:

Donor:

Ronny Kohavi and Barry Becker
Data Mining and Visualization
Silicon Graphics.
e-mail: ronnyk '@' live.com for questions.

Data Set Information:

Extraction was done by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0))

Prediction task is to determine whether a person makes over 50K a year.

Attribute Information:

Listing of attributes:

>50K, <=50K.

```
install.packages(c("dplyr", "ggplot2", "ISLR", "MASS","glmnet", "rpart","boot"))

library(dplyr)
library(ggplot2)
library(ISLR)
library(MASS)
library(glmnet)
library(rpart)
library(boot)
adult <- read.csv("C:/Users/miaeoh/Desktop/기계학습을이용한빅데이터분석강좌(2018)/R실습(2)/adult.txt",header=FALSE, strip.white=TRUE)

names(adult) <- c('age','workclass','fnlwgt',
'education','education-num','marital-status',
'occupation','relationship','race',
'sex','capital-gain','capital-loss','hours-per-week','native-country','wage')
head(adult)
```

glimpse(adult)

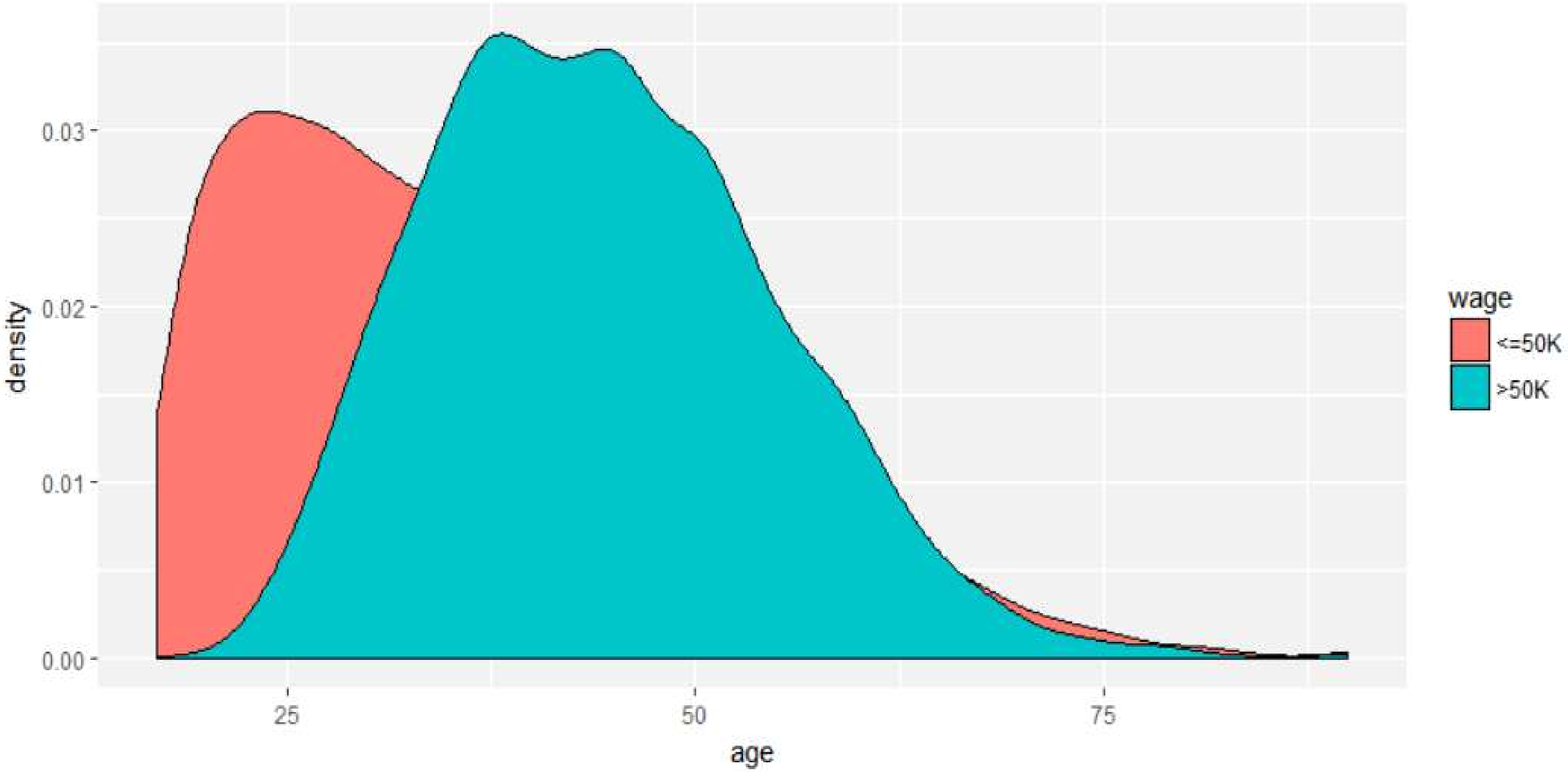
```
> glimpse(adult)
Observations: 32,561
Variables: 15
$ age            <int> 39, 50, 38, 53, 28, 37, 49, 52, 31, 42, 37, 30, 23, 32, 40, 34,
$ workclass      <fctr> State-gov, Self-emp-not-inc, Private, Private, Private, Privat
$ fnlwgt         <int> 77516, 83311, 215646, 234721, 338409, 284582, 160187, 209642, 4
$ education      <fctr> Bachelors, Bachelors, HS-grad, 11th, Bachelors, Masters, 9th,
$ education-num  <int> 13, 13, 9, 7, 13, 14, 5, 9, 14, 13, 10, 13, 13, 12, 11, 4, 9, 9
$ marital-status <fctr> Never-married, Married-civ-spouse, Divorced, Married-civ-spous
$ occupation     <fctr> Adm-clerical, Exec-managerial, Handlers-cleaners, Handlers-cle
$ relationship   <fctr> Not-in-family, Husband, Not-in-family, Husband, Wife, Wife, No
$ race           <fctr> White, White, White, Black, Black, White, Black, White, White,
$ sex            <fctr> Male, Male, Male, Male, Female, Female, Female, Male, Female,
$ capital-gain   <int> 2174, 0, 0, 0, 0, 0, 0, 0, 14084, 5178, 0, 0, 0, 0, 0, 0, 0, 0,
$ capital-loss   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
$ hours-per-week <int> 40, 13, 40, 40, 40, 40, 16, 45, 50, 40, 80, 40, 30, 50, 40, 45,
$ native-country <fctr> United-States, United-States, United-States, United-States, Cu
$ wage           <fctr> <=50K, <=50K, <=50K, <=50K, <=50K, <=50K, <=50K, >50K, >50K, >
```

summary(adult)

```
levels(adult$wage)   # 각 레벨은 내부적으로 수치값 1과 2에 대응

##  더미 변수 생성

levels(adult$race)
levels(adult$sex)

x <- model.matrix(~race + sex + age, adult)
glimpse(x)
head(x)
head(adult)

colnames(x)
```

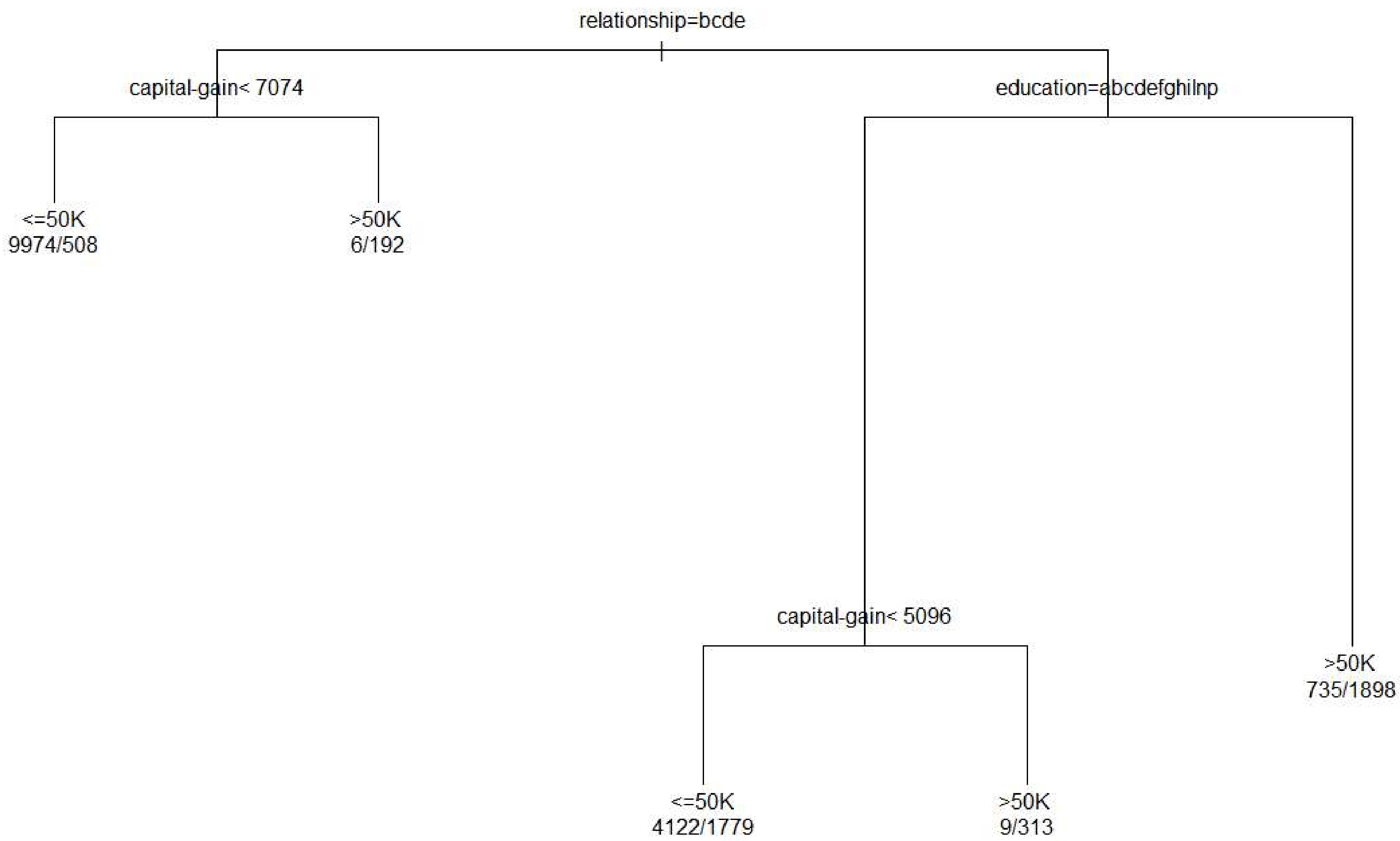
```
####################################
#   훈련,검증, 테스트셋 구분                    #
####################################
set.seed(1)
n=nrow(adult)
idx <- 1:n

training.idx <- sample(idx,n*.6)
idx <- setdiff(idx,training.idx)   # 나머지
validate.idx <- sample(idx,n*.2)
test.idx <- setdiff(idx,validate.idx)

length(training.idx) ;length(validate.idx) ; length(test.idx)

training <- adult[training.idx,]
validation <- adult[validate.idx,]
test <- adult[test.idx,]
```

```
#  시각화
training %>%
  ggplot(aes(age,fill=wage)) +
  geom_density(aplha=.5)
```



```
## Logistic

ad_glm_full <- glm(wage~., data=training, family=binomial)
summary(ad_glm_full)

summary.glm(ad_glm_full)



predict(ad_glm_full, newdata=adult[1:5,], type="response")

y_obs <- ifelse(validation$wage==">50K",1,0)

yhat_lm <- predict(ad_glm_full, newdata=validation, type="response")

pre_y_obs <-(yhat_lm >= 0.5)*1

table(y_obs, pre_y_obs)
```

```
#######################
#    Decision Tree           #
#######################

cvr_tr <- rpart(wage~. , data=training)
cvr_tr


summary(cvr_tr)


opar <- par(mfrow=c(1,1),xpd=NA)
plot(cvr_tr)
text(cvr_tr, use.n=TRUE)
par(opar)
```


relationship=bcde
capital-gain< 7074
education=abcdefghilnp
<=50K
9974/508
>50K
6/192
capital-gain< 5096
>50K
735/1898
<=50K
4122/1779
>50K
9/313