5장

# 분류분석(Classification Analysis I)

## 5.1 분류분석의 기본개념

▸ **분류분석**: 소속 집단을 알고 있는 데이터를 이용하여 모형을 만들어서 소속집단을 모르는 데이터들의 집단을 결정하는 기법

▸ 통계학에서는 판별분석(discriminant analysis) 또는 계획된 기계학습(supervised learning)이라 함.

▸ **군집분석(cluster analysis) 또는 비계획된 기계학습(unsupervised learning)**: 소속집단을 모르는 데이터를 이용하여 유사한 데이터들의 집단을 결정하는 기법

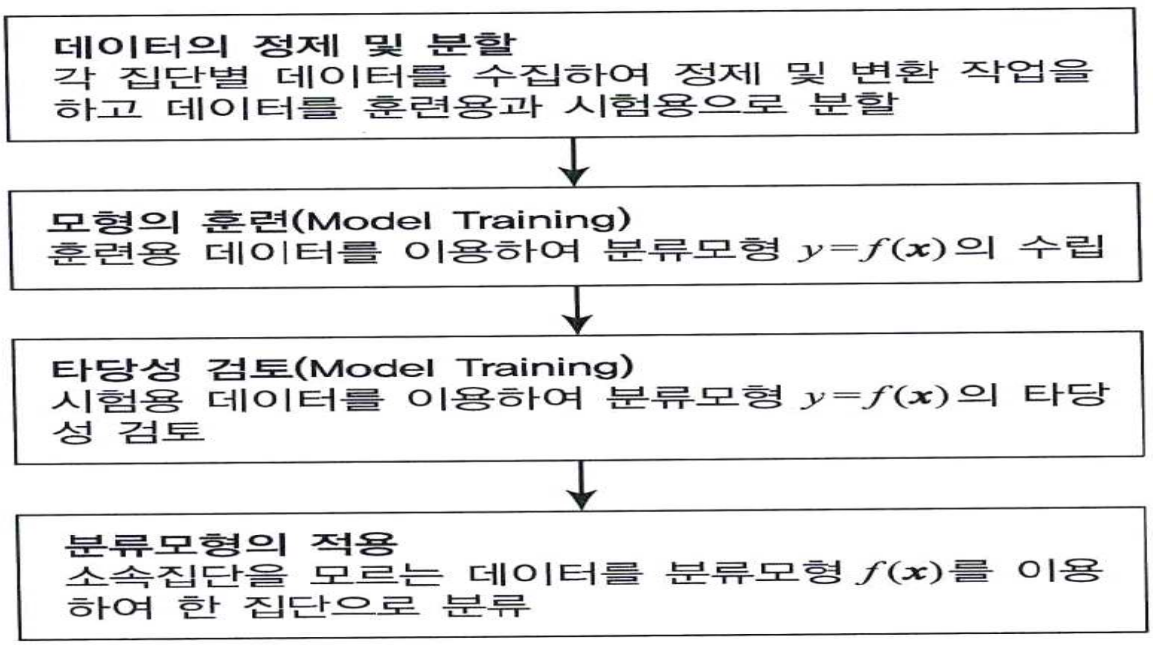
예) 의사가 환자를 진찰하여 그 진찰기록으로 어느 병인지 분류하는 일
이메일 제목을 이용하여 스팸메일을 가려내는 일
백화점의 매장을 방문한 고객이 상품을 구매할 고객인지 구별하는 일

▸분류모형: 베이즈분류(Bayes classficiation) 모형, 로지스틱회귀(logistic regression) 모형, 의사결정나무(decision tree)

### 5.1.1 분류분석의 절차

▸분류집단(group or class)가 k개 있다고 가정하자.
: $G_1, G_2, \cdots, G_k$ 또는 1, 2, ..., k로 표시.

▸확률변수 $X=(X_1, X_2, \cdots, X_m)$에 대해 관측 데이터를 $x=(x_1, x_2, \cdots, x_m)$이라 하고, 각 집단에서 $n_1, n_2, \cdots, n_k$개의 데이터가 관측되었다고 가정하자.



### 5.1.2 분류분석을 위한 데이터의 준비

분류의 정확성, 효율성, 확장 적용성 등을 향상시키기 위하여 다음과 같은 사전 준비를 하여야 함.

가. 데이터 정제

데이터에 잡음(noise)가 있으면 제거, 결측값(missing value)가 있으면 해당 변수의 평균 또는 최빈값을 사용하여 전처리 해 줌.

나. 관련성 분성

관련성을 분석하여 분류와 관련이 없는 변수나 중복 변수 제거

다. 데이터 변환

분류를 위해서는 연속형 데이터를 이산형화해야될 경우가 있음

### 5.1.3 분류모형의 평가 측도

두 개의 집단 $G_1, G_2$가 있고, n개의 소속을 아는 데이터가 있다면 분류모형을 이용하여 각 데이터를 분류한 후, 실제 집단과 모형에 의해 분류된 집단을 비교하여 그 결과를 요약

| | | 분류된 집단 | |
|---|---|---|---|
| | | $G_1$ | $G_2$ |
| 실제 집단 | $G_1$ | $f_{11}$ | $f_{12}$ |
| | $G_2$ | $f_{21}$ | $f_{22}$ |

▸ $f_{ij}$: 집단 $G_i$의 데이터가 집단 $G_j$로 분류된 수

▸ 전체 n개의 데이터 중 올바로 분류된 데이터의 수: $f_{11}+f_{22}$

▸ 전체 n개의 데이터 중 잘못 분류된 데이터의 수: $f_{12}+f_{21}$

▸ 분류 모형의 정확도(accuracy): 전체 데이터의 수 중 올바로 분류된 수의 비율 = $\frac{f_{11}+f_{22}}{n}$

▸ 오류율(error rate): 전체 데이터 중 오분류된 수의 비율 = $\frac{f_{12}+f_{21}}{n}$

▸ 일반적으로 분류모형은 정확도를 최대화하거나 오류율을 최소화하는 알고리즘을 찾도록 함

### 5.1.4 훈련용과 시험용 데이터의 분할 방법

▸ 분류모형을 객관적으로 평가하기 위해 전체 데이터의 집합을 훈련용 데이터(training data)와 시험용 데이터(testing data)로 분할

▸ 훈련용 데이터를 이용하여 모형을 수립하고, 시험용 데이터를 이용하여 모형의 정확도를 평가

▸ 데이터가 충분할 경우에는 모형의 성능을 개선하는데 사용하는 검증용 데이터(validating data)를 별도

로 두기도 함.

가. 예비법(Holdout Method)과 랜덤부표집(Random Subsampling)

예비법(Holdout Method)
▸전체 데이터를 서로 겹치지 않는 두 개의 데이터 집합으로 분할하여 그 중 하나를 훈련용 데이터, 나머지를 시험용 데이터로 예비(holdout)해 둠
▸훈련용 데이터를 이용하여 분류모형 수립하고, 이 모형을 시험용 데이터에 적용하여 정확도(오류율)을 측정하는 방법
(1/2 훈련용: 1/2 시험용) 또는 (2/3 훈련용: 1/3 시험용)으로 분할하는 방법이 많이 이용
▸결정한 비율로 데이터를 추출할 때는 편향(bias)를 줄이기 위해 비복원(without replacement) 단순확률추출(simple random sampling) 방법 이용

주의 사항
1) 훈련용 데이터의 수가 적은 경우, 훈련용 데이터만으로 만든 분류모형은 전체 데이터를 모두 사용하여 만든 모형보다 좋지 않을 수 있음.
2) 훈련용 데이터 수가 적을수록 모형의 정확도에 대한 분산은 커짐. 반면 훈련용 데이터 수가 커지면, 시험용 데이터에서 추정된 정확도는 신뢰도가 떨어짐.
3) 훈련용과 시험용 데이터는 전체 데이터의 부분집합이므로 서로 독립이 아님

▸이와 같은 문제점을 해결하고 분류모형 정확도에 대한 신뢰도를 높이기 위해 예비법을 반복하여 실시
▸각각의 비복원 추출된 표본을 부표집(subsampling)이라 부르고, 이를 반복하여 추출하는 방법을 랜덤부표집(random subsampling) 방법이라 부름
▸$accuracy_i$: i번째 부표집에 의한 분류모형의 정확도
▸이 실험을 r번 반복한다면 전체 분류 모형의 정확도(accuracy)는 각 정확도의 평균으로 정의함.

전체 정확도(accuracy)=$\frac{1}{r}\sum_{i=1}^{r} accuracy_i$
: 반복해서 모형의 정확도를 추정하기 때문에 신뢰도를 높일 수 있음

나. 교차타당성(cross validation) 방법
: 랜덤 부표집의 문제점을 해결하려는 방법

**▸2중 교차타당성 방법(two-fold cross validation):** 전체 데이터의 집합을 훈련용과 시험용으로 나눔 -> 훈련용 데이터로서 모형을 만들고 시험용 데이터를 이용하여 모형이 정확히 분류한 데이터의 수를 기록 -> 시험용 데이터를 훈련용으로, 훈련용 데이터는 시험용으로 역할을 교환하여 정확히 분류된 데이터의 수를 합산 -> 같은 방법으로 실험을 r번 반복하여 평균 정확도를 구함

▸2중 교차타당성을 확장하면 k중(k-fold) 교차타당성 방법을 만들 수 있음: 전체 데이터를 k개의 같은 크기의 부분집합으로 나눈 후, 이 중 한 부분집합을 시험용 데이터로 예비하여 두고 나머지 데이터를 훈련용으로 하여 분류함수를 구하는 것. 이 방법을 k번 반복하여 각 데이터 부분집합이 한 번씩 시험용으로 사용. 분류모형의 정확도는 k번 측정된 정확도의 평균으로 구함

▸**한 데이터 예비법(leave-one-out)**: 시험용 데이터가 1개, 나머지가 훈련용 데이터인 경우
훈련용 데이터를 최대로 할 수 있음. 그러나, 시험용 데이터가 1개이므로 추정된 정확도의 분산이 커지는 단점이 있음. 실험을 데이터 수만큼 반복함으로써 시간이 많이 걸림

**다. 붓스트랩(Bootstrap) 방법**
훈련용 데이터를 복원추출(with replacement) 함.
전체 데이터 수가 N개일 때, 붓스트랩으로 N개의 데이터를 추출하면 대략 전체 데이터의 63.2% 정도가 훈련용 데이터로 추출됨. 붓스트랩으로 추출되지 않은 표본은 시험용 데이터로 사용

## 5.2 베이즈 분류(Bayes classification) 모형

▸각 집단으로 분류되는 사전확률(prior probability)과 집단별 가능도확률(likelihood probability)을 알 때, 소속집단을 모르는 데이터에 대하여 베이즈 정리(Bayes theorem)를 이용한 사후확률(posterior probability)을 구하여 그 확률이 높은 집단으로 분류하는 방법

### 5.2.1 한 변수일 경우의 베이즈 분류

▸사전확률에 의한 분류: 사전확률을 비교하여 소속집단을 모르는 데이터를 확률이 높은 집단으로 분류
▸상품 구매 집단: $G_1$, 비구매 집단:$G_2$
▸사전확률: $P(G_1)$, $P(G_2)$

| 사전확률에 의한 분류규칙 |
|---|
| $P(G_1) \geq P(G_2)$이면 데이터를 집단 $G_1$으로 분류, 아니면 $G_2$로 분류 |

**가. 한 이산형 변수의 베이즈 분류**

▸구매 집단과 비구매 집단의 나이에 대한 분포를 구할 수 있으면 구매여부를 판단할 수 있는 유용한 정보가 됨: 이를 가능도확률분포(likelihood probability distribution) 또는 집단별 확률분포(class probability distribution)이라 함.
▸베이즈 정리(Bayes theorem)을 이용하여 각 집단의 사후확률을 구할 수 있음
▸사후확률을 이용한 분류를 베이즈 분류(Bayes classification)라고 함

**예제** 5.2.2

| | 구매 집단($G_1$) | 비구매 집단($G_2$) |
|---|---|---|
| 20대 | 2 | 8 |
| 30대 | 6 | 4 |
| 합계 | 8 | 12 |

어느 날 이 상품 매장을 방문한 고객이 20대라면 이 고객이 상품을 구매할 사람인지 아닌지 사후확률을 구하여 분류하여라.

(풀이) $P(G_1)=\frac{8}{20},\ P(G_2)=\frac{12}{20}$

각 집단의 나이분포를 가능도 확률분포라고 함.

$P(x=20\text{대}\,|G_1)=\frac{2}{8},\ P(x=20\text{대}\,|G_2)=\frac{8}{12}$

$$P(G_1|x=20\text{대})=\frac{P(x=20\text{대}\,|G_1)P(G_1)}{P(x=20\text{대}\,|G_1)P(G_1)+P(x=20\text{대}\,|G_2)P(G_2)}=\frac{\frac{2}{8}\times\frac{8}{20}}{\frac{2}{8}\times\frac{8}{20}+\frac{8}{12}\times\frac{12}{20}}=0.2$$

$$P(G_2|x=20\text{대})=\frac{P(x=20\text{대}\,|G_2)P(G_2)}{P(x=20\text{대}\,|G_1)P(G_1)+P(x=20\text{대}\,|G_2)P(G_2)}=\frac{\frac{8}{12}\times\frac{12}{20}}{\frac{2}{8}\times\frac{8}{20}+\frac{8}{12}\times\frac{12}{20}}=0.8$$

$P(G_1|\,x=20\text{대})<P(G_2|\,x=20\text{대})$이므로, 이 고객은 비구매 고객으로 분류함.

| 사후확률에 의한 분류규칙 |
|---|
| $P(G_1\|x)\geqq P(G_2\|x)$이면 $x$를 $G_1$으로 분류, 아니면 $G_2$로 분류 |

| 사후확률에 의한 분류규칙 |
|---|
| $\frac{P(x\|G_1)}{P(x\|G_2)}\geqq\frac{P(G_2)}{P(G_1)}$이면 $x$를 $G_1$으로 분류, 아니면 $G_2$로 분류 |

**나. 한 연속형 변수의 베이즈분류**

**예제 5.2.3**

나이변수가 연속형이고 구매 집단은 평균이 35세, 표준편차가 2세인 정규분포이며 비구매 집단은 평균이 25세, 표준편차가 2세인 정규분포라 가정하자. 어느 날 이 상품 매장을 방문한 고객이 30세라면 이 고객이 이 상품을 구매할 사람인지 아닌지 사후확률을 구하여 분류하여라.

(풀이) $P(x|\,G_1)=\frac{1}{\sqrt{2\pi}2}e^{-\frac{(x-35)^2}{2\times4}},\quad P(x|\,G_2)=\frac{1}{\sqrt{2\pi}2}e^{-\frac{(x-25)^2}{2\times4}}$

$P(G_1)=\frac{8}{20}=0.4,\qquad P(G_2)=\frac{12}{20}=0.6$

$\frac{P(x|\,G_1)}{P(x|\,G_2)}=\exp\left\{-\frac{(x-35)^2}{8}+\frac{(x-25)^2}{8}\right\}\geq\frac{0.6}{0.4}$이면, $x$를 $G_1$으로 분류, 아니면 $G_2$로 분류

$$-\frac{(x-35)^2}{8}+\frac{(x-25)^2}{8}\geq\log\left(\frac{3}{2}\right)\ \Rightarrow-(x^2-70x+35^2)+(x^2-50x+25^2)\geq8\log(1.5)$$

$$\Rightarrow x\geq30.16219$$

$\therefore\ x\geq30.16219$이면 데이터 $x$를 $G_1$으로 분류, 아니면 $G_2$로 분류

$\therefore$ 30세 고객은 비구매 고객($G_2$)으로 분류

## 5.2.2 다중변수의 베이즈 분류

| m개의 확률변수 $X=(X_1, X_2, \cdots, X_m)$, k개의 집단 $G_1, G_2, \cdots, G_k$의 사전확률을 $P(G_1), P(G_2), \cdots, P(G_k)$라 하고, 각 집단별 가능확률분포를 $P(X\|G_1), P(X\|G_2), \cdots, P(X\|G_k)$인 경우 |
|---|
| 데이터 $x$는 사후확률 $P(G_1\|x), P(G_2\|x), \cdots, P(G_k\|x)$가 가장 큰 집단으로 분류 |

$f_i(x)$: 각 집단별 가능도확률분포함수

| 베이즈분류규칙 - 여러 집단의 경우 |
|---|
| 모든 $k \neq i$에 대하여 $P(G_k)f_k(x) \geqq P(G_i)f_i(x)$이면 $x$를 집단 $G_k$으로 분류 |

| 베이즈분류규칙 - 두 집단의 경우 |
|---|
| $\frac{f_1(x)}{f_2(x)} \geqq \frac{P(G_2)}{P(G_1)}$이면 $x$를 $G_1$으로 분류, 아니면 $G_2$로 분류 |

### 가. 이산형 다차원분포의 단순베이즈 분류

두 개 이상의 이산형 변수가 있는 경우 변수들이 서로 독립인 경우의 베이즈 분류를 단순베이즈 분류(naive bayes classification)라 함

**예제 5.2.4**

나이를 20대와 30대, 월수입을 200만원 미만과 200만원 이상으로 이산형화하여 다차원 도수분포를 구하여라.

어느 날 이 상품 매장을 방문한 고객이 33세이고 월수입은 190만원, 신용상태는 양호하다면 이 고객이 상품을 구매할 사람인지 아닌지 사후확률을 구하여 분류하여라.

(풀이) $x_1$ = 30대, $x_2$ = 200만원 미만, $x_3$ = 양호

$P(x|G_1) = \frac{1}{8}$, $P(x|G_2) = \frac{0}{12} = 0 \Rightarrow$ 표본수가 충분치 않아서 가능도 확률분포를 올바르게 추정하기 어려움.

이런 경우, $X_1$, $X_2$, $X_3$가 독립이라 가정함.

$P(X=(X_1,X_2,X_3)\,|\,G_i) \approx P(X_1\,|\,G_i)P(X_2\,|\,G_i)P(X_3\,|\,G_i)$

$P(x=(\text{30대, 200만원 미만, 양호})\,|\,G_1) \approx \frac{6}{8}\times\frac{2}{8}\times\frac{4}{8} = 0.0938$

$P(x=(\text{30대, 200만원 미만, 양호})\,|\,G_2) \approx \frac{4}{12}\times\frac{4}{12}\times\frac{7}{12} = 0.0648$

$$P(G_1|x) = \frac{P(x|G_1)P(G_1)}{P(x|G_1)P(G_1)+P(x|G_2)P(G_2)} = \frac{0.0938\times0.4}{0.0938\times0.4+0.0648\times0.6} = 0.4911$$

$$P(G_2|x) = \frac{P(x|G_2)P(G_2)}{P(x|G_1)P(G_1)+P(x|G_2)P(G_2)} = \frac{0.0648\times0.6}{0.0938\times0.4+0.0648\times0.6} = 0.5089$$

$P(G_1\,|\,x) < P(G_2\,|\,x)$이므로, 이 고객은 비구매 고객($G_2$)으로 분류함.

**[표 5.3]** 각 고객의 나이, 월수입, 신용상태별 상품 구입여부

| 번호 | 나이($X_1$) | 월수입($X_2$) (단위: 만원) | 신용상태($X_3$) | 상품 구매여부($y$) |
|---|---|---|---|---|
| 1 | 25 | 150 | 양호 | 구매 |
| 2 | 34 | 220 | 우수 | 비구매 |
| 3 | 27 | 210 | 양호 | 비구매 |
| 4 | 28 | 250 | 양호 | 구매 |
| 5 | 21 | 100 | 나쁨 | 비구매 |
| 6 | 31 | 220 | 양호 | 비구매 |
| 7 | 36 | 300 | 우수 | 구매 |
| 8 | 20 | 100 | 양호 | 비구매 |
| 9 | 29 | 220 | 우수 | 비구매 |
| 10 | 32 | 250 | 양호 | 구매 |
| 11 | 37 | 400 | 우수 | 구매 |
| 12 | 24 | 120 | 양호 | 비구매 |
| 13 | 33 | 350 | 양호 | 비구매 |
| 14 | 30 | 180 | 양호 | 구매 |
| 15 | 38 | 350 | 우수 | 구매 |
| 16 | 32 | 250 | 양호 | 비구매 |
| 17 | 28 | 240 | 나쁨 | 비구매 |
| 18 | 22 | 220 | 나쁨 | 비구매 |
| 19 | 39 | 450 | 우수 | 구매 |
| 20 | 26 | 150 | 양호 | 비구매 |

**[표 5.4]** 각 고객의 나이, 월수입, 신용상태별 상품 구입여부

| 나이($X_1$) | 구매집단($G_1$) | 비구매집단($G_2$) |
|---|---|---|
| 20대 | 2 | 8 |
| 30대 | 6 | 4 |
| 총합계 | 8 | 12 |

| 월수입($X_2$) | 구매집단($G_1$) | 비구매집단($G_2$) |
|---|---|---|
| 200만원 미만 | 2 | 4 |
| 200만원 이상 | 6 | 8 |
| 총합계 | 8 | 12 |

| 신용상태($X_3$) | 구매집단($G_1$) | 비구매집단($G_2$) |
|---|---|---|
| 나쁨 | | 3 |
| 양호 | 4 | 7 |
| 우수 | 4 | 2 |
| 총합계 | 8 | 12 |

**[표 5.5]** 변수 나이, 월수입, 신용상태별 상품 구입여부 교차표

| 나이($X_1$) | 월수입($X_2$) | 신용상태($X_3$) | 구매집단($G_1$) | 비구매집단($G_2$) |
|---|---|---|---|---|
| 20대 | 200만원 미만 | 나쁨 | | 1 |
| | | 양호 | 1 | 3 |
| | | 우수 | | |
| | 200만원 이상 | 나쁨 | | 2 |
| | | 양호 | 1 | 1 |
| | | 우수 | | 1 |
| 30대 | 200만원 미만 | 나쁨 | | |
| | | 양호 | 1 | |
| | | 우수 | | |
| | 200만원 이상 | 나쁨 | | |
| | | 양호 | 1 | 3 |
| | | 우수 | 4 | 1 |
| 총합계 | | | 8 | 12 |

▸단순베이즈 분류: 각각의 변수를 독립이라 가정하고, 근사적으로 집단별 사후확률분포 구한 후 분류
▸변수들이 연관이 있을 경우 단순베이즈 분류는 정확치 않은 결과를 준다.

**나. 연속형 다변량 정규분포의 베이즈 분류**

$G_1 : f_1(x) \sim N(\mu_1, \Sigma_1)$ 여기서 $f_i(x)$: 그룹 $i$의 가능도 확률분포함수
$G_2 : f_2(x) \sim N(\mu_2, \Sigma_2)$

· 두 집단이 다변량 정규분포이고 공분산이 다른 경우 ($\Sigma_1 \neq \Sigma_2$)

$d^Q(x) = -\frac{1}{2} ln \frac{|\Sigma_1|}{|\Sigma_2|} - \frac{1}{2}(x-\mu_1)^t \Sigma_1^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)^t \Sigma_2^{-1}(x-\mu_2) \geq \ln \frac{P(G_2)}{P(G_1)}$ 이면

$x$를 $G_1$으로 분류하고 아니면 $G_2$로 분류

· 두 집단이 다변량 정규분포이고 공분산이 같은 경우 ($\Sigma_1 = \Sigma_2 = \Sigma$)

$d^L(x) = (\mu_1 - \mu_2)^t \Sigma^{-1} [x - \frac{1}{2}(\mu_1 + \mu_2)] \geq \ln \frac{P(G_2)}{P(G_1)}$ 이면 $x$를 $G_1$으로 분류하고 아니면 $G_2$로 분류

· 두 집단이 다변량 정규분포이고 공분산이 같은 경우 표본이용 분류 규칙 ($\Sigma_1 = \Sigma_2 = \Sigma$)

만약 $\mu_1, \mu_2, \Sigma$가 알려져 있지 않으면, 표본평균 $\overline{x_1}$, $\overline{x_2}$, 표본공분산행렬 $S$를 이용하여 추정한 분류함수

$\hat{d}^L(x) = (\overline{x_1} - \overline{x_2})^t S^{-1}[x - \frac{1}{2}(\overline{x_1} + \overline{x_2})] \geq \ln\frac{P(G_2)}{P(G_1)}$ 이면 $x$를 $G_1$으로 분류하고 아니면 $G_2$로 분류

**예제 5.2.5**

나이와 월수입이 공분산이 같은 다변량 정규분포를 따른다고 가정하자. 어느 날 이 상품 매장을 방문한 고객이 33세이고 월수입은 200만원인 사람이 상품을 구매할 사람인지 아닌지 베이즈 분류를 하라.

**[표 5.6] 각 고객의 나이, 월수입별 상품 구매여부**

| 번호 | 나이($X_1$) | 월수입($X_2$) (단위: 만원) | 상품 구매여부($y$) |
|---|---|---|---|
| 1 | 25 | 150 | 구매 |
| 2 | 34 | 220 | 비구매 |
| 3 | 27 | 210 | 비구매 |
| 4 | 28 | 250 | 구매 |
| 5 | 21 | 100 | 비구매 |
| 6 | 31 | 220 | 비구매 |
| 7 | 36 | 300 | 구매 |
| 8 | 20 | 100 | 비구매 |
| 9 | 29 | 220 | 비구매 |
| 10 | 32 | 250 | 구매 |
| 11 | 37 | 400 | 구매 |
| 12 | 24 | 120 | 비구매 |
| 13 | 33 | 350 | 비구매 |
| 14 | 30 | 180 | 구매 |
| 15 | 38 | 350 | 구매 |
| 16 | 32 | 250 | 비구매 |
| 17 | 28 | 240 | 비구매 |
| 18 | 22 | 220 | 비구매 |
| 19 | 39 | 450 | 구매 |
| 20 | 26 | 150 | 비구매 |

(풀이)

$\overline{x_1} = \begin{pmatrix}33.1250\\291.250\end{pmatrix}$: 구매 집단의 나이, 월수입의 평균, $\overline{x_2} = \begin{pmatrix}27.2500\\200.000\end{pmatrix}$: 비구매 집단의 나이, 월수입의 평균

$S = \begin{pmatrix}31.6211 & 470.1053\\470.1053 & 9129.2105\end{pmatrix}$: 표본공분산행렬

$P(G_1) = \frac{8}{20} = 0.4, \qquad P(G_2) = \frac{12}{20} = 0.6$

$(\overline{x_1} - \overline{x_2})^t S^{-1}[x - \frac{1}{2}(\overline{x_1} + \overline{x_2})] \geq \ln\frac{P(G_2)}{P(G_1)}$ , 여기서 $x = (x_1, x_2)$

$\Rightarrow 0.1587 \times x_1 + 0.0018 \times x_2 - 5.2377 \geq 0.4055$이면 $x$를 구매 고객, 아니면 비구매 고객으로 분류한다.

$x_1 = 33$, $x_2 = 200$를 식에 대입하면, $0.1587 \times 33 + 0.0018 \times 200 - 5.2377 = 0.36291$

$\therefore$ 33세이고 월수입은 200만원인 사람은 상품을 구매하지 않을 고객으로 분류한다.

### 5.2.3 변수의 선택 - 단계적 분류분석

연속형 변수 - 이산형화하여 베이즈 분류 적용
변수가 많은 경우 - 집단의 분류에 도움이 되는 변수만 선택해서 적용

단계적 분류분석(stepwise classification analysis): 적절한 변수를 단계적으로 선택하여 분류하는 방법

변수 선택 - 집단 변수를 가장 잘 설명할 수 있는 변수 사용. 즉, 판별력(discriminatory power)이 큰 변수 사용. 예) 전진선택법, 후진소거법, 단계적 방법

### 5.2.4 베이즈 분류의 특성

1) 베이즈 분류는 사후확률로 분류하기 때문에 모형의 과적합 위험성이 적고 견고(robust) 함
2) 베이즈 분류는 불완전한 데이터나 특이값 및 결측값이 있는 경우에도 안정된 분류를 할 수 있음.

## 5.3 로지스틱회귀(Logistic Regression) 모형

▸ 선형회귀(linear regression) 모형은 $m$개의 독립변수 $X_1,\cdots,X_m$와 1개의 종속변수 $Y$의 관계를 선형결합식으로 표현하는 모형

$$Y=\beta_0+\beta_1X_1+\cdots+\beta_mX_m+\epsilon, \quad \epsilon \overset{i.i.d}{\sim} N(0,\sigma^2)\text{가정}$$

▸ ㉮ $X$ : 월수입, $Y=1$ (구매), 0 (비구매)
단순선형회귀모형 $Y=\alpha+\beta X+\epsilon \Rightarrow$ 만약 $Y=0$ 또는 1인 경우, 회귀분석을 하면 $Y$의 예측값은 연속형 숫자가 되고, $[0,1]$의 범위를 벗어나는 경우도 발생하게 된다. 이러한 선형회귀 모형의 문제점을 두 집단의 분류에 적합하도록 변형한 것이 로지스틱 회귀모형이다.

▸ 오즈비(odds ratio): $X=x$값에서 $Y=1$일 확률과 $Y=0$일 확률의 비율

$$\frac{P(Y=1|X=x)}{P(Y=0|X=x)} \text{ = 오즈비}$$

▸ 로지스틱회귀(logistic regression) 모형은 오즈비의 로그값을 선형회귀식으로 표현한 모형

$$\log\left[\frac{P(Y=1|x)}{1-P(Y=1|x)}\right]=\alpha+\beta x$$

▸ $P(Y=1|x)=\bar{p}$ 라고 하자.

$\log\left(\frac{\bar{p}}{1-\bar{p}}\right)=p^*$라고 하면, $p^*=\alpha+\beta x$ (단순선형회귀모형이 됨)

$\alpha$와 $\beta$의 추정은 최소제곱법(least squared method)을 이용할 수 있다. $\Rightarrow \hat{p}^*=\hat{\alpha}+\hat{\beta}x$

$$\hat{P}(Y=1|x)=\frac{\exp(\hat{p}^*)}{1+\exp(\hat{p}^*)}=\frac{\exp(\hat{\alpha}+\hat{\beta}x)}{1+\exp(\hat{\alpha}+\hat{\beta}x)}$$

: 추정된 확률 $\hat{P}(Y=1|x)$를 로지스틱회귀 모형의 사후확률(posterior probability)이라 함.
이 사후확률 값이 분석자가 선정한 기준값보다 크면 집단 1로 분류하고 아니면 집단 0으로 분류한다.

▸ $\alpha$와 $\beta$에 대한 추정은 최대가능도추정법(maximum likelihood estimation)을 이용할 수 있다.

### 5.3.1 프로빗 모형과 곰페르츠 모형

▸ 로짓함수(logit function): $f(z)=\log\left(\frac{z}{1-z}\right)$

로짓함수에 확률 $P(Y=1|x)$를 대입한 것을 로지스틱회귀모형 또는 로짓모형이라고 한다.

즉, $f(P(Y=1|x))=\log\left[\frac{P(Y=1|x)}{1-P(Y=1|x)}\right]=\alpha+\beta x$

▸ 이 로짓함수를 연결함수(link function)라 부른다.

▸ 많이 이용되는 연결함수에는 로짓함수 이외에도, 프로빗함수(probit function) $f(z)=\Phi^{-1}(z)$, 곰페르츠함수(Gomperts function) $f(z)=\log[-\log(1-z)]$가 있다. (여기서, $\Phi^{-1}$는 누적표준정규분포함수의 역함수)

프로빗 모형 : $\Phi^{-1}[P(Y=1|x)]=\alpha+\beta x$
곰페르츠 모형: $\log[-\log(1-P(Y=1|x))]=\alpha+\beta x$ ⎫ $P(Y=1|x)$ 추정하여 데이터 분류

### 5.3.2 다변수 로지스틱회귀 모형

▸ 독립변수가 $m$개 있을 경우 : $X_1, X_2, \cdots, X_m$

$$\log\left\{\frac{P(Y=1|x_1,x_{2,}\cdots,x_m)}{1-P(Y=1|x_1,x_{2,}\cdots,x_m)}\right\}=\beta_0+\beta_1x_1+\cdots+\beta_mx_m$$

$$P(Y=1|x_1,x_2,\cdots,x_m)=\frac{\exp(\beta_0+\beta_1x_1+\cdots+\beta_mx_m)}{1+\exp(\beta_0+\beta_1x_1+\cdots+\beta_mx_m)}$$

추정된 $\hat{\beta_0}, \hat{\beta_1}, \cdots, \hat{\beta_m}$을 이용하여 $\hat{P}(Y=1|x_1,\cdots,x_m)$ 추정

⇒ 추정된 사후확률 $\hat{P}(Y=1|x_1,\cdots,x_m)$이 분석자가 선정한 기준값보다 크면 집단 1로 분류하고 아니면 집단 0으로 분류함

### 5.3.3 오즈비 증가비율

▸ 다른 모든 변수가 일정하고 변수값 $x_i$만 1 단위 증가한다면 (즉, $x_i$가 $x_i+1$로 증가)

오즈비 증가비율(incremental odds ratio)은 다음과 같다.

$$\text{오즈비 증가비율} = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_i(x_i+1) + \cdots + \beta_m x_m)}{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i + \cdots + \beta_m x_m)}$$

$$= \exp(\beta_i)$$

$\therefore$ 변수 $x_i$가 1 단위 증가할 경우,

$\beta_i$가 양수이면 오즈비 증가비율이 1보다 크므로 $P(Y=1|x_1, \cdots, x_m)$도 증가하게 됨.

반대로 $\beta_i$가 음수이면 오즈비 증가비율이 1보다 작게 되므로 $P(Y=1|x_1, \cdots, x_m)$는 감소하게 된다.

㉮ $\log\left\{\dfrac{\hat{P}(Y=1|x)}{1-\hat{P}(Y=1|x)}\right\} = 0.21 + 1.34x$

$Y=1$(구매), 0(비구매), $x$는 월수입 (단위: 백만원)

월수입 $x$가 1 단위(백만원) 증가하면, 오즈비 증가율은 $\exp(1.34) = 3.82$

월수입이 100만원 증가하면 상품을 구매하지 않을 확률에 비해 구매할 확률의 비율이 3.82배 증가한다.

## 5.34 변수의 선택

▸독립변수가 여러 개일 때 종속변수 $Y$를 잘 설명할 수 있는 변수를 선택하고, 변수 선택을 위해 여러 모형을 비교하여 변수를 선택한다.

▸선택기준: 아카이케정보기준(Akaike Information Criteria, AIC): AIC↓ ⇒ 선호하는 모형

▸선택방법: 전진선택법(forward selection), 후진소거법(backward selection), 단계적방법(stepwise method)

**예제 5.3.1**

[표 5.3]의 데이터에 대해 상품 구매여부를 목표변수로, 나이, 월수입, 신용상태를 독립변수로 하는 로지스틱회귀 모형을 구하라.

**[표 5.7]** 나이, 월수입, 신용상태별 상품 구입여부의 로지스틱회귀 결과

| Analysis of Maximum Likelihood Estimates | | | | | |
|---|---|---|---|---|---|
| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
| Intercept | 1 | -5.2205 | 7.9079 | 0.4358 | 0.5091 |
| x1 (나이) | 1 | 0.1607 | 0.2757 | 0.3400 | 0.5598 |
| x2 (월수입) | 1 | 0.0033 | 0.0130 | 0.0656 | 0.7978 |
| x3 (신용상태) | 1 | -0.4665 | 1.2836 | 0.1321 | 0.7163 |

**[표 5.8]** 각각의 나이, 월수입, 신용상태별 상품 구입여부의 로지스틱 회귀모형

| Analysis of Maximum Likelihood Estimates | | | | | |
|---|---|---|---|---|---|
| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
| Intercept | 1 | -8.0323 | 3.7887 | 4.49 | 0.0340 |
| x1 (나이) | 1 | 0.2521 | 0.1220 | 4.27 | 0.0388 |

| Analysis of Maximum Likelihood Estimates | | | | | |
|---|---|---|---|---|---|
| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
| Intercept | 1 | -3.5130 | 1.7335 | 4.1066 | 0.0427 |
| x2 (월수입) | 1 | 0.0129 | 0.0068 | 3.5432 | 0.0598 |

**[표 5.8]** 각각의 나이, 월수입, 신용상태별 상품 구입여부의 로지스틱 회귀모형(계속)

| Analysis of Maximum Likelihood Estimates | | | | | |
|---|---|---|---|---|---|
| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
| Intercept | 1 | 2.5316 | 1.6645 | 2.3133 | 0.0340 |
| x3 (신용상태) | 1 | -1.6460 | 0.9121 | 3.2564 | 0.0711 |

$$\log\left(\frac{\hat{P}(Y=1|x_1)}{1-\hat{P}(Y=1|x_1)}\right) = -8.0323 + 0.2521x_1$$

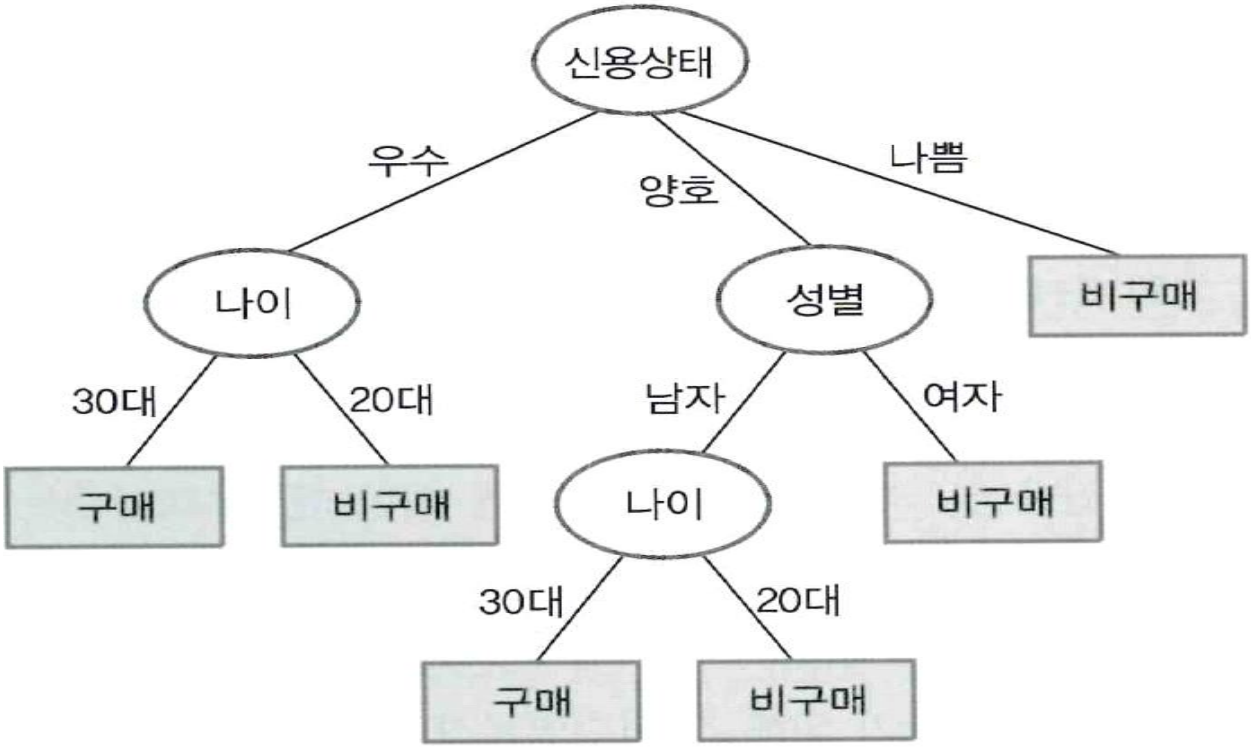
$$\Leftrightarrow \hat{P}(Y=1|x_1) = \frac{\exp(-8.0323+0.2521x_1)}{1+\exp(-8.0323+0.2521x_1)}$$

만약 $x_1 = 20$, $\hat{P}(Y=1|x_1=20) = \frac{\exp(-8.0323+0.2521\times 20)}{1+\exp(-8.0323+0.2521\times 20)} = 0.048$

$\Rightarrow$ 기준값을 0.5라고 정하면, 나이가 20세인 고객은 비구매 고객으로 분류한다.

## 5.4 의사결정나무(Decision tree) 모형

▸**의사결정나무**: 분류함수를 의사결정규칙(decision rule)으로 이루어진 나무 모양으로 그리는 것.
이산형 목표변수를 분류하기 위한 나무 (예) 구매집단 vs. 비구매집단




<그림 5.3> 한 고객이 상품을 구매할 것인지 아닌지를 분류하는 의사결정나무

▸의사결정나무
노드(node): 타원모양. 변수에 대한 검사를 의미 (예) 나이, 성별
뿌리노드(root node): 최상위 노드 (예) 신용도
가지(branch): 검사한 변수의 값
잎(leaf): 직사각형 모양. 최종 분류된 집단을 나타냄

(예) 의사결정나무 이용: 신용도가 양호, 남자, 30대인 사람 -> 구매집단으로 분류

▸회귀나무(regression tree): 연속형 목표변수인 경우에 회귀모형에 기초한 의사결정나무

### 5.4.1 의사결정나무 알고리즘

▸'어떻게 하면 정확도가 높으며 계산시간이 합리적인 알고리즘을 찾을 수 있는가?'

▸**합리적인 알고리즘:** 각 노드의 의사결정 지점에서 어느 변수를 사용할 것인지 결정할 때 국지적으로 최적인 의사결정(locally optimum decision)으로 데이터 집합을 분할하고, 이 방법을 분할된 데이터에 연속적으로 적용하여 모든 데이터 집합을 분할할 수 있는 의사결정나무를 완성하는 것.

### 5.4.2 가지분할을 위한 변수 선택

▸한 변수가 선택되어 가지를 쳤을 때 각 가지별로 분류가 더 정확히 되는 변수를 선택

**예제 5.4.1**
한 백화점의 어느 상품매장을 방문한 20명을 조사해보니 구매집단이 8명이고, 비구매집단이 12명이었다. 이들 20명의 성별과 신용상태를 분석하여 교차표를 작성하여 보니 다음과 같다. 의사결정나무에서 어떠한 변수의 가지분할이 더 좋은 것인가?

| 성별 | 구매집단 $G_1$ | 비구매집단 $G_2$ | 합계 |
|---|---|---|---|
| 남 | 4 | 6 | 10 |
| 여 | 4 | 6 | 10 |
| 합계 | 8 | 12 | 20 |

| 신용상태 | 구매집단 $G_1$ | 비구매집단 $G_2$ | 합계 |
|---|---|---|---|
| 양호 | 7 | 3 | 10 |
| 불량 | 1 | 9 | 10 |
| 합계 | 8 | 12 | 20 |

(풀이) 성별 변수의 경우, 남자이든 여자이든 구매집단과 비구매집단의 비율이 4대 6으로 전체 20명의 구매집단과 비구매집단 비율과 같다.
신용상태 변수의 경우, 불량인 경우와 양호인 경우의 구매 또는 비구매 비율의 차이가 크다. 즉, 신용상태가 불량이라면, 어느 고객은 비구매집단에 속할 가능성이 크고, 양호라면 구매집단에 속할 가능성이 크다. 따라서, 신용상태를 파악하게 되면 구매집단과 비구매집단을 어느 정도 구별할 수 있으므로, 의사결정나무에서 신용상태라는 변수의 가지분할이 더 좋다.

▸의사결정나무의 변수선택에 많이 이용되는 방법
- 카이제곱 독립성 검정(chi-square independence test)
- 엔트로피 계수(entropy coefficient)
- 지니계수(Gini coefficient)
- 분류오류율(classification error rate)

**가. 카이제곱 독립성검정 (Chi-square independence test)**
: 각 변수에 대한 집단의 분포가 독립인지 검정

**[표 5.10]** $2\times2$ 교차표에서 관찰도수 $n_{ij}$

| | 구매집단 $G_1$ | 비구매집단 $G_2$ | 행의 합 |
|---|---|---|---|
| 변수의 값 $A_1$ | $n_{11}$ | $n_{12}$ | $n_{1\cdot}$ |
| 변수의 값 $A_2$ | $n_{21}$ | $n_{22}$ | $n_{2\cdot}$ |
| 열의 합 | $n_{\cdot 1}$ | $n_{\cdot 2}$ | $n_{\cdot\cdot}$ |

**[표 5.11]** $2\times2$ 교차표에서 변수가 독립일 때의 기대도수 $E_{ij}$

| | 구매집단 $G_1$ | 비구매집단 $G_2$ | 행의 합 |
|---|---|---|---|
| 변수의 값 $A_1$ | $E_{11}=n_{1.}\times\frac{n_{.1}}{n_{..}}$ | $E_{12}=n_{1.}\times\frac{n_{.2}}{n_{..}}$ | $n_{1.}$ |
| 변수의 값 $A_2$ | $E_{21}=n_{2.}\times\frac{n_{.1}}{n_{..}}$ | $E_{22}=n_{2.}\times\frac{n_{.2}}{n_{..}}$ | $n_{2.}$ |
| 열의 합 | $n_{.1}$ | $n_{.2}$ | $n_{..}$ |

▸ 독립성 검정을 위한 $2\times2$ 교차표의 카이제곱 통계량

$$\chi^2=\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(n_{ij}-E_{ij})^2}{E_{ij}}\sim\chi^2(1)$$

▸ 카이제곱 통계량 값이 클수록 변수와 집단이 독립이라는 귀무가설이 기각된다.
▸ 기각의 정도가 강할수록 가지분할을 위한 더 좋은 변수라 할 수 있다.

**예제 5.4.2**
성별과 신용상태에 대해 카이제곱 독립성 검정으로 어느 변수가 가지분할에 더 좋은가?

| 성별 | 구매집단 $G_1$ | 비구매집단 $G_2$ | 합계 |
|---|---|---|---|
| 남 | 4 | 6 | 10 |
| 여 | 4 | 6 | 10 |
| 합계 | 8 | 12 | 20 |

| 신용상태 | 구매집단 $G_1$ | 비구매집단 $G_2$ | 합계 |
|---|---|---|---|
| 양호 | 7 | 3 | 10 |
| 불량 | 1 | 9 | 10 |
| 합계 | 8 | 12 | 20 |

(풀이) 성별 변수: $\chi^2=\frac{(4-4)^2}{4}+\frac{(6-6)^2}{6}+\frac{(4-4)^2}{4}+\frac{(6-6)^2}{6}=0<\chi^2_{0.05}(1)=3.84$

$\Rightarrow$ 귀무가설 기각 못함. 즉, 성별 변수와 구매여부는 독립이다.

신용상태 변수: $\chi^2=\frac{(7-4)^2}{4}+\frac{(3-6)^2}{6}+\frac{(1-4)^2}{4}+\frac{(9-6)^2}{6}=7.5>\chi^2_{0.05}(1)=3.84$

$\Rightarrow$ 귀무가설 기각. 즉, 신용상태 변수와 구매여부는 독립이 아니다.
따라서, 신용상태 변수가 가지분할에 더 좋다.

▸ 변수 A가 a개의 변수 값이 있고, 집단 수가 $K$개인 경우

$$\chi^2=\sum_{i=1}^{a}\sum_{j=1}^{K}\frac{(n_{ij}-E_{ij})^2}{E_{ij}}\sim\chi^2((a-1)(K-1)),\quad \text{여기서 } E_{ij}=n_{i.}\times\frac{n_{.j}}{n_{..}}$$

▸ p-값을 이용할 경우, p-값이 더 적은 변수를 의사결정나무에서 가지분할 변수로 선택한다.

**나. 엔트로피 계수, 지니계수, 분류오류율**

: 분포함수에 대한 불확실성(uncertainty) 또는 순수성(purity)을 측정하는 측도

▸분류하고자 하는 집단이 $G_1, G_2, \cdots, G_K$이고, 그 집단의 확률이 $p_1, p_2, \cdots, p_K$일 때

엔트로피계수 = $-\sum_{i=1}^{K} p_i \log_2 p_i$

지니계수 = $1-\sum_{i=1}^{K} p_i^2$

분류오류율 = $1-\max\{p_1, p_2, \cdots, p_K\}$

(예) 집단이 2개만 있는 경우: 집단 1의 확률이 $p$라면, 집단 2의 확률은 $1-p$이다.

엔트로피계수 = $-p\log_2 p-(1-p)\log_2(1-p)$

지니계수 = $1-p^2-(1-p)^2$

분류오류율 = $1-\max\{p,(1-p)\}$

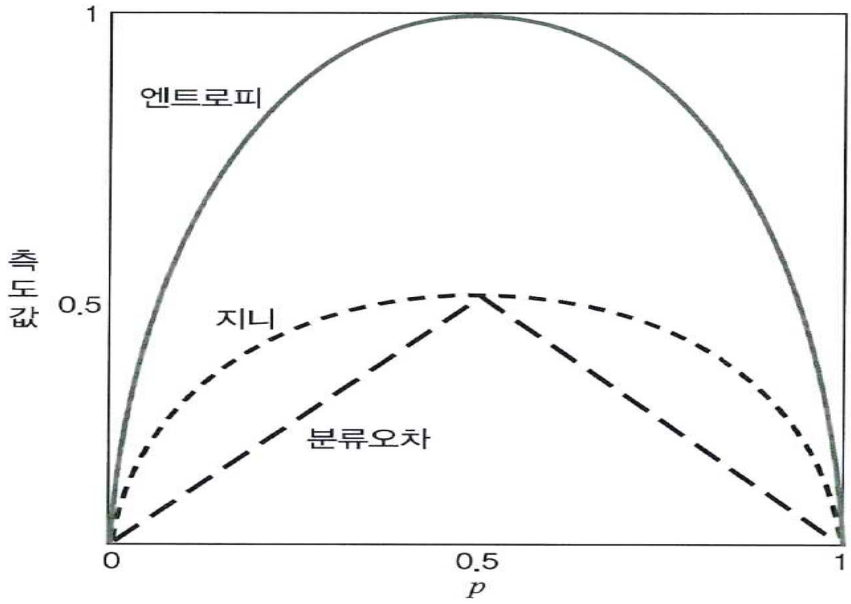



그림 5-5 분류집단이 두 개인 경우의 엔트로피계수, 지니계수, 분류오차 측도

▸세 측도 모두 $p=0.5$에서 최대값을 갖고, $p=0$ 또는 1에서 최소값 0을 가짐.

▸두 집단의 확률이 같은 경우($p=0.5$), 어느 집단으로 분류될지 모르기 때문에 불확실성(uncertainty)는 커지는데, 이 때 각 측도는 최대값을 갖는다.

▸어느 한 집단의 확률이 1이라면 불확실성이 없으므로(즉, 분류의 확실성이 100%임), 각 측도는 최소값 0을 갖는다.

▸가지분할은 불확실성이 적은 변수를 선택한다.

**예제 5.4.3**

성별과 신용상태에 대한 엔트로피 계수, 지니계수, 분류오류율을 구하고, 어느 변수로 가지분할을 하는 것이 좋은가?

(풀이)

| | 구매집단 $G_1$ | 비구매집단 $G_2$ | 합계 | 불확실성 측도 |
|---|---|---|---|---|
| 집단별 분포 | 8 | 12 | 20 | 엔트로피$=-0.4\log_2(0.4)-0.6\log_2(0.6)=0.9710$<br>지니$=1-0.4^2-0.6^2=0.48$<br>분류오류율$=1-\max(0.4,0.6)=0.4$ |

| 성별 | 구매집단 $G_1$ | 비구매집단 $G_2$ | 합계 | 불확실성 측도 |
|---|---|---|---|---|
| 남 | 4 | 6 | 10 | 엔트로피$=-0.4\log_2(0.4)-0.6\log_2(0.6)=0.9710$<br>지니$=1-0.4^2-0.6^2=0.48$<br>분류오류율$=1-\max(0.4,0.6)=0.4$ |
| 여 | 4 | 6 | 10 | 엔트로피$=-0.4\log_2(0.4)-0.6\log_2(0.6)=0.9710$<br>지니$=1-0.4^2-0.6^2=0.48$<br>분류오류율$=1-\max(0.4,0.6)=0.4$ |

| 신용상태 | 구매집단 $G_1$ | 비구매집단 $G_2$ | 합계 | 불확실성 측도 |
|---|---|---|---|---|
| 양호 | 7 | 3 | 10 | 엔트로피$=-0.7\log_2(0.7)-0.3\log_2(0.3)=0.8813$<br>지니$=1-0.7^2-0.3^2=0.42$<br>분류오류율$=1-\max(0.7,0.3)=0.3$ |
| 불량 | 1 | 9 | 10 | 엔트로피$=-0.1\log_2(0.1)-0.9\log_2(0.9)=0.4690$<br>지니$=1-0.1^2-0.9^2=0.18$<br>분류오류율$=1-\max(0.1,0.9)=0.1$ |

신용상태의 불확실성 측도값이 성별의 불확실성 측도값보다 작으므로, 신용상태를 이용한 가지치기가 합리적이다.

**[표 5.15]**

[표 5.15] $a\times K$ 교차표에서 관찰도수와 각 변수 값의 불확실성

| 변수 값 \ 집단 | 집단 $G_1$ | 집단 $G_2$ | … | 집단 $G_K$ | 행의 합 | 각 변수 값의 불확실성 |
|---|---|---|---|---|---|---|
| $A_1$ | $n_{11}$ | $n_{12}$ | … | $n_{1c}$ | $n_{1\bullet}$ | $I(A_1)$ |
| $A_2$ | $n_{21}$ | $n_{22}$ | … | $n_{2c}$ | $n_{2\bullet}$ | $I(A_2)$ |
| … | … | … | … | … | … | … |
| $A_a$ | $n_{a1}$ | $n_{a2}$ | … | $n_{ac}$ | $n_{a\bullet}$ | $I(A_a)$ |
| 열의 합 | $n_{\bullet 1}$ | $n_{\bullet 2}$ | … | $n_{\bullet c}$ | $n_{\bullet\bullet}$ | 변수 $A$의 불확실성 $I(A)$ |

▸ **현 노드 T의 불확실성 I(T)**

$I(T) = -\sum_{j=1}^{K}\left(\frac{n_{\cdot j}}{n_{\cdot\cdot}}\right)\log_2\left(\frac{n_{\cdot j}}{n_{\cdot\cdot}}\right)$ : 엔트로피 계수를 이용하였을 경우의 현 노드 T의 불확실성

▸ **변수 A의 기대불확실성(expected uncertainty)**

$I(A) = \frac{n_{1\cdot}}{n_{\cdot\cdot}} I(A_1) + \frac{n_{2\cdot}}{n_{\cdot\cdot}} I(A_2) + \cdots + \frac{n_{a\cdot}}{n_{\cdot\cdot}} I(A_a)$

: 각 변수값의 불확실성 $I(A_i)$를 관찰도수 $n_{i\cdot}$가 전체 데이터 수에서 차지하는 비율, 즉 $\frac{n_{i\cdot}}{n_{\cdot\cdot}}$를 가중치로 하여 계산한 기대값

▸ **정보이득**: 현 노드의 불확실성 I(T)와 어느 변수로 가지치기했을 때 얻어지는 기대불확실성 I(A)의 차이

$\Delta = I(T) - I(A)$

: 정보이득이 큰 변수 선택

▸ 한 변수의 정보이득이 많다는 것은 이 변수로 가지분할을 하면 불확실성을 더 많이 제거함으로써 분류를 더 정확히 할 수 있다는 것을 의미

▸ 엔트로피나 지니계수를 이용하여 구한 정보이득은 변수값의 수가 많은 변수를 선호하는 경향이 있음

이 문제를 극복하기 위하여 CART와 같은 알고리즘에서는 정보이득비율= $\frac{\Delta}{I(T)}$을 가지분할 기준으로 함.

**예제 5.4.4**

성별과 신용상태 변수의 각 측도별 정보이득을 구하여라.

(풀이) 성별 변수의 정보이득:

엔트로피 정보이득 $= 0.9710 - \left(\frac{10}{20}\times 0.9710 + \frac{10}{20}\times 0.9710\right) = 0$

지니 정보이득 $= 0.48 - \left(\frac{10}{20}\times 0.48 + \frac{10}{20}\times 0.48\right) = 0$

분류오류율 정보이득 $= 0.4 - \left(\frac{10}{20}\times 0.4 + \frac{10}{20}\times 0.4\right) = 0$

신용상태 변수의 정보이득:

엔트로피 정보이득 $= 0.9710 - \left(\frac{10}{20}\times 0.8813 + \frac{10}{20}\times 0.4690\right) = 0.2958$

지니 정보이득 $= 0.48 - \left(\frac{10}{20}\times 0.42 + \frac{10}{20}\times 0.18\right) = 0.18$

분류오류율 정보이득 $= 0.4 - \left(\frac{10}{20}\times 0.3 + \frac{10}{20}\times 0.1\right) = 0.2$

∴ 신용상태 변수의 정보이득이 더 많으므로, 현재 노드에서 가지분할 변수로는 신용상태 변수가 더 좋다.

### 5.4.3 의사결정나무의 예

**예제 5.4.5**

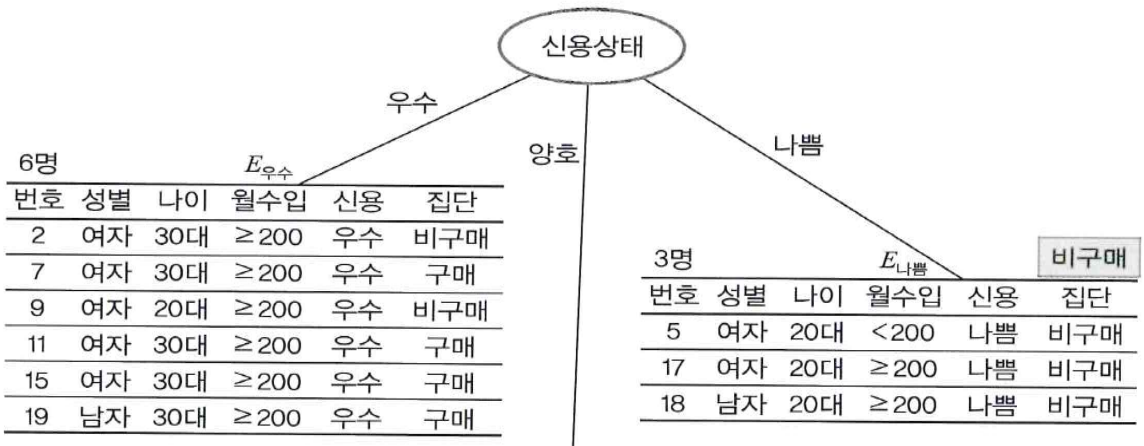
한 백화점의 상품매장을 방문한 20명을 조사해 보니 구매집단($G_1$)이 8명이고, 비구매 집단($G_2$)이 12명이였다. 의사결정나무를 이용한 분류모형을 구하여라. 변수 선택은 엔트로피 계수를 이용하고, 각 잎에서의 데이터 수가 5개 이하이면 더 이상 분할하지 않는다. 만일 각 잎의 데이터가 한 집단으로 분류되면 역시 분할하지 않는다. 각 잎의 집단에 대한 결정은 다수결로 한다.

[표 5.16] 한 백화점 상품매장 고객의 성별, 나이, 월수입, 신용상태, 구매여부

| 고객번호 | 성별 | 나이 | 월수입 | 신용상태 | 구매여부 |
|---|---|---|---|---|---|
| 1 | 남자 | 20대 | $<200$ | 양호 | 구매 |
| 2 | 여자 | 30대 | $\geq 200$ | 우수 | 비구매 |
| 3 | 여자 | 20대 | $\geq 200$ | 양호 | 비구매 |
| 4 | 여자 | 20대 | $\geq 200$ | 양호 | 구매 |
| 5 | 여자 | 20대 | $<200$ | 나쁨 | 비구매 |
| 6 | 여자 | 30대 | $\geq 200$ | 양호 | 비구매 |
| 7 | 여자 | 30대 | $\geq 200$ | 우수 | 구매 |
| 8 | 남자 | 20대 | $<200$ | 양호 | 비구매 |
| 9 | 여자 | 20대 | $\geq 200$ | 우수 | 비구매 |
| 10 | 남자 | 30대 | $\geq 200$ | 양호 | 구매 |
| 11 | 여자 | 30대 | $\geq 200$ | 우수 | 구매 |
| 12 | 여자 | 20대 | $<200$ | 양호 | 비구매 |
| 13 | 남자 | 30대 | $\geq 200$ | 양호 | 비구매 |
| 14 | 남자 | 30대 | $<200$ | 양호 | 구매 |
| 15 | 여자 | 30대 | $\geq 200$ | 우수 | 구매 |
| 16 | 여자 | 30대 | $\geq 200$ | 양호 | 비구매 |
| 17 | 여자 | 20대 | $\geq 200$ | 나쁨 | 비구매 |
| 18 | 남자 | 20대 | $\geq 200$ | 나쁨 | 비구매 |
| 19 | 남자 | 30대 | $\geq 200$ | 우수 | 구매 |
| 20 | 남자 | 20대 | $<200$ | 양호 | 비구매 |

[표 5.17] 각 변수별 기대정보와 정보이득

| 변수 | | 구매집단 $G_1$ | 비구매집단 $G_2$ | 합계 | 엔트로피계수 | 정보이득 $\Delta$ |
|---|---|---|---|---|---|---|
| 성별 | 남자 | 4 | 4 | 8 | 1.0000 | |
| | 여자 | 4 | 8 | 12 | 0.9183 | |
| | | | | 기대정보 | 0.9510 | 0.0200 |
| 나이 | 20대 | 2 | 8 | 10 | 0.7219 | |
| | 30대 | 6 | 4 | 10 | 0.9710 | |
| | | | | 기대정보 | 0.8464 | 0.1246 |
| 월수입 | <200 | 2 | 4 | 6 | 0.9183 | |
| | ≥200 | 6 | 8 | 14 | 0.9852 | |
| | | | | 기대정보 | 0.9651 | 0.0059 |
| 신용상태 | 우수 | 4 | 2 | 6 | 0.9183 | |
| | 양호 | 4 | 7 | 11 | 0.9457 | |
| | 나쁨 | 0 | 3 | 3 | 0.0000 | |
| | | | | 기대정보 | 0.7956 | 0.1754 |




6명 $E_{우수}$

| 번호 | 성별 | 나이 | 월수입 | 신용 | 집단 |
|---|---|---|---|---|---|
| 2 | 여자 | 30대 | ≥200 | 우수 | 비구매 |
| 7 | 여자 | 30대 | ≥200 | 우수 | 구매 |
| 9 | 여자 | 20대 | ≥200 | 우수 | 비구매 |
| 11 | 여자 | 30대 | ≥200 | 우수 | 구매 |
| 15 | 여자 | 30대 | ≥200 | 우수 | 구매 |
| 19 | 남자 | 30대 | ≥200 | 우수 | 구매 |

3명 $E_{나쁨}$ 비구매

| 번호 | 성별 | 나이 | 월수입 | 신용 | 집단 |
|---|---|---|---|---|---|
| 5 | 여자 | 20대 | <200 | 나쁨 | 비구매 |
| 17 | 여자 | 20대 | ≥200 | 나쁨 | 비구매 |
| 18 | 남자 | 20대 | ≥200 | 나쁨 | 비구매 |

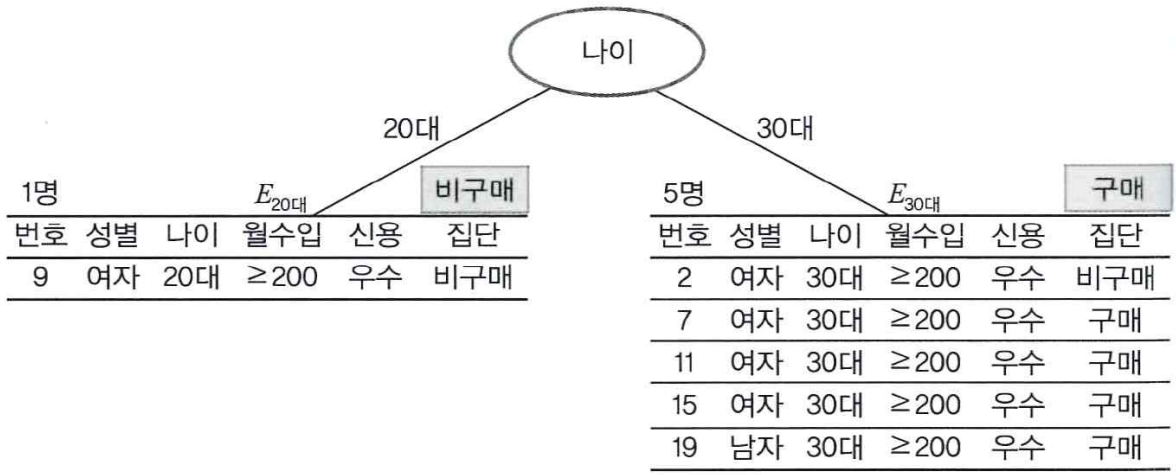
11명 $E_{양호}$

| 번호 | 성별 | 나이 | 월수입 | 신용 | 집단 |
|---|---|---|---|---|---|
| 1 | 남자 | 20대 | <200 | 양호 | 구매 |
| 3 | 여자 | 20대 | ≥200 | 양호 | 비구매 |
| 4 | 여자 | 20대 | ≥200 | 양호 | 구매 |
| 6 | 여자 | 30대 | ≥200 | 양호 | 비구매 |
| 8 | 남자 | 20대 | <200 | 양호 | 비구매 |
| 10 | 남자 | 30대 | ≥200 | 양호 | 구매 |
| 12 | 여자 | 20대 | <200 | 양호 | 비구매 |
| 13 | 남자 | 30대 | ≥200 | 양호 | 비구매 |
| 14 | 남자 | 30대 | <200 | 양호 | 구매 |
| 16 | 여자 | 30대 | ≥200 | 양호 | 비구매 |
| 20 | 남자 | 20대 | <200 | 양호 | 비구매 |

그림 5-6 신용상태에 따라 가지분할된 의사결정나무

[표 5.18] 신용상태가 우수한 $E_{우수}$에 대한 각 변수별 기대정보와 정보이득

| 변수 | | 구매집단 $G_1$ | 비구매집단 $G_2$ | 합계 | 엔트로피계수 | 정보이득 $\Delta$ |
|---|---|---|---|---|---|---|
| 성별 | 남자 | 1 | 0 | 1 | 0.0000 | |
| | 여자 | 3 | 2 | 5 | 0.9710 | |
| | | | | 기대정보 | 0.8091 | 0.1092 |
| 나이 | 20대 | 0 | 1 | 1 | 0.0000 | |
| | 30대 | 4 | 1 | 5 | 0.7219 | |
| | | | | 기대정보 | 0.6016 | 0.3167 |
| 월수입 | $<200$ | 0 | 0 | 0 | 0.9183 | |
| | $\geq 200$ | 4 | 2 | 6 | 0.9183 | |
| | | | | 기대정보 | 0.9183 | 0.0000 |



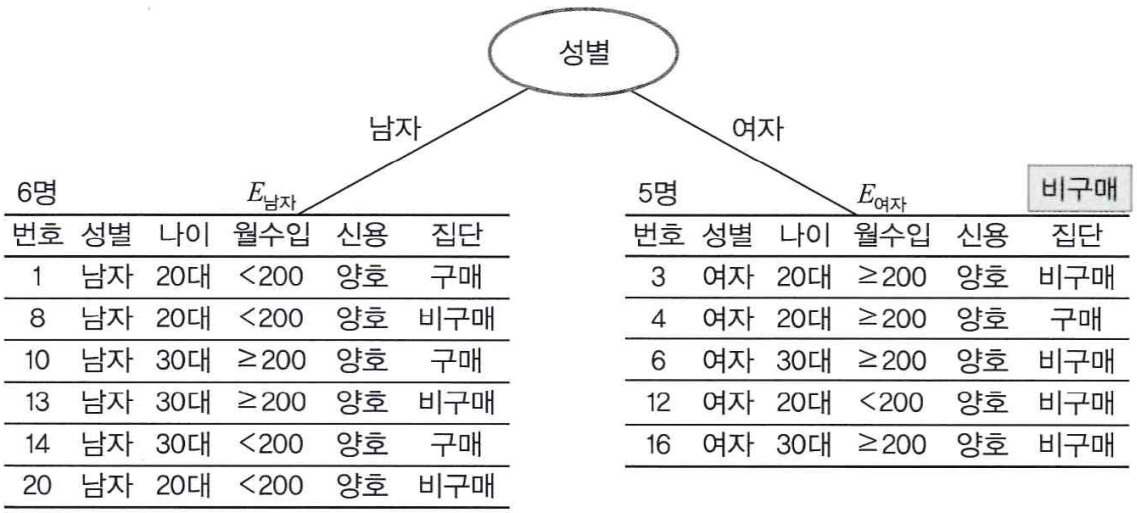

| 번호 | 성별 | 나이 | 월수입 | 신용 | 집단 |
|---|---|---|---|---|---|
| 9 | 여자 | 20대 | ≥200 | 우수 | 비구매 |

| 번호 | 성별 | 나이 | 월수입 | 신용 | 집단 |
|---|---|---|---|---|---|
| 2 | 여자 | 30대 | ≥200 | 우수 | 비구매 |
| 7 | 여자 | 30대 | ≥200 | 우수 | 구매 |
| 11 | 여자 | 30대 | ≥200 | 우수 | 구매 |
| 15 | 여자 | 30대 | ≥200 | 우수 | 구매 |
| 19 | 남자 | 30대 | ≥200 | 우수 | 구매 |

**그림 5-7** 신용상태가 우수인 $E_{우수}$에 대하여 나이로 분할된 의사결정나무

**[표 5.19]** 신용상태가 양호한 11명 $E_{양호}$에 대한 각 변수별 기대정보와 정보이득

| 변수 | | 구매집단 $G_1$ | 비구매집단 $G_2$ | 합계 | 엔트로피계수 | 정보이득 $\Delta$ |
|---|---|---|---|---|---|---|
| 성별 | 남자 | 3 | 3 | 6 | 1.0000 | |
| | 여자 | 1 | 4 | 5 | 0.7219 | |
| | | | | 기대정보 | 0.8736 | 0.0721 |
| 나이 | 20대 | 2 | 4 | 6 | 0.9183 | |
| | 30대 | 2 | 3 | 5 | 0.9710 | |
| | | | | 기대정보 | 0.9422 | 0.0034 |
| 월수입 | $<200$ | 2 | 3 | 5 | 0.9710 | |
| | $\geq 200$ | 2 | 4 | 6 | 0.9183 | |
| | | | | 기대정보 | 0.9422 | 0.0034 |




6명 $E_{남자}$

| 번호 | 성별 | 나이 | 월수입 | 신용 | 집단 |
|---|---|---|---|---|---|
| 1 | 남자 | 20대 | <200 | 양호 | 구매 |
| 8 | 남자 | 20대 | <200 | 양호 | 비구매 |
| 10 | 남자 | 30대 | ≥200 | 양호 | 구매 |
| 13 | 남자 | 30대 | ≥200 | 양호 | 비구매 |
| 14 | 남자 | 30대 | <200 | 양호 | 구매 |
| 20 | 남자 | 20대 | <200 | 양호 | 비구매 |

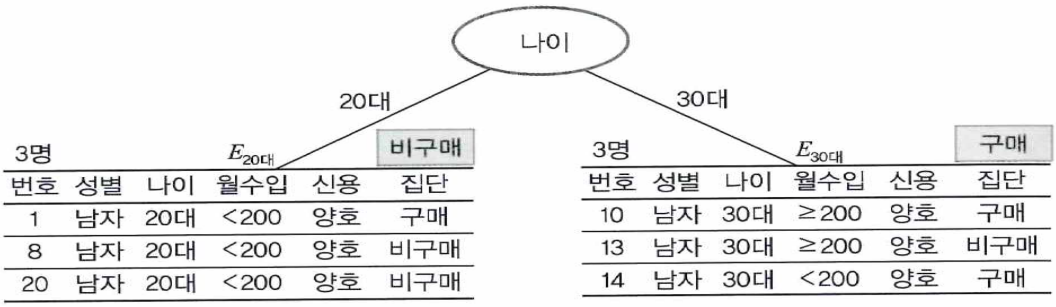
5명 $E_{여자}$

| 번호 | 성별 | 나이 | 월수입 | 신용 | 집단 |
|---|---|---|---|---|---|
| 3 | 여자 | 20대 | ≥200 | 양호 | 비구매 |
| 4 | 여자 | 20대 | ≥200 | 양호 | 구매 |
| 6 | 여자 | 30대 | ≥200 | 양호 | 비구매 |
| 12 | 여자 | 20대 | <200 | 양호 | 비구매 |
| 16 | 여자 | 30대 | ≥200 | 양호 | 비구매 |

**그림 5-8** 신용상태가 양호한 $E_{양호}$에 대하여 성별로 분할된 의사결정나무

[표 5.20] 신용상태가 양호한 $E_{남자}$에 대한 각 변수별 기대정보와 정보이득

| 변수 | | 구매집단 $G_1$ | 비구매집단 $G_2$ | 합계 | 엔트로피계수 | 정보이득 $\Delta$ |
|---|---|---|---|---|---|---|
| 나이 | 20대 | 1 | 2 | 3 | 0.9183 | |
| | 30대 | 2 | 1 | 3 | 0.9183 | |
| | | | | 기대정보 | 0.9183 | 0.0817 |
| 월수입 | <200 | 2 | 2 | 4 | 1.0000 | |
| | ≥200 | 1 | 1 | 2 | 1.0000 | |
| | | | | 기대정보 | 1.0000 | 0.0000 |



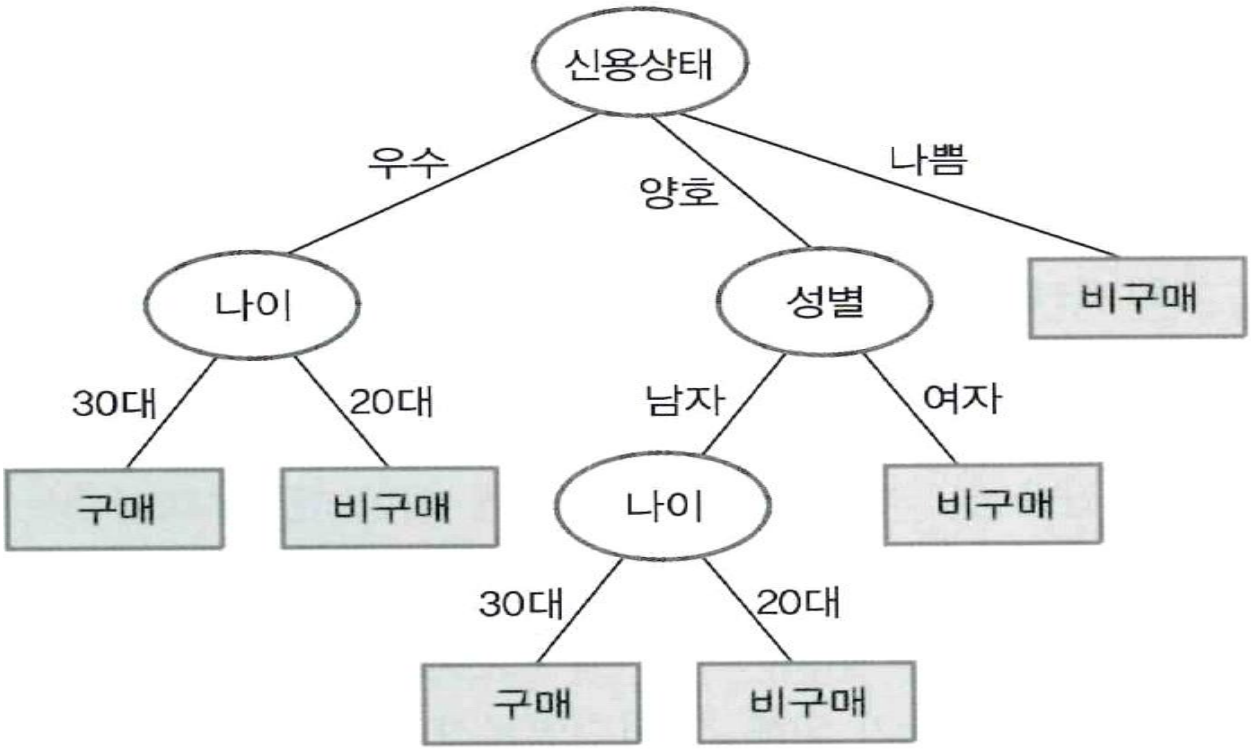

3명 $E_{20대}$ 비구매

| 번호 | 성별 | 나이 | 월수입 | 신용 | 집단 |
|---|---|---|---|---|---|
| 1 | 남자 | 20대 | <200 | 양호 | 구매 |
| 8 | 남자 | 20대 | <200 | 양호 | 비구매 |
| 20 | 남자 | 20대 | <200 | 양호 | 비구매 |

3명 $E_{30대}$ 구매

| 번호 | 성별 | 나이 | 월수입 | 신용 | 집단 |
|---|---|---|---|---|---|
| 10 | 남자 | 30대 | ≥200 | 양호 | 구매 |
| 13 | 남자 | 30대 | ≥200 | 양호 | 비구매 |
| 14 | 남자 | 30대 | <200 | 양호 | 구매 |

그림 5-9 신용상태가 양호한 $E_{남자}$에 대하여 나이로 분할된 의사결정나무



### 5.4.4 연속형 변수의 가지분할

▸연속형 변수는 범주형으로 변환하여 의사결정나무 모형에 적용
(예) 나이: 20대, 30대, 40대
월수입: <200, >=200

▸문제: '어떠한 경계값으로 연속형 값을 나누어야 좋은 가?'

▸경계값의 결정이 더 정확한 분류를 위해 필요하다면, 불확실성 측도를 이용하여 경계값 찾음
⇒ 불확실성 측도값이 가장 작은 경계값을 선택

**예제 5.4.6**
한 백화점의 어느 상품매장을 방문하는 사람 10명에 대해 상품 구매여부와 월수입을 조사해보니 다음과 같다. 구매하는 사람을 $Y$집단, 구매하지 않는 사람을 $N$집단으로 표시하고 월수입을 오름차순으로 정렬하였다. 의사결정나무 모형을 적용하기 위해 월수입을 두 개의 집단으로 나누고자 한다. 어떠한 경계값으로 나누어야 하는가?

| 구매 여부 | $N$ | $N$ | $N$ | $Y$ | $Y$ | $Y$ | $N$ | $N$ | $N$ | $N$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 월수입 | 100 | 120 | 160 | 180 | 186 | 190 | 210 | 250 | 270 | 300 |

(풀이) 중간값이 170만원인 경우

| 월수입 | $Y$ | $N$ | 합 |
|---|---|---|---|
| ≦170만원 | 0 | 3 | 3 |
| >170만원 | 3 | 4 | 7 |

지니계수를 이용한 기대불확실성$= \frac{3}{10}\times\left\{1-\left(\frac{0}{3}\right)^2-\left(\frac{3}{3}\right)^2\right\}+\frac{7}{10}\times\left\{1-\left(\frac{3}{7}\right)^2-\left(\frac{4}{7}\right)^2\right\}=0.343$

모든 중간값에 대해 지니계수를 이용한 기대불확실성 구하기

[표 5.22] 연속형 변수인 월수입의 분할 경계값 조사

| 상품구매 | $N$ | $N$ | $N$ | $Y$ | $Y$ | $Y$ | $N$ | $N$ | $N$ | $N$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 월수입 | 100 | 120 | 160 | 180 | 186 | 190 | 210 | 250 | 270 | 300 |
| 중간값 | 110 | 140 | 170 | 183 | 188 | 200 | 230 | 260 | 285 | |
| | ≤ > | ≤ > | ≤ > | ≤ > | ≤ > | ≤ > | ≤ > | ≤ > | ≤ > | ≤ > |
| $Y$분류 | 0 3 | 0 3 | 0 3 | 1 2 | 2 1 | 3 0 | 3 0 | 3 0 | 3 0 | |
| $N$분류 | 1 6 | 2 5 | 3 4 | 3 4 | 3 4 | 3 4 | 4 3 | 5 2 | 6 1 | |
| 지니계수 | 0.400 | 0.375 | 0.343 | 0.417 | 0.400 | 0.300 | 0.343 | 0.375 | 0.400 | |

∴ 중간값이 200만원일 때, 지니계수를 이용한 기대불확실성이 가장 작음. 즉, 200만원일 때 정보이득이 가장 크므로, 월수입 분할 경계값은 200만원이다.

**예제 5.4.7**

한 백화점의 어느 상품매장을 방문하는 20명을 조사해보니 구매하는 사람이 7명이고 구매하지 않는 사람이 13명이었다. 이들 20명의 나이에 대하여 25세 미만과 25세 이상 그리고 35세 미만과 35세 이상으로 구분하는 방법을 비교하려고 교차표를 다음과 같이 작성하였다. 어떠한 구간분할이 더 좋은 것인지 엔트로피 정보이득을 이용하여 결정하라.

| 구간분할 1 | 구매집단 $G_1$ | 비구매집단 $G_2$ | 합계 |
|---|---|---|---|
| 25세 미만 | 1 | 5 | 6 |
| 25세 이상 | 6 | 8 | 14 |
| 합계 | 7 | 13 | 20 |

| 구간분할 2 | 구매집단 $G_1$ | 비구매집단 $G_2$ | 합계 |
|---|---|---|---|
| 35세 미만 | 3 | 12 | 15 |
| 35세 이상 | 4 | 1 | 5 |
| 합계 | 7 | 13 | 20 |

(풀이) 20명 중 구매집단은 7명, 비구매집단은 13명이므로, $p_1 = \frac{7}{20}$, $p_2 = \frac{13}{20}$이다.

구매집단과 비구매집단에 대한 엔트로피 계수$= -\frac{7}{20}log_2\left(\frac{7}{20}\right) - \frac{13}{20}log_2\left(\frac{13}{20}\right) = 0.9341$

| 구간분할 1 | 구간별 엔트로피 | 기대엔트로피와 정보이득 |
|---|---|---|
| 25세 미만 | $-\frac{1}{6}log_2\left(\frac{1}{6}\right) - \frac{5}{6}log_2\left(\frac{5}{6}\right) = 0.65$ | 기대엔트로피$= \frac{6}{20}\times 0.65 + \frac{14}{20}\times 0.9852 = 0.8847$ |
| 25세 이상 | $-\frac{6}{14}log_2\left(\frac{6}{14}\right) - \frac{8}{14}log_2\left(\frac{8}{14}\right) = 0.9852$ | 정보이득 $\Delta = 0.9341 - 0.8847 = 0.0494$ |

| 구간분할 2 | 구간별 엔트로피 | 기대엔트로피와 정보이득 |
|---|---|---|
| 35세 미만 | $-\frac{3}{15}log_2\left(\frac{3}{15}\right) - \frac{12}{15}log_2\left(\frac{12}{15}\right) = 0.7219$ | 기대엔트로피$= \frac{15}{20}\times 0.7219 + \frac{5}{20}\times 0.7219 = 0.7219$ |
| 35세 이상 | $-\frac{4}{5}log_2\left(\frac{4}{5}\right) - \frac{1}{5}log_2\left(\frac{1}{5}\right) = 0.7219$ | 정보이득 $\Delta = 0.9341 - 0.7219 = 0.2121$ |

∴ 구간분할 2의 정보이득이 크므로, 35세 미만과 35세 이상으로 구간 분할하는 것이 더 좋다.

▸범주형 변수의 값이 여러 가지인 경우에 범주의 수를 축소하고자 할 때도 위와 같은 방법을 적용할 수 있음.

(예) 범주형 변수값이 3개인 경우($A_1, A_2, A_3$), 두 개로 축소하고자 함

→ 3가지 가능한 조합 $(A_1 + A_2, A_3)$, $(A_1 + A_3, A_2)$, $(A_2 + A_3, A_1)$에 대하여 정보이득을 조사하여 가장 큰 정보이득을 가지는 조합 선택

## 5.4.5 의사결정나무의 가지치기

▸의사결정나무 모형은 훈련용 데이터를 잘 분류하지만 시험용 데이터를 잘 분류하지 못하는 과잉적합(overfitting) 문제가 발생할 수 있음.

▸가지치기(pruning)를 통해 과잉적합 문제를 해결할 수 있음.
▸가지치기
사전가지치기(prepruning): 의미없는 분할이 계속되지 않도록 카이제곱, 정보이득 등을 이용하여 분할의 적합성을 조사. 임계값은 분석자가 결정. 임계값이 높으면 간단한 나무가 되고, 임계값이 낮으면 복잡한 나무가 될 수 있음
사후가지치기(postpruning): 완성된 나무에서 가지를 제거해 나가는 방법
(예) 각 노드에 대해 하위 나무들을 가지치기하였을 때, 예상 오류율을 계산해서 이 값이 최대 예상 오류율인 경우 하위 나무들은 유지되고 그렇지 않으면 가지치기 함.

### 5.4.6. 회귀나무 모형(Regression tree)

▸목표변수가 연속형 변수인 경우, 여러 개의 독립변수를 이용하여 목표변수를 예측하는 방법으로 회귀분석이 많이 이용.
▸회귀모형을 이용하여 의사결정나무를 만들어 분류하는 방법을 회귀나무 모형이라 부름
▸1984년 Breiman 등에 의해 이론이 정립됨.
▸회귀나무모형은 의사결정나무 모형과 유사한 과정을 거쳐 만들어짐. 단, 가지분할에 각 노드의 목표변수들의 관측값에 대한 분산을 사용하여 회귀나무모형을 만듦.
▸가지치기와 교차검증을 통해 최종 모형 선택

### 5.4.7 의사결정나무 모형의 특성

1) 각 집단의 분포함수를 가정하지 않는 비모수적 방법이다.
2) 설명이 용이. 모형의 정확도도 괜찮은 편.
3) 계산이 복잡하지 않기 때문에 대량 데이터에 대해서도 빠르게 만들 수 있다.
4) 비정상적인 잡음 데이터에 대해서도 분류가 가능
5) 한 변수와 매우 상관이 높은 다른 불필요한 변수가 있더라도 의사결정나무는 크게 영향을 받지 않는다. 불필요한 변수가 많아지면 의사결정나무가 너무 커지므로 불필요한 변수 제거하는 작업이 필요
6) 한 데이터에서 만들어질 수 있는 의사결정나무의 수는 매우 많기 때문에 탐색적 방법(heuristic search)으로 최적 의사결정나무를 찾는다.
7) 대부분의 의사결정나무 알고리즘은 하향식 반복분할 알고리즘이기 때문에 전체 데이터 집합을 계속 작은 데이터 집합으로 분할해 나간다. 이 작업을 반복해 나가면 어느 잎에는 데이터 수가 너무 적어 통계적으로 의미 있는 분류 결정을 내릴 수 없다. 이를 방지하기 위해 한 노드의 데이터 숫자가 일정한 수보다 작으면 더 이상 분할을 하지 않도록 하는 정지규칙을 만든다.
8) 전체 의사결정나무에는 같은 형태를 갖는 소규모 나무(subtree)가 여러 노드에 나타날 수 있어 의사결정나무를 복잡하게 만들 수 있다.
9) 의사결정나무는 한 노드에서 한 변수에 대한 조건만 조사한다. 따라서 의사결정나무의 분류규칙은 전체 의사결정 공간을 좌표 축(변수)에 평행인 직선으로 분할한다. <그림 5-11>
하지만 <그림 5-12>와 같은 데이터 집합은 축에 평행한 직선으로는 분류하기 쉽지 않다. 이 문제를 해결하기 위해서는 노드에서는 한 개 이상의 변수에 대한 것으로 변형할 수 있다. 두 개 이상의 변수에 대한 시험조건을 만들면 계산이 복잡하고 '어떻게 최적의 시험조건을 만드느냐?'하는 또 다른 문제가 발생한다.

10) 의사결정나무의 성능에 영향을 미치는 것은 가지치기(tree prunning) 방법의 선택이다.

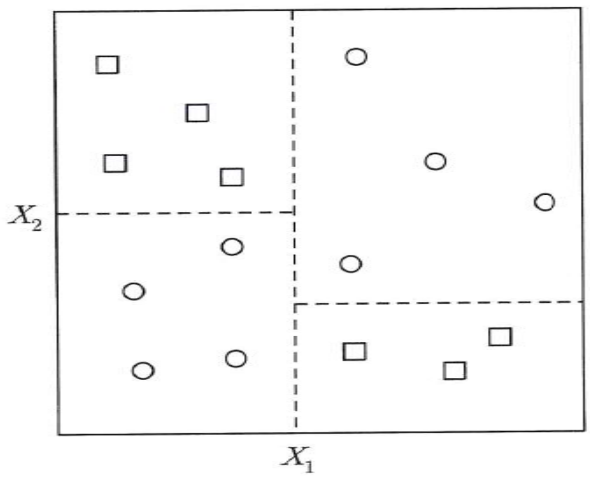



그림 5-11 의사결정나무 모형에서 변수 각각에 의한 의사결정공간의 분할

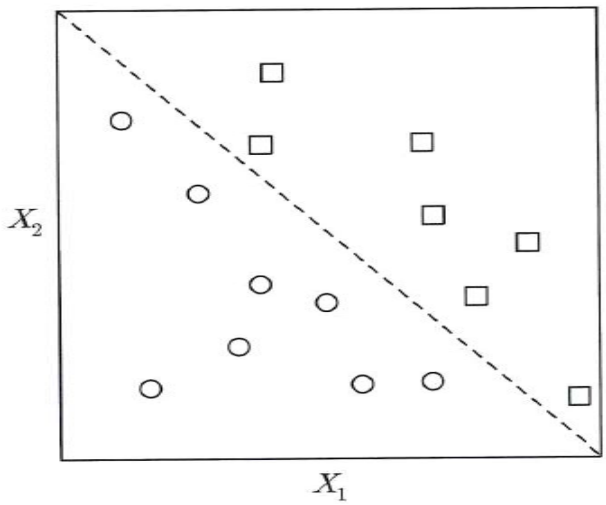



그림 5-12 의사결정나무 모형에서의 공간의 분할이 쉽지 않은 예

### 5.5 분류모형의 평가

▸훈련용 데이터(training data)에 의해 만들어진 모형함수를 시험용 데이터(testing data)에 적용하였을 때 나타나는 분류의 정확도(오류율) 이용.

#### 5.5.1 분류모형 평가를 위한 기타측도

| | | 분류된 집단 | |
|---|---|---|---|
| | | $G_1$ | $G_2$ |
| 실제 집단 | $G_1$ | $f_{11}$ | $f_{12}$ |
| | $G_2$ | $f_{21}$ | $f_{22}$ |

▸분류 모형의 정확도(accuracy) = $\frac{f_{11}+f_{22}}{n}$

▸오류율(error rate) = $\frac{f_{12}+f_{21}}{n}$

▸정확도와 오류율은 집단 $G_1$과 $G_2$의 오분류($f_{12}$,$f_{21}$)에 대한 위험성(risk)이 동일하다는 가정에서 합리적인 측도이다. 그러나 오분류에 대한 집단별 위험성이 다를 수 있다.
(예) 의사가 암환자-> 정상인, 정상인-> 암환자

▸오분류의 위험성이 서로 다른 경우: 민감도, 특이도, 정밀도 이용

▸민감도(sensitivity) = $\frac{f_{11}}{f_{11}+f_{12}}$ : 실제 집단 $G_1$을 집단 $G_1$으로 분류하는 비율

▸특이도(specificity) = $\frac{f_{22}}{f_{21}+f_{22}}$ : 실제 집단 $G_2$를 집단 $G_2$으로 분류하는 비율

▸정밀도(precision) = $\frac{f_{11}}{f_{11}+f_{21}}$ : 집단 $G_1$으로 분류된 사람들 중에서 실제 집단 $G_1$의 비율

▸정확도(accuracy) = $\left(\frac{f_{11}+f_{12}}{n}\right)$(민감도)+$\left(\frac{f_{21}+f_{22}}{n}\right)$(특이도)

= $\frac{f_{11}+f_{22}}{n}$ : 정확도는 민감도와 특이도의 가중합으로 표시

#### 가. 리프트차트(Lift Chart)

▸전체 데이터 중에서 집단 1의 비율: 기준선 반응률(baseline % response)
▸사후 확률의 내림차순으로 전체 데이터를 정리한 후, 정리된 데이터의 상위 10% 데이터에 대해 실제 집단 1을 집단 1로 분류하는 민감도: 상위 10% 반응률(upper 10% response)
▸상위 10%의 리프트(lift): 상위 10%의 반응률/기준선 반응률

**<그림 5-13>** 리프트 차트(lift chart)

x축: 데이터의 사후확률의 상위 백분위수

y축: 이 백분위수의 데이터를 집단 1로 간주하였을 때 실제로 집단 1인 데이터의 반응률

**[표 5.27]** 시험용 데이터의 실제집단과 분류된 집단의 리프트테이블

| 번호 | 데이터 군 | 데이터 수 | 집단 1 수 | 집단 2 수 | %반응률 | 리프트 | %포함률 |
|---|---|---|---|---|---|---|---|
| 1 | 상위 10% | 100 | 85 | 15 | 85.0% | 4.25 | 42.5% |
| 2 | 20% | 100 | 62 | 38 | 62.0% | 3.10 | 31.0% |
| 3 | 30% | 100 | 20 | 80 | 20.0% | 1.00 | 10.0% |
| 4 | 40% | 100 | 10 | 90 | 10.0% | 0.50 | 5.0% |
| 5 | 50% | 100 | 7 | 93 | 7.0% | 0.35 | 3.5% |
| 6 | 60% | 100 | 5 | 95 | 5.0% | 0.25 | 2.5% |
| 7 | 70% | 100 | 3 | 97 | 3.0% | 0.15 | 1.5% |
| 8 | 80% | 100 | 4 | 96 | 4.0% | 0.20 | 2.0% |
| 9 | 90% | 100 | 3 | 97 | 3.0% | 0.15 | 1.5% |
| 10 | 100% | 100 | 1 | 99 | 1.0% | 0.05 | 0.5% |
| 합계 | | 1000 | 200 | 800 | 20.0% | =기준선 반응률 | |

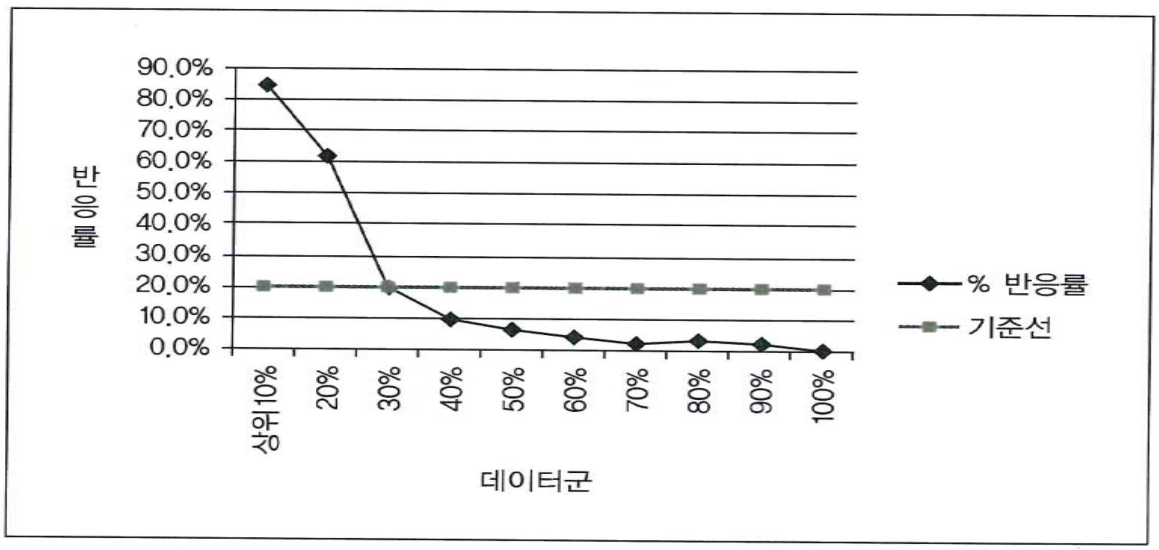



**그림 5-14** [표 5.27]의 리프트테이블에 대한 리프트차트

▸ **교차리프트차트(cross lift chart)**: 훈련용과 시험용 데이터 모두에 대하여 리프트차트 그림

▸ 안정성 있는 분류모형이라면 훈련용과 시험용 데이터의 리프트차트가 크게 다르지 않아야 됨

### 나. 정오분류표(confusion matrix)

▸사후확률의 값을 여러 기준값(cut-off value)로 나눈 후, 각 기준값에 대해 전체 데이터의 정분류, 오분류, 정확도, 민감도, 특이도 표시
▸집단을 결정하기 위한 사후확률의 기준값을 결정할 때 많이 사용

**[표 5.28]**

### 다. 기대이익(Expected Profit) 차트

▸각 데이터를 분류하였을 때 정확히 분류된 경우와 잘못 분류된 경우에 이익(profit)이 발생할 수 있다.

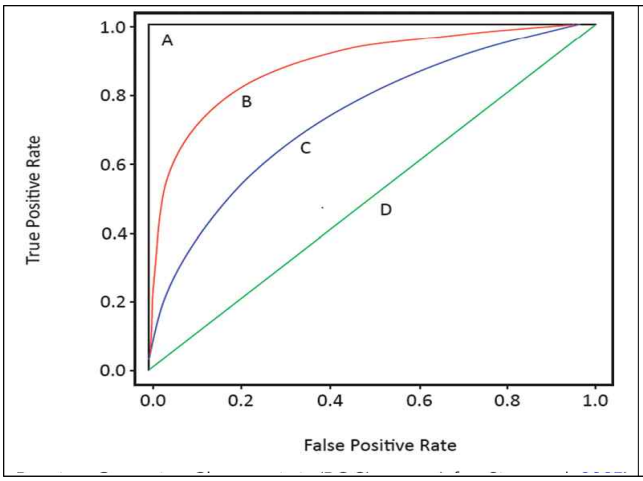
**[표 5.29]** 이익행렬(profit matrix)

| | | 분류된 집단 | |
|---|---|---|---|
| | | $G_1$ | $G_2$ |
| 실제 집단 | $G_1$ | $c_{11}$ | $c_{12}$ |
| | $G_2$ | $c_{21}$ | $c_{22}$ |

▸이익행렬을 알면 리프트테이블에서 각 데이터 군의 기대이익(expected profit) 계산할 수 있음
(예) 상위 10%의 집단 1의 기대이익 $= 0.85 \times c_{11} + 0.15 \times c_{12}$
▸손실행렬(loss matrix)을 알면 유사한 방법으로 기대손실(expected loss)에 대한 리프트차트 그릴 수 있음

### 라. ROC(Receiver Operating Characteristic) 그래프

▸ROC 그래프는 한 분류모형의 (1-특이도)를 $x$축으로 하고, 민감도를 $y$축으로 한 그림
▸민감도를 TPR(true positive rate)이라 하고, (1-특이도)를 FPR(false positive rate)이라 함.
▸ROC 그래프는 사후확률의 기준값을 변화시켜 가면서 각 기준값에 대한 TPR와 FPR을 그린 것.




TPR=0, FPR=0: 모든 데이터를 $G_2$로 분류한 경우
TPR=1, FPR=1: 모든 데이터를 $G_1$로 분류한 경우
TPR=1, FPR=0: 모든 데이터를 정확히 분류한 경우 ⇒ 이상적인 분류모형 의미
: 좋은 분류모형은 좌측 상단에 분류결과가 위치하여야 함.
TPR:FPR=$p$:$p$ (대각선): 데이터를 고정된 확률로 $G_1$과 $G_2$로 임의 분류하는 경우를 의미
: $G_1$의 데이터를 확률 $p$로서 $G_1$로 분류하고, $G_2$의 데이터를 확률 $p$로서 $G_1$로 분류한 경우

▸ROC 그래프는 여러 분류모형들을 비교하는 데 이용

▸ **c-통계량(c-statistics):** ROC 그래프 아래의 면적.
이를 이용하여 모형의 성능이 평균적으로 얼마나 좋은 지 비교할 수 있음.
▸ 이상적인 모형인 경우(FPR=0, TPR=1)에 면적은 1이다.
▸ 모형의 분류결과가 대각선에 위치하는 임의분류인 경우, 면적은 1/2이다.
▸ 한 모형의 ROC 그래프 아래의 면적이 다른 모형의 면적보다 크면 평균적으로 더 우수한 모형이라 함.

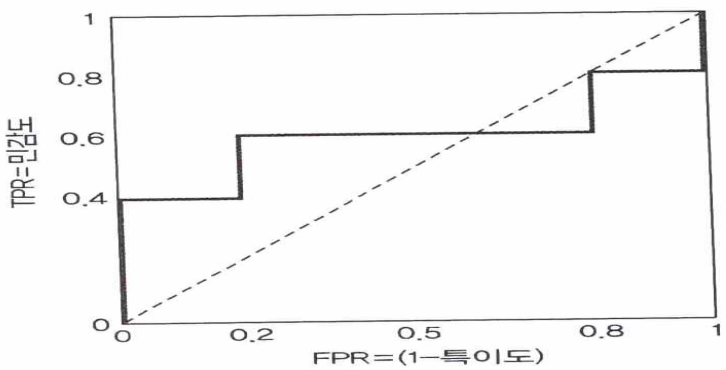
**예제 5.5.1**
베이즈분류 모형을 만든 후 시험용 데이터 10개에 대하여 사후확률을 구하여 오름차순으로 정렬한 결과가 다음과 같다. 이 분류모형의 ROC 그래프를 그리고, c-통계량을 구하여라.

| 데이터 번호 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 실제집단 | + | − | + | − | − | − | + | − | + | + |
| 사후확률 | 0.15 | 0.38 | 0.46 | 0.56 | 0.65 | 0.75 | 0.85 | 0.87 | 0.93 | 0.95 |

(풀이)
① 모든 데이터에 대해 사후확률 값을 구한 후, 오름차순으로 정렬
② 모든 데이터를 + 집단으로 분류하면 $f_{++}=5,\ f_{-+}=5 \Rightarrow$ TPR=$\dfrac{f_{++}}{f_{++}+f_{+-}}$=1, FPR=$\dfrac{f_{-+}}{f_{-+}+f_{--}}$=1
③ 첫 번째 데이터를 − 집단으로, 두 번째 데이터 이상은 + 집단으로 분류하면,
$f_{++}=4,\ f_{+-}=1,\ f_{-+}=5 \Rightarrow$ TPR=4/5=0.8, FPR=5/5=1
④ 첫 번째와 두 번째 데이터를 − 집단으로, 세 번째 데이터 이상은 + 집단으로 분류하면,
$f_{++}=4,\ f_{+-}=1,\ f_{--}=1,\ f_{-+}=4 \Rightarrow$ TPR=4/5=0.8, FPR=4/5=0.8
⑤ 유사한 방법으로 계속 진행하고, TPR과 FPR을 구함.

| 데이터 번호 | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 실제집단 | | + | − | + | − | − | − | + | − | + | + |
| 사후확률 | | 0.15 | 0.38 | 0.46 | 0.56 | 0.65 | 0.75 | 0.85 | 0.87 | 0.93 | 0.95 |
| $f_{++}$ | 5 | 4 | 4 | 3 | 3 | 3 | 3 | 2 | 2 | 1 | 0 |
| $f_{-+}$ | 5 | 5 | 4 | 4 | 3 | 2 | 1 | 1 | 0 | 0 | 0 |
| $f_{--}$ | 0 | 0 | 1 | 1 | 2 | 3 | 4 | 4 | 5 | 5 | 5 |
| $f_{+-}$ | 0 | 1 | 1 | 2 | 2 | 2 | 2 | 3 | 3 | 4 | 5 |
| TPR | 1 | 0.8 | 0.8 | 0.6 | 0.6 | 0.6 | 0.6 | 0.4 | 0.4 | 0.2 | 0 |
| FPR | 1 | 1 | 0.8 | 0.8 | 0.6 | 0.4 | 0.2 | 0.2 | 0 | 0 | 0 |

⑥ ROC 그래프 그림




▸ ROC 그래프 아래 면적을 c-통계량이라고 함 : c-통계량= $0.2\times0.8+0.6\times0.6+0.4\times0.2=0.6$

### 5.5.2 분류모형의 비교

▸데이터를 분류하기 위해서는 여러 모형을 시도한 후 제일 적합한 모형 선택

**가. 정확도의 신뢰구간 추정**

모형의 실제 정확도 = $p$

$X$ = $n$개의 시험용 데이터 집합에서 분류모형에 의해 올바로 분류된 데이터의 수

$X \sim B(n,p)$

분류모형의 정확도(accuracy) = $\frac{X}{n}$

$$E\left(\frac{X}{n}\right)=\frac{1}{n}E(X)=\frac{1}{n}np=p, \quad Var\left(\frac{X}{n}\right)=\frac{1}{n^2}Var(X)=\frac{1}{n^2}np(1-p)=\frac{p(1-p)}{n}$$

$n$이 충분히 크면 이항분포의 확률은 정규분포에 의해 근사적으로 구할 수 있음.

실제 정확도 $p$에 대한 $100(1-\alpha)\%$ 신뢰구간 확률은 다음을 이용함.

$$P\left(-Z_{\alpha/2} \le \frac{accuracy-p}{\sqrt{\frac{p(1-p)}{n}}} \le Z_{\alpha/2}\right)=1-\alpha$$

실제 정확도 $p$에 대한 $100(1-\alpha)\%$ 신뢰구간:

$$\frac{2n\times accuracy + Z_{\alpha/2}^2 \pm Z_{\alpha/2}\sqrt{Z_{\alpha/2}^2 + 4n\times accuracy - 4n\times accuracy^2}}{2n+2Z_{\alpha/2}^2}$$

**예제 5.5.2**

분류모형 $M_1$을 100개의 시험용 데이터에 적용하여 보니 80% 정확도를 가졌다. 실제 정확도를 95% 신뢰도로 구간 추정하여라.

(풀이) $n=100$, accuracy$=0.8$, 신뢰도 95%일 때, $z_{0.05/2}=1.96$이다. 위의 식에 대입하면

$$\frac{2\times100\times0.8+1.96^2 \pm 1.96\sqrt{1.96^2+4\times100\times0.8-4\times100\times0.8^2}}{2\times100+2\times1.96^2}=(71.1\%, 86.7\%)$$

**나. 두 모형의 정확도 비교**

두 분류모형 $M_1$과 $M_2$를 두 개의 서로 독립인 시험용 데이터 집합 $D_1$과 $D_2$에 각각 적용하여 정확도 $a_1$과 $a_2$를 측정하였다고 하자. 각 데이터 수가 $n_1$과 $n_2$일 때 정확도 $a_1$과 $a_2$가 통계적으로 유의한 지 검정하는 방법을 알아보자.

- $n_1$과 $n_2$가 충분히 크다면 정확도 $a_1$과 $a_2$는 근사적으로 정규분포 따름
- 정확도의 차 $d = a_1 - a_2$도 근사적으로 정규분포 따름
- 정확도의 차 $d$의 분산 추정량 : $\hat{\sigma_d}^2 = \frac{a_1(1-a_1)}{n_1} + \frac{a_2(1-a_2)}{n_2}$

- 정확도의 차 $d$에 대한 $100(1-\alpha)\%$ 신뢰구간:

$$d \pm Z_{\alpha/2}\sqrt{\frac{a_1(1-a_1)}{n_1} + \frac{a_2(1-a_2)}{n_2}}$$

**예제 5.5.3**

분류모형 $M_1$은 50개의 시험용 데이터에서 85%의 정확도를 가지고, 분류모형 $M_2$는 500개의 시험용 데이터에서 75%의 정확도를 보인다. 분류모형 $M_1$이 $M_2$보다 더 좋은 모형인가?

(풀이) 분류모형 $M_1$ : $n_1 = 50,\ a_1 = 0.85$

분류모형 $M_2$ : $n_2 = 500,\ a_2 = 0.75$

$d = a_1 - a_2 = 0.85 - 0.75 = 0.1$

$\hat{\sigma_d}^2 = \frac{0.85 \times 0.15}{50} + \frac{0.75 \times 0.25}{500} = 0.0029$

정확도의 차이에 대한 95% 신뢰구간:

$$0.1 \pm 1.96\sqrt{0.0029} = (-0.2060,\ 0.0060)$$

$\Rightarrow$ 신뢰구간이 0을 포함하고 있으므로, 두 모형의 정확도의 차는 통계적으로 유의하지 않음.
두 분류모형 $M_1$과 $M_2$ 중 어느 모형이 더 좋다고 말할 수 없다.