

## 제5장. 정형데이터 마이닝

### 1. 데이터마이닝 개요

1. 데이터마이닝 정의 : 대용량 데이터에서 \_\_\_\_\_ 을 파악하거나 예측하여 의사결정에 활용하는 방법
  - 통계분석과 데이터마이닝의 차이
    - A. 통계분석 : \_\_\_\_\_ 에 따른 분석이나 검증
    - B. 데이터 마이닝 : \_\_\_\_\_ 를 이용해 \_\_\_\_\_ 를 추출
  - 데이터마이닝 활용 : \_\_\_\_\_ , \_\_\_\_\_ , \_\_\_\_\_ , \_\_\_\_\_ , \_\_\_\_\_
  - 데이터마이닝 방법론 : \_\_\_\_\_ , \_\_\_\_\_ , \_\_\_\_\_ , \_\_\_\_\_ , \_\_\_\_\_ , \_\_\_\_\_ , \_\_\_\_\_ 등

### 2. 데이터마이닝 학습법

- A. \_\_\_\_\_ ( \_\_\_\_\_ ) : \_\_\_\_\_ 변수가 존재, \_\_\_\_\_ , \_\_\_\_\_ , \_\_\_\_\_
- B. \_\_\_\_\_ ( \_\_\_\_\_ ) : \_\_\_\_\_ 변수가 없이 설명을 위한 분석, \_\_\_\_\_ , \_\_\_\_\_ , \_\_\_\_\_

### 3. 데이터마이닝 추진단계

- A. \_\_\_\_\_ : 명확한 \_\_\_\_\_ 설정
- B. \_\_\_\_\_ : 다양한 \_\_\_\_\_ 를 준비 및 정제( \_\_\_\_\_ )
- C. \_\_\_\_\_ : 목적변수를 정의, 모델링을 위한 \_\_\_\_\_
- D. \_\_\_\_\_ : 기법을 \_\_\_\_\_ 정보 추출
- E. \_\_\_\_\_ : 결과를 \_\_\_\_\_ 하고 업무에 적용, \_\_\_\_\_ ( \_\_\_\_\_ ) 등의 기대효과 전파

### 4. 데이터 분할

- A. \_\_\_\_\_ ( \_\_\_\_\_ ) : \_\_\_\_\_ %의 데이터를 훈련용으로 활용
- B. \_\_\_\_\_ ( \_\_\_\_\_ ) : \_\_\_\_\_ %의 데이터를 과대/과소 추정의 판정 목적으로 사용
- C. \_\_\_\_\_ ( \_\_\_\_\_ ) : \_\_\_\_\_ %의 데이터를 테스트데이터나 과거 데이터를 활용하여 성능평가에 활용

5. 모델의 성능 평가 은행의 대출 문제로 연이율 20%로 가정, 100만원을 100명에게 대출한다고 가정

(EX) 두 모형에서 정확도가 85%로 같다면 은행 입장에선 어떤 모형이 더 좋은 모형인가?

모형1	A	B
a	65	10
b	5	20

모형2	A	B
a	75	0
b	15	10

a, b : 테스트 데이터의 예측 분류. a : 우량고객, b : 불량고객  
A, B : 테스트 데이터의 실제 분류. A : 우량고객, B : 불량고객  
: 연이율 20%로 100만원을 대출

#### A. 기대수익

a. 기대수익 = ( \_ \_ \_ \_ \_ 명 \_ \_ \_ \_ \_ 만원) - ( \_ \_ \_ \_ \_ 명 \_ \_ \_ \_ \_ 만원) =  
\_ \_ \_ \_ \_ 만원

b. 기대수익 = ( \_ \_ \_ \_ \_ 명 \_ \_ \_ \_ \_ 만원) - ( \_ \_ \_ \_ \_ 명 \_ \_ \_ \_ \_ 만원) =  
\_ \_ \_ \_ \_ 만원

#### B. 기대손실비용

a. 기대손실비용 = ( \_ \_ \_ \_ \_ 명 \_ \_ \_ \_ \_ 만원) + ( \_ \_ \_ \_ \_ 명 \_ \_ \_ \_ \_ 만  
원) = \_ \_ \_ \_ \_ 만원

b. 기대손실비용 = ( \_ \_ \_ \_ \_ 명 \_ \_ \_ \_ \_ 만원) - ( \_ \_ \_ \_ \_ 명 \_ \_ \_ \_ \_ 만  
원) = \_ \_ \_ \_ \_ 만원

C. 결과 기대수익과 기대손실비용으로 봤을 때 \_ \_ \_ \_ \_ 모형이 우수함

## 2. 의사결정분석 나무

### 1. 분류분석 vs 예측분석

A. 공통점 : 레코드의 \_ \_ \_ \_ \_ 을 미리 알아 맞히는 점

B. 차이점

- 분류 : \_ \_ \_ \_ \_ 속성의 값을 예측
- 예측 : \_ \_ \_ \_ \_ 속성의 값을 예측

C. 분류의 예 : 학생들의 국어, 영어, 수학 점수를 예측, 카드회사에서 회원들의 가입정보를 통해 1년 후 신  
용등급을 예측

D. 분류기법

a. \_ \_ \_ \_ \_ ( \_ \_ \_ \_ \_ )

b. \_ \_ \_ \_ \_ ( \_ \_ \_ \_ \_ ), \_ \_ \_ \_ \_ ( \_ \_ \_ \_ \_ ),  
C5.0

c. \_ \_ \_ \_ \_ ( \_ \_ \_ \_ \_ ), \_ \_ \_ \_ \_

d. \_ \_ \_ \_ \_ ( \_ \_ \_ \_ \_ )

e. \_ \_ \_ \_ \_ ( \_ \_ \_ \_ \_ )

f. \_ \_ \_ \_ \_ ( \_ \_ \_ \_ \_ )

g. 규칙기반 분류와 사례추론

### 1. 의사결정나무 특징

- 분류함수를 의사결정 규칙으로 이뤄진 나무 모양으로 그리는 방법
- 의사결정 문제를 시각화하여 \_ \_ \_ \_ \_ 과 \_ \_ \_ \_ \_ 를 한 눈에 볼 수 있음
- \_ \_ \_ \_ \_ 가 직접 나타나게 되어 분석이 간편함
- \_ \_ \_ \_ \_ 가 좋음
- \_ \_ \_ \_ \_ 에서도 빠르게 만들 수 있음
- \_ \_ \_ \_ \_ 에 대해서도 민감함 없이 분류
- \_ \_ \_ \_ \_ 이 높은 다른 불필요한 변수에 큰 영향을 받지 않음

### 2. 의사결정나무 활용

- \_\_\_\_\_ ( \_\_\_\_\_ ): 비슷한 특성을 갖는 그룹으로 분할
- \_\_\_\_\_ ( \_\_\_\_\_ ): 범주를 몇 개의 등급으로 나눔
- \_\_\_\_\_ ( \_\_\_\_\_ ): 규칙을 찾고, 미래의 사건을 예상
- \_\_\_\_\_ ( \_\_\_\_\_ ) 및 \_\_\_\_\_ ( \_\_\_\_\_ ): 목표변수에 큰 영향을 미치는 변수를 고름
- \_\_\_\_\_ ( \_\_\_\_\_ ): 여러개의 예측 변수들을 결합해 목표 변수 파악
- \_\_\_\_\_ ( \_\_\_\_\_ ): 범주형 변수를 소수의 몇 개로 병합, 또는 연속형 변수를 몇 개의 등급으로 이산화

### 3. 의사결정나무 분석

- 분석 단계 : \_\_\_\_\_ -> \_\_\_\_\_ -> \_\_\_\_\_  
\_\_\_\_\_ -> \_\_\_\_\_
- 나무의 가지치기 : 과적합, 과소적합을 예방하기 위해 마디의 자료가 일정수 이하일 경우 가지치기를 정지
- 분순도에 따른 분할 척도 : \_\_\_\_\_, \_\_\_\_\_, \_\_\_\_\_  
\_\_\_\_\_

### 4. 의사결정나무 분석의 종류

- \_\_\_\_\_ ( \_\_\_\_\_ )
  - 목적변수가 범주형인 경우 \_\_\_\_\_, 연속형인 경우 \_\_\_\_\_ 을 이용해 \_\_\_\_\_ 를 사용
  - 개별 입력변수 뿐만 아니라 입력변수들의 선형결합등 중 최적의 분리를 찾을 수있다.
- \_\_\_\_\_ 와 \_\_\_\_\_
  - 다지분리가 가능, 범부형 변수의 범주 수만큼 분리 가능
  - 불순도의 척도로 \_\_\_\_\_ 지수 사용
- \_\_\_\_\_ ( \_\_\_\_\_ )
  - 가지치기를 하지 않고 적당한 크기에서 나무모형을 성장을 중지
  - 입력변수가 반드시 \_\_\_\_\_ 변수이어야 함
  - 불순도 척도로 \_\_\_\_\_ 통계량 사용

## 3. 앙상블 기법

### 1. 앙상블 기법

- 주어진 자료들로부터 여러 개의 \_\_\_\_\_ 들을 만든후 조합하여 하나의 \_\_\_\_\_ 을 만드는 기법
- \_\_\_\_\_ ( \_\_\_\_\_ ), classifier Convination 등
- 학습방법의 \_\_\_\_\_ 을 해결하기 위해 고안
- 가장 불안정한 기법은 \_\_\_\_\_, 가장 안전한 기법은 \_\_\_\_\_  
\_\_\_\_\_
- 종류 :
  - \_\_\_\_\_ ( \_\_\_\_\_ )
    - 여러 \_\_\_\_\_ 을 생성하고 예측모형 결과를 결과를 결합

- 훈련자료를 \_\_\_\_\_으로 생각하고 \_\_\_\_\_을  
구한 것과 같음
  - \_\_\_\_\_을 줄이고, \_\_\_\_\_을 향상 시킬 수 있음
- B. \_\_\_\_\_(\_\_\_\_\_)
- \_\_\_\_\_모형들을 결합하여 \_\_\_\_\_을 만드는  
방법
  - \_\_\_\_\_를 빨리 쉽게 줄일 수 있고, \_\_\_\_\_의  
향상으로 \_\_\_\_\_에 비해 뛰어난 예측력을 보임
- C. \_\_\_\_\_(\_\_\_\_\_)
- 의사결정나무의 특징인 분산이 크다는 점을 고려, \_\_\_\_\_과 \_\_\_\_\_보다 더 많은  
\_\_\_\_\_을 주어 약한 학습기를 생성후 선형결합하여 \_\_\_\_\_  
\_\_\_\_\_를 만드는 방법
  - \_\_\_\_\_이나 해석이 어렵다는 단점, \_\_\_\_\_이  
매우 높은 장점
  - \_\_\_\_\_가 많은 경우 더 좋은 예측력을 보임
- D. \_\_\_\_\_(\_\_\_\_\_)
- 동일한 타입의 모델을 조합하는 \_\_\_\_\_, \_\_\_\_\_과는 달리 다양한 학습 모델을 통해 구성

## 2. 오분류표

n=165		Predicted: NO	Predicted: YES	
Actual: NO		TN = 50	FP = 10	60
Actual: YES		FN = 5	TP = 100	105
		55	110	

- A. [\_\_\_\_\_]:  $(TP+TN)/Total$ , 올바르게 검출 (실제 악성/정상을 예측)
- B. [\_\_\_\_\_]:  $(FP+FN)/Total$ , 잘못되게 검출 (잘못된 악성/정상 예측)
- C. [\_\_\_\_\_]:  $TP/Predicted\ YES$ , 참으로 분류한 것중 올바른 참의 비율 (악성으로  
예측한 것 중 실제 악성 샘플의 비율)
- D. [\_\_\_\_\_]:  $TP/Actual\ YES$ , 실제 참을 참으로 분류 (실제 악성 중에서 악  
성으로 예측)
- E. [\_\_\_\_\_]:  $FP/Actual\ NO$ , 실제 거짓을 거짓으로 분류 (실제 정상 중에서  
악성으로 예측)
- F. [\_\_\_\_\_]:  $TP/Actual\ YES$ , 예측과 실제 모두 참 (실  
제 악성 중에서 악성으로 예측)
- G. [\_\_\_\_\_]:  $TN/Actual\ NO$ , 예측과 실제 모두 거짓 (실  
제 정상 중에서 정상으로 예측)
- H. [\_\_\_\_\_]:  $FP/Actual\ NO$ , 실제 거짓인데 참으로 분  
류 (실제 정상을 악성으로 예측)
- I. [\_\_\_\_\_]:  $FN/Actual\ YES$ , 실제 참인데 거짓으로  
검출 (실제 악성을 정상으로 예측)

## 1. ROC

- \_\_\_\_\_ 와 \_\_\_\_\_ 를 활용하여 모형을 평가
- \_\_\_\_\_ ( \_\_\_\_\_ )
  - \_\_\_\_\_ =  $(AR+1)/2$
  - 90% 이상 excellent, 80% 이상 good, 70% 이상 Fair

## 4. 인공신경망 분석

### 1. 인공신경망 연구

- 1943년 \_\_\_\_\_ ( \_\_\_\_\_ )과 \_\_\_\_\_ ( \_\_\_\_\_ ): 인간의 뇌를 수많은 신경세포가 연결된 하나의 \_\_\_\_\_ 모형으로 간주, 신경세포의 신호처리 과정을 모형화 하여 단순 패턴분류 모형으로 개발
- 헵(Hebb): 신경세포(뉴런) 사이의 \_\_\_\_\_ ( \_\_\_\_\_ )를 조정하여 학습규칙을 개발
- 로젠블랫(Rosenblatt): \_\_\_\_\_ ( \_\_\_\_\_ )이라는 인공세포 개발, 비선형의 한계점 발생 ( \_\_\_\_\_ )문제
- 홉필드(Hopfield), 러멜하트(Rumelhart), 맥클린드(McClelland): \_\_\_\_\_ ( \_\_\_\_\_ )을 활용하여 \_\_\_\_\_ 을 극복한 \_\_\_\_\_ 퍼셉트론으로 새로운 인공신경망 모형 등장

### 2. 뉴런( \_\_\_\_\_ )

- 아주 단순하지만 복잡하게 연결된 프로세스로 이루어져 있음
- \_\_\_\_\_ 가 있는 \_\_\_\_\_ 로 연결됨
- 여러개의 \_\_\_\_\_ 를 받아 하나의 \_\_\_\_\_ 를 생성
- \_\_\_\_\_ 즉, \_\_\_\_\_ ( \_\_\_\_\_ )을 사용
  - 입력신호의 \_\_\_\_\_ 의 합을 계산하여 \_\_\_\_\_ 과 비교
  - \_\_\_\_\_ 의 합이 \_\_\_\_\_ 보다 작으면 뉴런의 출력은 \_\_\_\_\_, 같거나 크면 \_\_\_\_\_ 을 출력

### 3. 신경망모형 구축시 고려사항

#### A. 입력변수

- \_\_\_\_\_ 형 변수: \_\_\_\_\_ 가 일정수준 이상이고 빈도가 일정할 때
  - 가변수화하여 적용(성별[남여], 남성[1,0], 여성[0,1])
- \_\_\_\_\_ 형 변수: 범위가 변수들간에 큰 차이가 없을 때
  - 평균을 중심으로 \_\_\_\_\_ 가 대칭이 아니면 비효율적
  - \_\_\_\_\_ 또는 \_\_\_\_\_ 를 통해 활용하는 것이 적절

#### B. 가중치 초기값

- \_\_\_\_\_ 의 경우, 초기값에 따라 결과가 크게 달라질 수 있으므로 중요
- 가중치가 0이면 \_\_\_\_\_ 함수에서는 \_\_\_\_\_ 이 되고 신경망 모형은 \_\_\_\_\_ 모형이 됨.
- 초기값은 0 근처의 랜덤값으로 선정
- 초기에는 \_\_\_\_\_ 모형에서 \_\_\_\_\_ 가 증가하면서 \_\_\_\_\_ 으로 변경 됨

#### C. 신경망 모형의 \_\_\_\_\_ 함수는 비볼록하무이고, 여러 개의 국소 \_\_\_\_\_ ( \_\_\_\_\_ )를 가짐

- 랜덤하게 선택된 여러개의 초기값에 대한 신경망을 적합한 후 얻은 해들을 비교하여 가장 \_ \_ \_ \_ 가 적은 것을 선택
- 최종 예측값을 얻거나 \_ \_ \_ \_ (또는 \_ \_ \_ \_ )을 구하여 최종 예측값으로 선정
- 훈련자료에 대해서 \_ \_ \_ \_ ( \_ \_ \_ \_ )을 적용하여 최종 예측값을 선정

#### D. 학습률

- 처음에는 큰 값으로 정하고 반복이 진행될 수록 \_ \_ \_ \_ 에 가까워 짐

#### E. \_ \_ \_ \_ ( \_ \_ \_ \_ \_ \_ \_ \_ ), \_ \_ \_ \_ 노드( \_ \_ \_ \_ \_ \_ \_ \_ )의 수

- 많으면 가중치가 많아져 \_ \_ \_ \_ \_ \_ \_ \_ 문제 발생
- 적으면 \_ \_ \_ \_ \_ \_ \_ \_ 문제 발생
- 하나인 신경망은 범용근사자(Universal Approximator)이므로 가급적이면 하나로 선정
- 노드는 적절히 큰 값으로 설정하고 \_ \_ \_ \_ 를 감소하면서 모수에 대한 \_ \_ \_ \_ 를 적용

### 4. 로지스틱 회귀분석

- \_ \_ \_ \_ \_ \_ \_ \_ 가 \_ \_ \_ \_ \_ \_ \_ \_ 형인 경우에 적용하는 회귀분석모형
- 새로운 \_ \_ \_ \_ 변수가 주어질 때 \_ \_ \_ \_ 변수의 각 범주에 속할 확률이 얼마인지 추정, 추정 확률을 기준으로 분류하는 목적으로 활용
- 모형의 적합을 통해 추정될 확률을 \_ \_ \_ \_ \_ \_ \_ \_ ( \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ )이라고 함
- \_ \_ \_ \_ \_ \_ \_ \_ 함수를 활용하여 로짓분석 실행
- \_ \_ \_ \_ \_ \_ \_ \_ ( \_ \_ \_ \_ 변수 ~ \_ \_ \_ \_ 변수1 + \_ \_ \_ \_ 변수2 + ... , family=binomial, data=데이터셋)
- 결과 추정값이 5.14이면, 독립변수의 단위가 증가함에 따라 종속변수가 0에서 1로 바뀔 \_ \_ \_ \_ ( \_ \_ \_ \_ )가  $\exp(5.140) = 170$ 배 증가한다는 의미.

## 5. 군집분석

### 1. 군집분석

- 각 객체의 \_ \_ \_ \_ \_ \_ \_ \_ 을 측정하여 \_ \_ \_ \_ \_ \_ \_ \_ 이 높은 대상집단을 \_ \_ \_ \_
- 군집에 속한 객체들의 \_ \_ \_ \_ \_ \_ \_ \_ 과 서로 다른 군집에 속한 객체간의 \_ \_ \_ \_ \_ \_ \_ \_ 을 규명하는 분석 방법
- 특성에 따라 여러개의 \_ \_ \_ \_ \_ \_ \_ \_ 인 집단으로 나눔
- 군집의 \_ \_ \_ \_ \_ \_ \_ \_ 나 \_ \_ \_ \_ \_ \_ \_ \_ 에 대한 가정없이 \_ \_ \_ \_ \_ \_ \_ \_ 를 기준으로 군집화 유도

### 1. 군집분석 특징

- \_ \_ \_ \_ \_ \_ \_ \_ ( \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ )에 해당하며 \_ \_ \_ \_ \_ \_ \_ \_ 의 정의 없이 학습 가능
- 분석 목적에 따라 적절한 군집으로 정의 가능
- 요약분석과의 차이 : \_ \_ \_ \_ 변수를 묶는 것이 아닌 \_ \_ \_ \_ 를 묶어줌
- 판별분석과의 차이 : 판별분석은 사전에 집단이 나누어져 있어야 하지만, 군집분석은 \_ \_ \_ \_ 이 없는 상태에서 구분

### 2. 군집분석 거리 측정

- 데이터가 \_ \_ \_ \_ 인 경우 : \_ \_ \_ \_ \_ \_ \_ \_ 거리, \_ \_ \_ \_ \_ \_ \_ \_ 거리, \_ \_ \_ \_ \_ \_ \_ \_ 거리, \_ \_ \_ \_ \_ \_ \_ \_ 거리, \_ \_ \_ \_ \_ \_ \_ \_ 거리, \_ \_ \_ \_ \_ \_ \_ \_ 거리, \_ \_ \_ \_ \_ \_ \_ \_ 거리 등 활용
- 데이터가 \_ \_ \_ \_ 인 경우 : \_ \_ \_ \_ \_ \_ \_ \_ 거리 활용

### 3. 계층적 군집 분석

- n개의 군집으로 시작해 군집의 개수를 \_\_\_\_\_ 방법
  - A. \_\_\_\_\_ ( \_\_\_\_\_ , \_\_\_\_\_ )
    - $n \times n$  거리행렬에서 거리가 \_\_\_\_\_ 데이터를 묶어서 군집화
    - \_\_\_\_\_ 거리행렬에서 거리가 \_\_\_\_\_ 데이터 또는 군집을 새로운 군집으로 형성
  - B. \_\_\_\_\_ ( \_\_\_\_\_ , \_\_\_\_\_ )
    - 군집과 군집 또는 데이터와 거래를 계산시 \_\_\_\_\_ 를 계산하여 거리행렬 수정
  - C. \_\_\_\_\_ ( \_\_\_\_\_ )
    - 군집과 군집 또는 데이터와 거래를 계산시 \_\_\_\_\_ 를 거리로 계산하여 거리행렬 수정
  - D. \_\_\_\_\_ ( \_\_\_\_\_ )
    - 군집내 편차들의 \_\_\_\_\_ 을 고려한 방법
    - 군집간 정보 손실을 \_\_\_\_\_ 하기 위해 군집화를 진행

### 4. 비계층적 군집 분석

n개의 개체를 g개의 군집으로 나눌 수 있는 모든 \_\_\_\_\_ 방법을 점검해 최적화한 군집을 형성

- \_\_\_\_\_ ( \_\_\_\_\_ )
    - A. 원하는 군집의 개수와 초기값( \_\_\_\_\_ )들을 정해 \_\_\_\_\_ 를 중심으로 군집을 형성
    - B. 각 데이터를 \_\_\_\_\_ 가 가장 가까운 \_\_\_\_\_ 가 있는 군집으로 분류
    - C. 각 군집의 \_\_\_\_\_ 값을 다시 계산
    - D. \_\_\_\_\_ 값이 변화가 없고 \_\_\_\_\_ 가 군집으로 할당될 때까지 반복
  - \_\_\_\_\_의 특징
    - 거리 계산을 통해 군집화되므로 \_\_\_\_\_ 변수에 활용 가능
    - k개의 초기 중심값은 임의로 선택 가능, 가급적이면 \_\_\_\_\_ 것이 바람직하다
    - 초기값을 일러로 선택하지 않는 것이 좋다
    - 초기 중심으로부터 \_\_\_\_\_ 을 최소화하는 방향으로 군집이 형성되는 \_\_\_\_\_ ( \_\_\_\_\_ ) 알고리즘이므로 안정된 군집은 보장하나 \_\_\_\_\_ 이라는 보장은 없다.
  - \_\_\_\_\_의 장점
    - 1. 알고리즘이 \_\_\_\_\_ 하며, 빠르게 수행되어 분석 방법 적용이 용이
    - 1. 계층적 군집분석에 비해 \_\_\_\_\_ 양의 데이터를 다룰 수 있다.
  - \_\_\_\_\_의 단점
    - 1. 군집의 \_\_\_\_\_ , \_\_\_\_\_ 와 \_\_\_\_\_ 정의가 어렵다
    - 1. 사전에 주어진 \_\_\_\_\_ 이 없으므로 결과 해석이 어렵다.
    - 1. \_\_\_\_\_ 이나 \_\_\_\_\_ 에 영향을 많이 받는다.
    - 1. \_\_\_\_\_ 형태가 아닌( \_\_\_\_\_ ) 군집이 존재할 경우 성능이 떨어진다
- <br><br>

#### 1. 혼합분포 군집

모형기반( \_\_\_\_\_ )의 군집 방법이며, 데이터가 k개의 \_\_\_\_\_ (흔히 정규분포 또는 다변량 정규분포를 가정함)의 가중합으로 표현되는 모집단 모형으로부터 나왔다는 가정하에서 \_\_\_\_\_ 와 함께 \_\_\_\_\_ 를 자료로부터 추정하는 방법

- A. \_\_\_\_\_ : 전체 거래 중 항목 A와 B를 \_\_\_\_\_ 하는 거래의 비율.  
 • \_\_\_\_\_ :  $P(A \cap B) = ( \text{_____} ) / ( \text{_____} )$



- B. \_\_\_\_\_ : 항목 A를 포함하는 거래 중에서 항목 A와 항목 B가 같이 포함될 확률. \_\_\_\_\_ 의 정도를 파악
- \_\_\_\_\_ :  $P(A \cap B) / P(A) = ( \text{_____} ) / ( \text{_____} )$
- C. \_\_\_\_\_ : A가 주어지지 않았을 때의 품목 B의 확률에 비해 A가 주어졌을 때의 품목 B가 \_\_\_\_\_ 비율.
- D. \_\_\_\_\_ : \_\_\_\_\_ /  $P(B) = P(A \cap B) / P(A)P(B) = ( \text{_____} ) / ( \text{_____} \times \text{_____} )$

#### 4. 연관분석 특징

##### A. 절차

- \_\_\_\_\_ ( \_\_\_\_\_ )를 선정 - 5% 시작
- 품목 중 \_\_\_\_\_ 를 넘는 품목을 분류
- 2가지 품목 집합 생성
- 반복적으로 수행해 빈발품목 집합을 선정

##### B. 장점

- \_\_\_\_\_ 기법 : \_\_\_\_\_ 으로 표현되는 연관성 분석의 결과를 쉽게 이해 가능
- 강력한 \_\_\_\_\_ 분석 기법 : 분석 \_\_\_\_\_ 이나 \_\_\_\_\_ 이 없을 경우 유용
- 사용인 편리한 분석 데이터 \_\_\_\_\_ : 거래 내용에 대한 데이터를 변환 없이 그 자체로 이용
- 계산의 \_\_\_\_\_ : 계산이 간단함

##### C. 단점

- 상당한 수의 \_\_\_\_\_ : 품목수가 늘어나면 계산은 기하급수적으로 늘어남 -> 유사한 품목을 범주로 한 \_\_\_\_\_
- 연관 규칙의 신뢰도 하한을 새롭게 정의, 의미가 적은 연관규칙을 제외
- \_\_\_\_\_ 품목의 결정 : 너무 세분화한 품목으로 연관성을 찾으려면 의미 없는 분석이 될 수도 있음 -> 적절히 구분되는 큰 범주로 구분 전체 분석후 연관성 작업 수행
- 품목의 \_\_\_\_\_ : 전체자료에서 동일한 빈도를 갖는 경우, 연관성 분석은 가장 좋은 결과를 얻음. 그러나 \_\_\_\_\_ 이 적은 품목은 당연히 포함된 거래수가 적을 것이고 규칙 발견 과정에서 제외되기 쉬움.

#### 1. 평가기준 적용시 주의점

- 두 항목의 \_\_\_\_\_ 가 높다고 해서 꼭 두 항목이 높은 \_\_\_\_\_ 가 있는 것은 아님. ( \_\_\_\_\_ 를 함께 고려)
- 만일 두 항목의 \_\_\_\_\_ 가 높게 나왔어도 전체 항목 중 두 항목의

#### 1. Apriori 가장 많이 사용하는 알고리즘

