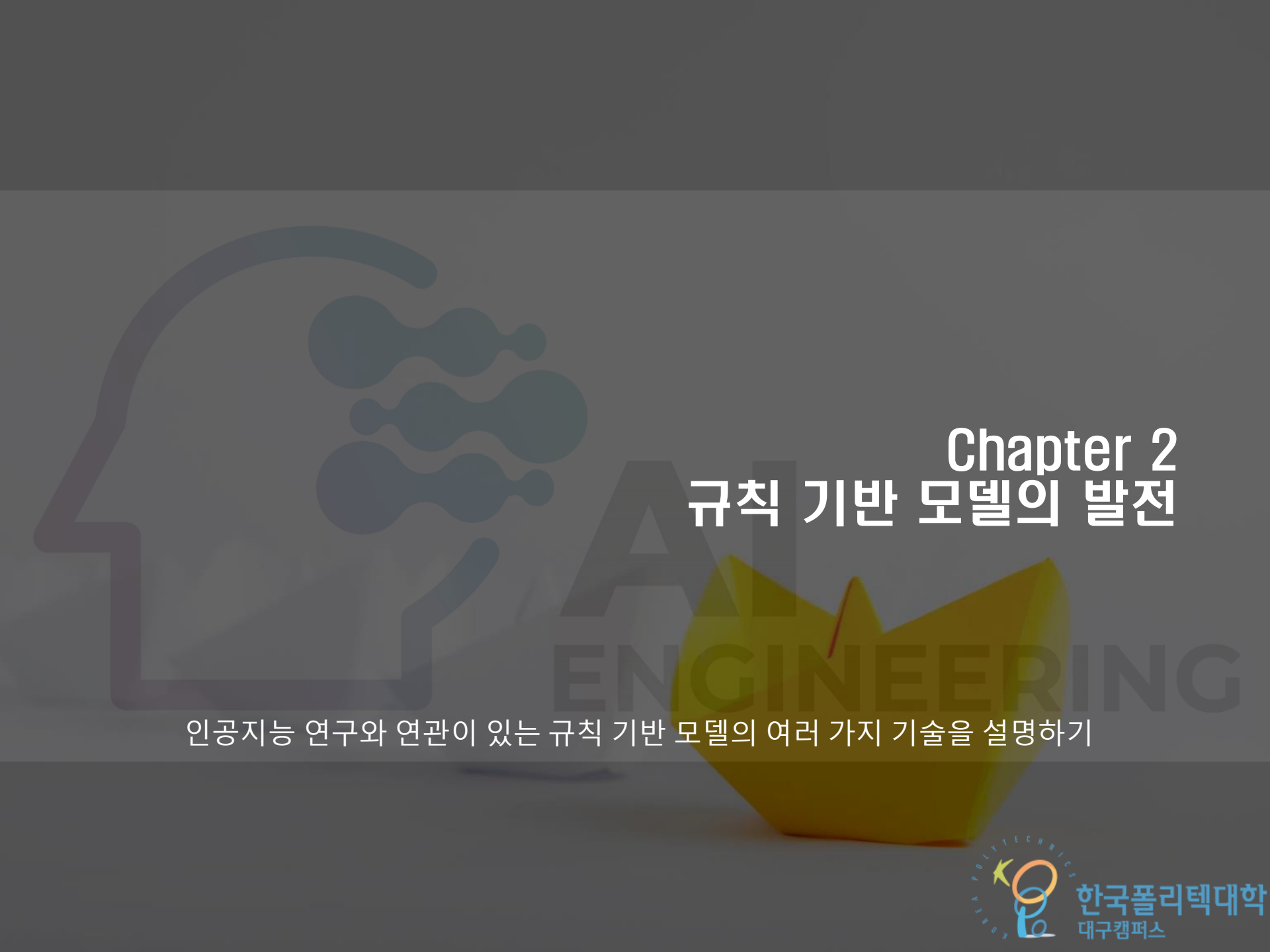
인공지능 개론

한국폴리텍대학 대구캠퍼스
AI엔지니어링학과 강현우



Chapter 2
규칙 기반 모델의 발전

인공지능 연구와 연관이 있는 규칙 기반 모델의 여러 가지 기술을 설명하기

한국폴리텍대학
대구캠퍼스

SECTION 01 규칙 기반 모델

◆1.1조건 분기 프로그램과 규칙 기반 시스템

- 특정 조건을 비교해서 처리할 일을 나누는 것을 조건 분기라고 함.
- 컴퓨터를 이용한 문제 해결은 조건 분기를 구현한 프로그램을 실행해 답을 끌어냄.
- 규칙(조건 설정)을 사용해 조건 분기 프로그램을 실행하는 시스템을 규칙 기반 시스템이라고 함.
- 기반 시스템을 만들기 전 순서도를 이용해 규칙을 설정하면 좋음.

SECTION 01 규칙 기반 모델

그림 2-1 조건 분기를 이용해 '선택'을 하는 사람과 컴퓨터

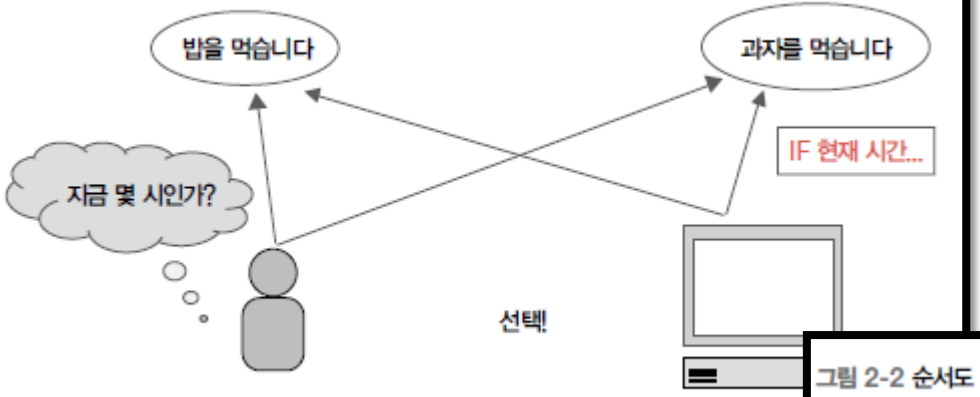
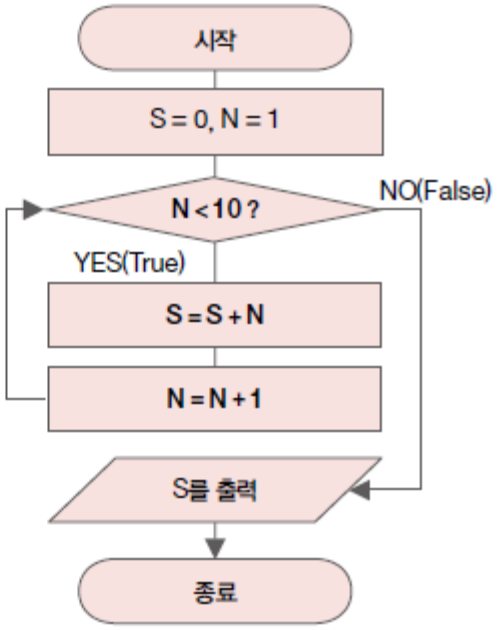


그림 2-2 순서도 예

S와 N의 초깃값을 각각 0과 1로 설정합니다. N이 10 미만이면 S에 N을 더해주고 N을 1만큼 증가시키는 작업을 반복합니다. N이 10 이상이 되면 S를 출력하고 종료합니다. 즉, S에 1에서 9까지 더한 총합을 계산해서 출력하는 순서도입니다



SECTION 01 규칙 기반 모델

◆1.1규칙 설계와 문제의 공식화

➢조건 분기의 기반이 되는 규칙

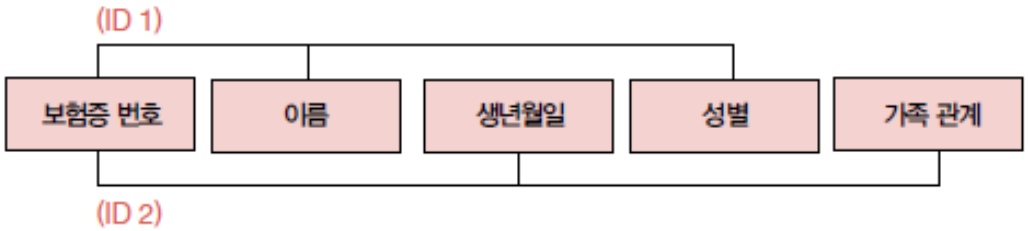
✓사람이 미리 결정해 두어야 함.

➢ 문제의 공식화

✓규칙을 설계해 나가는 단계에서 문제와 해법을 명확히 하는 것

그림 2-3 의료비 청구서의 사용자 정보 수집 작업

- ID 1: 보험증 번호, 이름, 성별 등을 조합한 정보의 해시값(해시 변수를 통해 얻은 값). 병원에서 발행하는 의료비 청구서와 조제약 청구서를 통해 얻습니다.
- ID 2: 보험증 번호, 생년월일, 가족 관계 등을 조합한 정보의 해시값. 의료 관계 데이터베이스에서 정보를 조회해서 얻습니다.
- ID 1과 ID 2의 정보를 서로 비교하면 ID 1만으로는 구분하기 어려웠던 정보(이름이 잘못 표기되었다 등의 이유)들을 판별할 수 있습니다. 그 결과 ID를 통해 비교하는 정보의 정확도를 높일 수 있게 되었습니다. 단, ID나 규칙을 늘리는 것은 수작업에 의존합니다.

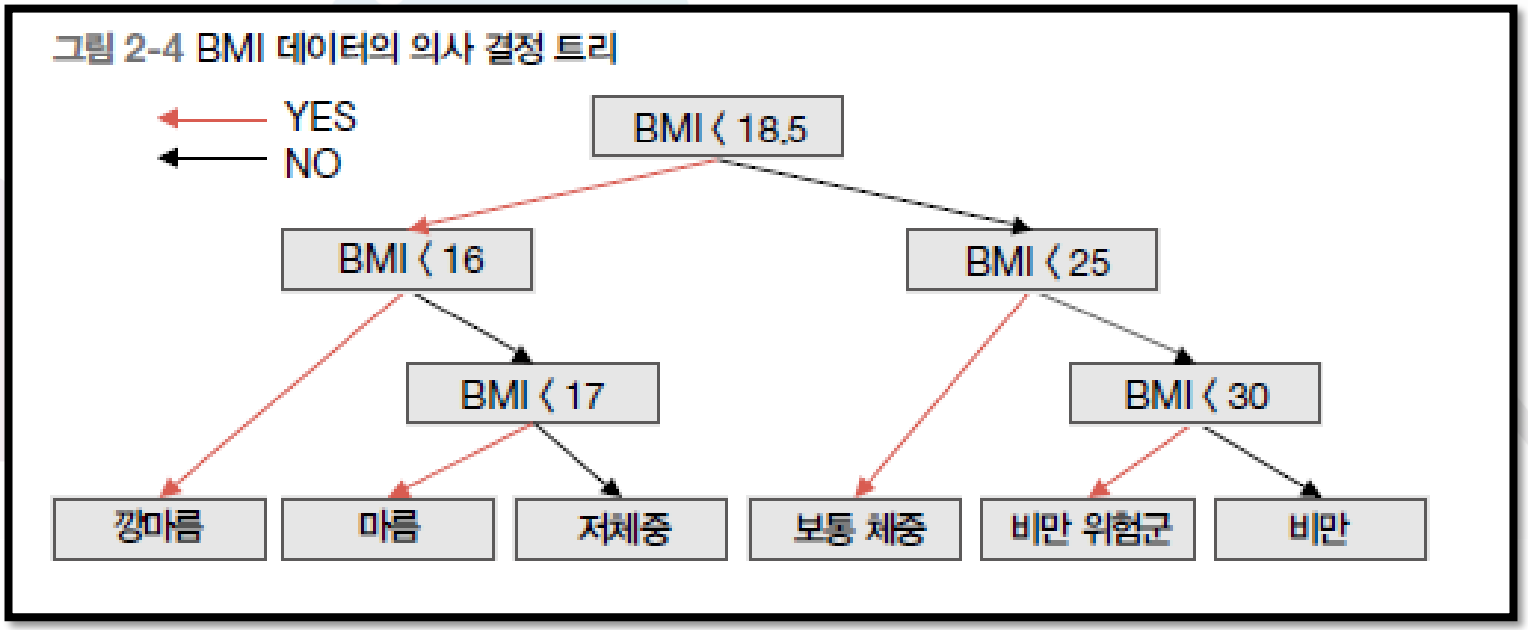


SECTION 01 규칙 기반 모델

◆1.1의사 결정 트리의 구축

➢규칙을 바탕으로 그린 순서도로 구축한 이진 트리를 의사 결정 트리라고도 함

그림 2-4 BMI 데이터의 의사 결정 트리

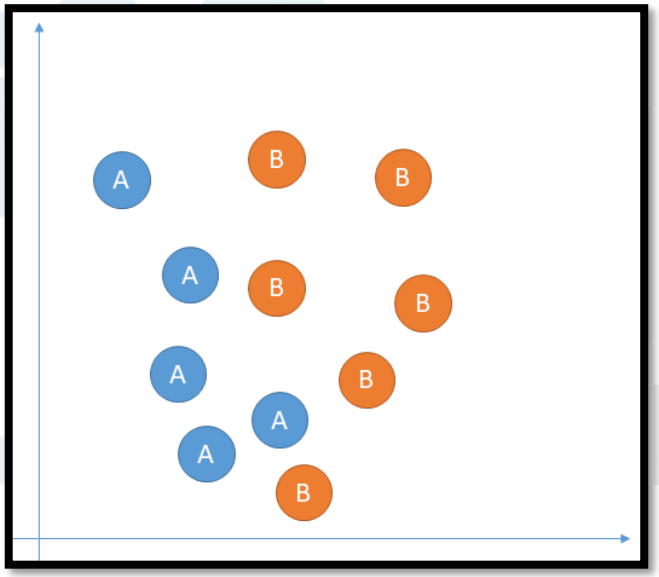


한국폴리텍대학
대구캠퍼스

Decision Tree

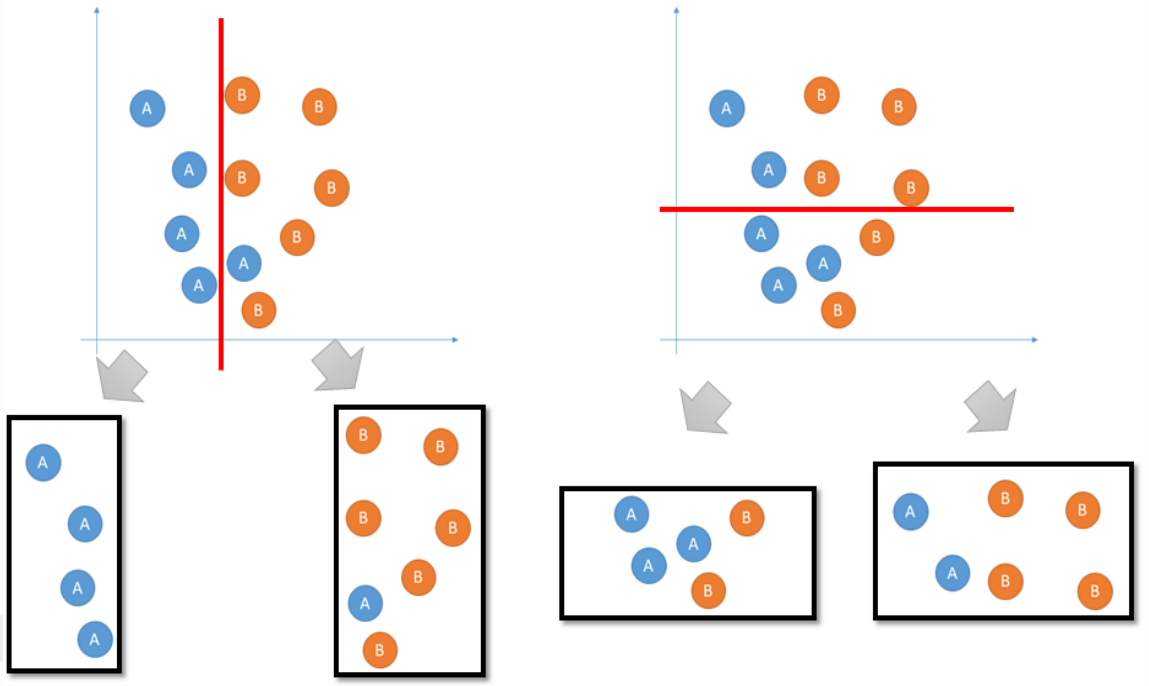
◆ Decision Tree 활용 예

- ➢ 구분 대상 Class 는 2개로 한정
- ➢ feature 도 2개로 한정하고 진행



Decision Tree

◆트리 분류 규칙



◆ 어느 것이 더 잘 분류한 것인가?

Decision Tree

◆ Entropy

- 불순도를 수치로 나타냄
- 엔트로피가 높다 = 불순도가 높다

- $\text{Entropy} = \sum_i (P_i) \log_2 (P_i)$

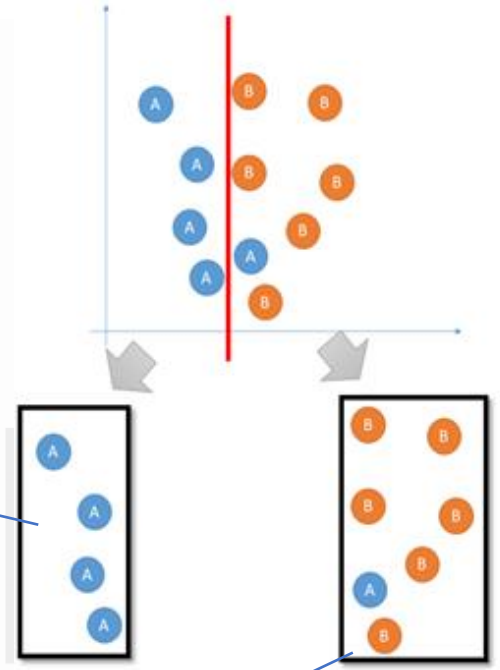
- P_i = 범주 i에 속하는 데이터의 비율

Entropy 계산

$$-\{(P_a)\log_2(P_a) + (P_b)\log_2(P_b)\}$$

좌: $(P_a = 4/4)\ (P_b = 0/4)$

Entropy(좌) = 0



우: $(P_a = 1/7)\ (P_b = 6/7)$

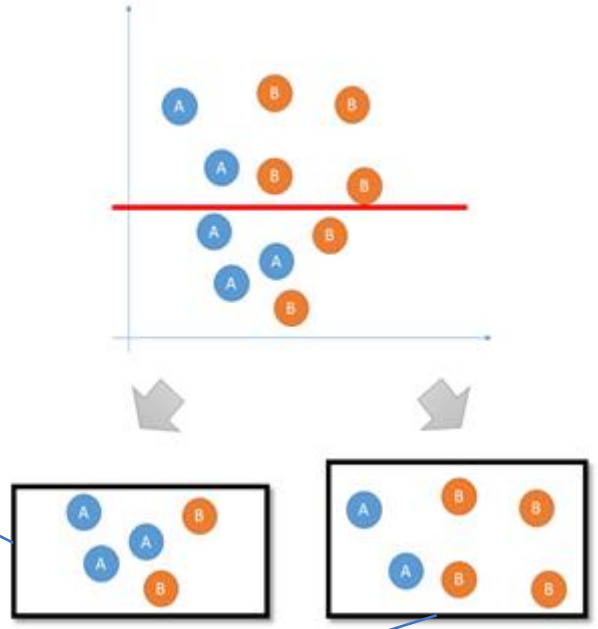
Entropy(우) = 0.59

Entropy 계산

$$-\{(P_a)\log_2(P_a) + (P_b)\log_2(P_b)\}$$

좌: $(P_a = 3/5)\ (P_b = 2/5)$

Entropy(좌)= 0.97



우: $(P_a = 2/6)\ (P_b = 4/6)$

Entropy(우) = 0.92

Information Gain

◆Information Gain =
- ➢ 현재 불순도 – 분류된 두 노드의 불순도의 합

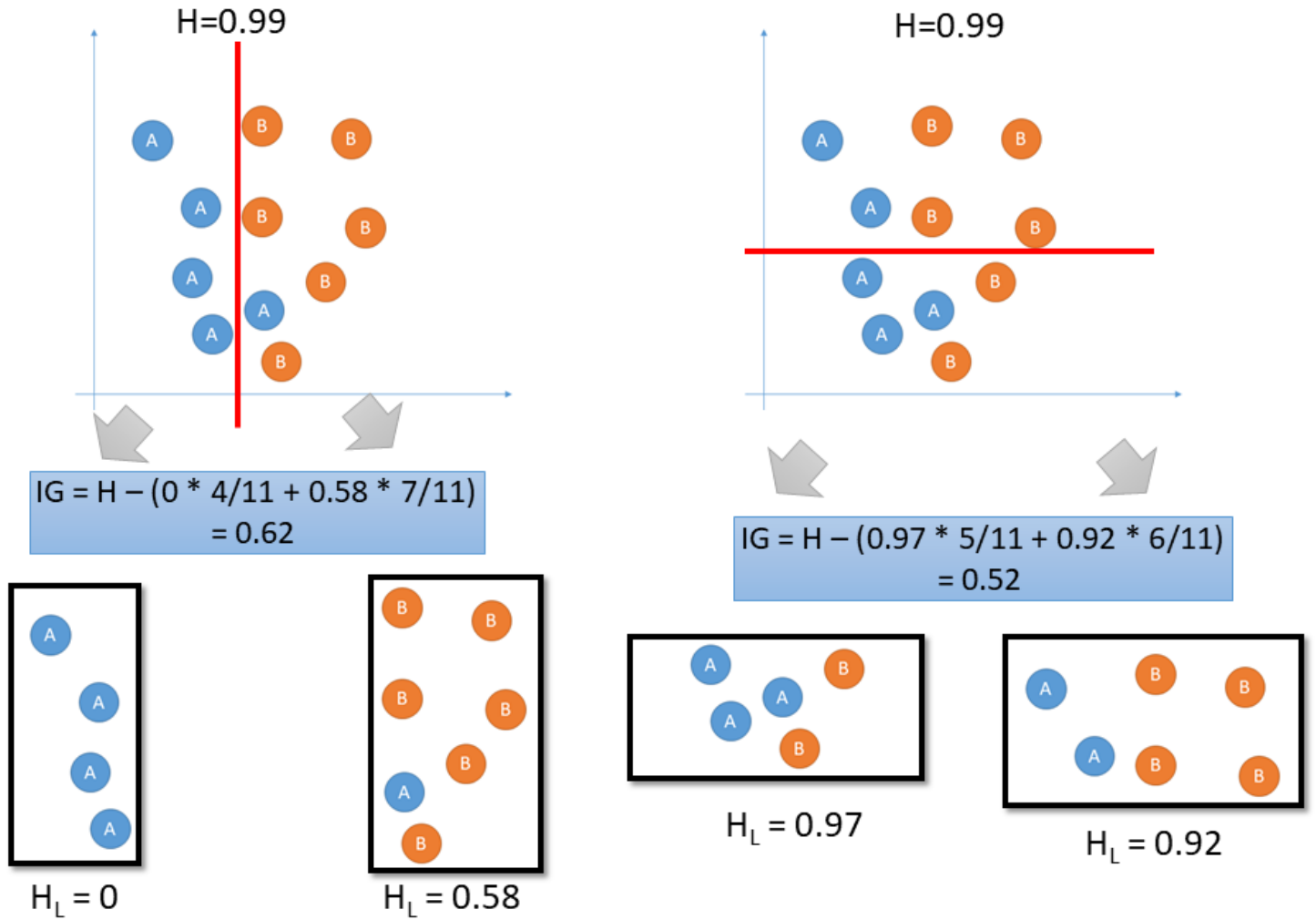
◆현재의 불순도를 H(x) 라고 하면

◆ Information Gain IG 는
- ➢ $IG = H - (H_L \times P_L + H_R \times P_R)$

◆ 즉, IG가 크면 클 수록 얻을 수 있는 정보가 많다는 의미가 됨.

◆ 따라서 IG가 최대가 되도록 분류!

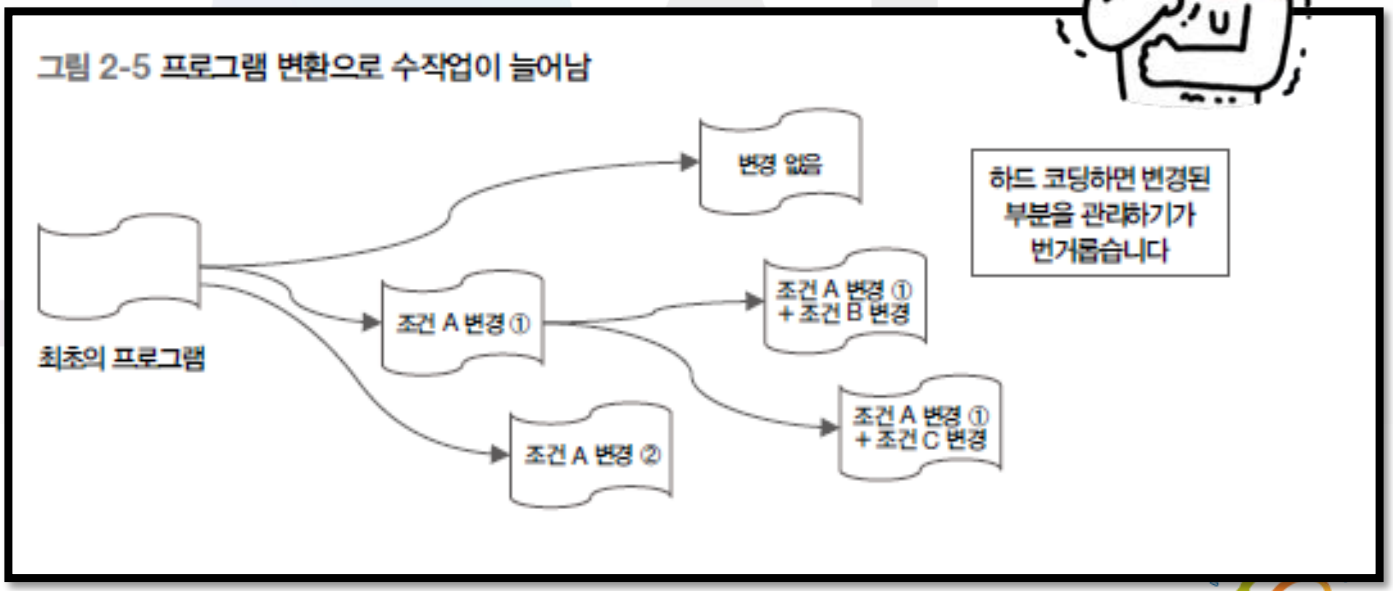
Information Gain



SECTION 02 지식 기반 모델

◆2.1 규칙이 늘거나 변하는 경우

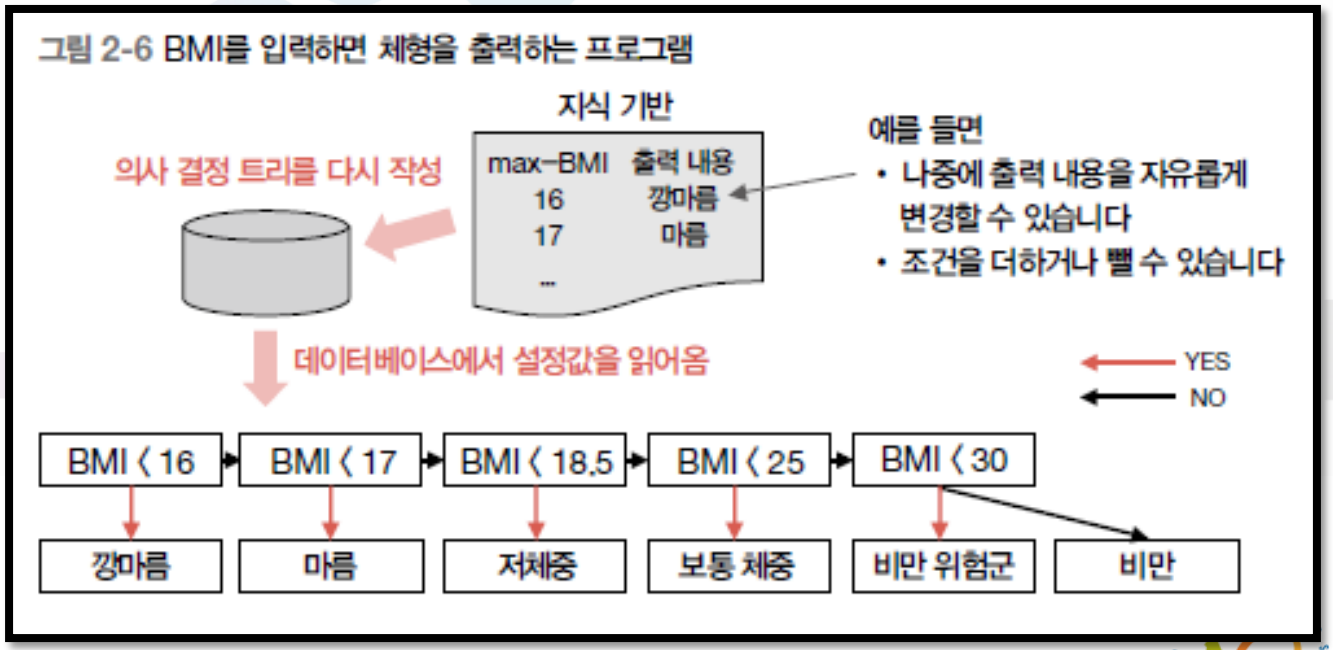
➢규칙이 바뀌면 프로그램을 다시?!



SECTION 02 지식 기반 모델

◆2.2 사람도 검색할 수 있는 지식 기반 시스템

- ➢ 방대한 정보 데이터를 저장
- ➢ 사람도 검색이 가능
- ➢ UniProtKB



SECTION 03 전문가 시스템

◆3.1 전문가 시스템

- ➢전문가 시스템은 규칙 기반 모델을 이용하는 추론 엔진에 기반
- ➢ 초기 전문가 시스템 Dendral

- ➢Dendral에서 파생된 MYCIN
- ➢환자의 전염성 혈액질환을 진단한 후
- ➢투약해야 하는 항생제, 투약량 등을 제시

SECTION 03 전문가 시스템

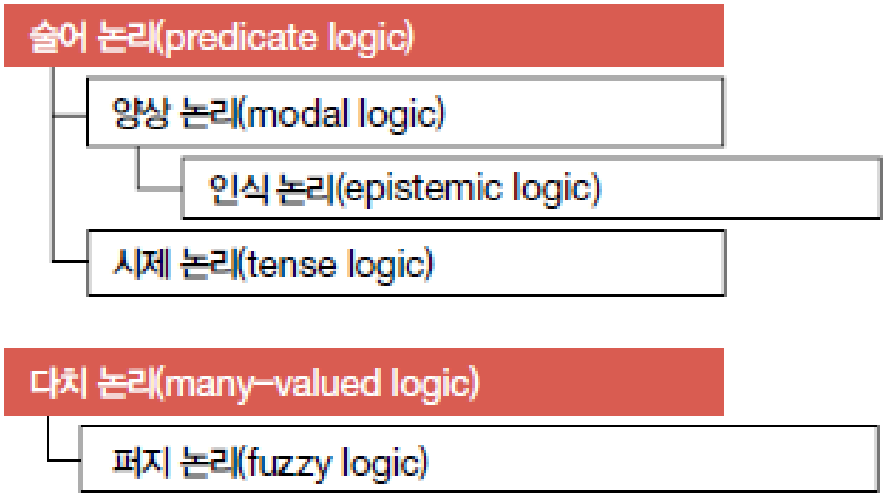
◆3.2 추론 엔진의 종류와 기법

➢추론엔진

✓전문가시스템이 규칙을 사용해 결과를 추론하는 프로그램

✓주로 수리논리학을 이용함

그림 2-9 추론 엔진 논리 구성 예



SECTION 03 전문가 시스템

표 2-1 명제 논리의 기호 종류

항목	내용
논리식	원자 논리식 혹은 원자 논리식과 명제 결합 기호의 조합으로 표현
원자 논리식(원자식)	명제 변수로 표현
명제 변수	P, Q, p, q, Φ, Ψ 등
명제 결합 기호(결합 기호)	¬, ~(부정, NOT), ∧(연언, 논리곱 AND), ∨(선언, 논리합, OR), ⇒(함축, implication), ⇔(동치, equivalence), NOT과 OR 이외의 기호는 NOT과 OR로 풀어서 표현 가능
보조 기호	( )는 기호법에 따라 없는 경우도 있음
논리적 동치	≡는 2개의 논리식이 같은 값인 경우를 표시

표 2-2 술어 논리의 기호 종류

항목	내용
논리식	원자 논리식 혹은 원자 논리식과 논리 기호의 조합으로 표현
원자 논리식(원자식)	원자 논리식 혹은 원자 논리식과 항의 조합으로 표현
항	정수 기호, 변수 기호, 함수 기호의 조합으로 표현
정수 기호	TRUE, FALSE, X, Y, apple, Tommy 등
변수 기호	P, Q, p, q, Φ, Ψ 등
함수 기호	FATHER( ) 등 관계를 표시
술어 기호	cold( ) 등 성질과 상태를 표시
논리 기호	명제 결합 기호와 한정 기호로 표현
한정 기호	∀(전칭 기호), ∃(존재 기호)

표 2-3 술어 논리식의 예

술어 논리식	의미
MOTHER(Tom)	Tom의 어머니
cold(x)	x가 차갑다.
∃x(have(I, x)∧book(x))	나는 책이 있다.
∀x(girl(x) ⇒ ∃y(loves(x, y)∧cake(y)))	모든 여자는 케이크를 좋아한다.
¬∃x(human(x)∧touch(x, BACK(x)))	아무도 자신의 등을 만지지 않는다.

SECTION 04 추천 엔진

◆4.1 추천 엔진의 개념

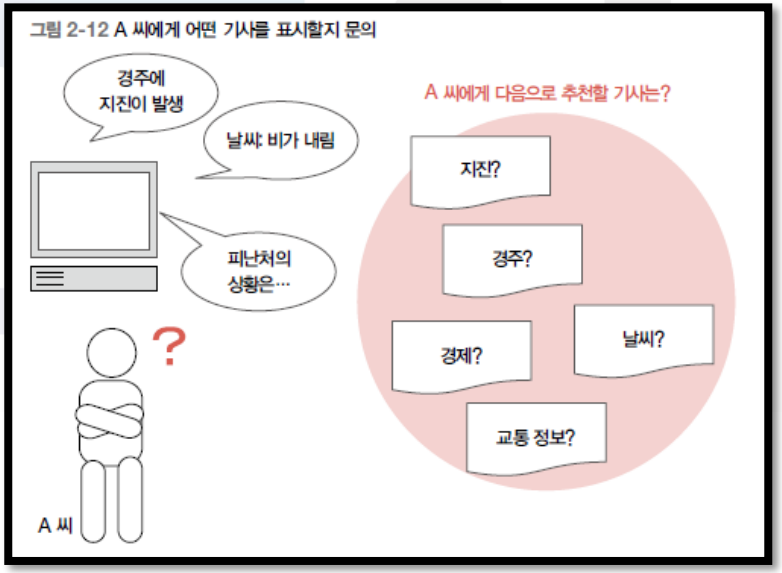
- ➢널리 사용하는 전문가 시스템의 예
- ➢추천 엔진은 쇼핑몰 등의 사이트 방문자에게
- ➢비슷한 정보를 추천하는 시스템임

- ➢ **추천 엔진은 크게 두 가지 방법으로 실행함.**
  - ✓콘텐츠 내용에서 비슷한 정보를 찾아 정보를 추천
  - ✓문자의 검색 이력, 구매 이력 등
  - ✓사이트 방문자 고유 정보를 이용해 연관된 정보를 추천

SECTION 04 추천 엔진

◆4.2 콘텐츠 내용을 분석하는 추천 엔진

- **방문자 정보를 제외한 콘텐츠 자체의 정보**
  - ✓쇼핑몰 사이트라면 상품 정보
  - ✓뉴스 사이트라면 기사 정보
- **에서 관련 내용을 찾아 추천**



SECTION 04 추천 엔진

◆4.3 협업 필터링을 이용하는 추천 엔진

- ➢ 사이트 방문자와 상품의 구매 기록으로
- ➢ 상관계수를 계산

그림 2-13 상관 계수 계산

X 씨의 구매 기록 $\{x_1, x_2, x_3, x_4, x_5\} = \{1, 0, 0, 0, 1\}$

A 씨의 구매 기록 $\{y_1, y_2, y_3, y_4, y_5\} = \{1, 1, 0, 0, 0\}$

상관 계수 $$r = \frac{\sum_{i=1}^{5}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum_{i=1}^{5}(x_i - \bar{x})^2\right)\left(\sum_{i=1}^{5}(y_i - \bar{y})^2\right)}}$$

$$= \frac{\sum_{i=1}^{5}(x_i - 0.4)(y_i - 0.4)}{\sqrt{\left(\sum_{i=1}^{5}(x_i - 0.4)^2\right)\left(\sum_{i=1}^{5}(y_i - 0.4)^2\right)}}$$

$$= \frac{0.6 \times 0.6 + (-0.4) \times 0.6 + 0.4 \times 0.4 + 0.4 \times 0.4 + 0.6 \times (-0.4)}{\sqrt{(0.6^2 + 0.4^2 + 0.4^2 + 0.4^2 + 0.6^2)(0.6^2 + 0.6^2 + 0.4^2 + 0.4^2 + 0.4^2)}}$$

$$= \frac{0.36 + 0.16 \times 2 - 0.24 \times 2}{0.36 \times 2 + 0.16 \times 3} = \frac{0.2}{1.2} \cong 0.167$$

SECTION 04 추천 엔진

표 2-9 사이트 방문자, 각 상품의 구매 기록, 상관 계수 표

		상품										X 씨와의 상관 계수
		1	2	3	4	5	6	7	8	9	10	
방문자	X	–	1	0	–	–	–	–	0	0	1	1.000
	A	1	1	1	–	–	–	–	0	0	0	0.167
	B	–	–	–	0	0	0	1	1	1	0	−1.000
	C	0	1	0	0	–	1	1	0	0	1	1.000
	D	0	–	–	0	1	1	0	0	1	1	0.500
	E	–	1	0	–	1	0	–	0	0	0	0.612
추천 정도												

표 2-10 사이트 방문자, 각 상품의 구매 기록, 상관 계수, 추천 정도 표

		상품										상관 계수
		1	2	3	4	5	6	7	8	9	10	
방문자	X	–	1	0	–	–	–	–	0	0	1	1.000
	A	1	1	1	–	–	–	–	0	0	0	0.167
	B	–	–	–	0	0	0	1	1	1	0	−1.000
	C	0	1	0	0	–	1	1	0	0	1	1.000
	D	0	–	–	0	1	1	0	0	1	1	0.500
	E	–	1	0	–	1	0	–	0	0	0	0.612
추천 정도		0.00			0.00	1.00	0.67	0.50				

Contents Based 방식의 단점

- **DB는 어떻게 구성할 것인가?**
  - 제품을 카테고리화 하는 것 부터
  - 비슷한 색상, 디자인의 제품이 있는지
  - 재고는 있는지 어디에 있는지
  - 이걸 어떻게 다 만드는가? → ~~사람이…~~

- **대용량의 DB를 구축했다고 치고…**
  - 비용은 엄청나게 들었겠지만
  - 상품이 수 천만/ 수억 개라면 검색은 어떻게?
    - ✓ Hadoop, Map Reduce 이런 기술이 결국 검색을 빠르게 하기 위한 것

Collaborative Filter

◆ 협업 필터는 그럼 만능?

- ➢ **좋긴 좋은데… CB도 잘 못하고 있는걸?**
- ➢ **수학 모델 (통계 모델) 을 잘 만들어야 된다.**
  - ✓ 그런 사람이 잘 있나?
- ➢ **쉽게 말해서 CB 보다도 더 DB 설계가 어렵다.**
  - ✓ 우리 나라에 사용하는 회사가 없다는 이야기도 들었다.
- ➢ **우리 예제에 나온 상관도는 시간 순서가 없지만**
  - ✓ 실제로는 시간 순서도 고려한 모델을 짠다.

Latent Matrix Factorization

- 그 유명한 Netflix 의 추천 시스템
  - Kaggle에서 받았다는 건 안 비밀

- Latent Matrix Factorization??
  - 잠재 행렬 인수 분해법
  - 사용자들이 아이템에 점수를 등록해둔 것을
  - 2개의 메트릭스로 분해한다.
    - ✓ 잠재 요인을 분해해낸다. 정도. ~~(나도 잘 모르겠다)~~

- 그래서…
  - 좋아 보이지만, 넷플릭스니까 가능하다.

Summary

- **규칙 기반 모델**
  - 규칙 기반 모델은 전문가 시스템을 거쳐
  - 추천 엔진으로 발전하였다

- **협업 필터링 – 추천 시스템을 더 알고 싶다면**
  - Collaborative Filtering – 추천시스템의 핵심기술
  - http://www.oss.kr/oss_repository14/658203

  - 후에 Latent Factorization 으로 발전 (Netflix)
  - Introduction to Latent Matrix Factorization Recommender Systems | by Tumas Rackaitis | Towards Data Science