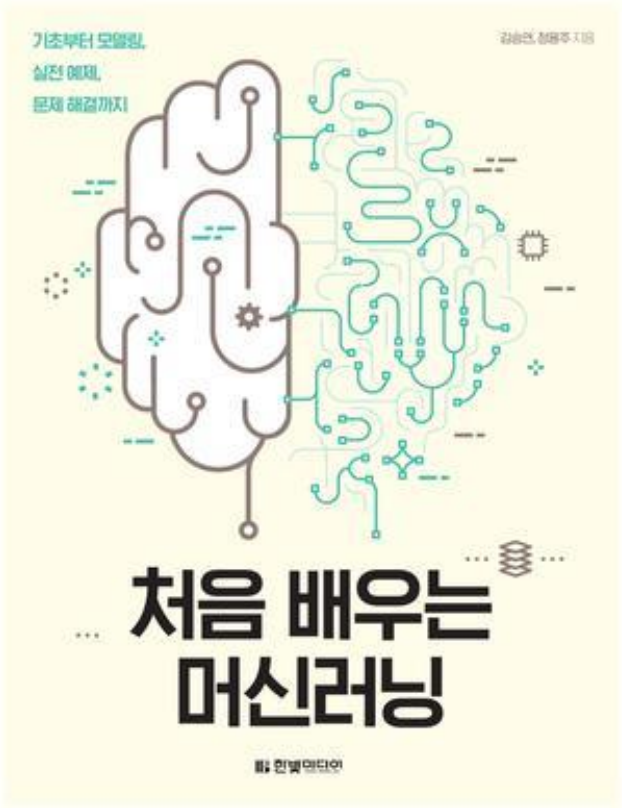
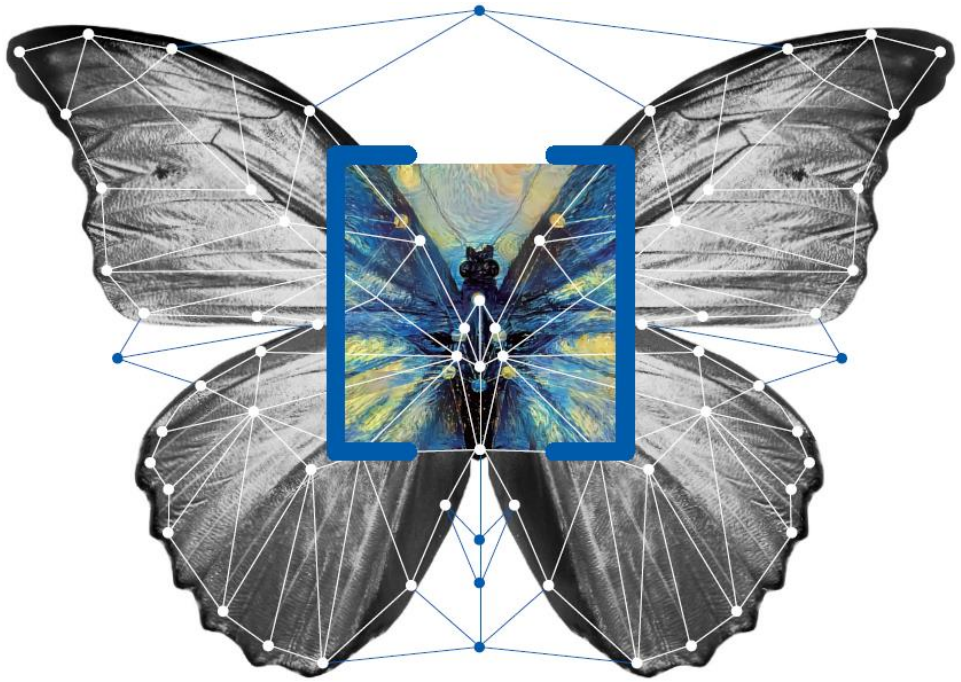




MACHINE LEARNING 기계 학습

분류 문제

PREVIEW

- 분류
  - 회귀만큼이나 기본적인 머신러닝 문제
- 세상에는 참으로 많은 데이터가 있다.
  - 계량 데이터
    - 점수, 매출액, GDP, BOD, 속도, 마찰계수, 토끼 개체수 등
    - 거리 개념 있다. 5는 31보다 크다. 5는 10보다 7에 가깝다.
  - 비계량 데이터
    - 직업, 행정 구역, 혈액형, 성씨, PC 브랜드 등
    - 거리 개념 없다. 'O형은 B형보다 A형에 가깝다'는 성립 안한다.
- 분류 방법
  - 의사 결정 트리
  - KNN
  - 나이브 베이지안(Naïve Bayes) 분류기

Phase I: ClassifierConstruction

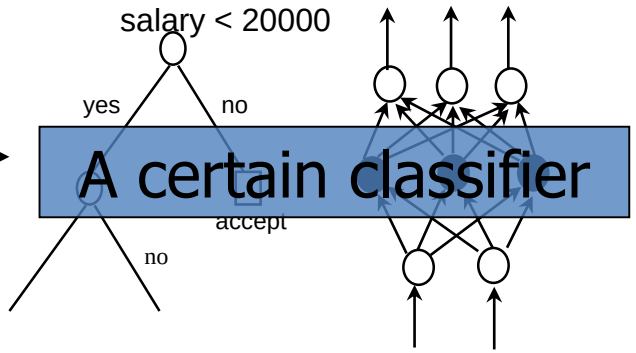
Database (training set)

salary	education	label
10000	high school	reject
40000	under graduate	accept
15000	under graduate	reject
75000	graduate	accept
18000	graduate	accept

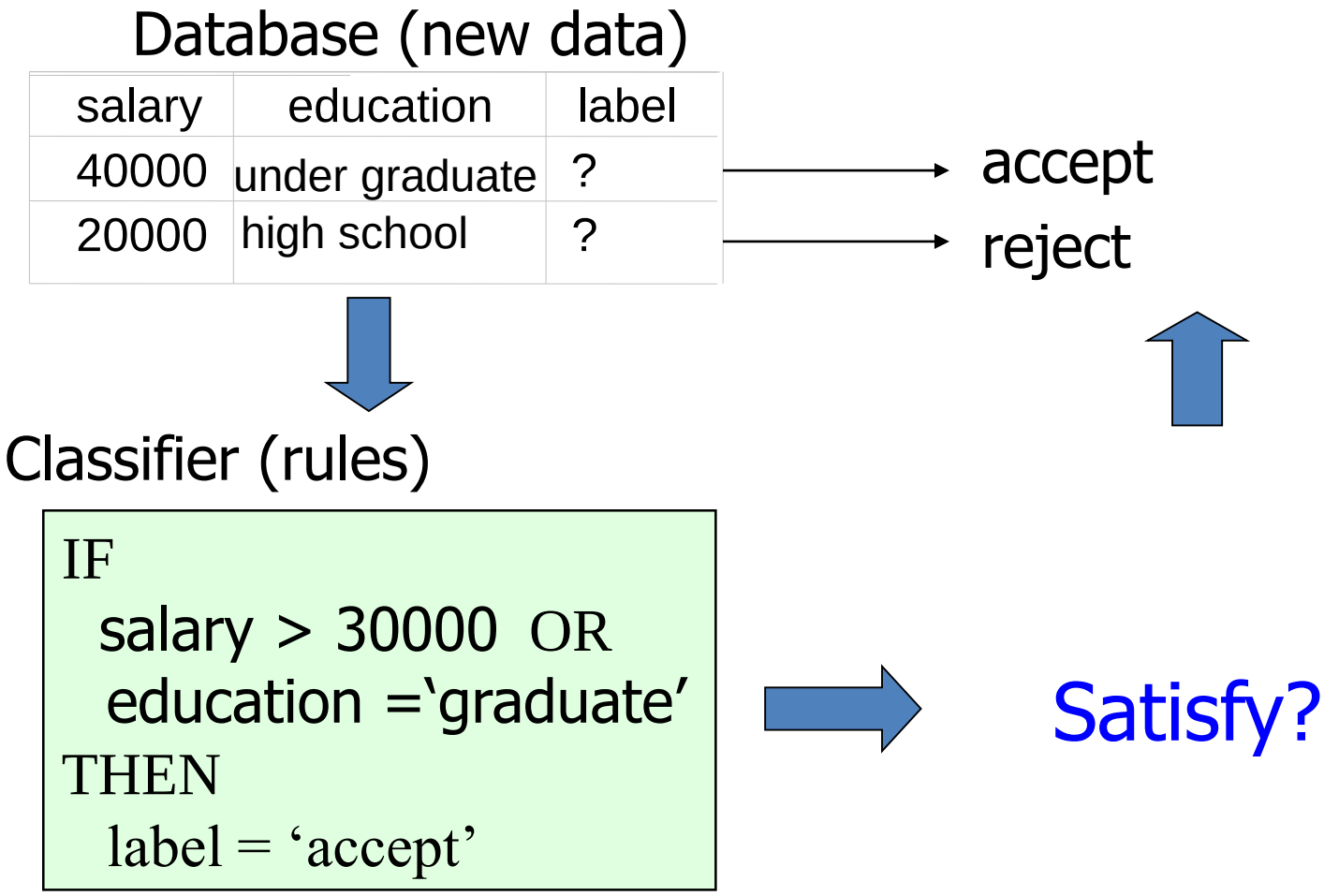
IF
salary > 30000 OR
education ='graduate'
THEN
label = 'accept'

Classification Algorithms

create

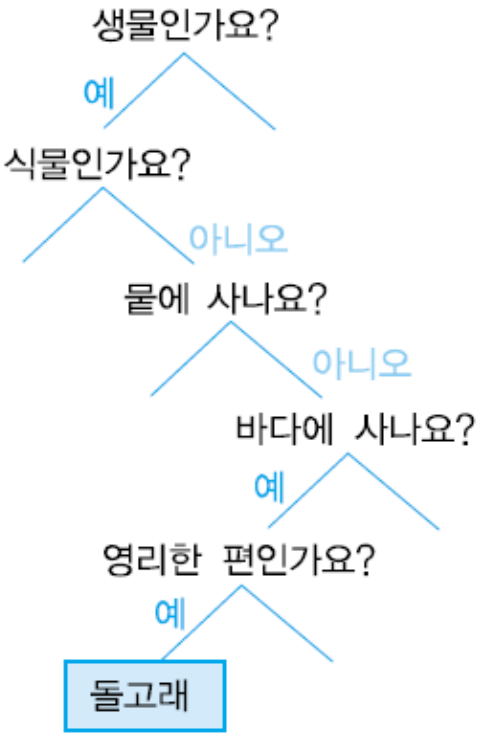


Phase II: Prediction



결정 트리

- 결정 트리의 원리
  - 스무고개와 개념이 비슷
  - 최적 기준에 따라 자동으로 질문을 만들어야 함

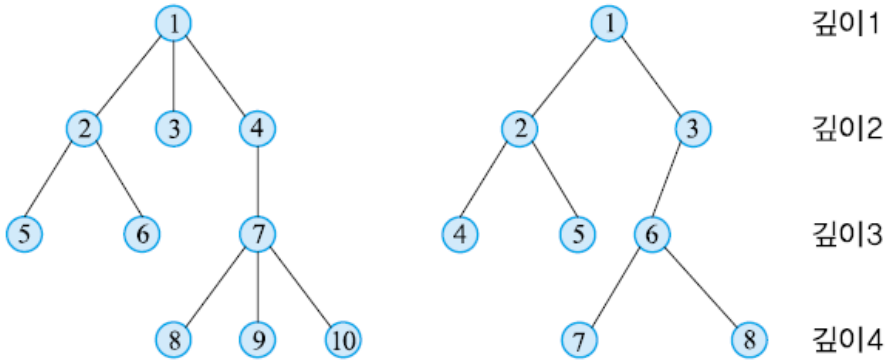


스무고개 놀이

- 몇 가지 고려 사항
  1. 노드에서 몇 개의 가지로 나눌 것인가?
  2. 각 노드의 질문을 어떻게 만들 것인가?
  3. 언제 멈출 것인가?
  4. 잎 노드를 어느 부류에 할당할 것인가?

결정 트리의 원리

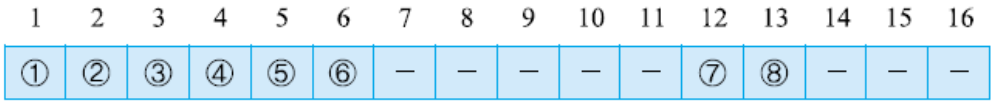
- 결정 트리의 표현
  - 트리 또는 이진 트리 사용



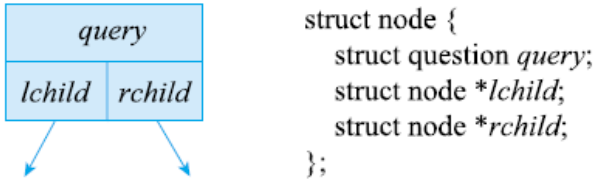
(a) 트리 (b) 이진 트리

트리와 이진 트리

- 이진 트리의 구현



(a) 1 차원 배열 표현



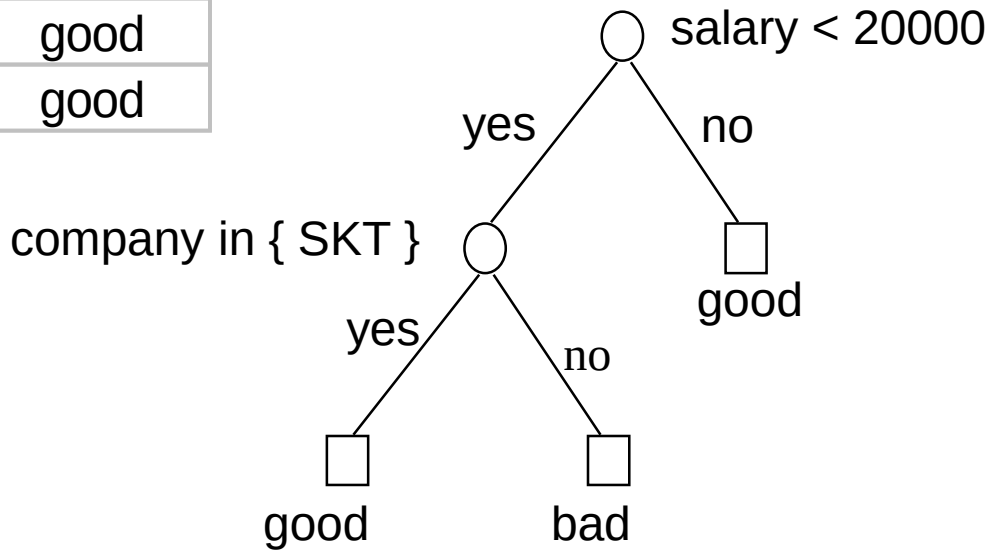
(b) 연결 리스트 표현

이진 트리 표현 방법

결정 트리의 예

salary	company	label
10000	KTF	bad
40000	LGT	good
15000	LGT	bad
75000	SKT	good
18000	SKT	good

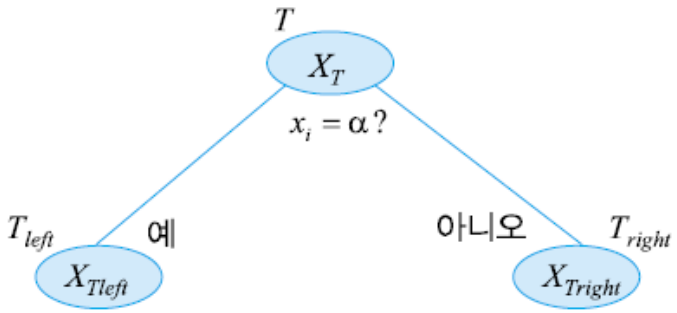
Credit Analysis



결정 트리에서의 노드 분기

- 결정 트리의 노드
  - 노드의 분기

$$\left.\begin{aligned} X_{T\,left} \cup X_{T\,right} &= X_T \\ X_{T\,left} \cap X_{T\,right} &= \varnothing \end{aligned}\right\}$$



노드의 분기

  - 질문 x_j=α? 어떻게 만들 것인가?
    - *d* 개의 특징이 있고 그들이 평균 *n* 개의 값을 가진다면 *dn* 개의 후보 질문
    - 그들 중 어느 것을 취해야 가장 유리한가?

노드에서의 질문

- 유리한 정도의 판단 기준은?
  - X_{Tleft}와 X_{Tright}가 동질일 수록 좋다.

- 불순도 측정 기준
  - 엔트로피

$$im(T) = -\sum_{i=1}^{M} P(\omega_i \mid T) \log_2 P(\omega_i \mid T)$$

  - 지니 불순도

$$im(T) = 1 - \sum_{i=1}^{M} P(\omega_i \mid T)^2 = \sum_{i \neq j} P(\omega_i \mid T) P(\omega_j \mid T)$$

  - 오분류 불순도

$$im(T) = 1 - \max_i P(\omega_i \mid T)$$

  - 노드 T에서 ω_i가 발생할 확률은

$$P(\omega_i \mid T) = \frac{X_T\text{에서 } \omega_i\text{에 속한 샘플의 수}}{|X_T|}$$

불순도 측정

- 불순도 측정

노드 T의 샘플 집합 X_T가 아래와 같다고 하자.

$$X_T = \{(\mathbf{x}_1,\omega_2),\ (\mathbf{x}_2,\omega_1),\ (\mathbf{x}_3,\omega_3),\ (\mathbf{x}_4,\omega_2),\ (\mathbf{x}_5,\omega_2),\ (\mathbf{x}_6,\omega_2),\ (\mathbf{x}_7,\omega_1),\ (\mathbf{x}_8,\omega_3),\ (\mathbf{x}_9,\omega_1)\}$$

$$P(\omega_1 \mid T) = 3/9,\ P(\omega_2 \mid T) = 4/9,\ P(\omega_3 \mid T) = 2/9$$

엔트로피 불순도: $im(T) = -(\frac{3}{9}\log_2\frac{3}{9} + \frac{4}{9}\log_2\frac{4}{9} + \frac{2}{9}\log_2\frac{2}{9}) = 1.5305$

지니 불순도: $im(T) = 1 - (\frac{3^2}{9^2} + \frac{4^2}{9^2} + \frac{2^2}{9^2}) = 0.642$

오분류 불순도: $im(T) = 1 - \frac{4}{9} = 0.556$

엔트로피 예제

$$\text{entropy}(T) = -\Sigma p_j \times \log_2(p_j)$$

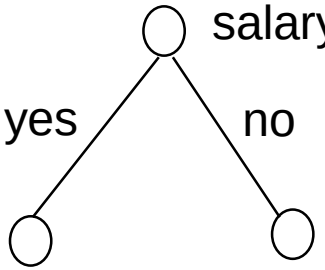
salary	company	label
10000	KTF	reject
40000	LGT	accept
15000	LGT	reject
75000	SKT	accept
18000	SKT	accept

$$Pobablity(class = "reject") = \frac{2}{5}$$

$$Pobablity(class = "accept") = \frac{3}{5}$$

$$Entropy(S) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$$

$$= 0.970951$$



$$Entropy(S_{left}) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$$

$$= 0.918296$$

$$Entropy(S_{right}) = 0$$

salary	company	label
10000	LGT	reject
15000	LGT	reject
18000	SKT	accept

salary	company	label
40000	LGT	accept
75000	SKT	accept

$$E_{split}(S) = \frac{N_{left}}{N}E(S_{left}) + \frac{N_{right}}{N}E(S_{right}) \longrightarrow Entropy_{split}(S) = \frac{3}{5}\times 0.918296 + \frac{2}{5}\times 0 = 0.550978$$

분기에 필요한 최적의 질문

- 불순도를 하나의 노드 대상으로 측정 (후보 질문에 대한 평가)
  - 분기 결과로 만들어지는 새로운 샘플 X_{treft}와 X_{tright}는 가급적 불순도가 낮은 것이 좋다
  - 예) 왼쪽과 오른쪽 자식 노드 모두 한 부류의 샘플만을 가지게 되어 불순도가 둘다 0이된다면 제일 좋음.
  - 불순도 감소량 $$\Delta im(T) = im(T) - \frac{|X_{Tleft}|}{|X_T|} im(T_{left}) - \frac{|X_{Tright}|}{|X_T|} im(T_{right})$$
  - 불순도 감소량이 최대값이 되는 질문을 취하면 됨

- 후보 질문을 어떻게 생성
  - 비계량 특징 (한정된 개수의 값)
    - 예) 혈액형의 경우 A, B, O, AB의 네 가지 경우만 존재
  - 계량 특징
    - 몇 개의 이산 값만을 가지면 '$x_i<\alpha$' 식의 질문을 만들면 됨
    - 제일 까다로운 경우: 실수의 경우
      - 실수 범위를 등간격으로 나누어 구간화
      - 샘플들이 갖는 값의 분포를 보고 인접한 두 값의 가운데를 α로 사용

예제: 후보 질문 생성

- 아래와 같은 3개의 특징으로 표현되는 데이터(2개는 비계량, 1개는 계량)

직업 (x_1): [1,7]의 정수 (1 = 디자이너, 2 = 스포츠맨, 3 = 교수, 4 = 의사, 5 = 공무원, 6 = NGO, 7 = 무직)

선호 품목 (x_2): [1,5]의 정수 (1 = 의류, 2 = 전자 제품, 3 = 스포츠 용품, 4 = 책, 5 = 음식)

몸무게 (x_3): 실수

x_1에 의한 후보 질문: x_1=1?, x_1=2?, x_1=3?, x_1=4?, x_1=5?, x_1=6?, x_1=7?

x_2에 의한 후보 질문: x_2=1?, x_2=2?, x_2=3?, x_2=4?, x_2=5?

- X_3의 값의 분포가 다음과 같다면

45.6, 47.8, 50.6, 65.3, 67.8, 72.8, 88.7, 92.3, 102.2

x_3에 의한 후보 질문: x_3<46.7?, x_3<49.2?, x_3<57.95?, x_3<66.55?, x_3<70.3?, x_3<80.75?, x_3<90.5?, x_3<97.25?

예) 계량 데이터+엔트로피

- 5개의 데이터

[Histogram for salary]

		accept	*reject*
[Position 1]	C_{below}	0	1
	C_{above}	3	1

$$Entropy_{split}(S) = \frac{1}{5}\times\frac{1}{1}\log 1 + \frac{4}{5}\times\left(-\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4}\right) = 0.811278$$

[Position 2]	C_{below}	1	1
	C_{above}	2	1

$$Entropy_{split}(S) = 0.950978$$

[Position 3]	C_{below}	1	2
	C_{above}	2	0

$$Entropy_{split}(S) = 0.550978$$

[Position 4]	C_{below}	2	2
	C_{above}	1	0

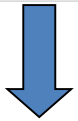
$$Entropy_{split}(S) = 0.8$$

파티션 3이 가장 낮은 엔트로피를 가짐!

예) 비계량 데이터 + 엔트로피

[Attribute List]

company	label	rid
KTF	reject	0
LGT	accept	1
LGT	reject	2
SKT	accept	3
SKT	accept	4



[Histogram for education]

	accept	reject
KTF	0	1
LGT	1	1
SKT	2	0

3 개의 값 → 2^3-2 분기 조건!

{KTF}

$$Entropy_{split}(S) = \frac{1}{5}\times\frac{1}{1}\log 1 + \frac{4}{5}\times\left(-\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4}\right)$$

$$= 0.811278$$

{LGT}

$$Entropy_{split}(S) = 0.950978$$

{SKT}

$$Entropy_{split}(S) = \frac{3}{5}\times 0.918296 + \frac{2}{5}\times 0 = 0.550978$$

{KTF, LGT}

$$Entropy_{split}(S) = \frac{3}{5}\times 0.918296 + \frac{2}{5}\times 0 = 0.550978$$

{LGT, SKT}

$$Entropy_{split}(S) = 0.811278$$

{KTF, SKT}

$$Entropy_{split}(S) = 0.950978$$

{SKT}가 가장 낮은 엔트로피**!**

학습 알고리즘

- 결정 트리 학습 알고리즘
  - 언제 멈출 것인가?
    - 과적합 vs. 설익은 수렴
  - 잎 노드의 부류 할당

알고리즘

입력: 훈련 집합 $X = \{(\mathbf{x}_1, \mathbf{t}_1), \cdots, (\mathbf{x}_N, \mathbf{t}_N)\}$

출력: 결정 트리 R

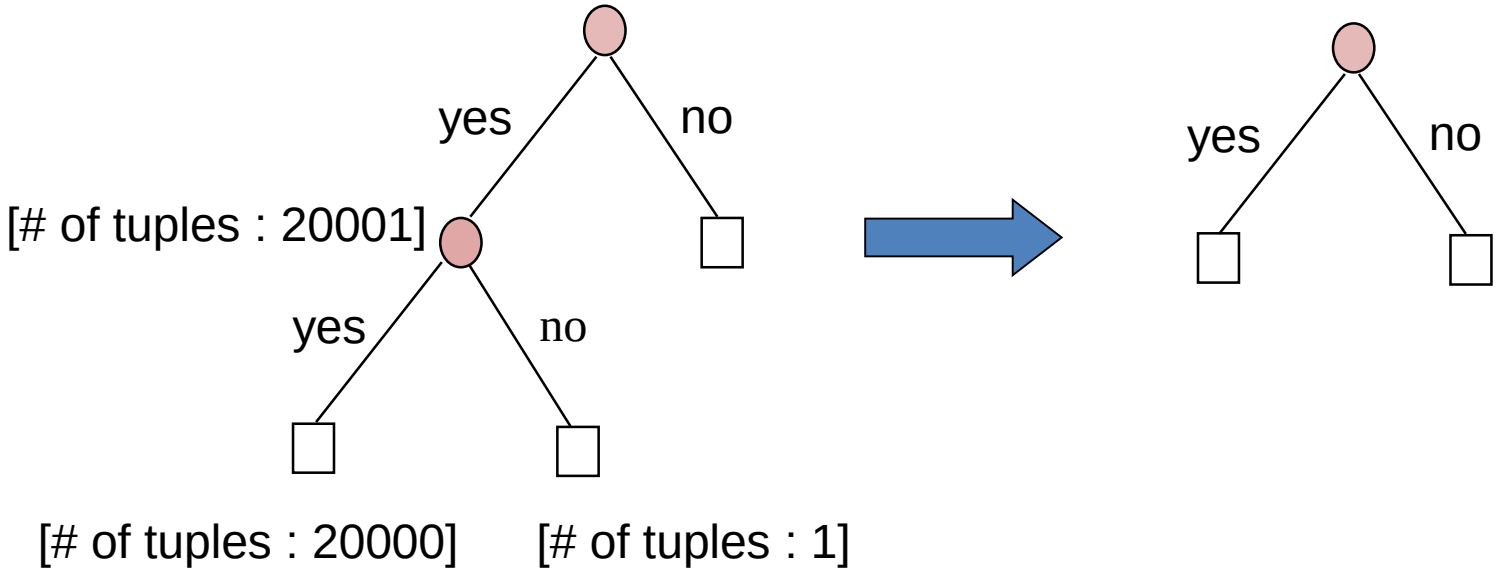
알고리즘:

```
1. 노드 하나를 생성하고 그것을 R이라 한다.  // 이것이 루트 노드이다.
2. T = R;
3. X_T = X;
4. split_node(T, X_T);    // 루트 노드를 시작점으로 하여 순환 함수를 호출한다.

5. split_node(T, X_T) {   // 순환 함수
6.     노드 T에서 후보 질문을 생성한다.
7.     모든 후보 질문의 불순도 감소량을 측정한다.
8.     불순도 감소량이 최대인 질문 q를 선택한다.
9.     if (T가 멈춤 조건을 만족) {
10.        T에 부류를 할당한다.
11.        return;
12.    }
13.    else {
14.        q로 X_T를 X_Tleft와 X_Tright로 나눈다.
15.        새로운 노드 T_left와 T_right를 생성한다.
16.        split_node(T_left, X_Tleft);
17.        split_node(T_right, X_Tright);
18.    }
19. }
```

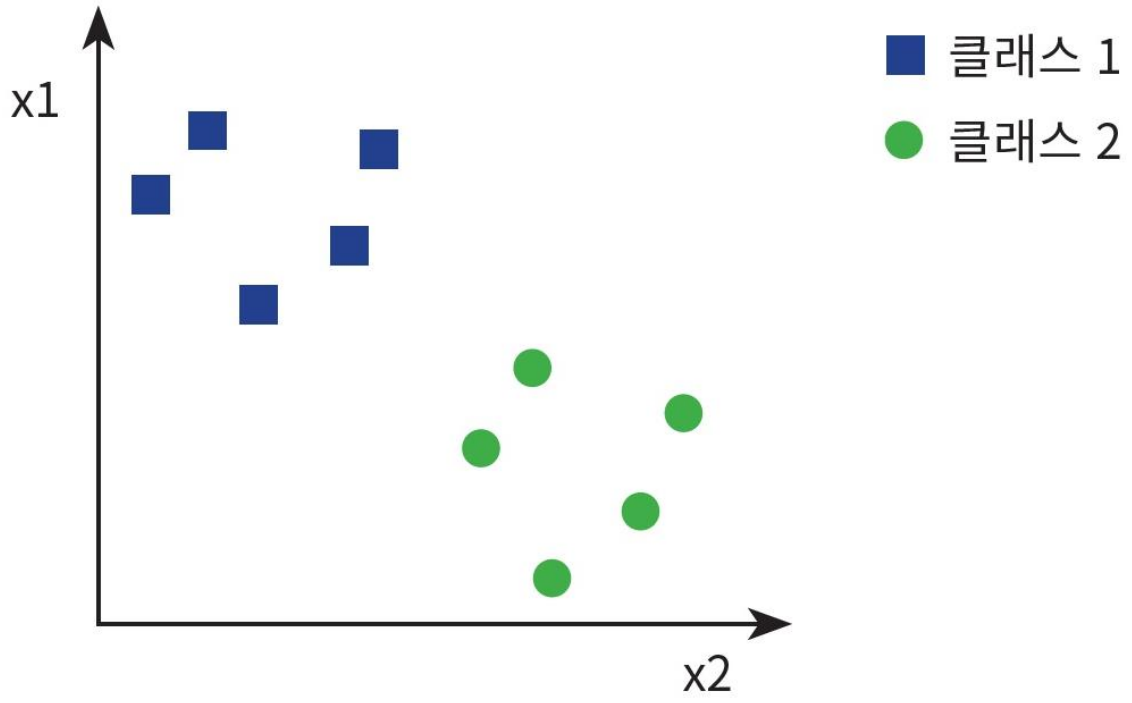
결정 트리에서의 가지치기

- 가지치기
  - 데이터의 과잉적합(overfitting): 잡음과 이상치로 인하여 훈련 데이터의 이상이 반영
  - 사전 가지치기
    - 트리가 생성되는 초기에서 정지 (엔트로피 활용)
  - 사후 가지치기
    - 완성된 트리에서 가지를 제거해 나가는 방법



KNN 알고리즘

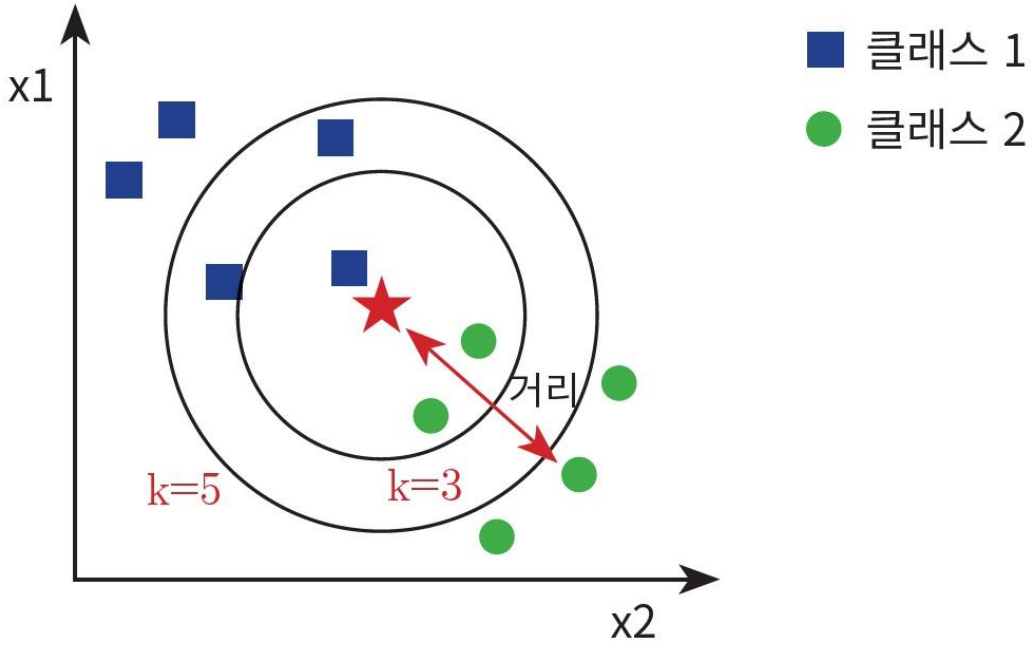
- K-Nearest Neighbor (KNN) 알고리즘은 모든 기계 학습 알고리즘 중에서도 가장 간단하고 이해하기 쉬운 분류 알고리즘



2개의 특징으로 구성된 학습 데이터

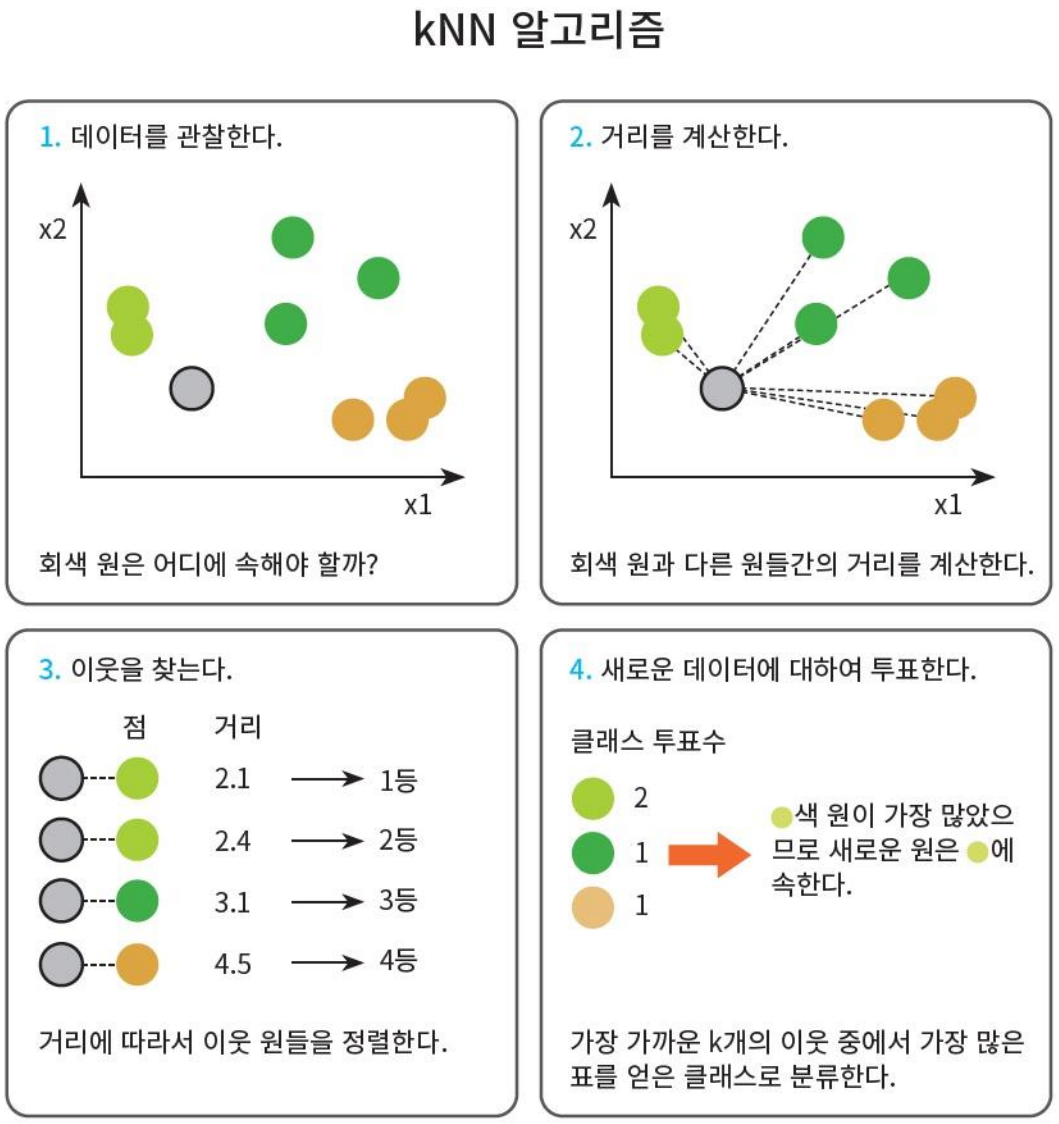
KNN 알고리즘

- 이제 새로운 데이터가 입력되어서 그래프 상에 별표로 표시되었다고 하자. 별표는 파랑색 사각형과 빨강색 원 중에서 하나에 속해야 한다. 이것을 분류(classification)라고 한다.



- KNN 알고리즘은 가장 가까운 이웃에 의존 (가장 가까운 도형은 파란색)
- 가장 가까운 것을 확인하는 것만으로는 충분하지 않음
  - 가장 가까운 K개의 도형을 확인한 다음 그들 중 다수인 쪽으로 선택

KNN 알고리즘


kNN 알고리즘
1. 데이터를 관찰한다.
x2
x1
회색 원은 어디에 속해야 할까?
2. 거리를 계산한다.
x2
x1
회색 원과 다른 원들간의 거리를 계산한다.
3. 이웃을 찾는다.
점
거리
2.1
1등
2.4
2등
3.1
3등
4.5
4등
거리에 따라서 이웃 원들을 정렬한다.
4. 새로운 데이터에 대하여 투표한다.
클래스 투표수
2
1
1
색 원이 가장 많았으므로 새로운 원은 에 속한다.
가장 가까운 k개의 이웃 중에서 가장 많은 표를 얻은 클래스로 분류한다.

KNN 알고리즘의 장점과 단점

- 특징 공간에 있는 모든 데이터에 대한 정보가 필요하다. 왜냐하면, 가장 가까운 이웃을 찾기 위해 새로운 데이터에서 모든 기존 데이터까지의 거리를 확인해야 하기 때문이다. 데이터와 클래스가 많이 있다면, 많은 메모리 공간과 계산 시간이 필요하다.

- 어떤 종류의 학습이나 준비 시간이 필요 없다.

나이브 베이즈 분류

- 통계량의 조건부 확률 사이의 관계를 나타내는 방정식인 베이즈 정리를 기반
  - 관측된 특징(feature)이 주어졌을 때, 레이블(label)의 확률을 구하는 데 관심
  - Naïve: (경험, 지식부족 등으로) 순진해빠진
    - 분류 방법에 나이브가 붙은 이유는 분류를 쉽고 빠르게 하기 위해 특징들이 서로 "확률적으로 독립 " 이라는 가정이 들어갔기 때문
    - 확률적으로 독립이라는 가정에 위반되는 경우에는 에러가 발생할 수 있음
    - 특징(feature)이 너무 많은 경우에는 이 특징들 간의 연관 관계를 모두 고려하면 복잡해지므로, 빠르게 판단을 내릴 때 많이 사용
    - 문서 분류, 질병 진단, 스팸 메일 분류 등에 사용

복습 : 통계

- 체인 규칙 (Chain Rule)

$$P(V_1, V_2, \ldots, V_k) = \prod_{i=1}^{k} P(V_i \mid V_{i-1}, \ldots, V_1)$$

  - 예) P (A=a, B=b, C=c)
    - P(abc) = P(a)P(b|a)P(c|ab)

- 베이지안 (Bayes Theorem)

$$P(V_i \mid V_j) = \frac{P(V_j \mid V_i) P(V_i)}{P(V_j)}$$

  - e.g., *P* (A=a|B=b)
    - P(a|b) = P(b|a) P(a) / P(b)

Bayes Theorem에서 유도

- Bayes 법칙
  - P(C_i| X)는 특정 개체 X가 특정 그룹 C_i에 속할 사후 확률 (우리가 구하고자 하는 것)
  - P(X | C_i)는 특정 그룹 C_i인 경우에 개체 X가 거기에 속할 조건부 확률 (가능도)
  - P(C_i)는 특정 그룹 C_i가 발생할 빈도 (클래스 사전 고유 확률)
  - P(X)는 특정 개체 X가 발생할 확률 (모든 그룹에 동일하기 때문에 보통 무시)

$$P(C_i|\mathbf{X}) = \frac{P(\mathbf{X}|C_i)P(C_i)}{P(\mathbf{X})}$$

  - 특정 개체 X가 다음과 같이 n개의 특징을 가졌다면 X= (x_1, x_2, …, x_n)

$$P(\mathbf{X}|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(x_n|C_i)$$

나이브 베이지안 분류 예제#1

- 문제: "drew라는 이름을 가진 사람은 여자일까? 남자일까?"

Drew Barrymore

Drew Carey

  - 클래스는 2개로 나누어짐 : 남자(male), 여자(female)
  - 남성일 확률을 나이브 베이지안 분류기로 계산

남자들 중에서 "drew"이름을 가질 확률

$$p(\text{male} \mid drew) = \frac{p(drew \mid \text{male})\, p(\text{male})}{p(drew)}$$

← 남성일 확률

← 무시해도 됨

나이브 베이지안 분류 예제#1

- 다음 "drew" 이름을 가진 경찰관이 남자인지 여자인지 알아내는 문제
  - 오른쪽 표와 같은 표본 데이터는 가지고 있음

Officer Drew

Name	Sex
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

$$p(c_j \mid d) = \frac{p(d \mid c_j)\, p(c_j)}{p(d)}$$

예제 출처: http://www.cs.ucr.edu/~eamonn/CE/Bayesian%20Classification%20withInsect_examples.pdf

나이브 베이지안 분류 예제#1

Officer Drew

$$p(c_j \mid d) = \frac{p(d \mid c_j)\, p(c_j)}{p(d)}$$

Name	Sex
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

$$p(\text{male} \mid drew) = \frac{1/3 * 3/8}{3/8} = \frac{0.125}{3/8}$$

$$p(\text{female} \mid drew) = \frac{2/5 * 5/8}{3/8} = \frac{0.250}{3/8}$$

Officer Drew is more likely to be a Female.

예제 출처: http://www.cs.ucr.edu/~eamonn/CE/Bayesian%20Classification%20withInsect_examples.pdf

나이브 베이지안 분류 예제#2

- 다음과 같이 컴퓨터를 구매하거나 구매하지 않는 이력이 주어졌을 때

- 클래스
  - C1:buys_computer = 'yes'
  - C2:buys_computer = 'no'

- X=(age<30, Income-medium, Student = yes, Credit_rating = Fair) 일때, 컴퓨터를 구매할지 안할지 분류하는 문제

age	income	student	credit_rating	ys_compu
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

From Jiawei Han's Slide for Data Mining: Concepts and Techniques

나이브 베이지안 분류 예제#2

- P(C_i) : 컴퓨터를 구매할지 안할지 확률

  P(buys_computer = "yes") = 9/14 = 0.643
  P(buys_computer = "no") = 5/14= 0.357

- P(X | C_i) 를 계산 (해당되는 속성들을 2개의 클래스별로 계산)

  P(age = "<=30" | buys_computer = "yes") = 2/9 = 0.222
  P(age = "<= 30" | buys_computer = "no") = 3/5 = 0.6
  P(income = "medium" | buys_computer = "yes") = 4/9 = 0.444
  P(income = "medium" | buys_computer = "no") = 2/5 = 0.4
  P(student = "yes" | buys_computer = "yes) = 6/9 = 0.667
  P(student = "yes" | buys_computer = "no") = 1/5 = 0.2
  P(credit_rating = "fair" | buys_computer = "yes") = 6/9 = 0.667
  P(credit_rating = "fair" | buys_computer = "no") = 2/5 = 0.4

$$P(\mathbf{X} \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i) = P(x_1 \mid C_i) \times P(x_2 \mid C_i) \times \ldots \times P(x_n \mid C_i)$$

- X=(age <=30, income=medium, Student=yes, credit_rating=fair)

  P(X|C_i) : P(X|buys_computer = "yes") = 0.222 x 0.444 x 0.667 x 0.667 = 0.044
  P(X|buys_computer = "no") = 0.6 x 0.4 x 0.2 x 0.4 = 0.019

  P(X|C_i)*P(C_i) : P(X|buys_computer = "yes") * P(buys_computer = "yes") = 0.028
  P(X|buys_computer = "no") * P(buys_computer = "no") = 0.007

따라서, X 는 ("buys_computer = yes")에 속한다