Decision Tree : 의사결정 규칙을 나무구조로 나타내 전체 데이터를 소집단으로 분류하거나 예측하는 방법

↳ 좋은 Decision Tree : 정확도를 높이면서, simple한 것을 선호한다! → 깊이가 최대한 한 개씩 케이스만 가지도록

↳ 좋은 기준은 어떻게 정하는 것일까? ┌ 불순도 : Entropy, Gini index
                                    └ 어떤 기준으로 노드를 놓을지 : ID3, CART 알고리즘

Entropy $\left(-\sum_{k=1}^{m} p_k \log_2(p_k)\right)$

: 데이터의 불확실성을 나타낸다. 즉, Entropy 값 = 불순도 = 순도의 ┌ 불확실성 ↓ = 순도 ↑ : 엔트로피 ↓
                                                                  └ 불확실성 ↑ = 순도 ↓ : 엔트로피 ↑

ID3 알고리즘

Entropy 차이를 통해 Information Gain도를 : 즉, Information Gain 값이 클수록, 엔트로피를 많이 줄였다는 의미 = 엔트로피가 작아졌다

상위 노드의 Entropy에서 하위 노드의 Entropy를 뺀 값

$Gain(S,A) = E(S) - \underline{I(S,A)} = E(S) - \sum_i \frac{|S_i|}{|S|} \cdot E(S_i)$

↳ $I(S,A) = \sum_i \frac{|S_i|}{|S|} \cdot E(S_i)$

Gini index $\left(\sum_{j=1}^{2} \frac{|D_i|}{|D|} \cdot Gini(D_i)\right)$

데이터의 통계적 불순 정도를 정량화 해서 표현한 값, Gini index 값↓ = 불순도 값↓ = 순도↑

$Gini(A) = \sum_{j=1}^{2} \frac{|D_i|}{|D|} * Gini(D_i)$ → $Gini(D_i) = 1 - \sum_{j=1}^{x} P_j^2$

CART 알고리즘

Gini index를 이용한 알고리즘, Binary split만을 장려함

▷ feature가 연속형이라면?

1. 전체 데이터를 오름차순으로 분할후 불순도를 계산한다.
2. 종속변수 범위수를 가정으로 한다.
3. Label의 class가 바뀌는 수를 가정으로 한다.

Step1. 각 Feature에 대해 오름차순으로 정렬 → STEP2. Label의 class가 바뀌는 지점을 찾기

→ STEP3. 정제의 평균값을 가능값으로 잡기 → STEP4. 각 가정에 대해 분할후, Gini index 혹은 Entropy 계산

↳ 위와같은 과정을 통해 첫번째 가정을 정한다. 이 과정을 반복하면 최종 Decision Tree model 만들수있다.

▷ 가지치기

모든 terminal node의 순수도 100%인 상태를 Full tree라고 한다. 분기가 너무 많아 과적합 위험이 발생한다

┌ Pre pruning (사전 가지치기) : 트리의 최대 depth나 분할의 최소 개수를 미리 지정
└ Post pruning (사후 가지치기) : 트리를 만든 후 데이터 포인트가 적은 노드를 삭제 / 병합

Naive Bayes Classifier

▷ 확률 기초

- 확률 : 특정한 사건이 일어날 가능성을 나타내는 것

- 조건부확률 : 어떤 사건이 일어난 조건에서, 다른 사건이 일어날 확률

$P(B|A) = P(B \cap A) = P(A \cap B) = P(A|B)P(B)$

- 독립과 조건부독립 :
  - 독립 : 한 사건이 일어날 확률이 다른 사건이 일어날 확률에 영향을 미치지 않는 상태
    $P(A \cap B) = P(A)P(B)$
  - 조건부독립 : 한 사건이 일어났다는 가정하에서, 서로 다른 두 사건은 독립인 상황
    $P(A, B|C) = P(A|C) \cdot P(B|C)$

▷ 베이즈 정리

"두 확률변수의 사전 확률과 사후 확률 사이의 관계를 나타내는 정리"

$$P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D)}$$

- $P(H)$: 사전확률, 과거의 경험을 토대로 내놓은 확정된 파라미터 (H)의 확률
- $P(H|D)$: 사후확률, 관측결과 사건 D가 일어난 경우 파라미터의 확률
- $P(D|H)$: 사전확률의 과거 경향을 잘 설명하는 정도, 모델 파라미터를 바탕으로 하는 관측치의 확률
- $P(D)$: 사건 D의 발생가능성

▷ 계산의 한계

변수가 늘어날수록 가능경우의 연산량이 증가 → 나이브 : 조건부 독립을 가정 !!

가정 : 종속변수가 주어졌을때, 입력 변수들이 모두 독립이다 !! (조건부독립 가정)

결과가 주어졌을 때, 여러 변수 벡터의 결합된 조건부 확률은 각 조건부 확률의 곱으로 충분히 잘 추정할 수 있다는 단순한 가정을 기초로 한다.

$$P(A \cap B|C) = P(A|C)P(B|C)$$

알아내야 할 파라메터의 수가 대폭 줄어들게 된다. $P(X=x|Y=y) \Rightarrow dk$

Feature들의 곱으로 베리언씨 계산이 수월해진다

- 라플라스 스무딩 : likelihood가 0이 되는 것을 방지하여 최적의 확률을 찾아내자

$$P_{LAP} = \frac{C(x)+1}{\sum_x [C(x)+1]}, \quad P(x|c) = \frac{Count(x,c)+1}{\sum_{x \in V} Count(x,c) + V}$$

실제보다 한번씩 더 관측되었다고 가정하기

분모에 1 대신 분모의 입력변수들의 개수 V 더해줄