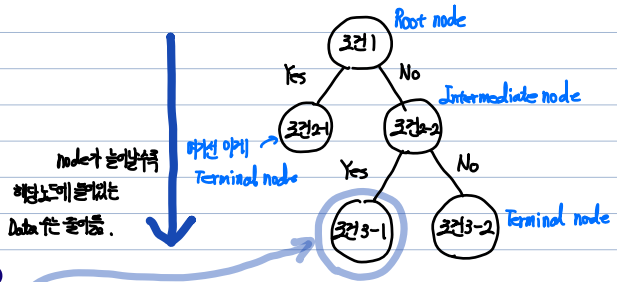
두빅스 3주차 정리 Decision Tree

Unit 1. 의사결정 나무란?

- 의사결정 규칙을 나무구조로 나타내어 데이터를 분류 / 예측.
  - ↳ Data 속에 존재하는 패턴을 예측가능한 규칙들의 조합으로 나타냄.
  - 그리고 이 모양이 나무와 비슷해 의사결정 나무라는 이름이 붙여짐.



What is good decision tree? → ① (똑같은 정확도를 가졌다는 전제하 :) 더 단순한 좋은 Decision Tree!
② 어떤 노드들었을때 한눈에 Data가 비슷해보다? 좋은 Decision Tree!

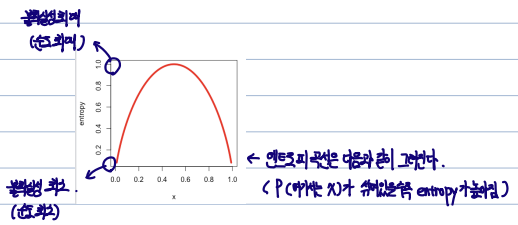
좋은 decision tree 만드는 좋은 기준은?

좋은 기준 ① Entropy - ID3 (Unit 2) ② Gini index - CART 알고리즘 (Unit 3)

Unit 2. ID3 알고리즘

Entropy → (in Data,) 데이터의 불확실성.

- → entropy 높다 : 집단의 특징을 찾기어렵다. (분류가 어렵다는 뜻인가...?)
- → ∴ entropy를 낮추는 방향으로 분류해야 한다.
- → $-\sum_{k=1}^{m} P_k \log_2(P_k)$



← 엔트피 곡선은 대략 이런 모양 그린다.
(P(어떤 X)가 50%일때 entropy 가장높음)

어떤 entropy 차를 이용해 불순도를 줄여주는 기준을 찾는 알고리즘...

ID3 알고리즘 entropy 계산 → Information Gain 도출 → 제일 크게 값이나오는 A를 기준으로 선택.

Information Gain = 상위노드 entropy − 하위노드 entropy.

∴ Information Gain 값이 크다 : 엔트로피를 많이 줄였다.

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

1. 각 변수의 고유값 구성을 살펴보고, 각 class 집합에 대해 entropy 구하기
2. → Information Gain 구하기
3. → Information Gain 중에서 크게 하는 변수를 첫번째 기준으로 선택.
4. → 조건별 Data에 대해 위과정 반복하기.

Unit 3. chart 알고리즘.

Gini Index - Data의 통계적 분산정도를 정량화 해서 표현한값.

- ∴ gini index를 감소하는 방향으로 분류하기.
- → $Gini(A) = \sum_{j=1}^{n} \frac{|Dj|}{|D|} * Gini(D_j)$
  - ↳ $Gini(D_j) = 1 - \sum_{j=1}^{n} P_j^2$

어떤 Gini를 이용해 분류하는 알고리즘....

Cart 알고리즘

- → 데이터의 대상속성을 얼마나 잘분류하는지를 계산한다.
- → Gini Index를 각변수마다 계산하여 가장 작은 Gini index 값을 가지는 변수가 최초의 split이된다.
- → 대신 ID3와달리 Binary split을 전제로 한다.

Unit 4. feature가 연속형이라면?

step1. 각 feature에 대해 오름차순으로 정렬
step2. label의 class가 변하는 지점을 찾기
step3. 경계의 평균값을 기준값으로 잡기
step4. 각 기준에 대해 분할후, Gini index혹은 Entropy 계산.

⇒ 과정 반복!!

Unit 5. 가지치기

왜 가지치기를 하는가? → Full tree 인경우, 분기가 너무 많아 과적합 위험이 발생함. (분기가 너무많으면 일반화 능력이 떨어진다.)

- → ∴ 이를 방지하기위해 적절한 수준에서 terminal node를 결정해준다.
- → 가지치기종류 ┌ 사전가지치기 (pre pruning) : Tree의 최대 깊이를 제한하여 leaf node의 최대 개수를 제한. 노드를 분할하기 위해 필요한 최소한의 Data개수 정해야함.
  - └ 사후가지치기 (Post pruning) : 트리를 만든후 하위노드를 제거 or 병합. / 하위 노드의 불순도의 감소가 특정값이하인 경우 하위 노드를 병합함.
- → 이밖에 Cross-Validation 이나 Independent Validation Set 등의 방법을 사용하기도 한다.!