9강. Decision Tree

◈ 담당교수 : 김 동 하

■ 학습개요

지도 학습법의 일종으로 if-then 규칙으로부터 나무 형태의 모형을 가지고 있는 의사결정나무에 대해 학습한다. 의사결정나무의 구성요소와 다양한 분리규칙에 대해서 다루고, 의사결정나무를 성장시키는 방법과 과적합을 방지하기 위해 가지치기를 하는 방법에 대해서도 배운다. 대표적인 의사결정나무 알고리즘인 CART에 대해서도 배워보도록 한다.

■ 학습목표

1	의사결정나무의 개념에 대해 학습한다.
2	나무의 성장과 가지치기에 대해 학습한다.
3	CART 알고리즘에 대해 학습한다.

■ 주요용어

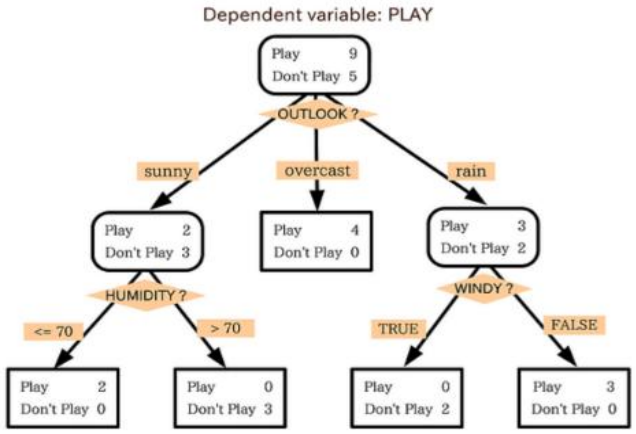
용어	해설
분리 규칙	나무를 성장시킬 때 부모 마디를 자식마디로 분리하는 규칙을 뜻하며, 자식마디의 불순도의 합을 최소로 하는 방향으로 분리 규칙을 정한다.
불순도	각 마디에서 종속 변수의 분포를 가장 잘 구별해주는 변수와 분리 기준을 설정하는데 필요한 측도이다. 불순도를 최소로 하는 분리 규칙을 설정한다.
성장하기	분리 규칙에 따라서 나무를 성장시키는 것을 뜻하며, 모든 끝마디가 정지 규칙을 만족할 때까지 성장시킨다.
가지치기	나무의 과적합을 막기 위한 방법으로 성장이 끝난 나무의 가지를 적당히 제거하여 적당한 크기를 갖는 나무 모형을 만드는 것을 뜻한다.

■ 학습하기

01. 의사결정나무

의사결정나무 개요

- 지도 학습 기법 중 한 가지.
- 적용 결과에 의해 if-then으로 표현되는 규칙 생성.
- 규칙의 이해가 쉽고 우수한 해석력.
- 의사결정나무의 예:



의사결정나무의 구성요소

- 뿌리마디 (Root node)
  -> 나무구조가 시작되는 마디
- 자식마디 (Child node)
  -> 하나의 마디로부터 분리된 2개 이상의 마디들
- 부모마디 (Parent node)
  -> 주어진 마디의 상위 마디
- 끝마디 (Terminal or leaf node)
  -> 자식마디가 없는 마디.
- 중간마디 (Internal node)
  -> 부모마디와 자식마디가 모두 있는 마디.
- 가지 (Branch)
  -> 뿌리마디로부터 끝마디까지 연결된 마디들
- 깊이 (Depth)
  -> 뿌리마디로부터 끝마디까지 분리한 횟수

의사결정나무의 장점

- 이해하기 쉬운 규칙 (if-then) 이용해 생성된다.
- 연속형, 범주형 자료를 모두 다 취급할 수 있다.
- 이상치에 덜 민감하다.

- 모형의 가정 (예: 선형성, 등분산성 등)이 필요 없다.

의사결정나무의 단점
- 회귀 모형에서는 그 예측력이 떨어진다.
- 나무가 너무 깊은 경우에는 예측력이 나쁠 뿐만 아니라 해석 또한 쉽지 않다.
- 계산량이 많을 수 있다.
- 결과가 불안정하다.

02. 의사결정나무의 형성

의사결정나무의 형성과정
- 나무의 성장 (Growing)
  -> 각 마디에서 적절한 최적의 분리규칙 (splitting rule)을 찾아서 나무를 성장시킨다.
  -> 정지규칙 (stopping rule)을 만족하면 성장을 중단한다.
- 가지치기 (Pruning)
  -> 오분류율을 크게 할 위험이 높거나 부적절한 추론규칙을 가지고 있는 가지를 제거.
  -> 불필요한 가지를 제거하는 과정.
- 타당성 평가
  -> 각 끝마디에 예측값을 할당
  -> 이익도표 (Lift chart), 검증 자료 (validation data)의 사용, 또는 교차 타당성 (cross-validation) 등을 이용하여 의사결정나무를 평가.
- 해석 및 예측
  -> 구축된 나무모형을 해석하고 예측.

의사결정나무로 예측하기
- 입력값은 뿌리마디에서 출발
- 분리 조건에 따라 자식마디로 내려감.
- 끝마디에 도착할 때까지 계속 내려감.
- 회귀 문제일 경우
  -> 입력값이 도착한 끝마디에 속하는 모든 훈련자료 출력값의 평균으로 예측.
- 분류 문제일 경우
  -> 입력값이 도착한 끝마디에 속하는 모든 훈련자료 출력값의 최빈값으로 예측.

분리 규칙
- 각 마디에서의 분리규칙
  -> 입력 변수와 분리 기준을 정해야 함.
- 연속 변수의 경우
  -> 변수 X와 분리 기준 c
  -> 변수 X의 값이 c보다 작으면 왼쪽 자식마디, 크면 오른쪽 자식마디
- 범주형 변수의 경우
  -> 전체 범주를 두 개의 부분집합으로 나눔.

-> 예: 전체 범주가 1,2,3,4일 때
-> 1,2,4 중 하나 -> 왼쪽 자식마디
-> 3 -> 오른쪽 자식마디

03. 의사결정나무 만들기

분리 규칙의 선정

- 각 마디에서는 목표 변수의 분포를 가장 잘 구별해주는 변수와 분리 기준을 설정.
- 불순도 (impurity)를 사용
- 불순도를 최소화하는 분리 규칙을 사용
  -> 생성된 두 개의 자식마디의 불순도의 합이 최소
- 불순도 측정량
- 분류 모형
  -> 카이제곱 통계량 (Chi-square statistic)
  -> 지니 지수 (Gini index)
  -> 엔트로피 지수 (Entropy index)
- 회귀 모형
  -> 분산 분석에 의한 F-통계량 (F-statistic)
  -> 분산의 감소량

카이제곱통계량

- 특정 분리 변수와 분리 기준에 의해 다음과 같이 노드를 분리했다고 하자.

	Good	Bad	Total
Left	32	48	80
Right	178	42	220
Total	210	90	300

- 앞의 표에서 각 셀에 대한 기대도수를 구할 수 있다.

	Good	Bad	Total
Left	80/300*210/300*300=54	80/300*90/300*300=24	80
Right	154	66	220
Total	210	90	300

- 실제 도수와 기대 도수를 이용.

$$\text{카이제곱통계량} = \sum \frac{(\text{기대도수} - \text{실제도수})^2}{\text{기대도수}}$$

- 앞의 표에서 카이제곱통계량을 구하면 다음과 같다.

$$\frac{(56-32)^2}{56} + \frac{(24-48)^2}{24} + \frac{(154-178)^2}{154} + \frac{(66-42)^2}{66} = 46.75$$

지니 지수

- 지니 지수는 다음과 같이 계산.
  지니지수=왼쪽에서 **good**일 확률∗왼쪽에서 **bad**일 확률+오른쪽에서 **good**일 확률∗오른쪽에서 **bad**일 확률
- 앞의 표에서 지니 지수는 다음과 같다.

$$\frac{32}{80}\times\frac{48}{80}+\frac{178}{220}\times\frac{42}{220}=0.3944$$

- 최소의 지니 지수를 갖는 분리기준을 선택

엔트로피 지수

- 엔트로피 지수는 다음과 같이 계산.
  엔트로피=왼쪽에서 **good**일 확률∗log(왼쪽에서 **good**일 확률)
  +왼쪽에서 **bad**일 확률∗log(왼쪽에서 **bad**일 확률)
  +오른쪽에서 **good**일 확률∗log(오른쪽에서 **good**일 확률)
  +오른쪽에서**bad**일 확률∗log(오른쪽에서**bad**일 확률)
- 앞의 표에서 엔트로피를 구하면 약 0.4796
- 엔트로피를 가장 작게 하는 분리 기준을 탐색.

회귀 모형에서의 불순도

- 왼쪽 자식 마디와 오른쪽 자식 마디의 평균의 차이를 검정하는 F-통계량의 유의 확률이 가장 작은 분리 변수와 분리 기준을 사용하여 분리를 수행.
- 왼쪽 자식 마디의 자료의 분산과 오른쪽 자식 마디의 자료의 분산의 합이 가장 작은 분리를 선택하여 분리를 수행.

정지 규칙

- 현재의 마디가 더 이상 분리가 일어나지 못하게 하는 규칙.
- 대게 다음의 규칙 중 하나를 정지 규칙으로 사용한다.-> 모든 자료가 한 그룹에 속할 때 (목표 변수가 범주형일 때에만 해당).
  -> 마디에 속하는 자료가 일정 수 이하일 때.
  -> 불순도의 감소량이 아주 작을 때.
  -> 뿌리 마디로부터의 깊이가 일정 수 이상일 때.

가지치기

- 지나치게 많은 마디를 가지는 의사결정나무는 새로운 자료에 적용했을 때 예측 오차가 매우 클 가능성이 있다. -> 과적합
- 성장이 끝난 나무의 가지를 적당히 제거하여 적당한 크기를 갖는 나무 모형을 최종적인 예측모형으로 선택하는 것이 예측력의 향상에 도움이 된다.
- 적당한 크기를 결정하는 방법은 평가용 자료를 사용하거나 교차 확인을 이용하여 예측에러를 구하고 이 예측에러가 가장 작은 나무 모형을 선택한다.

04. CART 알고리즘

CART

- Classification and Regression Tree
- 1984년 Breiman 과 그의 동료들이 발명
- 가장 널리 사용되는 의사결정나무 알고리즘

CART의 나무 성장

- 이진 분류 (Binary split) 을 이용.
- 분류 문제의 경우 불순도를 지니 지수를 이용하고 회귀 문제의 경우 분산을 이용.
- 각 마디의 자료의 수가 일정 수보다 작거나 불순도의 감소량이 일정 양 이하이면 성장을 정지.

CART의 가지치기

- 주어진 나무 T와 양수 α에 대해 비용 복잡도 (cost complexity) 를 다음과 같이 정의한다:

  $C_\alpha(T) = Err(T) + \alpha \cdot |T|,$

  여기서, $Err(T)$는 나무 T의 학습 데이터의 오분류율, $|T|$는 나무 T의 끝마디 개수.
- 나무 성장과정을 통해 생성된 큰 나무 T_0에 대하여, 주어진 α에 대해 $C_\alpha(\cdot)$를 최소로 만드는 T_0의 부분나무를 $T(\alpha)$라 하자.
- α를 0에서 시작해서 계속 증가시키면서 그에 대응되는 나무 $T(\alpha)$를 찾아나간다.
- 평가용 자료나 교차 확인 방법을 이용하여 $T(\alpha)$의 오분류율을 계산한다.
- 이 중에서 오분류율이 가장 작은 나무를 최종 의사결정나무 모형으로 선택한다.

■ 연습문제

(객관식)1. 다음 보기 중 의사결정나무에 대한 설명으로 올바르지 않은 것을 고르시오.

① 이해하기 쉬운 if-then 규칙을 사용한다.

② 나무가 매우 깊더라도 설명력은 언제나 우수하다.

③ 분류 문제 뿐만 아니라 회귀 문제도 해결할 수 있다.

④ 다른 방법론에 비해 결과가 불안정하다.

정답 : ②

해설 : 나무가 너무 깊을 때에는 해석이 쉽지 않다.

(객관식)2. 다음 중 분류 문제에서 마디를 분리할 때 사용하는 불순도로 올바르지 않은 것은?

① F-통계량

② 엔트로피
③ 카이 제곱 통계량
④ 지니 지수

정답) ①
해설) F-통계량은 회귀 문제에서 사용되는 불순도이다.

(서술형)3. 가지치기를 하는 이유를 서술하시오.

정답 및 해설 : 지나치게 많은 마디를 갖는 의사결정나무는 과적합되었기에 예측 성능이 나쁠 수 있다. 따라서 불필요한 마디를 잘라내어 적당한 크기를 갖게 할 필요가 있다.

■ 정리하기

1. 의사결정나무는 지도 학습법의 한 가지로 규칙의 이해가 쉽고 우수한 해석력을 갖는다.

2. 의사결정나무의 형성은 분리규칙을 이용해 나무를 성장시키는 성장하기와 성장한 나무의 가지를 잘라내어 성능을 향상시키는 가지치기로 이루어져 있다.

3. 마디를 분리시켜 나무를 성장시킬 때는 미리 정의된 불순도를 최소화하는 방향으로 진행한다.

■ 참고자료 (참고도서, 참고논문, 참고사이트 등)

박창이, 김용대, 김진석, 송종우, 최호식. 『R을 이용한 데이터마이닝』. 서울:교우사, 2018.