의사결정 트리

4.3 가지치기
4.4 연속값과 결측값
4.5 다변량 의사결정트리

4.1 기본 프로세스
4.2 분할선택

→ 어떻게 나누냐

4.3 가지치기
4.4 연속값과 결측값
4.5 다변량 의사결정트리

→ 더 잘 나누기 위해서

가지치기란?

과적합에 대응하기 위한 주요수단

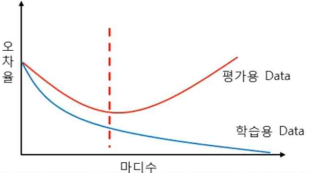
4.3 가지치기

과적합(Overfitting)

- 학습용 데이터에 완전히 적합
- 학습용 데이터의 노이즈(noise)도 모형화 -> 테스트 데이터 검증시 전체 오차 일반적으로 증가

모델 개발의 목적

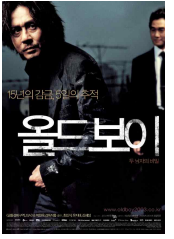
- 학습용 데이터에서 높은 성과 -> 테스트 데이터에서 낮은 성과 : X
- 현재 데이터의 설명 -> 미래 데이터 예측 : O



4.3 가지치기

과적합의 예시

- 교과서 **수박** 데이터 -> 직관적X
- 데이터 수가 같은 임의의 데이터 선정 -> 어떤 영화가 대중성이 있는가?
- 평점 높은 **한국영화 17편** 선정, **1000만명** 이상의 관객 동원 = 대중성 O



번호	영화명	개봉연도	장르	상영시간	제작비	출연자
1	괴물	2006	SF,가족	119	110	송강호,배두나,박해일
2	태극기 휘날리며	2004	전쟁,드라마	148	170	장동건,원빈,최민식
3	광해	2012	사극,드라마	131	65	이병헌,류승룡,한효주
4	베테랑	2015	액션,드라마	123	90	황정민,유아인,오달수
5	왕의남자	2005	사극,드라마	119	44	감우성,이준기,정진영
6	기생충	2019	드라마	131	150	송강호,최우식,박소담
7	극한직업	2018	코미디	111	85	류승룡,진선규,신하균
8	부산행	2016	액션,스릴러	118	115	공유,최우식,정유미
9	김종욱찾기	2010	로맨스	112	45	공유,임수정
10	달콤한인생	2005	액션	120	65	이병헌,황정민,오달수
11	내머리속의지우개	2004	로맨스	117	30	정우성,손예진
12	올드보이	2003	스릴러	120	33	최민식,유지태,오달수
13	타짜	2006	드라마	139	53	조승우,유해진,김윤석
14	김씨표류기	2008	코미디	116	50	정재영,정려원
15	오아시스	2002	드라마	132	28	설경구,문소리
16	지금은맞고그때는틀리다	2015	드라마	120	1	정재영,김민희,고아성
17	빈집	2004	드라마	88	10	재희,이승연

영화 데이터 세트 1.0

번호	영화명	개봉시기	장르	상영시간	제작비	청소년관람	종등장	대중성
1	괴물	2000년대	액션	두시간이내	80억 이상	가능	예	예
2	태극기 휘날리며	2000년대	액션	두시간초과	80억 이상	가능	예	예
3	광해	2010년대	사극	두시간초과	80억 미만	가능	아니오	예
4	베테랑	2010년대	액션	두시간초과	80억 이상	가능	예	예
5	왕의남자	2000년대	사극	두시간이내	80억 미만	가능	아니오	예
6	기생충	2010년대	드라마	두시간초과	80억 이상	가능	아니오	예
7	극한직업	2010년대	코미디	두시간이내	80억 이상	가능	예	예
8	부산행	2010년대	액션	두시간이내	80억 이상	가능	예	예
9	김종욱찾기	2010년대	로맨스	두시간이내	80억 미만	가능	아니오	아니오
10	달콤한인생	2000년대	액션	두시간초과	80억 미만	불가능	예	아니오
11	내머리속의지우개	2000년대	로맨스	두시간이내	80억 미만	가능	아니오	아니오
12	올드보이	2000년대	액션	두시간초과	80억 미만	불가능	예	아니오
13	타짜	2000년대	액션	두시간초과	80억 미만	불가능	예	아니오
14	김씨표류기	2000년대	코미디	두시간이내	80억 미만	가능	아니오	아니오
15	오아시스	2000년대	드라마	두시간초과	80억 미만	불가능	아니오	아니오
16	지금은맞고그때는틀리다	2010년대	드라마	두시간초과	80억 미만	불가능	아니오	아니오
17	빈집	2000년대	드라마	두시간이내	80억 미만	가능	아니오	아니오

영화 데이터 세트 2.0

번호	영화명	대중성
1	괴물	예
2	태극기 휘날리며	예
3	광해	예
4	베테랑	예
5	왕의남자	예
6	기생충	예
7	극한직업	예
8	부산행	예
9	김종욱찾기	아니오
10	달콤한인생	아니오
11	내머리속의지우개	아니오
12	올드보이	아니오
13	타짜	아니오
14	김씨표류기	아니오
15	오아시스	아니오
16	지금은맞고그때는틀리다	아니오
17	빈집	아니오

루트노드 정보 엔트로피

양성 샘플 비율 $p_1 = \frac{8}{17}$

번호	영화명	대중성
1	괴물	예
2	태극기 휘날리며	예
3	광해	예
4	베테랑	예
5	왕의남자	예
6	기생충	예
7	극한직업	예
8	부산행	예
9	김종욱찾기	아니오
10	달콤한인생	아니오
11	내머리속의지우개	아니오
12	올드보이	아니오
13	타짜	아니오
14	김씨표류기	아니오
15	오아시스	아니오
16	지금은맞고그때는틀리다	아니오
17	빈집	아니오

루트노드 정보 엔트로피

양성 샘플 비율 $p_1 = \frac{8}{17}$

음성 샘플 비율 $p_2 = \frac{9}{17}$

번호	영화명	대중성
1	괴물	예
2	태극기 휘날리며	예
3	광해	예
4	베테랑	예
5	왕의남자	예
6	기생충	예
7	극한직업	예
8	부산행	예
9	김종욱찾기	아니오
10	달콤한인생	아니오
11	내머리속의지우개	아니오
12	올드보이	아니오
13	타짜	아니오
14	김씨표류기	아니오
15	오아시스	아니오
16	지금은맞고그때는틀리다	아니오
17	빈집	아니오

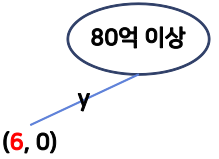
루트노드 정보 엔트로피

양성 샘플 비율 $p_1 = \frac{8}{17}$

음성 샘플 비율 $p_2 = \frac{9}{17}$

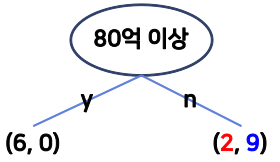
$$\text{Ent(D)} = -\sum_{k=1}^{2} p_k log_2 p_k = -\left(\frac{8}{17} log_2 \frac{8}{17} + \frac{9}{17} log_2 \frac{9}{17}\right) = 0.998$$

번호	영화명	제작비 80억 이상	대중성
1	괴물	예	예
2	태극기 휘날리며	예	예
3	광해	아니오	예
4	베테랑	예	예
5	왕의남자	아니오	예
6	기생충	예	예
7	극한직업	예	예
8	부산행	예	예
9	김종욱찾기	아니오	아니오
10	달콤한인생	아니오	아니오
11	내머리속의지우개	아니오	아니오
12	올드보이	아니오	아니오
13	타짜	아니오	아니오
14	김씨표류기	아니오	아니오
15	오아시스	아니오	아니오
16	지금은맞고그때는틀리다	아니오	아니오
17	빈집	아니오	아니오



$$\mathrm{Ent}(D^1) = -\left(\frac{6}{6}log_2\frac{6}{6} + \frac{0}{6}log_2\frac{0}{6}\right) = 0$$

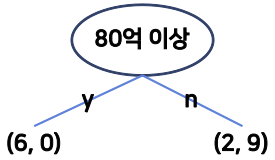
번호	영화명	제작비 80억 이상	대중성
1	괴물	예	예
2	태극기 휘날리며	예	예
3	광해	아니오	예
4	베테랑	예	예
5	왕의남자	아니오	예
6	기생충	예	예
7	극한직업	예	예
8	부산행	예	예
9	김종욱찾기	아니오	아니오
10	달콤한인생	아니오	아니오
11	내머리속의지우개	아니오	아니오
12	올드보이	아니오	아니오
13	타짜	아니오	아니오
14	김씨표류기	아니오	아니오
15	오아시스	아니오	아니오
16	지금은맞고그때는틀리다	아니오	아니오
17	빈집	아니오	아니오



$$\mathrm{Ent}(D^1) = -\left(\frac{6}{6}log_2\frac{6}{6} + \frac{0}{6}log_2\frac{0}{6}\right) = 0$$

$$\mathrm{Ent}(D^2) = -\left(\frac{2}{11}log_2\frac{2}{11} + \frac{9}{11}log_2\frac{9}{11}\right) = 0.684$$

번호	영화명	제작비 80억 이상	대중성
1	괴물	예	예
2	태극기 휘날리며	예	예
3	광해	아니오	예
4	베테랑	예	예
5	왕의남자	아니오	예
6	기생충	예	예
7	극한직업	예	예
8	부산행	예	예
9	김종욱찾기	아니오	아니오
10	달콤한인생	아니오	아니오
11	내머리속의지우개	아니오	아니오
12	올드보이	아니오	아니오
13	타짜	아니오	아니오
14	김씨표류기	아니오	아니오
15	오아시스	아니오	아니오
16	지금은맞고그때는틀리다	아니오	아니오
17	빈집	아니오	아니오

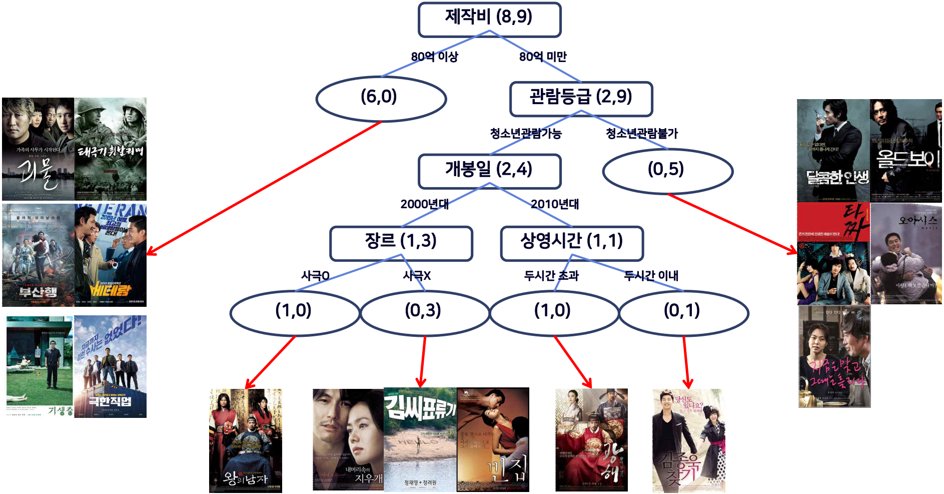


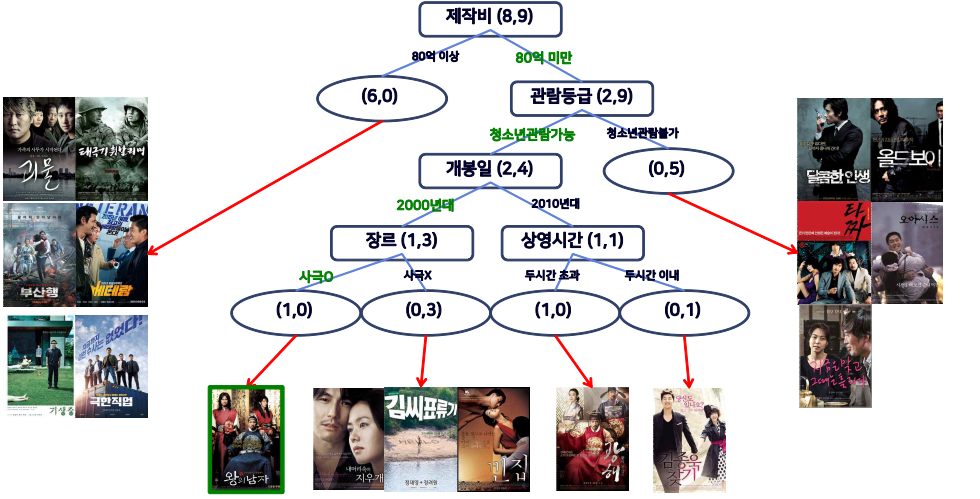
$$\mathrm{Ent}(D^1) = -\left(\frac{6}{6}log_2\frac{6}{6} + \frac{0}{6}log_2\frac{0}{6}\right) = 0$$

$$\mathrm{Ent}(D^2) = -\left(\frac{2}{11}log_2\frac{2}{11} + \frac{9}{11}log_2\frac{9}{11}\right) = 0.684$$

$$Gain(\mathrm{D}, \text{제작비}) = Ent(D) - \sum_{v=1}^{V}\frac{D^v}{D}Ent(D^v)$$

$$= 0.998 - (\frac{6}{17}\times 0 + \frac{11}{17}\times 0.684) = 0.555$$


제작비 (8,9)
80억 이상
80억 미만
(6,0)
관람등급 (2,9)
청소년관람가능
청소년관람불가
개봉일 (2,4)
(0,5)
2000년대
2010년대
장르 (1,3)
상영시간 (1,1)
사극O
사극X
두시간 초과
두시간 이내
(1,0)
(0,3)
(1,0)
(0,1)
괴물
태극기 휘날리며
부산행
베테랑
기생충
극한직업
왕의 남자
내머리속의 지우개
김씨표류기
빈집
광해
김종욱 찾기
달콤한 인생
올드보이
타짜
오아시스



80억 미만 제작비에 청소년 관람가능한 2000년대 사극장르의 영화 = 대중성?

4.3 가지치기

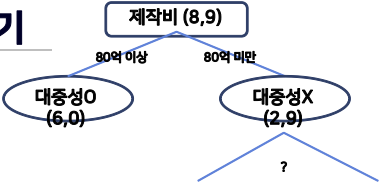
과적합의 문제

- "왕의 남자" -> 노이즈(Noise)
- train data 에서 정확성 : 100% / 그러나 test data에서 정확성 오히려 감소
- 과적합 문제 해결 : 가지치기

가지치기 종류

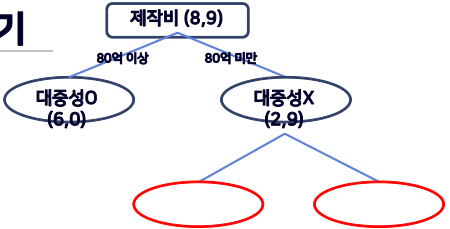
- 사전 가지치기(pre-pruning): 분할 전 미리 예측 -> 분할중지
- 사후 가지치기(post-pruning): 완전한 의사결정 트리 -> 가지를 터미널 노드로

4.3.1 사전 가지치기



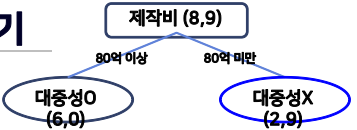
1. 가지를 뻗을까 말까

4.3.1 사전 가지치기



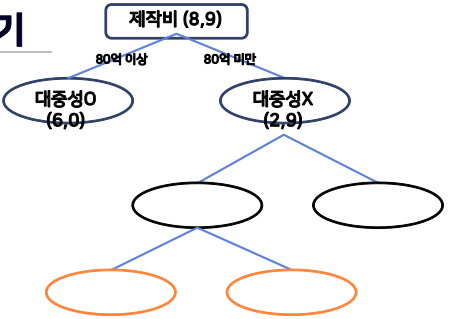
1. 가지를 뻗을까 말까
2-1). 일반화 성능 향상 O -> 가지 뻗기

4.3.1 사전 가지치기



1. 가지를 뻗을까 말까
2-1). 일반화 성능 향상 O -> 가지 뻗기
2-2). 일반화 성능 향상 X -> 분할 중지 -> 해당노드 -> 터미널노드

4.3.1 사전 가지치기



1. 가지를 뻗을까 말까
2-1). 일반화 성능 향상 O -> 가지 뻗기
2-2). 일반화 성능 향상 X -> 분할 중지 -> 해당노드 -> 터미널노드
3. 1.반복

일반화 성능측정

1. 홀드 아웃(Hold-out)
2. train data 일부를 test data로 제외 -> 나중 성능평가

번호	영화명	개봉시기	장르	상영시간	제작비	청소년관람	흥행작	대중작
1	괴물	2000년대	액션	두시간이내	80억 이상	가능	예	예
2	태극기 휘날리며	2000년대	액션	두시간초과	80억 이상	가능	예	예
3	광해	2010년대	사극	두시간초과	80억 미만	가능	아니오	예
4	베테랑	2010년대	액션	두시간초과	80억 이상	가능	예	예
5	왕의남자	2000년대	사극	두시간이내	80억 미만	가능	아니오	예
6	기생충	2010년대	드라마	두시간초과	80억 이상	가능	아니오	예
7	극한직업	2010년대	코미디	두시간이내	80억 이상	가능	예	예
8	부산행	2010년대	액션	두시간이내	80억 이상	가능	예	예
9	김종욱찾기	2010년대	로맨스	두시간이내	80억 미만	가능	아니오	아니오
10	달콤한인생	2000년대	액션	두시간초과	80억 미만	불가능	예	아니오
11	내머리속의지우개	2000년대	로맨스	두시간이내	80억 미만	가능	아니오	아니오
12	올드보이	2000년대	액션	두시간초과	80억 미만	불가능	예	아니오
13	타짜	2000년대	액션	두시간초과	80억 미만	불가능	예	아니오
14	김씨표류기	2000년대	코미디	두시간이내	80억 미만	가능	아니오	아니오
15	오아시스	2000년대	드라마	두시간초과	80억 미만	불가능	아니오	아니오
16	지금은맞고그때는틀리다	2010년대	드라마	두시간초과	80억 미만	불가능	아니오	아니오
17	빈집	2000년대	드라마	두시간이내	80억 미만	가능	아니오	아니오

4.3.1 사전 가지치기

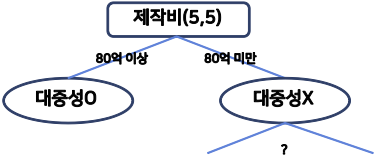
대중성 O(5,5)

검정 정확도 : $\frac{3}{7} \times 100\% = 42.9\%$

4.3.1 사전 가지치기

검정 정확도 : $\frac{6}{7} \times 100\% = 85.7\% > 42.9\%$

4.3.1 사전 가지치기



현재 검정 정확도 : 85.7%

1) 분할 후 : 85.7% 미만 -> 분할종료

2) 분할 후 : 85.7% -> 분할 종료

3) 분할 후 : 85.7% 초과 -> 분할진행

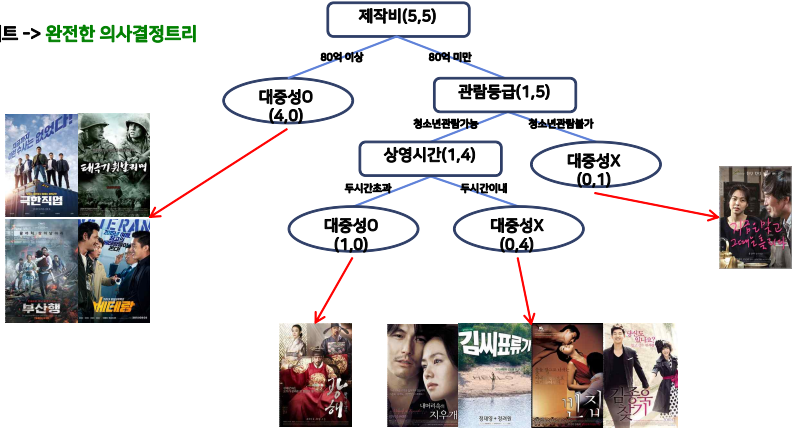
4.3.1 사전 가지치기

분할전 검정 정확도 : 85.7%

분할후 검정 정확도 : 100%

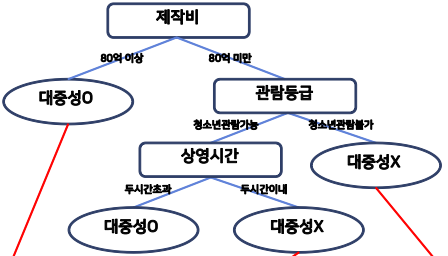
4.3.2 사후 가지치기

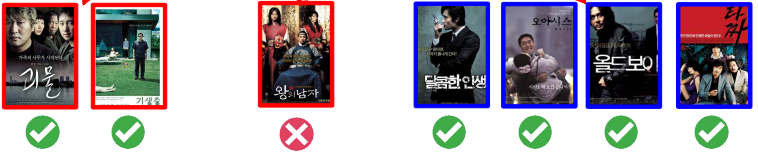
훈련세트 -> 완전한 의사결정트리



4.3.2 사후 가지치기

테스트세트 -> 최하위노드부터 검증
현재 검정정확도 : 85.7%

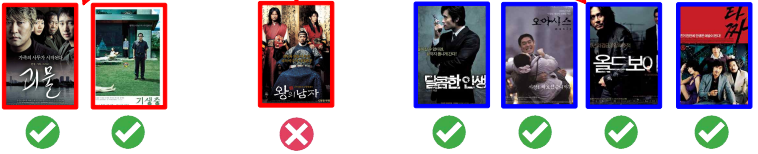




4.3.2 사후 가지치기

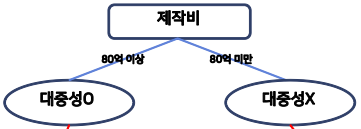
테스트세트 -> 최하위노드부터 검증
현재 검정정확도 : 85.7%

1) 가지치기 후 : 85.7% 미만 -> 중지
2) 가지치기 후 : 85.7% -> 오컴의 면도날
3) 가지치기 후 : 85.7% 초과 -> 가지치기 진행

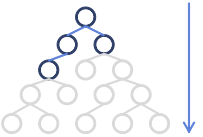


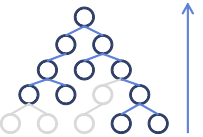
4.3.2 사후 가지치기

테스트세트 -> 최하위노드부터 검증
현재 검정정확도 : 85.7%



1) 가지치기 후 : 85.7% 미만 -> 중지
2) 가지치기 후 : 85.7% -> 오컴의 면도날
3) 가지치기 후 : 85.7% 초과 -> 가지치기 진행





사전 가지치기	사후 가지치기
가지수 적다	가지수 많다
훈련 및 테스트 시간 짧다	훈련 및 테스트 시간 길다
성능향상 기회차단	일반화 성능 좋다
과소적합 위험 높음	과소적합 위험 낮음
과적합 위험 낮음	과적합 위험 높음

4.4 연속값과 결측값

연속값 처리

- 지금까지 : 이산 속성
- 현실문제 : 연속 속성
- 연속 속성의 이산화 작업 -> C4.5 의사결정 트리 알고리즘(이분법)
- example : 영화 data -> 개봉시기, 제작비, 상영시간

번호	영화명	개봉시기(년)	장르	상영시간(분)	제작비(억원)	청소년관람	흥행성	대중성
1	괴물	2006	액션	119	110	가능	예	예
2	태극기 휘날리며	2004	액션	148	170	가능	예	예
3	광해	2012	사극	131	65	가능	아니오	예
4	베테랑	2015	액션	123	90	가능	예	예

Method 1

번호	영화명	제작비(억원)	대중성	
2	태극기 휘날리며	170	예	
6	기생충	150	예	160
8	부산행	115	예	132.5
1	괴물	110	예	112.5
4	베테랑	90	예	100
7	극한직업	85	예	87.5
3	광해	66	예	75
10	달콤한인생	64	아니오	65
13	타짜	53	아니오	58.5
14	김씨표류기	50	아니오	51.5
9	김종욱찾기	45	아니오	47.5
5	왕의남자	44	예	44.5
12	올드보이	33	아니오	38.5
11	내머리속의지우개	30	아니오	31.5
15	오아시스	28	아니오	29
17	빈집	10	아니오	19
16	지금은맞고그때는틀리다	1	아니오	5.5

Method 2

번호	영화명	제작비(억원)	대중성	
2	태극기 휘날리며	170	예	
6	기생충	150	예	
8	부산행	115	예	
1	괴물	110	예	
4	베테랑	90	예	
7	극한직업	85	예	
3	광해	66	예	65
10	달콤한인생	64	아니오	
13	타짜	53	아니오	
14	김씨표류기	50	아니오	
9	김종욱찾기	45	아니오	44.5
5	왕의남자	44	예	38.5
12	올드보이	33	아니오	
11	내머리속의지우개	30	아니오	
15	오아시스	28	아니오	
17	빈집	10	아니오	
16	지금은맞고그때는틀리다	1	아니오	

Method 1

번호	영화명	제작비(억원)	대중성	
2	태극기 휘날리며	170	예	
6	기생충	150	예	160
8	부산행	115	예	132.5
1	과물	110	예	112.5
4	베테랑	90	예	100
7	극한직업	85	예	87.5
3	광해	66	예	75
10	달콤한인생	64	아니오	65
13	타짜	53	아니오	58.5
14	김씨표류기	50	아니오	51.5
9	김종욱찾기	45	아니오	47.5
5	왕의남자	44	예	44.5
12	올드보이	33	아니오	38.5
11	내머리속의지우개	30	아니오	31.5
15	오아시스	28	아니오	29
17	빈집	10	아니오	19
16	지금은맞고그때는틀리다	1	아니오	5.5

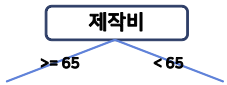
$$T_a = \left\{ \frac{a^i + a^{i+1}}{2} \mid 1 \ll i \ll n-1 \right\}$$

$$T_a = \{5.5, 19, 29, 31.5 \dots 160\}$$

$$Gain(\mathrm{D}, \mathrm{a}) = Ent(D) - \sum_{\lambda \in \{-,+\}} \frac{|D_t^\lambda|}{|D|} Ent(D_t^\lambda)$$

$$Gain(\mathrm{D}, \mathrm{Cost}(5.5)), Gain(\mathrm{D}, \mathrm{Cost}(19)), \dots Gain(\mathrm{D}, \mathrm{Cost}(160))$$

$$Gain(\mathrm{D}, \mathrm{Cost}(65))$$



4.4 연속값과 결측값

결측값 처리

- Entropy는 Non-missing value로만 계산

$$\mathrm{Ent}(\tilde{D}) = -\sum_{k=1}^{2} p_k \log_2 p_k$$

$$= -\left(\frac{6}{14}\log_2\frac{6}{14} + \frac{8}{14}\log_2\frac{8}{14}\right) = 0.985$$

번호	영화명	관람등급(세)	대중성
2	태극기 휘날리며	15	예
6	기생충	15	예
8	부산행	15	예
1	괴물	12	예
4	베테랑	15	예
7	극한직업	12	예
3	광해	-	예
10	달콤한인생	-	아니오
13	타짜	18	아니오
14	김씨표류기	15	아니오
9	김종욱찾기	12	아니오
5	왕의남자	-	예
12	올드보이	18	아니오
11	내머리속의지우개	12	아니오
15	오아시스	18	아니오
17	빈집	15	아니오
16	지금은맞고그때는틀리다	18	아니오

4.4 연속값과 결측값

결측값 처리

- Entropy는 Non-missing value로만 계산
- Information Gain는 Weighted Information Gain으로 변경

$$\mathrm{Ent}(\tilde{D}^1) = -\left(\frac{2}{4}log_2\frac{2}{4} + \frac{2}{4}log_2\frac{2}{4}\right) = 1.000$$

$$\mathrm{Ent}(\tilde{D}^2) = -\left(\frac{4}{6}log_2\frac{4}{6} + \frac{2}{6}log_2\frac{2}{6}\right) = 0.918$$

$$\mathrm{Ent}(\tilde{D}^3) = -\left(\frac{0}{4}log_2\frac{0}{4} + \frac{4}{4}log_2\frac{4}{4}\right) = 0.000$$

$$Gain(\tilde{D}, \text{관람등급}) = Ent(\tilde{D}) - \sum_{v=1}^{3}\tilde{r}_v\, Ent(\tilde{D}^v) = 0.306$$

$$Gain(\mathrm{D}, \text{관람등급}) = \rho \times Gain(\tilde{D}, \text{관람등급}) = \frac{14}{17} \times 0.306 = 0.252$$

번호	영화명	관람등급(세)	대중성
2	태극기 휘날리며	15	예
6	기생충	15	예
8	부산행	15	예
1	과물	12	예
4	베테랑	15	예
7	극한직업	12	예
3	광해	-	예
10	달콤한인생	-	아니오
13	타짜	18	아니오
14	김씨표류기	15	아니오
9	김종욱찾기	12	아니오
5	왕의남자	-	예
12	울드보이	18	아니오
11	내머리속의지우개	12	아니오
15	오아시스	18	아니오
17	빈집	15	아니오
16	지금은맞고그때는틀리다	18	아니오

4.4 연속값과 결측값

결측값 처리

- Entropy는 Non-missing value로만 계산
- Information Gain는 Weighted Information Gain으로 변경
- Intrinsic Value는 missing value를 하나의 클래스로 보고 계산

$$IV(a) = -\left(\frac{4}{17}log_2\frac{4}{17} + \frac{6}{17}log_2\frac{6}{17} + \frac{4}{17}log_2\frac{4}{17} + \frac{3}{17}log_2\frac{3}{17}\right) = 1.955$$

$$Gain\_ratio(\mathrm{D,a}) = \frac{Gain(D,a)}{IV(a)}$$

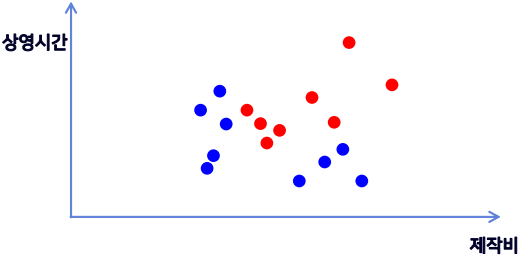
$$= \frac{0.252}{1.955} = 0.128$$

번호	영화명	관람등급(세)	대중성
2	태극기 휘날리며	15	예
6	기생충	15	예
8	부산행	15	예
1	과물	12	예
4	베테랑	15	예
7	극한직업	12	예
3	광해	-	예
10	달콤한인생	-	아니오
13	타짜	18	아니오
14	김씨표류기	15	아니오
9	김종욱찾기	12	아니오
5	왕의남자	-	예
12	올드보이	18	아니오
11	내머리속의지우개	12	아니오
15	오아시스	18	아니오
17	빈집	15	아니오
16	지금은맞고그때는틀리다	18	아니오

4.5 다변량 의사결정 트리

데이터 세트에서 생성된 의사결정트리

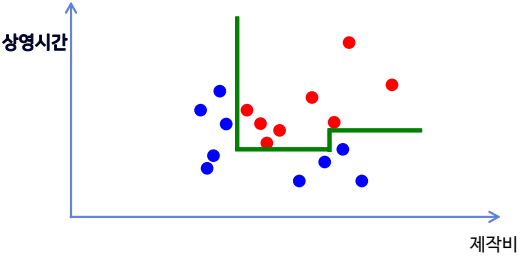
- d개 속성 -> d차원 공간 하나의 데이터 포인트
- 좌표 공간 -> 분류하는 **"경계"**를 찾는 것
- 특징: 축에 **평행**



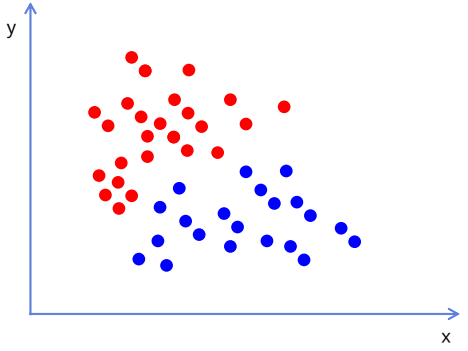
4.5 다변량 의사결정 트리

데이터 세트에서 생성된 의사결정트리

- d개 속성 -> d차원 공간 하나의 데이터 포인트
- 좌표 공간 -> 분류하는 **"경계"**를 찾는 것
- 특징: 축에 **평행**

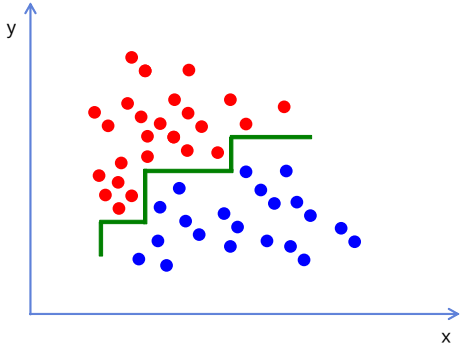


4.5 다변량 의사결정 트리



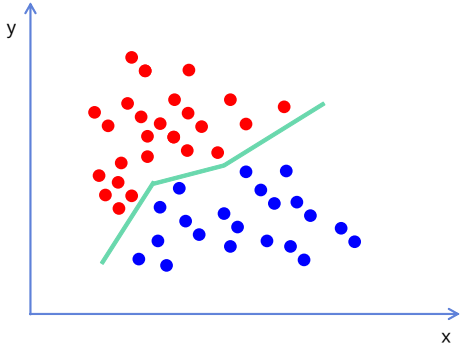
- 현실의 분류 문제 -> 매우 복잡

4.5 다변량 의사결정 트리



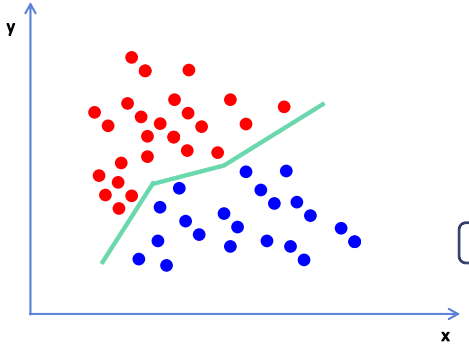
- 현실의 분류 문제 -> 매우 복잡
- 복잡한 분류 경계의 선

4.5 다변량 의사결정 트리

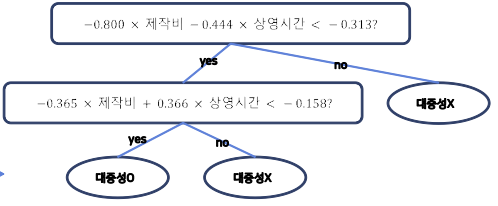


- 현실의 분류 문제 -> 매우 복잡
- 복잡한 분류 경계의 선
- 대각선 분할 => 다변량 의사결정트리

4.5 다변량 의사결정 트리



- 현실의 분류 문제 -> 매우 복잡
- 복잡한 분류 경계의 선
- 대각선 분할 => 다변량 의사결정트리



4.6 더 읽을거리

트리 알고리즘	CART	C4.5	CHAID
분류나무(분류)	O	O	O
회귀나무(예측)	O	O	X
예측변수	범주,수치	범주,수치	범주형 Only
불순도 알고리즘	Gini 지수	Entropy	Chi-squre 통계량
분리	이지분리	다지분리	다지분리
나무성장	완전모형개발(가지치기)		최적모형 개발
가지치기	학습data->검증data	학습data	X
개발자	Breiman	Quinlan	