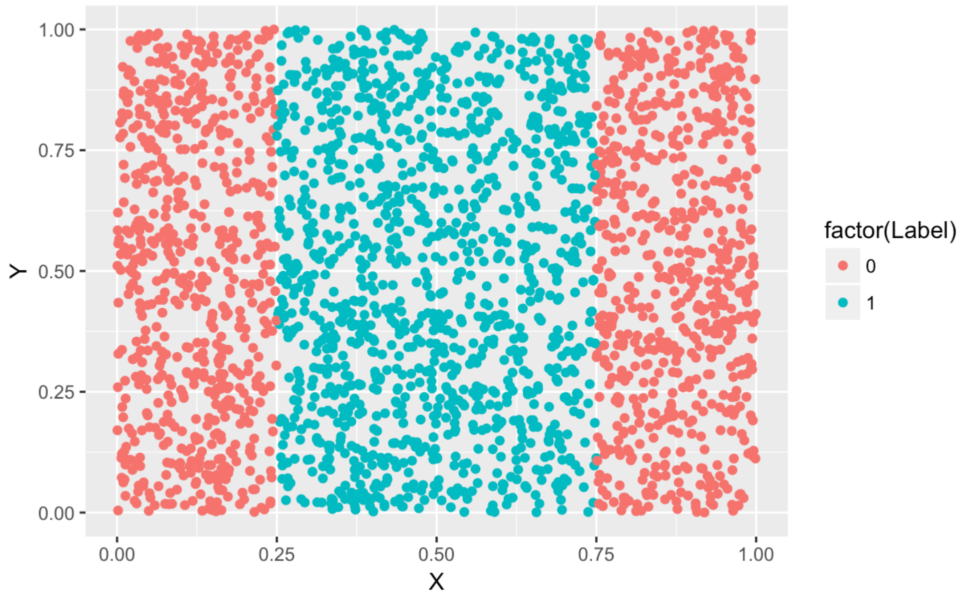
의사결정나무

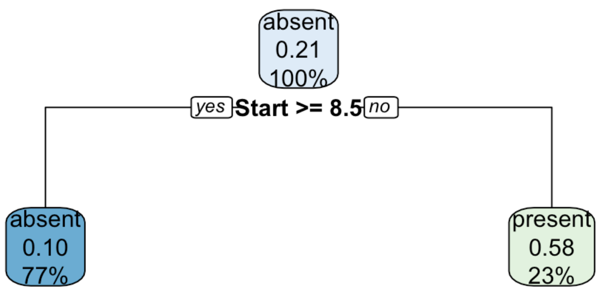
허석진

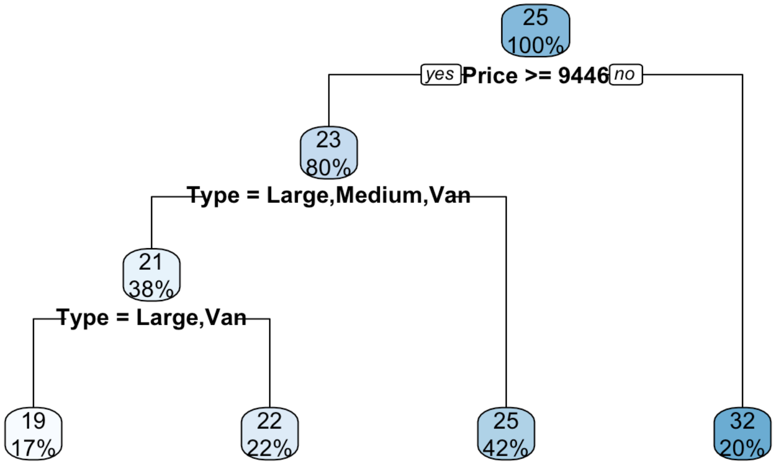
로지스틱 회귀로 가능하지 않은 분류



- 독립변수와 종속변수 사이의 관계가 복잡하거나 매우 비선형적일 때

의사결정나무 (Decision Tree)





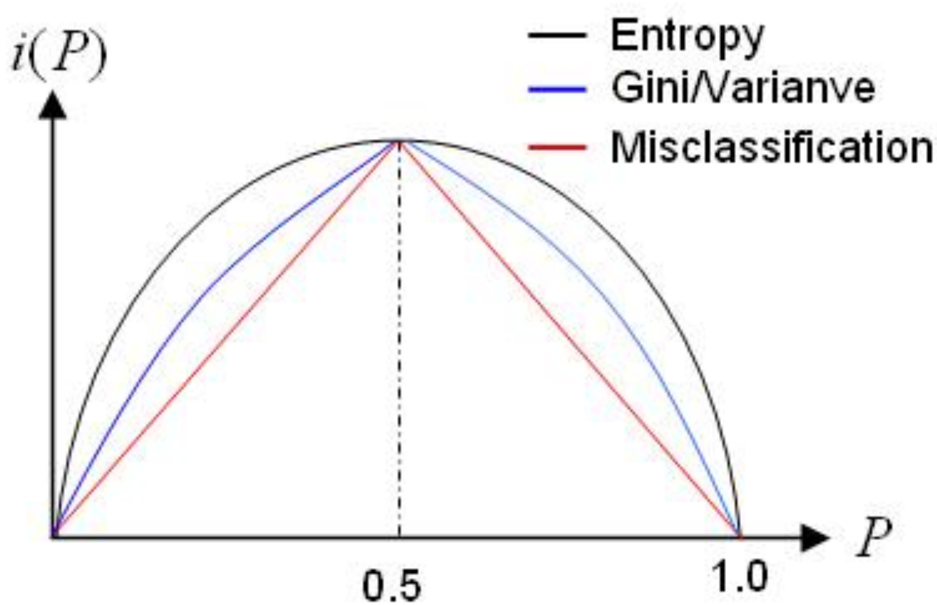
- 조건에 따른 데이터 분류 방법
- 종류
  - 분류나무(classification tree): 범주형 목적 변수
  - 회귀나무(regression tree): 연속형 목적 변수

장단점

- 장점
  - 간단하고 직관적
  - 웬만하면 적당한 결과 산출
  - 비선형인 경우에도 적용
  - 의사결정에 사용하기 편리
  - 범주형(분류나무)과 연속형(회귀나무) 목적값 가능
- 단점
  - 과적합의 우려
  - 정확성이 떨어짐

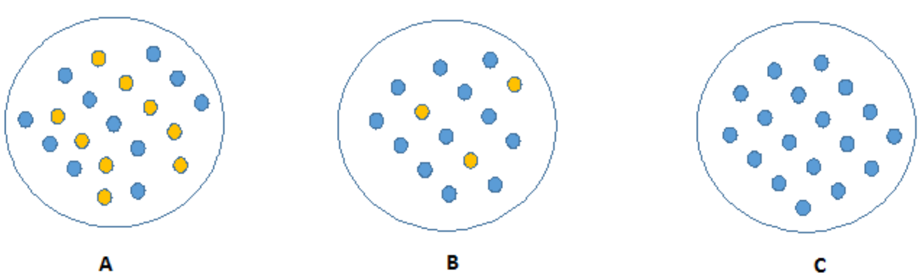
불순도(Impurity)

- $P = 0$ 또는 $P = 1$에서 0
- $P = 1/2$에 대해 대칭 (P와 $1 - P$를 교환해도 식이 바뀌지 않음)
- $P = 1/2$에서 최대



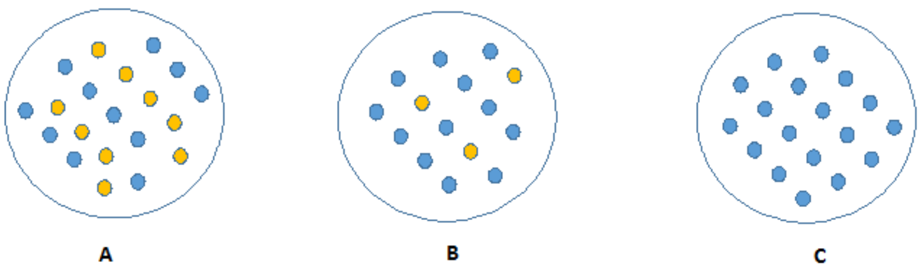
불순도(Impurity)

- 분류에러 (Misclassification error)
  - $\min(p, 1-p)$ (최대값은 0.5)
- 지니계수 (Gini Index)
  - $1-(p^2+q^2)=1-(p^2+(1-p)^2)=2p(1-p)$ (최대값은 0.5)
- 엔트로피 (Entropy)
  - $-p\ \log_2 p-\ (1-p)\log_2(1-p)$ (최대값은 1)



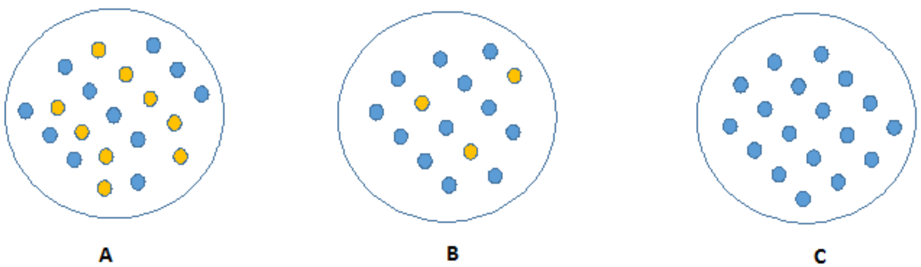
불순도(Impurity)

- 분류에러 (Misclassification error)
  - $\min(p, 1-p)$
  - A: $\min\left(\frac{9}{20}, 1-\frac{9}{20}\right) = 0.45$
  - B: $\min\left(\frac{3}{15}, 1-\frac{3}{15}\right) = 0.2$
  - C: $\min\left(\frac{0}{18}, 1-\frac{0}{18}\right) = 0$



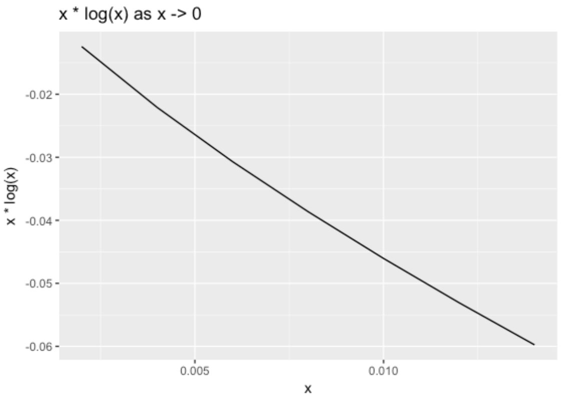
불순도(Impurity)

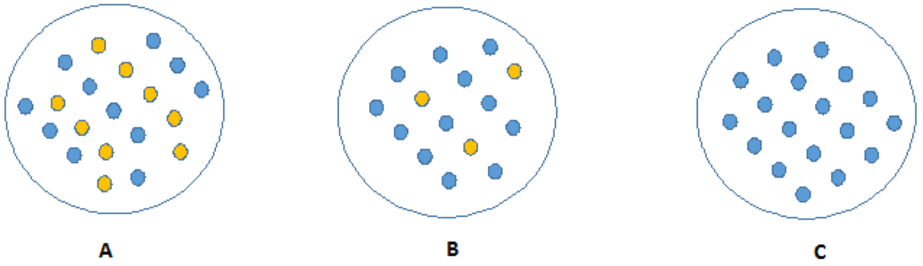
- 지니계수 (Gini Index)
  - $1 - (p^2 + q^2) = 1 - (p^2 + (1-p)^2) = 2p(1-p)$
  - A: $2 \cdot \frac{9}{20} \cdot \left(1 - \frac{9}{20}\right) = 0.495$
  - B: $2 \cdot \frac{3}{15} \cdot \left(1 - \frac{3}{15}\right) = 0.32$
  - C: $2 \cdot \frac{0}{18} \cdot \left(1 - \frac{0}{18}\right) = 0$



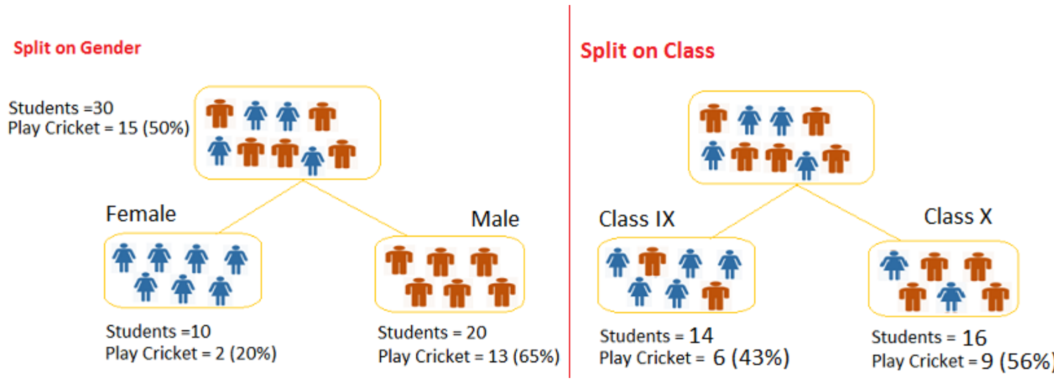
불순도(Impurity)

- 엔트로피 (Entropy)
  - $-p\ \log_2 p - \ (1-p)\log_2(1-p)$
  - A: $-\frac{9}{20}\log_2\frac{9}{20} - \left(1-\frac{9}{20}\right)\log_2\left(1-\frac{9}{20}\right) = 0.9927745$
  - B: $-\frac{3}{15}\log_2\frac{3}{15} - \left(1-\frac{3}{15}\right)\log_2\left(1-\frac{3}{15}\right) = 0.7219281$
  - C:$-\frac{0}{18}\log_2\frac{0}{18} - \left(1-\frac{0}{18}\right)\log_2\left(1-\frac{0}{18}\right) \to 0$





분류 나무 분기 알고리듬

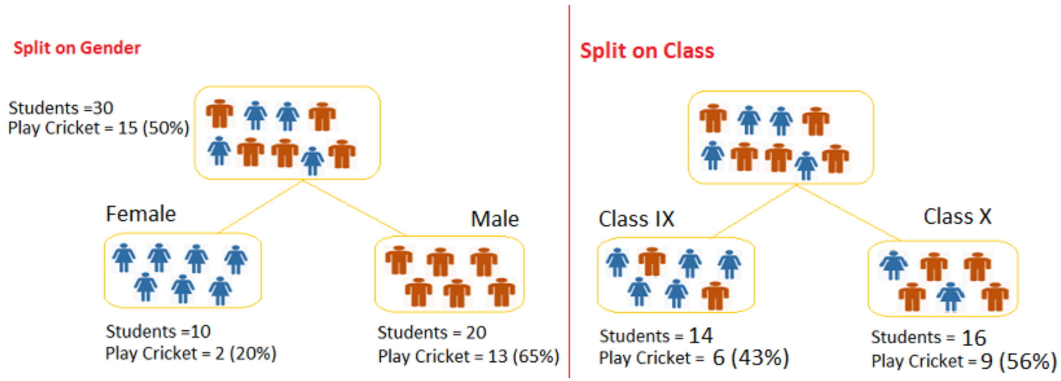


1. 불순도(impurity)가 크게 감소하는 기준을 선택하여 분기(split)
   - 기존 불순도에 비해 분기된 하위 노드들의 불순도가 가장 작아지도록
   - 정보이득(Information Gain): 불순도의 차이

$$\text{Gain} = P(N) - P(N_L, N_R)$$

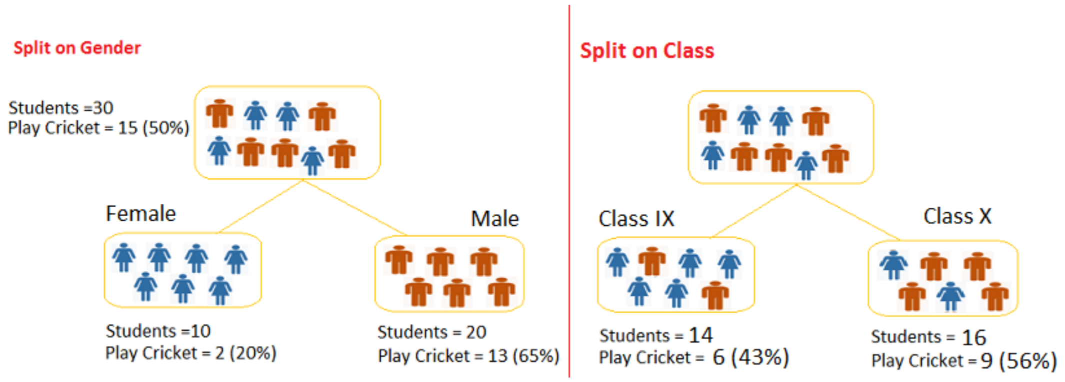
   - $P(N_L, N_R)$: 왼쪽 노드 N_L의 불순도와 오른쪽 노드 N_R의 불순도의 가중 평균

분류 나무 분기 알고리듬



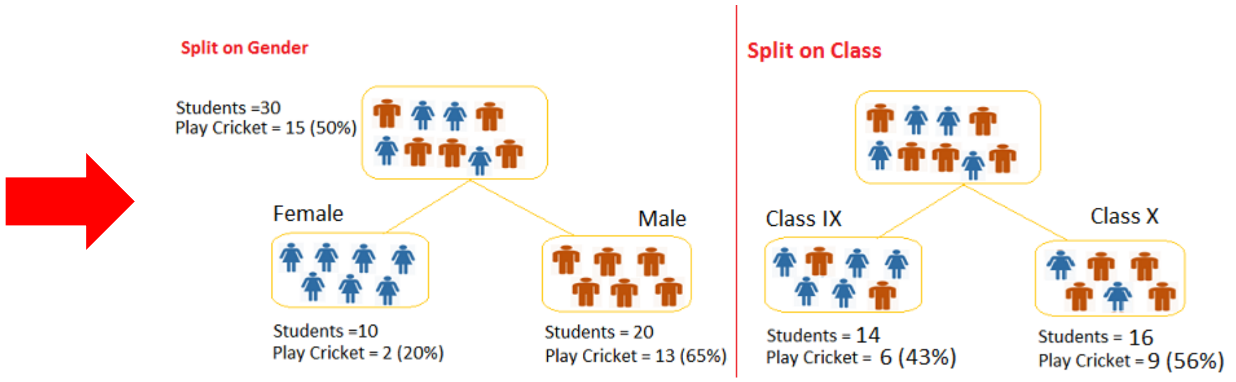
- 엔트로피를 사용하여 Gender 기준으로 분기하는 경우 Gain
  - $P(N) = -\frac{15}{30}\log_2\frac{15}{30} - \left(1-\frac{15}{30}\right)\log_2\left(1-\frac{15}{30}\right) = 1$
  - $P(N_L) = -\frac{2}{10}\log_2\frac{2}{10} - \left(1-\frac{2}{10}\right)\log_2\left(1-\frac{2}{10}\right) = 0.7219281$
  - $P(N_R) = -\frac{13}{20}\log_2\frac{13}{20} - \left(1-\frac{13}{20}\right)\log_2\left(1-\frac{13}{20}\right) = 0.9340681$
  - $P(N_L, N_R) = 0.8633547$
  - $P(N) - P(N_L, N_R) = 1 - 0.8633547 = 0.1366453$

분류 나무 분기 알고리듬



- 엔트로피를 사용하여 Class 기준으로 분기하는 경우 Gain
  - $P(N) = -\frac{15}{30}\log_2\frac{15}{30} - \left(1-\frac{15}{30}\right)\log_2\left(1-\frac{15}{30}\right) = 1$
  - $P(N_L) = -\frac{6}{14}\log_2\frac{6}{14} - \left(1-\frac{6}{14}\right)\log_2\left(1-\frac{6}{14}\right) = 0.9852281$
  - $P(N_R) = -\frac{9}{16}\log_2\frac{9}{16} - \left(1-\frac{9}{16}\right)\log_2\left(1-\frac{9}{16}\right) = 0.9886994$
  - $P(N_L, N_R) = 0.9870795$
  - $P(N) - P(N_L, N_R) = 1 - 0.9870795 = 0.01292052$

분류 나무 분기 알고리듬



2. 어떤 변수로 대해서도 분기에 의해 불순도가 변하지 않을 때까지 재귀적으로 하위 노드를 분기

가지치기(Pruning)

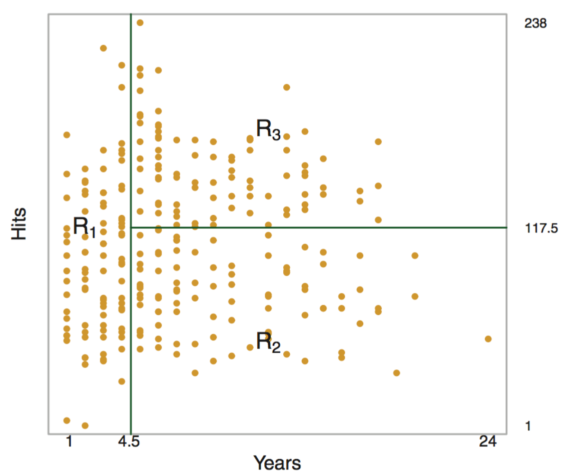
- 과적합을 방지하기 위함
  - 가지가 많다는 것은 훈련 데이터를 상세하게 적합했다는 의미
- 기본 절차
  - 중요도가 낮은 분기를 제거
- 조정 기준
  - cp(complexity parameter): 얼마나 작은 차이를 위해 분기할지
  - 단말 노드의 크기
  - 나무의 높이

가지치기(Pruning)

- cp는 어떻게 결정?
  - cp별 cross validation error (xerror) 사용
  - cross validation error: Leave-One-Out Validation
  - cp를 증가시킬 때 cross validation error가 처음으로 '최솟값 + 표준오차'보다 작아지는 cp 선택
- 처음부터 기준을 정해서 작은 나무를 만들면?
  - 중요한 변수의 영향이 무시될 수 있음
  - 따라서 최대한 큰 나무를 만들고 가지치기를 해야 함

회귀 나무 분기 알고리듬

1. 모든 독립 변수에 대해 오차 제곱의 합(SSE)의 감소가 최대가 되도록 두 영역으로 분기
   - SSE: $\sum_{i \in S_1}(y_i - \bar{y}_1)^2 + \sum_{i \in S_2}(y_i - \bar{y}_2)^2$
   - S_1: 왼쪽으로 분기된 관측값들, $\bar{y}_1$: S_1에 속한 관측값들의 평균
     S_2: 오른쪽으로 분기된 관측값들, $\bar{y}_2$: S_2에 속한 관측값들의 평균
2. SSE의 감소량이 어떤 문턱(threshold) 이하가 될 때까지 하위 노드에 대해 반복



회귀 나무 분기 알고리듬

독립변수(x) 하나와 종속변수(y) 하나로 이루어진 데이터 예시

- S={(2, 7), (3, 12), (4, 15), (5, 16), (6, 15), (7, 12), (8, 7)}
- 전체의 SSE

$$\text{SSE} = \sum_{i \in S} (y_i - \bar{y}_2)^2 = \sum_{i \in S} (y_i - 12)^2 = 84$$

- $S_1 = \{2\}, \; S_2 = \{3, 4, 5, 6, 7, 8\}$

$$\text{SSE} = \sum_{i \in S_1} (y_i - \bar{y}_2)^2 + \sum_{i \in S_2} (y_i - \bar{y}_2)^2 = \sum_{i \in S_2} (y_i - 12.83333)^2 = 54.83333$$

- $S_1 = \{2, 3\}, \; S_2 = \{4, 5, 6, 7, 8\}$

$$\text{SSE} = \sum_{i \in S_1} (y_i - \bar{y}_2)^2 + \sum_{i \in S_2} (y_i - \bar{y}_2)^2 = 66.5$$

- $S_1 = \{2, 3, 4\}, \; S_2 = \{5, 6, 7, 8\}$

$$\text{SSE} = \sum_{i \in S_1} (y_i - \bar{y}_2)^2 + \sum_{i \in S_2} (y_i - \bar{y}_2)^2 = 81.66667$$

회귀 나무 분기 알고리듬

독립변수(x) 하나와 종속변수(y) 하나로 이루어진 데이터 예시

- S={(2, 7), (3, 12), (4, 15), (5, 16), (6, 15), (7, 12), (8, 7)}
- $S_1 = \{2, 3, 4, 5\},\ S_2 = \{6, 7, 8\}$

$$\mathrm{SSE} = \sum_{i \in S_1} (y_i - \bar{y}_2)^2 + \sum_{i \in S_2} (y_i - \bar{y}_2)^2 = 81.66667$$

- $S_1 = \{2, 3, 4, 5, 6\},\ S_2 = \{7, 8\}$

$$\mathrm{SSE} = \sum_{i \in S_1} (y_i - \bar{y}_2)^2 + \sum_{i \in S_2} (y_i - \bar{y}_2)^2 = 66.5$$

- $S_1 = \{2, 3, 4, 5, 6, 7\},\ S_2 = \{8\}$

$$\mathrm{SSE} = \sum_{i \in S_1} (y_i - \bar{y}_2)^2 + \sum_{i \in S_2} (y_i - \bar{y}_2)^2 = 54.83333$$