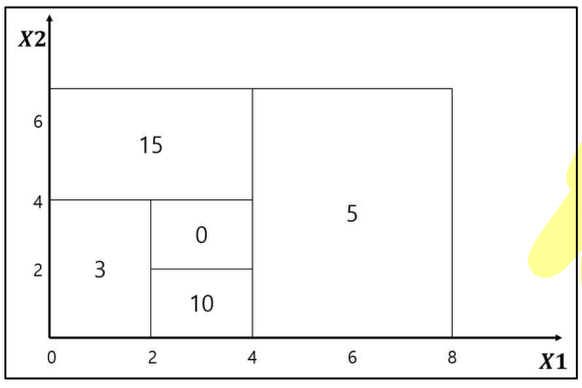
1. scoredEX.csv 자료는 어느 훈련 데이터를 이용하여 이진 분류 알고리즘을 학습한 다음, 평가 데이터(test data)에 대해 각 관찰치 별 특징변수값 (x1 ~ x29), 목표변수값(y_test, 0과 1로 입력됨)과 알고리즘을 적용했을 때의 y=1에 대한 예측확률 추정치(p_pred)를 기록한 것이다.

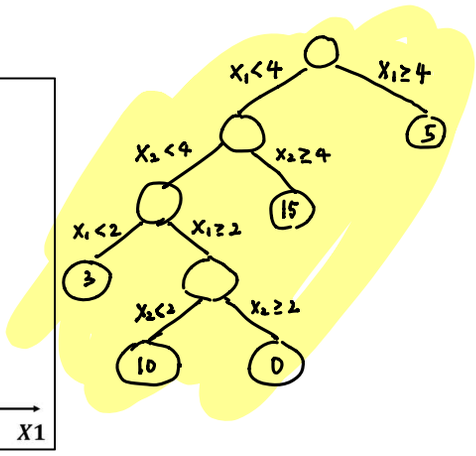
   (1) 이 자료를 이용하여, 가능한 분류임계치(threshold) 별 정밀도와 재현율을 계산한 뒤, 그 관계를 나타내는 정밀도-재현율 그림(가로축:재현율, 세로축:정밀도)을 그려라. 단, sklearn의 confusion_matrix 함수를 제외하고는 sklearn의 기능을 이용하지 말 것.

   (2) (1)에서 구한 각 분류임계치 별 정밀도과 재현율을 이용하여, 각 분류임계치 별 f1_score를 계산하여라. f1_score로 평가한다고 했을 때 주어진 데이터를 이진분류하기 위한 최선의 분류임계치는 무엇인가?

   (3) (2)에서 찾은 분류임계치로 분류한 결과를 y_pred라는 변수로 기존의 데이터 셋의 마지막 열로 추가하여라. 또한 y_test와 y_pred를 이용하여 정확도(accuracy)를 계산하여라.

2. 다음은 2개의 특징변수 $X1$과 $X2$가 만들어내는 공간이 어떤 이진(binary) 분할 트리 알고리즘에 의해 어떻게 분할되어 있는지를 요약하는 그림이다. 이 특징변수 공간의 분할에 대응하는 트리를 스케치하여라. 각 부분공간 안의 숫자들은 각 영역 내 목표변수 Y의 평균값에 해당한다.





3. 다음과 같은 자료가 부모마디에 주어졌다고 하자. X 변수를 기준으로, $X < c$와 $X \geq c$의 분리기준을 적용하여 왼쪽자식마디와 오른쪽 자식마디로 구분하려고 한다. c=5 경우와 c=7인 경우 중 지니 불순도가 어느 쪽이 더 개선되는지 판단하여라.

X	Y
1	0

3	0
3	1
4	0
5	0
5	1
6	0
6	1
7	1
8	1
9	1

4. Churn_Modelling.csv 자료는 일정 기간 동안 어느 은행의 고객에 대한 정보와 고객이탈 여부를 포함하고 있다. 목표변수는 Exited로 0이면 유지, 1이면 이탈을 의미한다. 나머지 변수들 중 CustomerID, RowNumber, Surname 등은 개별 샘플의 고유값에 해당하므로 분석에서 제외하고 나머지 Age, Balance, CreditScore, EstimatedSalary, NumberOfProducts, Tenure, Gender, HasCrCard, IsActiveMember를 특징변수로 둘 수 있다.

   (1) 주어진 자료는 모두 훈련용 자료로 둘 것. 그 중 70%는 모델의 적합(training), 30%는 검증 및 튜닝(validation) 용으로 랜덤하게 분리하여라.

   (2) (1)에서의 training 자료를 이용하여 CART 의사결정나무 알고리즘을 적합하여라.

   - 적합 시 나무의 최대깊이(max_depth)를 얼마로 하는 것이 좋을지 validation 자료를 이용하여 결정할 것.
   - Validation을 위한 분류 성능의 평가는 F1스코어를 이용할 것.
   - 의사결정나무 적합 결과를 그림으로 시각화한 뒤 해석해 볼 것.

   (3) (2)의 결과에서 각 특징변수 별 (불순도의 개선정도 측면에서) 중요도를 구하고, 가장 중요한 변수부터 순서대로 3개를 선택하여라.