정규세션 4주차

ToBig's 11기 김유민

Decision Tree

Assignment

Assignment

[과제 1]

DT_Assignment1.ipynb를 켜주세요.

본인이 구현한 함수를 통해 다음 문제를 풀어주세요!

문제 1) 변수 'income'의 이진분류 결과를 보여주세요.

문제 2) 분류를 하는 데 가장 중요한 변수를 선정하고, 해당 변수의 Gini index를 제시해주세요.

문제3) 문제 2에서 제시한 feature로 DataFrame을 split한 후,
나눠진 2개의 DataFrame에서 각각 다음으로 중요한 변수를 선정하고 해당 변수의 Gini index를 제시해주세요.

Assignment

[과제 1_Hint]

get_gini(df, label) → $Gini(D_i) = 1 - \sum_{j=1}^{3} P_j$

get_attribute_gini_index(df, attribute, label) → $Gini(A) = \sum_{j=1}^{2} \frac{|D_j|}{|D|} * Gini(D_i)$

Assignment

[과제 2] **ID3 알고리즘을 통해 아래 데이터의 최초 split feature가 무엇인지 도출하는 과정을 계산해주세요!**

age	*income*	*student*	*credit_rating*	*Class: buys_computer*
youth	high	no	fair	no
youth	high	no	excellent	no
middle_aged	high	no	fair	yes
senior	medium	no	fair	yes
senior	low	yes	fair	yes
senior	low	yes	excellent	no
middle_aged	low	yes	excellent	yes
youth	medium	no	fair	no
youth	low	yes	fair	yes
senior	medium	yes	fair	yes
youth	medium	yes	excellent	yes
middle_aged	medium	no	excellent	yes
middle_aged	high	yes	fair	yes
senior	medium	no	excellent	no

Q. 어떻게 제출해요?

1안: 손으로 공책에 푼다

-> 카메라로 찍는다

-> 이미지를 첨부한다

2안: 손으로 공책에 푼다

-> 풀어낸 수식을 파이썬 코드로 표현한다

-> 컬러스크립터를 첨부한다

Assignment

[주의사항]

/과제 1/

✓ 본인이 직접 구현한 함수임을 증명하기 위해 줄 별로 어떤 과정을 의미하는지 꼭 주석을 달아주세요!

✓ 이 데이터 셋에만 적용되는 함수가 아닌, 데이터에 상관없이 적용 가능하도록 함수를 구현해주세요!

(ex) a, b, c, d => ({a},{b, c, d})/({a, b},{c, d})…. 등 변수의 class가 3개 이상일 수도 있습니다.

✓ 함수에 들어가는 변수나 flow는 본인이 변경해도 무관하며, 결과만 똑같이 나오면 됩니다!

/과제 2/

✓ 제출 방식은 1안이나 2안 둘 중 편한 방법을 선택해 제출하시면 됩니다.

✓ ID3 알고리즘 설명에 사용되었던 예제 데이터와 다르게, 과제2 데이터는 class가 3개 이상인 변수도 있다는 점 주의하세요!