Decision Tree Node Explanation

# 1. 조건 (Condition): marital\_status\_Married-civ-spouse <= 0.5

이론:

결정 트리에서의 조건은 특정 특징(피처)을 기준으로 데이터를 분할하는 역할을 합니다. 각 노드는 하나의 조건을 가지고 있으며, 이 조건은 데이터셋을 두 개 이상의 하위 집합으로 나눕니다. 이 조건은 특정 피처의 값을 기준으로 트리의 가지(branch)를 선택하는 기준이 됩니다.

수식:

분할 기준은 일반적으로 엔트로피(Entropy), 지니 계수(Gini Index), 혹은 정보 이득(Information Gain)을 사용하여 최적의 분할을 찾습니다. 예를 들어, 지니 계수는 다음과 같이 계산됩니다:

Gini(t) = 1 - Σ(p\_i^2)

여기서 p\_i는 특정 클래스에 속할 확률입니다. 결정 트리는 이러한 수식을 기반으로 각 노드에서 최적의 조건을 찾아 데이터를 분할합니다.

사례:

만약 데이터에서 marital\_status\_Married-civ-spouse라는 피처가 결혼 여부를 나타내며, 1은 결혼, 0은 미혼을 의미한다면, 이 조건은 "해당 개인이 결혼했는가?"를 묻는 것과 같습니다. 트리는 이 질문에 따라 결혼한 사람과 결혼하지 않은 사람으로 데이터를 분리하게 됩니다.

# 2. 샘플 (Samples): samples = 100.0%

이론:

샘플은 현재 노드에서 처리 중인 전체 데이터의 비율을 나타냅니다. 초기 루트 노드에서는 전체 데이터가 100%이며, 이후 노드에서 데이터가 분할될수록 샘플의 비율은 줄어듭니다.

수식:

노드에서의 샘플 비율은 간단히 전체 데이터 중 해당 노드에 속하는 데이터의 수를 비율로 표현한 것입니다.

Sample percentage = (Number of samples at node / Total number of samples) × 100

사례:

만약 루트 노드에서 100개의 샘플이 시작되었다고 가정하면, 특정 조건을 통해 40개의 샘플이 한 노드로 분할되고, 60개의 샘플이 다른 노드로 분할될 수 있습니다. 이 경우, 샘플 비율은 각각 40%와 60%로 나타날 것입니다.

# 3. 값 (Value): value = [0.239, 0.761]

이론:

이 값은 노드에 도달한 데이터가 각 클래스에 속하는 비율을 나타냅니다. 각 클래스의 비율은 해당 노드의 샘플 중 특정 클래스에 속하는 데이터의 비율로 계산됩니다.

수식:

클래스 비율은 다음과 같이 계산됩니다:

Class probability = (Number of samples in class / Total number of samples at node)

예를 들어, 노드에 100개의 샘플이 있고 그 중 24개가 첫 번째 클래스에, 76개가 두 번째 클래스에 속한다면, 비율은 [0.24, 0.76]이 됩니다.

사례:

예를 들어, 의료 데이터에서 특정 노드에 도달한 환자들 중 23.9%가 질병 A를 가지고 있고, 76.1%가 질병 B를 가지고 있는 경우, value = [0.239, 0.761]로 표현됩니다. 이 값은 현재 노드의 데이터를 분석하여 어떤 클래스가 더 지배적인지를 파악하는 데 유용합니다.

# 4. 클래스 (Class): class = low

이론:

클래스는 해당 노드에서 예측된 최종 결과를 나타냅니다. 노드에 있는 데이터의 대부분이 어느 클래스에 속하는지를 기반으로 클래스가 결정됩니다.

수식:

클래스는 노드에서 가장 높은 비율을 차지하는 클래스가 선택됩니다. 만약 value = [0.239, 0.761]이라면, 76.1%로 더 큰 비율을 차지하는 두 번째 클래스가 최종 클래스가 됩니다.

사례:

예를 들어, 소비자의 구매 행동을 예측하는 모델에서 class = low는 "이 소비자는 낮은 가격대의 상품을 선호할 가능성이 크다"라는 결론을 나타낼 수 있습니다. 이는 노드에 도달한 데이터가 주로 낮은 가격대를 선호하는 소비자들로 구성되어 있다는 의미입니다.