교육빅데이터의 이해와 분석(14주차)

- 머신러닝 맛보기

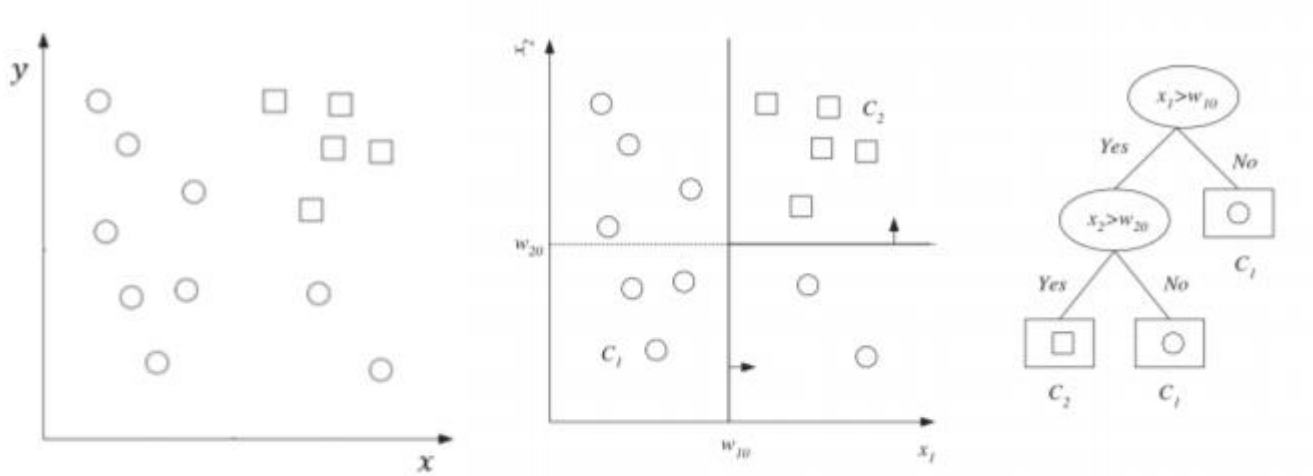
읽기 자료 읽어보기

의사결정나무분석(Decision Tree)과 앙상블 방법(Ensanble)

IT기술의 발달로 데이터의 크기, 다양성, 생성속도가 증가함에 따라 과거 할 수 없었던 대량의 데이터를 학습하여 미래를 예측하는 머신러닝이 많은 주목을 받고 있다. 머신러닝은 학습 데이터에 레이블이 있는 경우와 그렇지 않은 경우에 따라 각각 지도학습(Supervised Learning)과 비지도학습(Unsupervised Learning)으로 구분한다. 레이블(Label)이라는 것은 학습 데이터의 속성을 연구자가 분석하고자 하는 관점에 따라 정의하는 것이다. 예를 들어, 우리 주변에 있는 사물을 찍은 사진 중에서 어떤 사물이 있는지 구분하는 과제가 있다면, 갖고 있는 사진을 학습 데이터라 하고 사진 속에 있는 사물의 이름을 레이블이라고 한다. 레이블은 연구자가 사진을 보고 정의한 것이기 때문에, 컴퓨터 입장에서 레이블된 사진을 읽어 학습하는 것이 사람에게 지도(Supervised)를 받아 학습하는 것이 된다. 반면 학습 데이터에 레이블이 없다면 컴퓨터가 사람으로부터 지도받지 않았기 때문에 비지도학습이라 한다. 지도학습에는 분류와 회귀가, 비지도학습에는 군집이 대표적이다. 그중 지도학습의 의사결정나무(Decision Tree)는 대표적인 분류 모형으로서 해석력이 높으며, 나무 형태의 시각화된 그래프가 학습 결과로 도출되기 때문에 사회 전 분야에 널리 활용되고 있다.

의사결정나무분석_읽기 자료.pdf

읽기 자료 읽어보기



이 부분 손으로 계산 꼭 해보기

- 계산하신 부분 사진 찍어 올려주세요.

$$Gini(w_{10}) = [1-(\frac{7}{7})^2-(\frac{0}{7})^2]\times\frac{7}{14}+[1-(\frac{2}{7})^2-(\frac{5}{7})^2]\times\frac{7}{14}=0.204$$

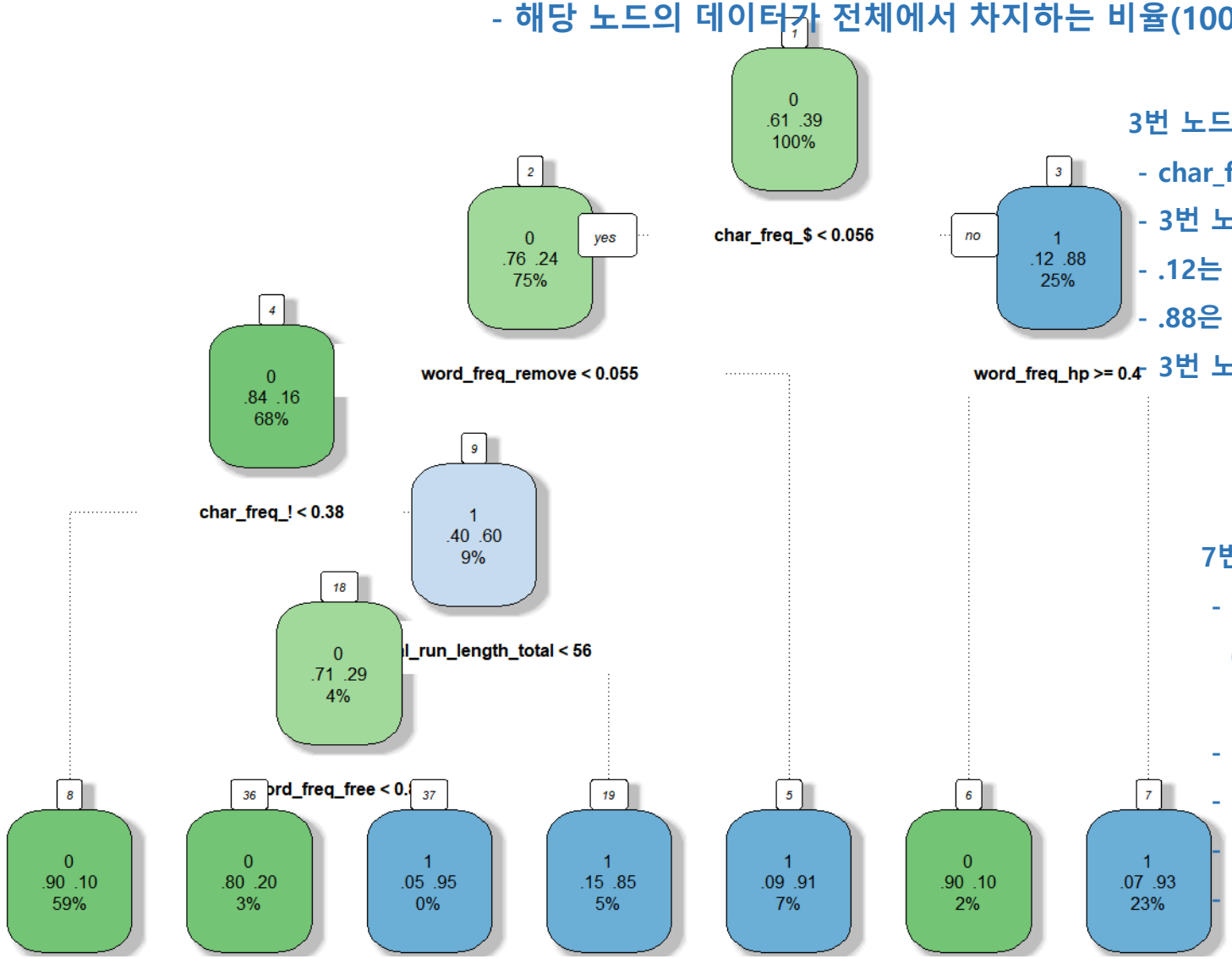
$$Gini(w_{20}) = [1-(\frac{5}{5})^2-(\frac{0}{5})^2]\times\frac{5}{14}+[1-(\frac{4}{9})^2-(\frac{5}{9})^2]\times\frac{9}{14}=0.317$$

동영상 보며 따라가보기

영상 보시면서 한번 쓰윽 정도로 들으세요~

결과 해석하기

1번 노드: 기본 클래스 라벨은 0(스펨메일 아님)
- 왼쪽이 해당 노드에서 클래스 라벨 0을 갖는 데이터의 비율(.61)
- 오른쪽이 해당 노드에서 클래스 라벨 1을 갖는 데이터의 비율(.39)
- 해당 노드의 데이터가 전체에서 차지하는 비율(100%)



3번 노드: 기본 클래스 라벨은 1(스펨메일 맞음)
- char_freq_$ < 0.056 (메일에서 $문자의 개수가 0.056 보다 크거나 같으면)
- 3번 노드로 분류
- .12는 3번 노드에서 클래스 라벨 0을 갖는 데이터의 비율
- .88은 3번 노드에서 클래스 라벨 1을 갖는 데이터의 비율
- 3번 노드의 데이터가 전체에서 차지하는 비율(25%)

7번 노드: 기본 클래스 라벨은 1(스펨메일 맞음)
- char_freq_$ < 0.056: No and word_freq_hp >= 0.4: No
  (메일에서 $문자의 개수가 0.056 보다 크거나 같고,
  hp라는 단어의 개수가 0.4보다 작으면)
- 7번 노드로 분류
- .07은 7번 노드에서 클래스 라벨 0을 갖는 데이터의 비율
- .93은 7번 노드에서 클래스 라벨 1을 갖는 데이터의 비율
- 7번 노드의 데이터가 전체에서 차지하는 비율(23%)