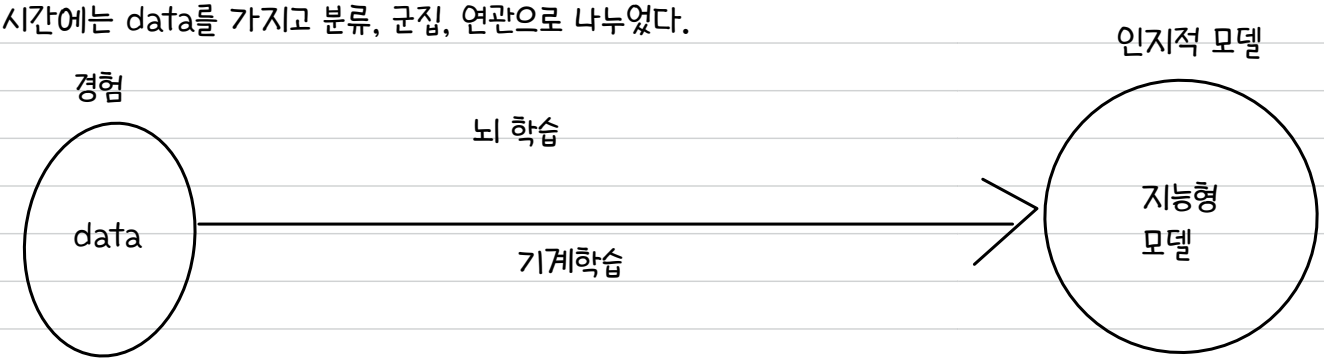
실습한 내용은 퀴즈로 나누지 않는다.
수업은 12/15일까지 있다.

지난 시간에는 data를 가지고 분류, 군집, 연관으로 나누었다.

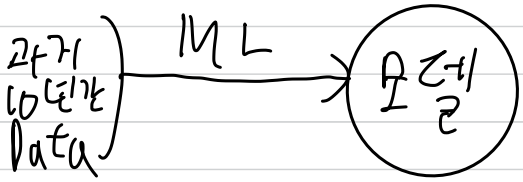


- 예측(prediction)
- 분류(classification)
- 클러스터링 (군집)
- 연관(association)

예) 통신 고객
졸업을 반도 못하는 학교가 있더라.
이유가 무엇인가?
그리고 학생이 들어오면 학생에 대한 정보를 바탕으로
졸업을 할 수 있을지, 없을지 뭐가 필요한 지를 찾는 것이다.

속성	분류
~	O
~	X
~	O
~	X

1. 새로운 학생의 미래를, 졸업을 할지, 퇴학을 당할지를 예측하고 싶다.
2. 졸업을 하는 학생들의 특징을 알고 싶다. 퇴학을 당하는 학생들의 특징을 알고 싶다.
(즉, 원인을 설명하고 싶은 것이다.)



이 학생을 상담하게 되는 것이다.
모델은 이런 특징들을 가진 학생이 2학년때에 재적당할 확률이 75%이다.
라는 것을 알려주는 것이다.

교수님은 분류 모델을 만든 것이다.

문제는 상당히 민감한 정보들이 들어갈 수 있다는 것이다.
누군가가 나의 정보를 수집해서 그것으로 나의 졸업 가능성을 판단하는 것은 문제가 될 수 있다.
이는 기계학습 모델의 공정성 문제도 발생가능하다.

교수님의 모델은 인종을 가지고 판단을 하였다.
이는 공정성 문제로, 매우 민감하다.

군집 :
속성의 관점에서, 특정한 그룹이 생기더라.

연관:
속성들 간의 숨겨진 연관성을 찾아내는 것이다.
물리 성적이 높은 애들이 거의 예외 없이 기숙사에 살더라 등

숙제:
학교급식에서 모든 학생들이 매일 어떠한 메뉴를 선택했는지에 대해
각 학생에 대한 정보들이 수집되어 있다.
특정 학생의 관점에서 보면 1학년 입학 때부터 선택했던 메뉴들이 수집되어 있다.
그 학생은 이제 3학년임. 여전히 급식이 한 학기가 남아 있음

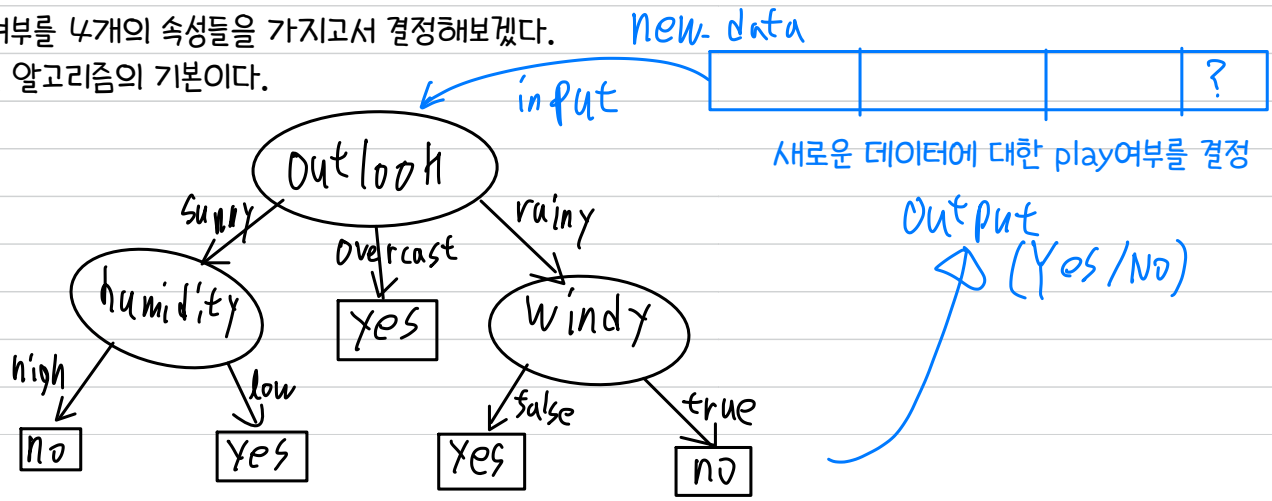
데이터가 쌓여있음
각 학생이 선택한 메뉴 데이터가 있음
각 메뉴에 대한 재료 데이터가 있고,
각 재료에 대한 영양 성분 데이터가 있다.

문제:
이 데이터를 가지고 상사가
"당신은 이 데이터를 가지고 분류/군집/연관의 방법을 사용하여
어떤 것을 만들어 올 수 있는지 시나리오/스토리/아이디어를 내서
내게 제출해주세요."
급식의 효율성과 학생들의 건강에 도움을 주는지 여부에 활용하고자 합니다.

매일 약 3~5가지 정도의 메뉴가 제공되고,
일년 동안 나가는 메뉴의 전체 종류수는 100가지 라고 가정합시다.

기계학습 알고리즘
분류화 : 결정트리 (decision tree)

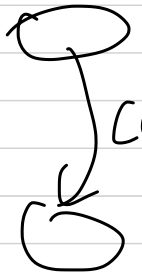
게임을 할지 안할지의 여부를 4개의 속성들을 가지고서 결정해보겠다.
: 이것이 기계학습의 모든 알고리즘의 기본이다.



그렇다면 이런 decid'ion tree를 어떻게 만드는가?
이는 스무 고개의 원리와 같다.
yes or no에 대한 질문을 20번만하면 도메인을 2^{20} 으로 나눌 수 있다.

여기서 중요한 것은 domain을 잘 나누는 질문을 해야 한다는 것이다.
outlook, humid 둘 중 무엇을 물어보는 것이 더 많은 정보를 우리에게 주는가?

entropy를 계산하는 것이다.
class로 나누는데, size, color, surface 중 어떤 것이 가장 큰 도움을 줄까?
surface가 smooth하다 : A:3/5, B:2/5이다.
color가 red이다. : A:3/3이다.
이는 혼란스럽지 않다. color가 red이면 무조건 A가 된다.



: entropy의 차이만큼 정보를 얻는 것이다.

각각의 속성에 대한 entropy를 계산하여, entropy가 가장 작은 것으로 node를 만드는 것이다.

아무런 정보도 없이 학생의 졸업/제적 여부를 물어본다면?
lg(2) bit= 1 bit이다.

정보 : 80% 졸업, 20% 제적
lg(80/100) + lg(20/100) =0.6xxxx

정보 : 물리성적
entropy가 낮춰지는지의 여부를 본다.