머신러닝

Week-05. 데이터 표현과 특성 공학

Jungwon Seo, 2021-Spring

원본 데이터와 모델은 동일한데 성능이 다르다?

범주형 변수
Categorical Variable

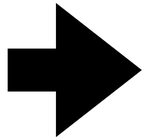
- 모든 모델은 수학적인 연산을 통해서 학습이 된다
- 그렇다면 범주형 변수가 숫자가 아닌 경우에는 어떻게 학습이 될까?
  - 숫자로 만들어준다
- 단순히 A, B, AB, O => 0, 1, 2, 3?
- 인코딩 : 사용자가 입력한 문자나 기호를 컴퓨터가 이해할 수 있는 형태로 만드는 것
- Ordinal Encoding, One-hot Encoding, Binary Encoding

범주형 변수 인코딩

Ordinal Encoding

- 인코딩 : 사용자가 입력한 문자나 기호를 컴퓨터가 이해할 수 있는 형태로 만드는 것
- Ordinal Encoding: 순서가 있는 범주형 데이터
  - 인턴:0, 사원: 1, 주임:2, 대리: 3, …
- 가장 간단하고 기존 속성과 1:1 mapping이 가능하기 때문에 불필요한 데이터 사이즈 증가가 발생하지 않음

ID	성별	직급
1	여	인턴
2	남	사원
3	여	주임
4	남	대리



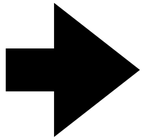
ID	성별	직급
1	여	0
2	남	1
3	여	2
4	남	3

범주형 변수 인코딩

One-hot Encoding

- 순서가 없는 범주형 데이터를 인코딩 할때 주로 사용
  - 혈액형
  - A,B, AB, O
  - [1,0,0,0] , [0,1,0,0], [0,0,1,0], [0,0,0,1]
- 해당 속성에 존재하는 속성 값의 종류 만큼 데이터 속성이 추가됨

ID	성별	혈액형
1	여	A
2	남	B
3	여	AB
4	남	O



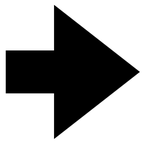
ID	성별	혈액형0	혈액형1	혈액형2	혈액형3
1	여	1	0	0	0
2	남	0	1	0	0
3	여	0	0	1	0
4	남	0	0	0	1

범주형 변수 인코딩

Binary Encoding

- Ordinal 인코딩한 결과를 2진수로 표현
- One-hot Encoding 처럼 feature 수가 급격히 늘어나는걸 방지
  - 200개의 속성값을 표현하기 위해 one-hot encoding은 200개의 feature를 생성
  - 2진수 형태로 표현 할 경우 8개의 feature로 표현 가능

ID	성별	혈액형
1	여	A
2	남	B
3	여	AB
4	남	O



ID	성별	혈액형0	혈액형1
1	여	0	0
2	남	0	1
3	여	1	0
4	남	1	1

범주형 변수 인코딩

모든 속성 값을 반드시 인코딩 해야할까?

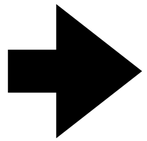
- 만약 속성 값들이 균등하게 분포가 되어 있는 데이터라면 해야한다!
- 그런데 실제 데이터의 경우 아닌 경우가 많다
  - 연세대학교에 다니는 학생들의 데이터 중 국적 feature 중에 모든 나라를 표시하는게 반드시 좋을까?
  - 한국인/외국인 정도로만 구별해도 되지 않을까?
- 이 과정에서 빈도수를 기준으로 새로운 속성을 생성
  - Feature engineering!

특성공학

기존의 특성을 활용한 특성 생성 및 대체

- 성적관련 데이터가 있을때, 추가적인 특성을 어떻게 만들 수 있을까?
- 특성 공학은 데이터/테스크 도메인 지식에 따라 다양하게 진행을 할 수 있습니다

	과목	문제	정답여부
1	수학	1	0
2	영어	1	0
3	수학	3	1
4	영어	2	1



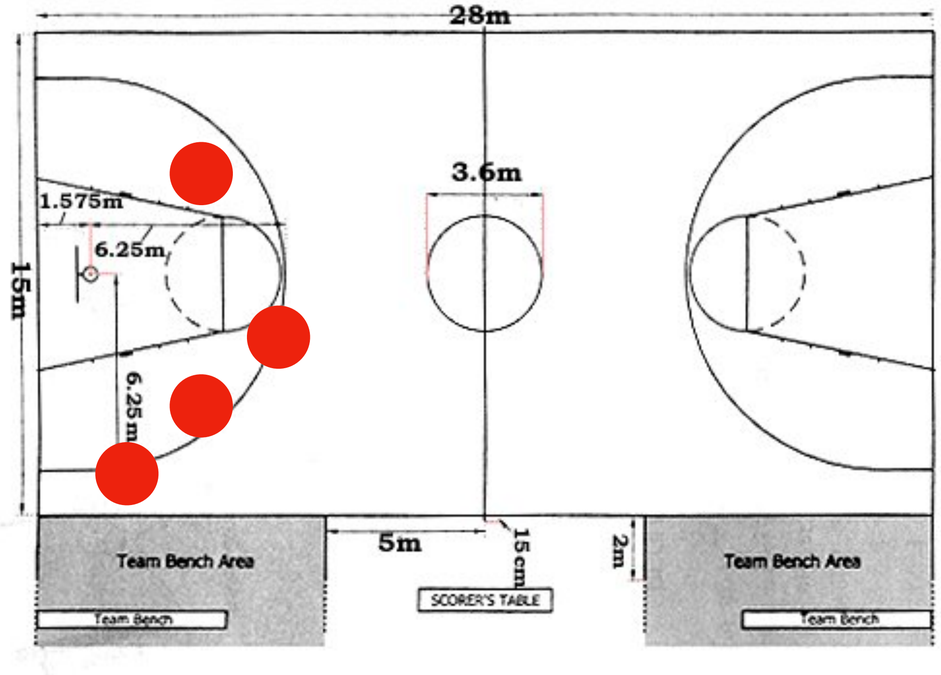
	과목	문제	정답여부	정답률
1	수학	1	0	0
2	영어	1	0	0.5
3	수학	3	1	1
4	영어	1	1	0.5

특성공학

기존의 특성을 활용한 특성 생성 및 대체

- 슛을 쏜 위치의 데이터가 있을때 이를 기반으로 어떠한 특성을 생성 할 수 있을까?

	선수	X	Y
1	A	1.5	0.1
2	B	2.3	1
3	C	4.4	3.3
4	D	2	10



데이터 분리

데이터를 어떻게 분리해야 효과적일까?

- 훈련/테스트 셋을 분리를 어떻게 하느냐도 훈련시 모델 성능에 크게 영향을 미칠 수 있음
  - 더 정확하게 표현하면, 일반화된 모델을 얼마나 잘 만드느냐와 관련이 있음
- 데이터 분리 관련 기술
  - Random Sampling: 임의의 n개를 전체 데이터에서 선택
  - Stratified Sampling: 특정 특성을 기준으로 비율을 유지한 채로 선택 (각 도시의 10% 인구 데이터)
  - 데이터가 충분하고 고르게 분포가 되어있다면, 결국엔 비슷한 비율을 유지

원본데이터 (Black: Class1, Red: Class2)

Random

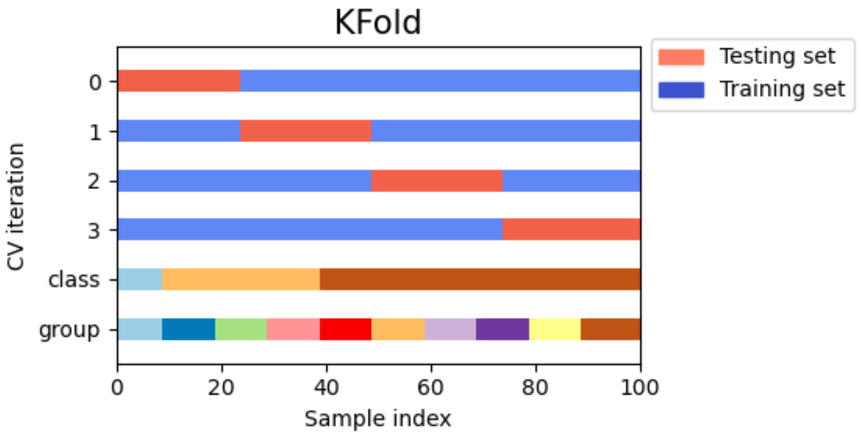
Stratified

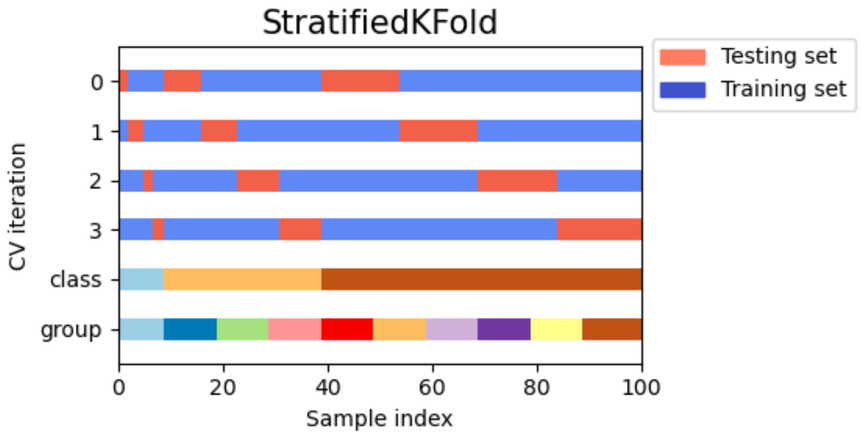
Test 셋

모델 검증

모델 검증을 편향되지 않게 하기 위해서는?

- 데이터 셋을 여러번 나눠서 검증을 진행
  - 테스트 데이터셋의 분포를 다르게 가져감으로써 더욱 일반화된 모델을 선택
  - K-fold Cross Validation: K-겹 교차검증
    - 전체 데이터를 K 번 다른 방식으로 훈련/테스트 데이터 셋을 분리하여, 모델의 훈련 / 검증을 진행
    - 전체 평균이 높은 모델을 선택하는 식으로 일반화된 모델을 선별





https://scikit-learn.org/stable/modules/cross_validation.html

E.O.D