Chapter. 18

문자보다는 숫자: 범주형 변수 문제

문제 정의 및 해결 방법

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

I 문제 정의

- 데이터에 범주형 변수가 포함되어 있어, 대다수의 지도 학습 모델이 학습되지 않거나 **비정상적**으로 학습되는 문제를 의미
  - ➢ str 타입의 범주형 변수가 포함되면 대다수의 지도 학습 모델 자체가 학습이 되지 않음
  - ➢ int 혹은 float 타입의 범주형 변수는 모델 학습은 되나, 비정상적으로 학습이 되지만, 입문자는 이를 놓치는 경우가 종종 있음

- 모델 학습을 위해 범주형 변수는 **반드시 숫자로 변환**되어야 하지만, 임의로 설정하는 것은 매우 부적절함
  - ➢ (예시) 종교 변수: 기독교 = 1, 불교 = 2, 천주교 = 3
    불교는 기독교의 2배라는 등의 대수 관계가 실제로 존재하지 않지만, 이처럼 변환하면 비정상적인 관계가 생성됨

- 특히, **코드화된 범주형 변수**도 적절한 숫자로 변환해줘야 함

I 범주형 변수 판별

- 범주형 변수는 **상태 공간의 크기가 유한**한 변수를 의미하며, 반드시 도메인이나 변수의 상태 공간을 바탕으로 판단해야 함
- int 혹은 float 타입으로 정의된 변수는 반드시 연속형 변수가 아닐 수 있다는 점에 주의해야 함
- (예시) 월(month)은 비록 숫자지만 범주형 변수임

범주형 변수 변환 방법 (1) 더미화

- 가장 일반적인 범주형 변수를 변환하는 방법으로, 범주형 변수가 특정 값을 취하는지 여부를 나타내는 더미 변수를 생성하는 방법

#1의 종교 변수가 기독교 값을 취하므로, 기독교 변수가 1을 가짐

불교 변수는 나머지 변수로 완벽히 추론 가능하므로 변수간 상관성 제거 및 계산량 감소를 위해 제거

레코드	종교
#1	기독교
#2	천주교
#3	불교
#4	기독교
#5	기독교
#6	천주교

더미화 →

레코드	기독교	천주교	불교
#1	1	0	0
#2	0	1	0
#3	0	0	1
#4	1	0	0
#5	1	0	0
#6	0	1	0

#1의 종교 변수가 기독교 값을 취하지 않으므로, 기독교 변수가 0을 가짐

I 범주형 변수 변환 방법 (2) 연속형 변수로 치환

- 범주형 변수의 상태 공간 크기가 클 때, 더미화는 과하게 많은 변수를 추가해서 차원의 저주 문제로 이어질 수 있음

- 라벨 정보를 활용하여 범주 변수를 연속형 변수로 치환하면 기존 변수가 가지는 정보가 일부 손실될 수 있고 활용이 어렵다는 단점이 있으나, 차원의 크기가 변하지 않으며 더 효율적인 변수로 변환할 수 있다는 장점이 있음

X	Y
A	100
B	200
A	300
C	400
B	500
C	600

X에 따른 Y의 평균 계산 →

X	Y
A	200
B	350
C	500

치환 →

X	Y
200	100
350	200
200	300
500	400
350	500
500	600

Chapter. 18

문자보다는 숫자: 범주형 변수 문제

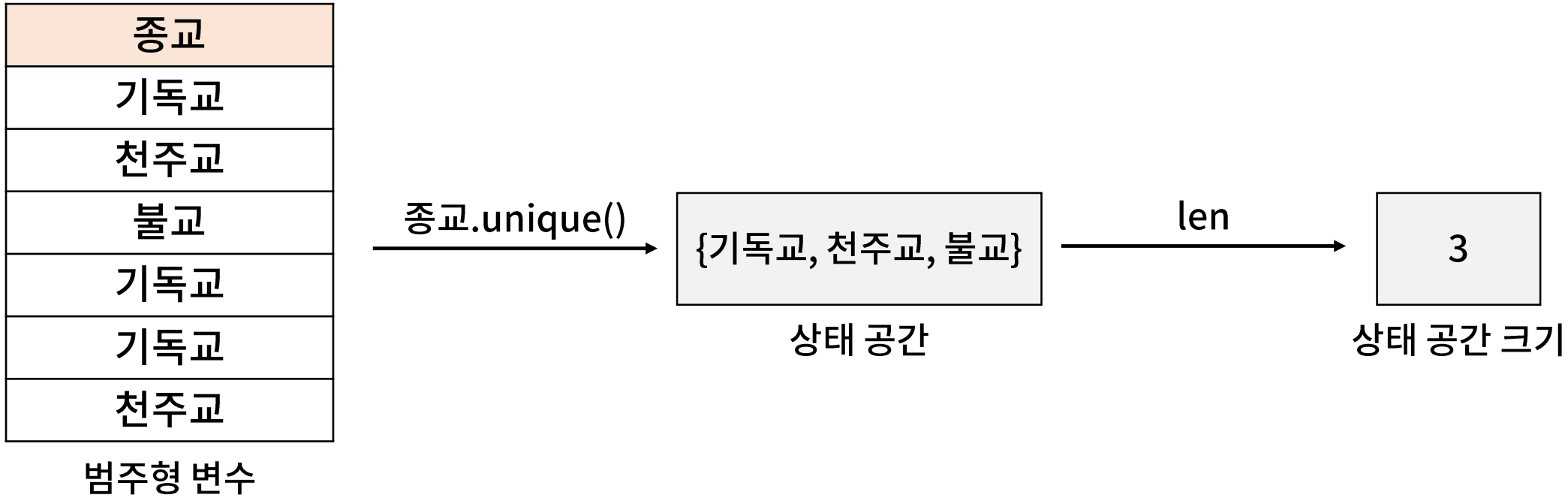
관련 문법 및 실습

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

Series.unique()

- Series에 포함된 유니크한 값을 반환해주는 함수로, **상태 공간을 확인**하는데 사용



feature_engine.categorical_encoders.OneHotCategoricalEncoder

- 더미화를 하기 위한 함수로, 활용 방법은 sklearn의 인스턴스의 활용 방법과 유사함

- 주요 입력
  - variables: 더미화 대상이 되는 범주형 변수의 이름 목록 (주의: 해당 변수는 반드시 str 타입이어야 함)
  - drop_last: 한 범주 변수로부터 만든 더미 변수 가운데 마지막 더미 변수를 제거할 지를 결정
  - top_categories: 한 범주 변수로부터 만드는 더미 변수 개수를 설정하며, 빈도 기준으로 자름

- 참고: pandas.get_dummies()는 이 함수보다 사용이 훨씬 간단하지만, 학습 데이터에 포함된 범주형 변수를 처리한 방식으로 새로 들어온 데이터에 적용이 불가능하기 때문에, 실제적으로 활용이 어려움

Chapter.

문자보다는 숫자: 범주형 변수 문제

감사합니다

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승