데이터 전처리 - 인코딩과 스케일링

[데이터 전처리(Preprocessing)]

- 데이터 클린징
- 결손값 처리(Null/NaN 처리)
- 데이터 인코딩(레이블, 원-핫 인코딩)
- 데이터 스케일링
- 이상치 제거
- Feature 선택, 추출 및 가공

[데이터 인코딩]

머신러닝 알고리즘은 문자열 데이터 속성 입력 받지 않음. 모든 데이터는 숫자형으로 표현되어야 함. 문자형 카테고리형 속성은 모두 숫자 값으로 변환/인코딩 되어야 함.

- 레이블 인코딩

원본 데이터

상품 분류	가격
TV	1,000,000
냉장고	1,500,000
전자렌지	200,000
컴퓨터	800,000
선풍기	100,000
선풍기	100,000
믹서	50,000
믹서	50,000

상품 분류를 레이블 인코딩 한 데이터

상품 분류	가격
0	1,000,000
1	1,500,000
4	200,000
5	800,000
3	100,000
3	100,000
2	50,000
2	50,000

[TV, 냉장고, 전자레인지, 컴퓨터, 선풍기, 믹서] ➡ [0, 1, 4, 5, 3, 2]

- 원-핫(One-Hot) 인코딩
  - 피처 값의 유형에 따라 새로운 피처를 추가해 고유값에 해당하는 컬럼에만 1을 표시하고 나머지 컬럼에는 0을 표시하는 방식

TV:0, 냉장고:1, 전자레인지:4, 컴퓨터:5, 선풍기: 3, 믹서:2 를 원-핫 인코딩

원본 데이터

상품 분류
TV
냉장고
전자렌지
컴퓨터
선풍기
선풍기
믹서
믹서

원-핫 인코딩

상품분류_TV	상품분류_냉장고	상품분류_믹서	상품분류_선풍기	상품분류_전자렌지	상품분류_컴퓨터
1	0	0	0	0	0
0	1	0	0	0	0
0	0	0	0	1	0
0	0	0	0	0	1
0	0	0	1	0	0
0	0	0	1	0	0
0	0	1	0	0	0
0	0	1	0	0	0

[사이킷런 - 원-핫 인코딩]

원본 데이터 -> 숫자로 인코딩 -> 원-핫 인코딩

[판다스 get_dummies()를 이용한 원-핫 인코딩]

pd.get_dummies(DataFrame)

[피처 스케일링]

- 표준화
  - 데이터의 피처 각각이 평균이 0, 분산이 1인 가우시안 정규 분포를 가진 값으로 변환
- 정규화
  - 서로 다른 피처의 크기를 통일하기 위해 크기를 변환

[사이킷런 피처 스케일링 지원]

- StandardScaler
  - 평균이 0이고, 분산이 1인 정규 분포 형태로 변환
- MinMaxScaler
  - 데이터값을 0과 1 사이의 범위 값으로 변환(음수 값이 있으면 -1에서 1값으로 변환)