Chapter04.

데이터 전처리(Data Preprocessing)

목차

1. 전처리 방안

- **머신러닝 모델의 예측력/신뢰성을 향상시키기 위한 전처리 방안**

1. 결측치 처리 : 결측치 제거, 0또는 상수 대체, 대표값(평균, 중위수) 대체
2. 이상치 처리 : 이상치 탐색(IQR 이용) 이상치 제거, 상수 대체
3. 데이터 인코딩 : 레이블(lebel) 인코딩, 원-핫(one-hot) 인코딩
4. 스케일링 : 최소-최대 정규화, 표준화, 로그화
5. 초매개변수(hyper parameger) 최적화 : 알고리즘에서 제공하는 매개변수 조정

2. 결측치 처리

- 결측치(NaN)?
  - ✓ 누락된 값, 비어 있는 값, NULL 자료 의미
  - ✓ 데이터 수집 과정에서 발생한 오류 등으로 인해 결측치가 포함되어 있는 경우 이 결측치를 정제하는 과정을 거쳐야 분석 결과에 왜곡이 없다.

2. 결측치 처리

- 결측치 발견과 제거

전체 칼럼 단위 결측치 확인

✓ df.isnull().any() # 결측치 유무 확인

✓ df.isnull().sum() # 결측치 개수 확인

결측치 제거(특정 칼럼 기준 결측치 포함 행 제거) : 결측치가 적은 경우

✓ new_df = df.dropna(subset=['칼럼명']) # 해당 칼럼의 결측치 행 제거

✓ new_df.shape # 결측치 제거 확인

3. 결측치(NA) 처리

- 결측치(NA) 대체
  - 상수 대체 : 0 또는 특정 상수로 채우기
  - 중위수 대체
  - 평균 대체
  - 최빈수 대체

2. 결측치 처리

- 결측치 대체

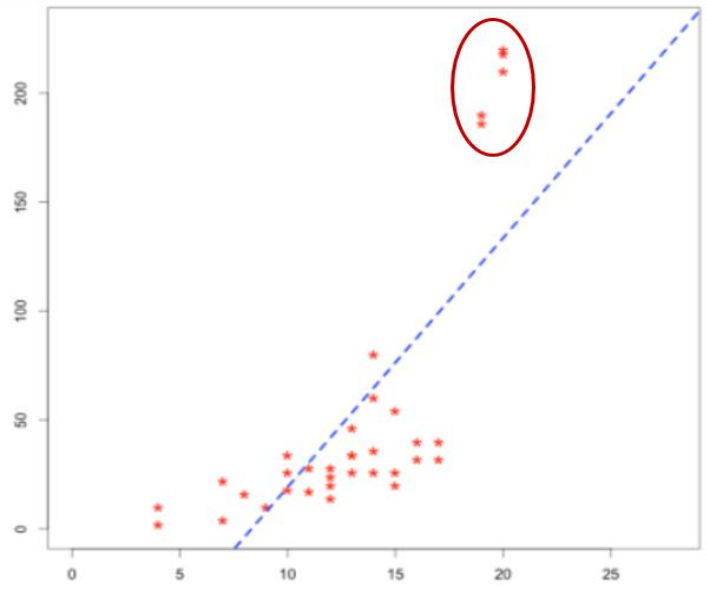
```
# 1) 결측치 대체 : 0 또는 상수
df['bare_nuclei'] = df['bare_nuclei'].fillna(0) # 0으로 대체


# 2) 결측치 대체 : 숫자형 변환 -> 통계
df['bare_nuclei'].fillna(0, inplace=True) # [1] 결측치 0 교체
df['bare_nuclei'] = df['bare_nuclei'].astype('int64') # [2] 자료형 변경
df['bare_nuclei'].fillna(df['bare_nuclei'].mean(), inplace=True) #[3]평균 대체
```

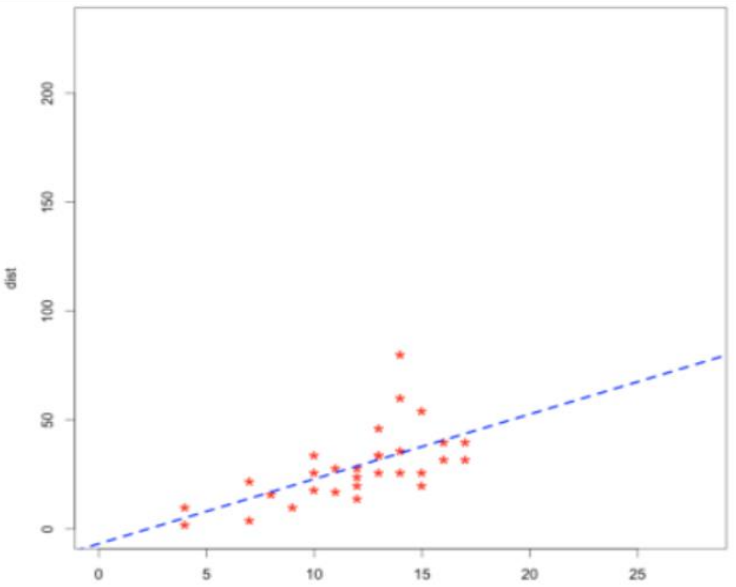
3. 이상치 처리

- 이상치(outlier)가 미치는 영향

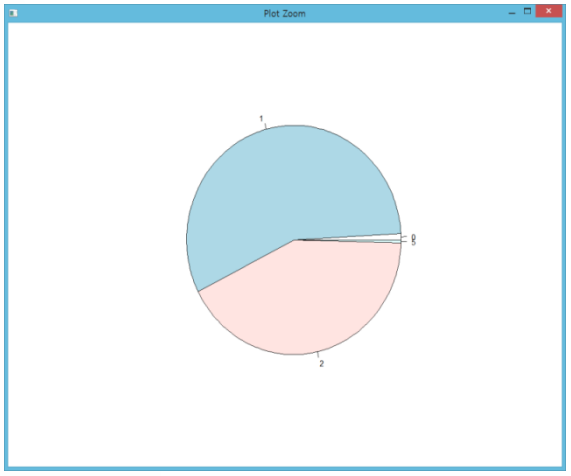
이상치가 있는 경우



이상치가 없는 경우



3. 이상치 처리



- 이상치(극단치) 발견과 정제

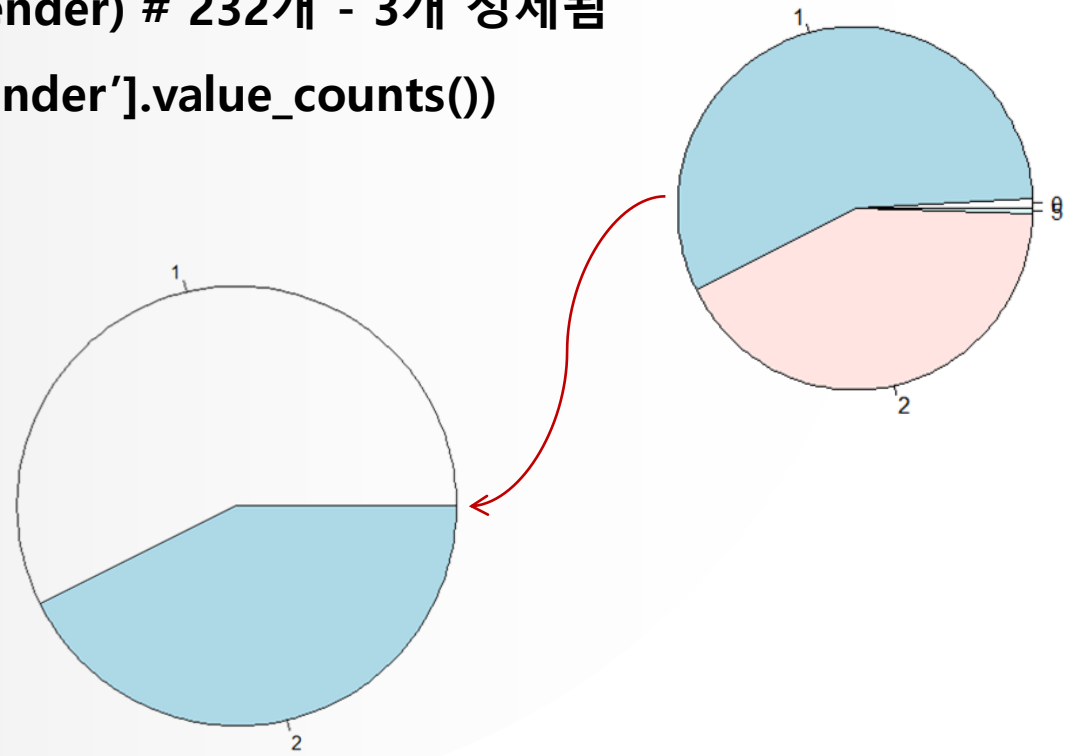
(1) 범주형 변수 극단치 처리

```
# gender 변수 outlier 확인
gender = dataset.gender
plt.hist(gender) # 히스토그램으로 outlier 확인
plt.pie(gender.value_counts())) # 파이 차트로 outlier 확인
```

3. 이상치 처리

- 성별 데이터 정제 - subset() 함수 이용

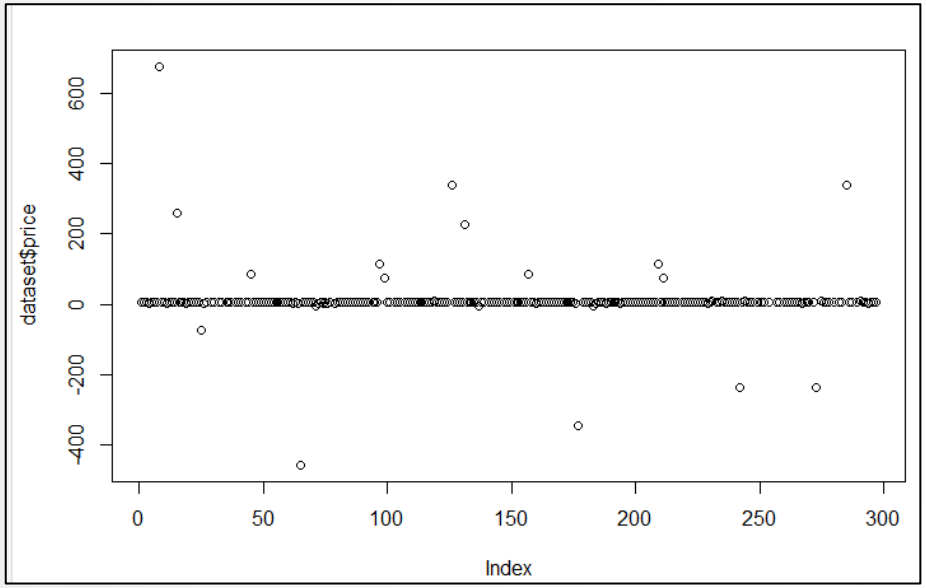
```
data = data[data['gender'] == 1 | data['gender'] == 2]
data # gender변수 데이터 정제
length(data$gender) # 232개 - 3개 정제됨
plt.pie(data['gender'].value_counts())
```



3. 이상치 처리

2) 연속형 변수 이상치 처리

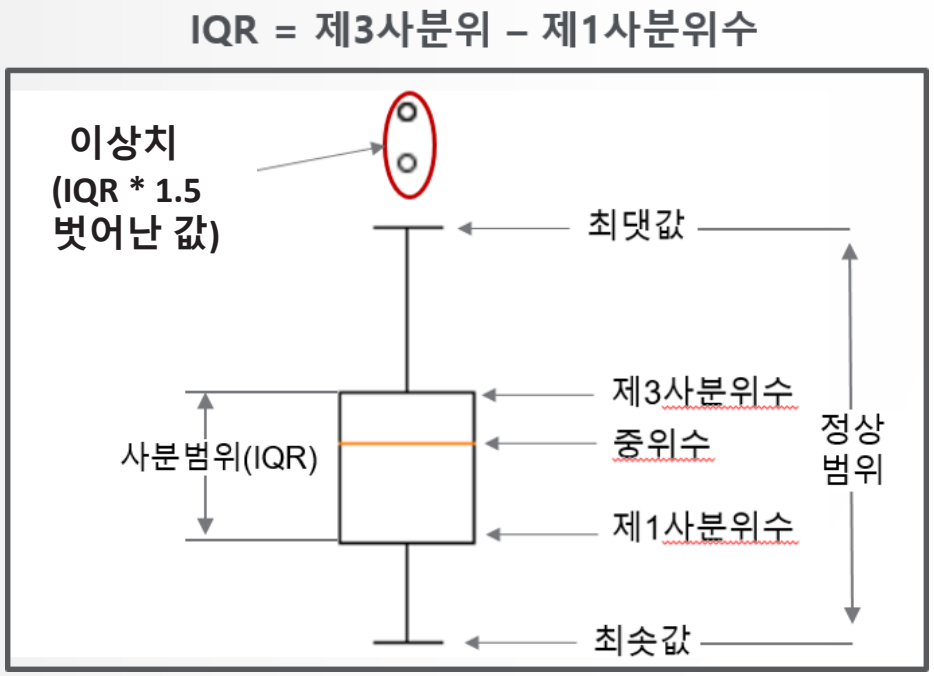
① 시각화 도구 이용 이상치 발견



3. 이상치 처리

② 사분범위(IQR) 이용 이상치 발견

- IQR : Inter Quartile Range = 제3사분위수(Q3) - 제1사분위수(Q1)
- 이상치 : 사분범위(IQR) * 1.5를 벗어나는 경우



3. 이상치 처리

- 이상치 처리
  1. 이상치 제거 : 이상치를 제거하는 방법
  2. 이상치(outlier) 대체 : 다른 값으로 대체하는 방법
     - 하한값, 상한값 이용 : 정상값의 하한값과 상한값으로 대체
     - 평균과 표준편차 이용 : 하한값=평균 - n*표준편차,
       상한값=평균 + n*표준편차
       (n=3 : threshold)

4. 데이터 인코딩

- **데이터 인코딩(encoding)?**
  - ✓ 머신러닝 모델은 숫자형 변수를 대상으로 한다.
  - ✓ 범주형 변수를 대상으로 숫자형의 목록으로 변환해주는 전처리 작업
  - ✓ 방법 : 레이블 인코딩(label encoding), 원-핫 인코딩(one-hot encoding)

4. 데이터 인코딩

1. 레이블 인코딩(label encoding)

- ✓ 레이블 인코딩(label encoding) : 10진수 인코딩
- ✓ 문자형(범주형) -> 영문자 오름차순으로 0 ~ N-1 형식의 10진수 변환
- ✓ 숫자의 차이가 모델에 영향을 주지 않은 트리계열모델(Tree,앙상블) 적용
- ✓ 선형회귀모델(로지스틱, SVM, 신경망)에서 X변수 인코딩 시 가중치 발생

예) DecesionTree 모델에서 독립변수(X)와 종속변수(Y) 대상

상품유형		상품유형
TV		0
냉장고		1
컴퓨터		2
에어콘		3

비유무		비유무
Yes		1
No		0

4. 데이터 인코딩

2. 원-핫 인코딩(one hot encoding)

- ✓ N개의 범주를 갖는 범주형 변수를 대상으로 N차원의 2진수 인코딩
- ✓ 정답에 해당하는 열은 1로 나머지는 0으로 표시
- ✓ 숫자의 차이가 모델에 영향을 미치는 선형계열모델(선형/로지스틱,SVM, 신경망)에서 X변수에 적용
- ✓ 방법) k개 가변수, k-1개 가변수

예) 선형/로지스틱회귀모델에서 독립변수(x) 인코딩

상품유형	가격
TV	350
냉장고	80
컴퓨터	150
에어콘	180

TV	냉장고	컴퓨터	에어콘	가격
1	0	0	0	350
0	1	0	0	80
0	0	1	0	150
0	0	0	1	180

4. 데이터 인코딩

ex) 혈액형(A,B,O,AB) -> K개 가변수

A, AB, B, O

1 0 0 0 - A

0 1 0 0 - AB

0 0 1 0 - B

0 0 0 1 - O

선형종속으로 회귀계수 문제로 k-1개 가변수 이용

ex) 혈액형(A,B,O,AB) -> K-1개 가변수

A, AB, B, O

0 0 0 - base(A)

1 0 0 - AB

0 1 0 - B

0 0 1 - O

5. 특징 스케일링

- **특징 스케일링(feature scaling)?**
  - ✓ 서로 다른 크기(단위)를 갖는 X변수(feature)를 대상으로 일정한 범위로 조정하는 전처리 작업
  - ✓ 방법 : 표준화(StandardScaler), 정규화((MinMaxScaler), 로그변환

5. 특징 스케일링

1. 표준화(StandardScaler)

- ✓ X변수를 대상으로 정규분포가 될 수 있도록 평균=0, 표준편차=1로 통일시킴
- ✓ 회귀모델, SVM 계열은 X변수가 정규분포라고 가정하에 학습이 진행되므로 표준화를 적용

5. 특징 스케일링

2. 최소-최대 정규화(MinMaxScaler)

- ✓ 서로 다른 척도(값의 범위)를 갖는 X변수를 대상으로 최솟값=0, 최댓값=1로 통일 시킴
- ✓ 트리모델 계열(회귀모델 계열이 아닌 경우)에서 서로 다른 척도를 갖는 경우 적용

5. 특징 스케일링

3. 로그 변환(Log-transformation)

- ✓ 로그변환 : log()함수 이용하여 로그변환
- ✓ 비선형(곡선) -> 선형(직선)으로 변환
- ✓ 왜곡을 갖는 분포 -> 좌우대칭 형태의 정규분포로 변환
- ✓ 회귀모델에서 Y변수 적용
  - X변수를 표준화 또는 정규화 할 경우 Y변수는 로그변환
  - 입력변수 X가 조정되면 출력변수 Y도 변환이 필요함