Chapter. 17

왜 여기엔 값이 없을까: 결측치 문제

문제 정의

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

I 문제 정의

- 데이터에 결측치가 있어, 모델 학습 자체가 되지 않는 문제

- 결측치는 크게 NaN과 None으로 구분됨
  - NaN: 값이 있어야 하는데 없는 결측으로, 대체, 추정, 예측 등으로 처리
  - None: 값이 없는게 값인 결측 (e.g., 직업 – 백수)으로 새로운 값으로 정의하는 방식으로 처리

- 결측치 처리 방법 자체는 매우 간단하나, 상황에 따른 처리 방법 선택이 매우 중요

I 용어 정의

- 결측 레코드: 결측치를 포함하는 레코드
- 결측치 비율: 결측 레코드 수 / 전체 레코드 개수

ID	V1	V2	V3	V4	V5
#1		X		X	
#2					
#3	X		X		X
#4					
#5		X			
#6					
#7					
#8					X
#9					
#10					

- 결측 레코드: #1, #3, #5, #8
- 결측치 비율: 4 / 10 = 0.4
- 변수 별 결측치 비율
  - ✓ V1: 1 / 10 = 0.1
  - ✓ V2: 2 / 10 = 0.2
  - ✓ V3: 1 / 10 = 0.1
  - ✓ V4: 1 / 10 = 0.1
  - ✓ V5: 2 / 10 = 0.2

Chapter. 17

왜 여기엔 값이 없을까: 결측치 문제

해결 방법 (1) 삭제

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

I 행 단위 결측 삭제

- 행 단위 결측 삭제는 **결측 레코드를 삭제하는 매우 간단**한 방법이지만, 두 가지 조건을 만족하는 경우에만 수행할 수 있음

조건	근거	결정 방법
결측이 없는 레코드가 모델을 학습하는데 **충분**해야 한다	결측을 제거한 데이터로 모델을 학습하기 때문	학습 샘플 개수에 따른 성능의 **수렴 여부** 확인
결측이 **새로운 데이터**에는 **없어야** 한다	결측을 제거한 데이터로 학습한 모델은 결측이 포함된 레코드를 분류 및 예측할 수 없으므로	**도메인 지식** 기반의 판단

I 열 단위 결측 삭제

- 열 단위 결측 삭제는 **결측 레코드를 포함하는 열을 삭제하는 매우 간단**한 방법이지만, 두 가지 조건을 만족하는 경우에만 사용 가능
  - ➢ **소수 변수**에 결측이 많이 포함되어 있음
  - ➢ 해당 변수들이 크게 중요하지 않음 (by 도메인 지식)

I 관련 문법: Series / DataFrame.isnull

- 값이 결측이면 True를, 그렇지 않으면 False를 반환 (notnull 함수와 반대로 작동)
- sum 함수와 같이 사용하여 결측치 분포를 확인하는데 주로 사용

Index	V1	V2	V3
a	1	NaN	3
b	2	3	5
c	NaN	5	NaN
d	3	4	1
e	4	NaN	2

df

df.isnull()

Index	V1	V2	V3
a	False	True	False
b	False	False	False
c	True	False	True
d	False	False	False
e	False	True	False

sum(axis = 1)

Index	
a	1
b	0
c	2
d	0
e	1

sum(axis = 0)

	V1	V2	V3
	1	2	1

I 관련 문법: DataFrame.dropna

- 결측치가 포함된 행이나 열을 제거하는데 사용

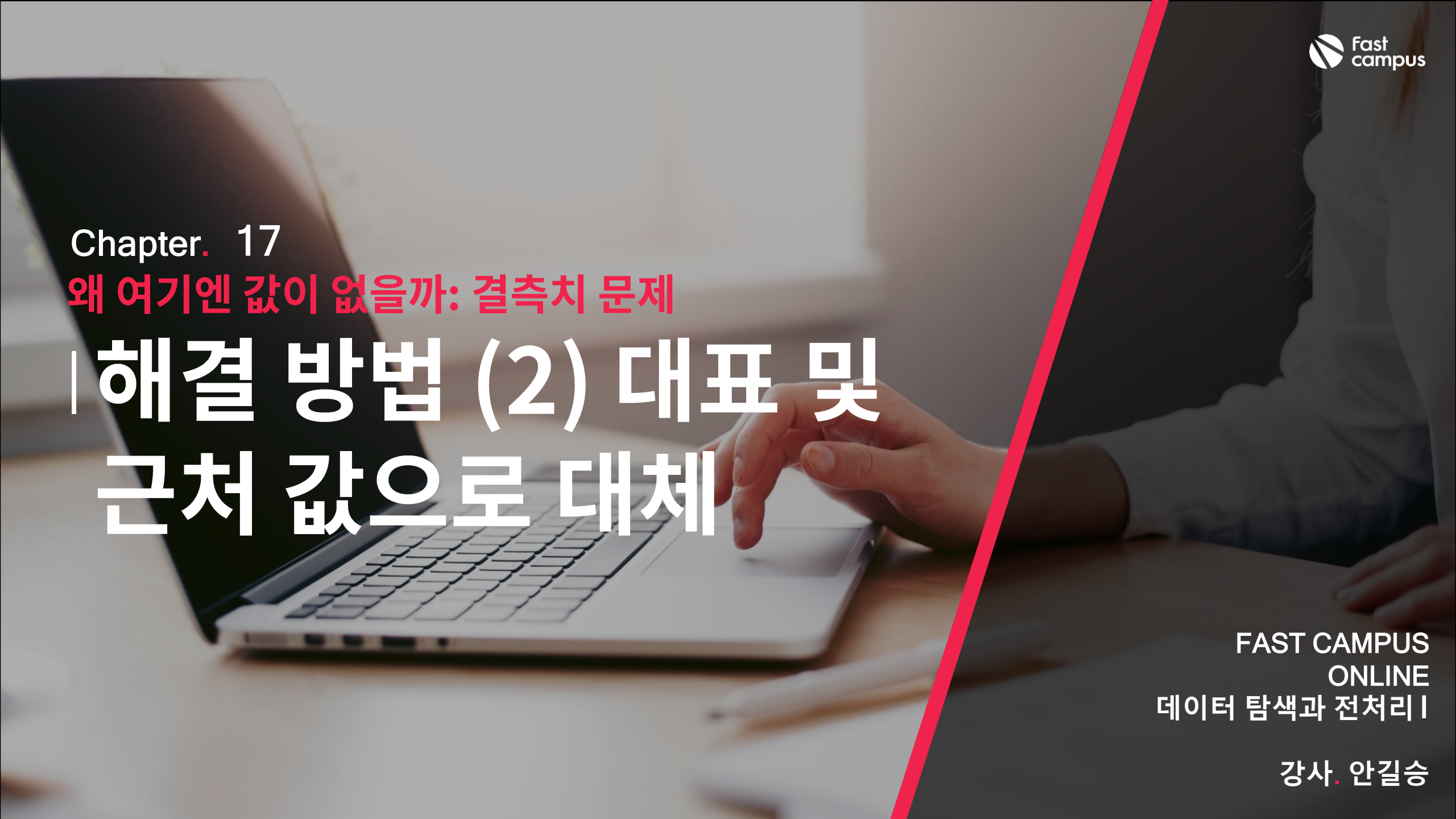
- 주요 입력
  - ➢ axis: 1이면 결측이 포함된 열을 삭제하며, 0이면 결측이 포함된 행을 삭제
  - ➢ how: ‘any’면 결측이 하나라도 포함되면 삭제하며, ‘all’이면 모든 값이 결측인 경우만 삭제 (주로 any로 설정)

Index	V1	V2	V3
a	1	NaN	3
b	2	3	5
c	7	5	4
d	3	4	1
e	4	NaN	2

df

df.dropna()

Index	V1	V2	V3
b	2	3	5
c	7	5	4
d	3	4	1



Chapter. 17

왜 여기엔 값이 없을까: 결측치 문제

해결 방법 (2) 대표 및 근처 값으로 대체

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

I 대표 값으로 대체 (SimpleImpute)

- 가장 널리 사용되는 방법이지만, (1) 소수 특징에 결측이 쏠린 경우와 (2) 특징 간 상관성이 큰 경우에는 활용하기 부적절함

V1	V2
1	4
NaN	4
NaN	5
NaN	7
NaN	2

V1은 결측이 너무 많아,
대표 값인 1이 대표성을 띄지 않음

V1	V2
0	1
0	1
0	1
0	1
1	NaN
1	0
1	0
1	0
0	1
NaN	0

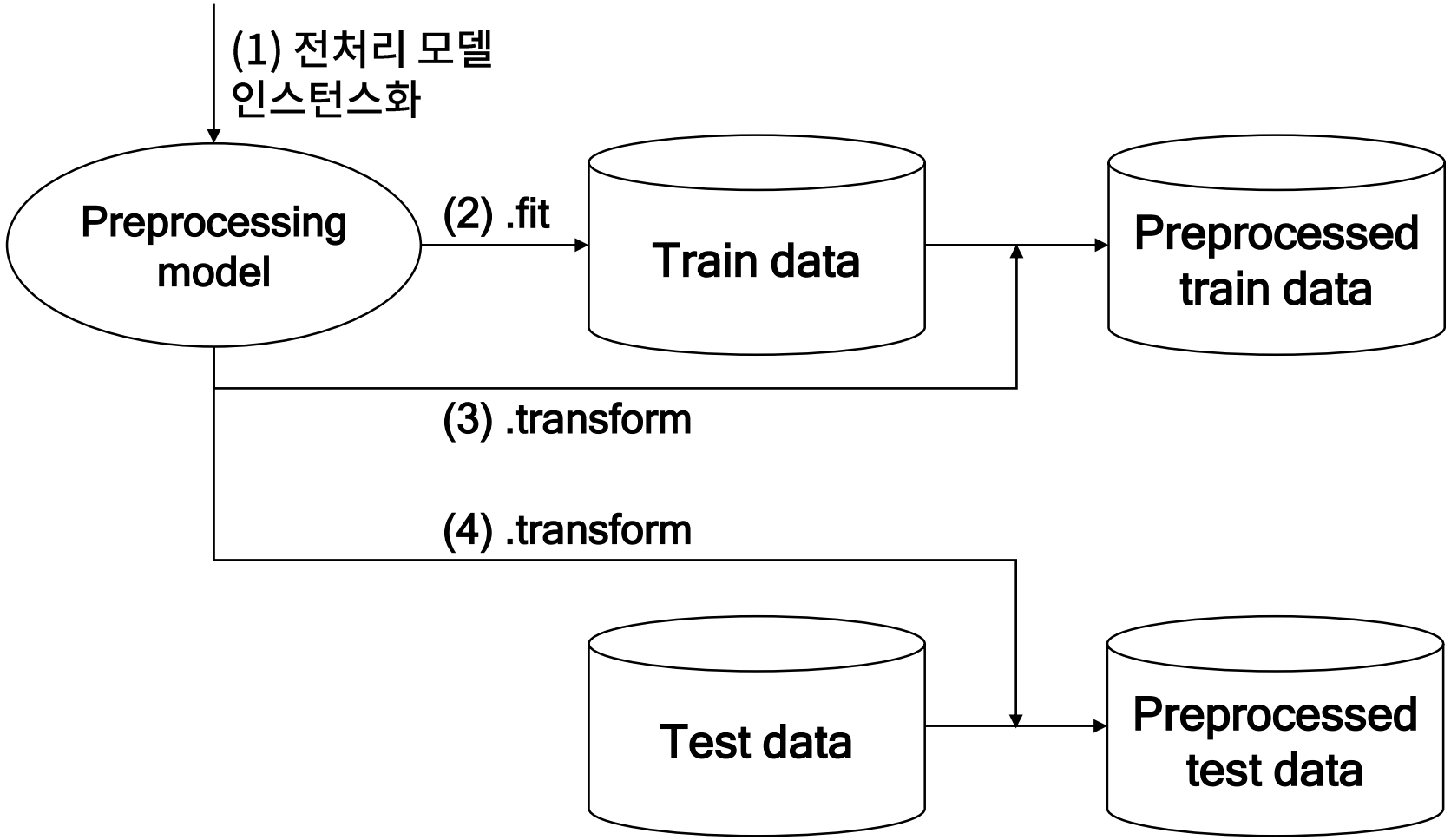
simple impute →

V1	V2
0	1
0	1
0	1
0	1
1	1
1	0
1	0
1	0
0	1
0	0

V1과 V2 간에 V1 + V2 = 1이라는
명확한 관계가 있지만, 이를 무시함

I 관련 문법: sklearn을 이용한 전처리 모델

- sklearn을 이용한 대부분의 전처리 모델의 활용 과정의 이해는 매우 중요하며, 특히 **평가 데이터는 전처리 모델을 학습하는데 사용하지 않음**에 주목해야 함



l 관련 문법: sklearn.impute.SimpleImputer

- 결측이 있는 변수의 대표값으로 결측을 대체하는 인스턴스

- 주요 입력
  - ➢ strategy: 대표 통계량을 지정 (‘mean’, ‘most_frequent’, ‘median’)

Index	V1	V2
a	1	NaN
b	2	4
c	NaN	5
d	NaN	6
e	3	NaN

fit →

V1	2
V2	5

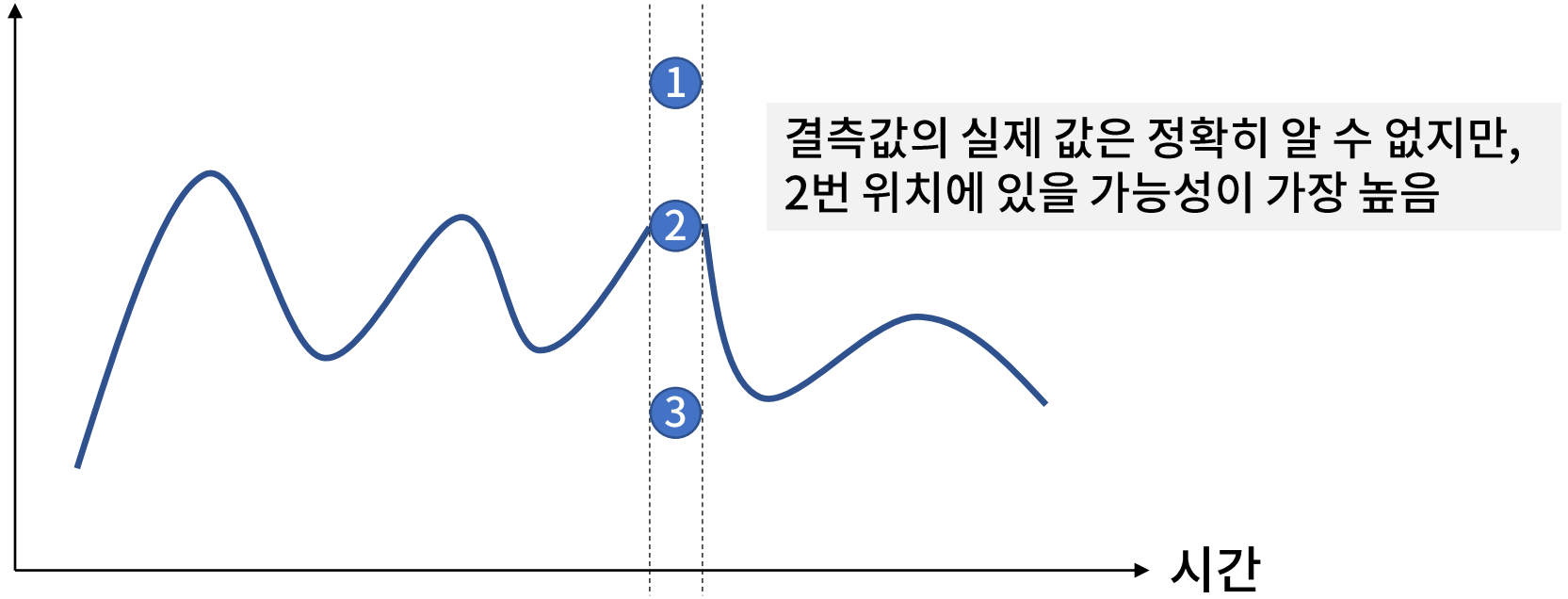
transform →

Index	V1	V2
a	1	5
b	2	4
c	2	5
d	2	6
e	3	5

- 변수 타입에 따라 두 개의 인스턴스를 같이 적용해야 할 수 있음

I 근처 값으로 대체

- 시계열 변수인 경우에는 결측이 바로 이전 값 혹은 이후 값과 유사할 가능성이 높음



결측값의 실제 값은 정확히 알 수 없지만,
2번 위치에 있을 가능성이 가장 높음

I 관련 문법: DataFrame.fillna

- 결측치를 특정 값이나 방법으로 채우는 함수

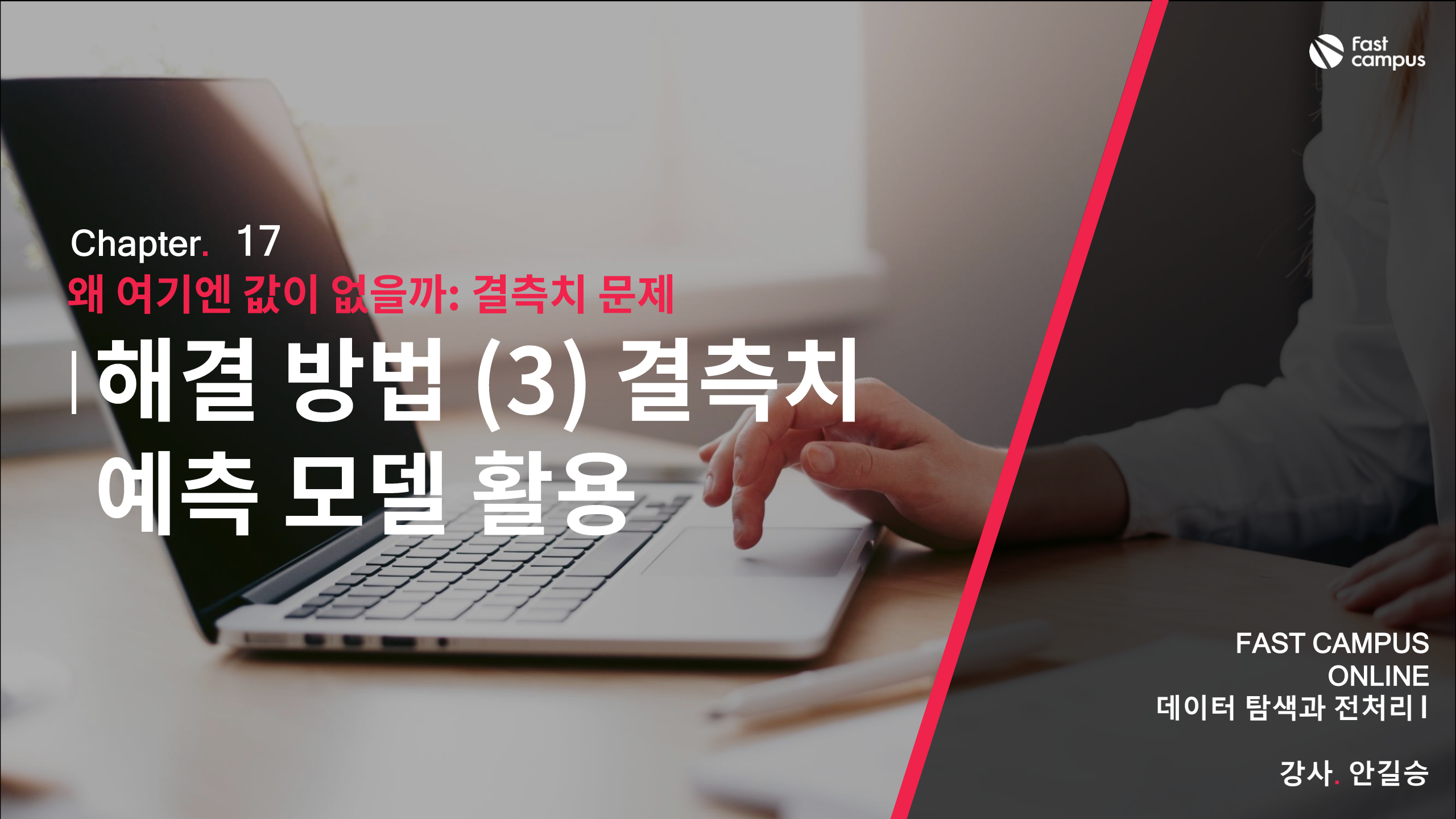
- 주요 입력
  - ➢ value: 결측치를 대체할 값
  - ➢ method: 결측치를 대체할 방법
    - ffill: 결측치 이전의 유효한 값 가운데 가장 가까운 값으로 채움
    - bfill: 결측치 이후의 유효한 값 가운데 가장 가까운 값으로 채움

Index	V1	V2	V3
a	1	NaN	3
b	2	3	5
c	NaN	5	NaN
d	NaN	4	1
e	4	NaN	2

df

df.fillna(metod = 'ffill')

Index	V1	V2	V3
a	1	NaN	3
b	2	3	5
c	2	5	5
d	2	4	1
e	4	4	2



Chapter. 17

왜 여기엔 값이 없을까: 결측치 문제

해결 방법 (3) 결측치 예측 모델 활용

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

I 결측치 예측 모델 정의

- 결측이 발생하지 않은 컬럼을 바탕으로 결측치를 예측하는 모델을 학습하고 활용하는 방법
- (예시) V2 열에 포함된 결측 값을 추정

ID	V1	V2	V3	V4	V5
#1		X	X		
#2					
#3			X		
#4					
#5		X			
#6					
#7					
#8					
#9		X		X	X
#10					

원본 데이터

V2가 결측인 레코드와 V2와 동시에 결측이 발생한 컬럼 삭제 →

ID	V1	V2
#2		
#3		
#4		
#6		
#7		
#8		
#10		

모델 학습 →

$V2 = f(V1)$

결측치 예측 →

ID	V2
#1	$f(V_1^1)$
#5	$f(V_1^5)$
#9	$f(V_1^9)$

I 결측치 예측 모델 활용

- 결측치 예측 모델은 어느 상황에서도 무난하게 활용할 수 있으나, 사용 조건 및 단점을 반드시 숙지해야 함

- 사용 조건 및 단점
  - ➢ 조건 1. 결측이 소수 컬럼에 쏠리면 안 된다
  - ➢ 조건 2. 특징 간에 관계가 존재해야 한다.
  - ➢ 단점: 다른 결측치 처리 방법에 비해 시간이 오래 소요된다.

I 관련 문법: sklearn.impute.KNNImputer

- 결측이 아닌 값만 사용하여 이웃을 구한 뒤, **이웃들의 값의 대표값**으로 결측을 대체하는 결측치 예측 모델

- 주요 입력
  - ➢ n_neighbors: 이웃 수 (주의: 너무 적으면 결측 대체가 정상적으로 이뤄지지 않을 수 있으므로, 5 정도가 적절)

Index	V1	V2
a	1	NaN
b	2	4
c	NaN	5
d	NaN	7
e	3	2

fit →

Index	이웃1	이웃2
a	b	e
c	b	e
d	c	b

transform →

Index	V1	V2
a	1	(4+2) / 2
b	2	4
c	(2+3) / 2	5
d	2	7
e	3	2

	a	b	c	e
d	NaN	3	2	5

거리 행렬

V1: $|7 - 2|$

Chapter.

왜 여기엔 값이 없을까: 결측치 문제

감사합니다

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승