정형데이터 다루기

데이터 전처리

학습 주제

정형데이터 다루기에서 데이터 전처리는 ?

01 결측값(결측치) 처리

02 이상값(이상치) 처리

03 원핫 인코딩(One Hot Encoding)

04 스케일링

학습 목표

1. 인공지능 학습을 위한 데이터 처리법을 이해할 수 있다.

2. 데이터 처리능력을 갖추기 위한 소양을 함양할 수 있다.

결측값이란,

정형데이터에서 결측값은 규칙에 따라 고정된 틀에 저장되지 않은 데이터를 의미하며, 빈칸, Nan, Null 등으로 표현 됩니다. 시간별로 온도를 자동으로 저장하는 데이터의 경우, 기기의 문제나 통신의 문제로 데이터가 저장되지 않는 경우나, 사람들이 수작업으로 데이터를 생성하면서 실수로 빠뜨리거나 삭제하였을 때 발생할 수 있습니다.

01 결측값(결측치) 처리

인공지능은 저장된 데이터들을 가지고 값을 분류하거나 예측을 하는데 이러한 값들이 하나라도 없으면, 동작을 할 수가 없습니다. 따라서 모델을 학습하기 전에 우리는 이러한 결측값을 처리해야합니다. 결측치를 처리하는데 있어 주의할 점은 삭제를 통해 데이터의 수가 줄어들 수 있으며, 대치나 추정의 경우 틀린 값으로 대체될 수 있다는 가정을 항상 하고 있어야 합니다. 결측값을 처리할 때 주로 “삭제”, “대치”, “추정” 3가지 방법을 활용합니다.

결측값 처리 : 삭제

데이터의 행 중에 어느 한 변수라도 결측값이 존재한다면 해당 행을 삭제하는 것!

결측값이 해당열에 많이 존재한다면 해당열 자체를 삭제 할 수도 있습니다.

가장 간단한 방법이지는 인공지능이 보고 학습할 수 있는 데이터가 줄어드는 단점을 가집니다.

또한 결측된 데이터가 학습결과에 크게 영향을 주지 않는다면

해당 행을 삭제 하지 않는 것이 좋습니다.

결측값 처리 : 대치

결측값을 결측값이 존재하는 열이 가지고 있는 통계값(최빈값, 중앙값, 평균값 등)으로 바꾸는 것

최빈값으로 대치하는 경우는 주로 범주형 데이터에서, 결측값이 발생하는 경우 활용하며,

범주별로 빈도가 가장 높은 값으로 결측값을 바꿔줍니다.

이 경우에는 가장 빈도가 높은 값에대한 의존도가 높아지는 문제가 발생할 수 있습니다.

중앙값이나 평균값은 숫자형 데이터에서, 데이터들의 중앙값이나 평균값으로 결측값를 바꿔줍니다.

이때, 결측값은 0으로 인식되는데 이러한 결측값들은 중앙값이나 평균값을 계산할 때 제외시켜줘야 합니다.

결측값 처리 : 추정

결측값이 없는 데이터셋을 활용하여 결측값을 예측하게 하여 그 값으로 바꾸는 방법

이때, 데이터열 간의 상관관계가 높은 데이터를 활용하면 좋습니다.

결측값을 처리하는데 일반적인 방법

결측값이 10% 미만인 경우 : “삭제” or “대치”

결측값이 10~50% 인경우 : “추정”

결측값이 50% 이상인 경우 : 해당열 삭제

이상값이란,

정형데이터에서 이상값이란 해당열의 숫자형 데이터들의 값들의 통계값을 기준으로 정상 범주에서 벗어난 값을 가지는 데이터를 의미합니다. 보통 관측값이 기기 및 통신의 오류 또는 데이터를 수작업으로 작성하면서 실수로 값을 잘못 기입한 경우에 발생하는데요. 그러나 이상치가 항상 의미 없는 값이라고는 할수 없기에, 해당데이터와 관련한 전문가가 이러한 이상값에 대해 검토하는 것이 좋습니다. 이상값의 처리 방법 및 기준은 결측값과 유사하며 “삭제”와 “대치” 이외에 “분리”하는 방법이 있습니다.

02 이상값(이상치) 처리

이상값처리 : 삭제

데이터의 행 중에 해당열의 데이터들과 비교하였을 때, 오타, 오류, 비상식적 반응과 같은 경우의 데이터를 단순히 제거합니다. 이러한 경우 데이터의 수가 줄어드는 단점이 존재합니다.

이상값처리 : 대치

삭제가 어려운 경우에는 평균, 최빈값, 중앙값, 예측값 등으로 치환합니다.

단, 결측값의 경우와 같이 신뢰도 문제가 발생합니다.

이상값처리 : 분리

독립변수가 충분히 세분되지 않은 경우 이상치가 발생할 수 있습니다. 이러한 경우에는 변수를 세분하여 이상치를 분리하게 되는데요. 예를 들어 시계열로 관측된 데이터셋의 경우 특정 기간과 동떨어진 시점에서의 데이터의 값에 이상치가 발생 할 수 있으며, 이때, 특정 기간을 기준으로 데이터셋으로 분리하면 인공지능이 데이터를 학습하는데 유리합니다.

03 원핫 인코딩 : One Hot Encoding

원핫 인코딩이란,

컴퓨터는 사람과 다르게 문자열을 있는 그대로 이해할 수 없답니다.

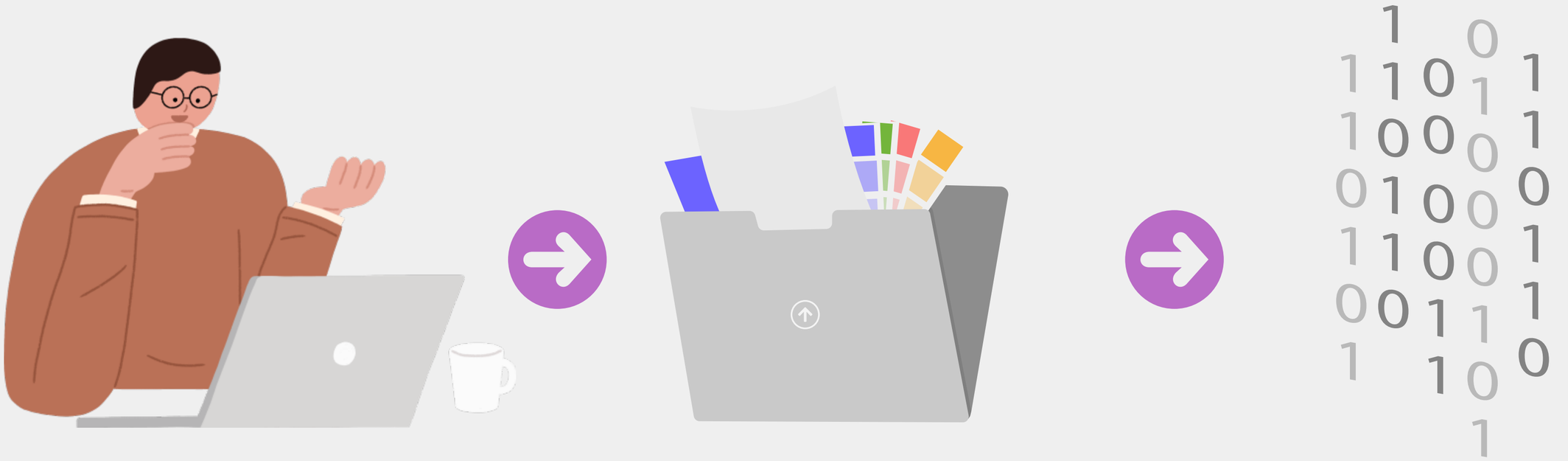
데이터를 이해하기 위해서는 전기적 신호를 나탈 낼 수 있는

“0” 과 “1”으로만 이루어진 이진법으로 변환하여야 하는데요.

각 문자마다 특정한 값으로 1대1 매칭하여 7비트로 표현한

아스키코드(ASCII)가 대표적인 예라고 할 수 있습니다.

03 원핫 인코딩 : One Hot Encoding



데이터셋에서 명목형 데이터를 처리 하는 방법을 원핫 인코딩이라고 하며, 데이터를 수많은 0과 한개의 1의 값으로 데이터를 구별하는 인코딩 방식입니다. 아래와 같은 가장 좋아하는 과목들을 컴퓨터가 어떻게 이해 할 수 있는지에 대해 이야기 해 보겠습니다.

03 원핫 인코딩 : One Hot Encoding

이름	좋아하는 과목
가영	수학
나영	영어
다영	수학
라영	국어

가장 좋아하는 과목은 “수학”, “영어”, “국어” 의 값들을 가지고 있습니다.

컴퓨터는 이러한 문자열을 이해하기 각 항목들을 구분하고 맞으면 1 틀리면 0이라고 표현 해줍니다.

가영이와 다영이는 수학을 가장 좋아하기 때문에 수학(1), 영어(0), 국어(0)이라고 표현 할 수 있습니다.

마찬가지로 나영이는 수학(0), 영어(1), 국어(0)으로 표현 할 수 있습니다.

03 원핫 인코딩 : One Hot Encoding

이름	과목-수학	과목-영어	과목-국어
가영	1	0	0
나영	0	1	0
다영	1	0	0
라영	0	0	1

앞서 내용을 정리하면 위의 표와 같아지게 되는데요. 이러한 원핫 인코딩 방식은 흔히 사용되는 기법으로

인공지능을 학습시키는 것 뿐만 아니라 데이터 마이닝, 자연어 처리 등 많은 분야에서 사용되고 있는 기법입니다.

04 스케일링

스케일링이란,

인공지능 학습을 위해 데이터를 입력할 때, 데이터별로 그 데이터 값들의 범위가 다르다면 컴퓨터가 이것을 이해하기 어렵습니다. 예를 들어 변수1은 0~1의 사이의 값, 변수2는 100~1000의 사이의 값 결과값은 100~1000의 값을 가진다면 변수1은 결과를 도출하는데 큰 영향을 주지 않는 것으로 이해할 수 있습니다.

하지만 실제 변수1이 변수2 보다 더 큰 영향을 줄 수도 있기 때문에 인공지능 학습을 방해할 수 있습니다.
또한 모델이 학습하는데 있어 0으로 수렴 또는 무한 발산이 발생할 수도 있습니다. 이러한 오류를 방지하기 위한 방법으로 데이터 스케일링은 각 변수들의 범위 혹은 분포를 같게 만드는 작업입니다. 이 때 주의할 점은 입력 변수만 스케일링을 하고 결과 변수는 스케일링 하지 말아야 합니다.

04 스케일링

스케일링의 대표적인 두 가지 방법은 정규화와 표준화가 있습니다.

데이터의 특성이나 모델의 특성에 따라 두 가지 방법 중 한 가지를 선택하여 활용하면 됩니다.

정규화

숫자형 데이터들을 0과 1의 사이 값으로 변경하는 것으로 분류보다 회귀모델에 유용하게 사용 됩니다.

데이터를 정규화 하는 방법은 각 행의 최소값과 최대값을 구하여 데이터에서 최소값 빼고,

이를 최대값에서 최소값을 뺀 값으로 나누는 것입니다. 수식은 아래와 같습니다.

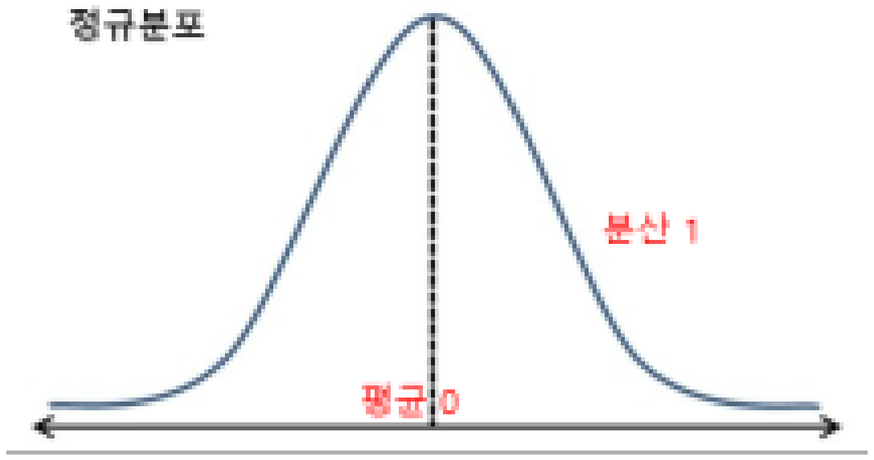
$$\frac{\text{데이터} - \text{최소값}}{\text{최대값} - \text{최소값}}$$

데이터에 이상치가 존재하여 최소값과 최대값이 너무 작거나 커지면서

실 제 데이터의 값을 구분하기 어려워 질 수도 있기 때문에,

미리 결측값, 이상값을 처리하는 것이 중요합니다.

04 스케일링



표준화

평균과 분산을 활용하여 데이터를 변경하는 하는 것입니다. 가장 기본적인 방법은 평균을 0, 분산을 1로 스케일링하여 정규분포형태로 만듭니다. 이때 하한값과 상한값이 존재하지 않아, 이상값이 존재하는 데이터의 경우 학습에 문제가 있을 수 있습니다.