수능 문해력 리터러블
생각의 숲

고3_평가원_2024_11_기술_[데이터 결측치와 이상치 보완법]

Lit. 문해력 리터러블
너희들이 더 잘 읽고 더 넓은 세상을 보도록

날짜	이름

데이터를 처리할 때 데이터의 정확성은 매우 중요하다. 그런데 데이터에 결측치와 이상치가 포함되면 데이터의 특징을 제대로 ⓐ 나타내기 어렵다.

결측치는 데이터 값이 ⓑ빠져 있는 것이다. 결측치를 처리하는 방법 중 하나인 대체는 다른 값으로 결측치를 채우는 것인데, 대체하는 값으로는 평균, 중앙값, 최빈값을 많이 사용한다. 중앙 값은 데이터를 크기순으로 정렬했을 때 중앙에 위치한 값이다. 크기가 같은 값이 복수일 경우에도 순위를 매겨 중앙값을 찾고, 데이터의 개수가 짝수이면 중앙에 있는 두 값의 평균이 중앙값이다. 또 최빈값은 데이터에 가장 많이 나타나는 값을 이른다. 일반적으로 데이터 값이 연속적인 수치이면 평균으로, 석차처럼 순위가 있는 값에는 중앙값으로, 직업과 같이 문자인 경우에는 최빈값으로 결측치를 대체한다.

이상치는 데이터의 다른 값에 비해 유달리 크거나 작은 값으로, 데이터를 수집할 때 측정 오류 등에 의해 주로 ⓒ생긴다. 그러나 정상적인 데이터라도 데이터의 특징을 왜곡하는 데이터 값이 있을 수 있다. 예를 들어, 데이터가 어떤 프로 선수들의 연봉이고 그중 한 명의 연봉이 유달리 많다면, 이상치가 포함된 데이터에 해당한다. 이런 데이터의 특징을 하나의 수치로 나타내려는 경우 ㉠ 대푯값으로 평균보다 중앙값을 주로 사용한다.

평면상에 있는 점들의 위치를 나타내는 데이터에서도 이상치를 발견할 수 있다. 대부분의 점들이 가상의 직선 주위에 모여 있다면 이 직선은 데이터의 특징을 잘 나타낸다고 할 수 있다. 이 직선을 직선 l이라고 하자. 그런데 직선 l로부터 멀리 떨어진 위치에도 몇 개의 점이 있다. 이 점들이 이상치이다.

㉡ 이상치를 포함하는 데이터에서 직선 l을 찾는다고 하자. 이때 사용할 수 있는 기법의 하나인 A기법은 두 점을 무작위로 골라 정상치 집합으로 가정하고, 이 두 점을 ⓓ 지나는 후보 직선을 그어 나머지 점들과 후보 직선 사이의 거리를 구한다. 이 거리가 허용 범위 이내인 점들을 정상치 집합에 추가한다. 정상치 집합의 점의 개수가 미리 정해 둔 기준, 즉 문턱값보다 많으면 후보 직선을 최종 후보군에 넣는다. 반대로 점의 개수가 문턱값보다 적으면 후보 직선을 버린다. 만약 처음에 고른 점이 이상치이면, 대부분의 점들은 해당 후보 직선과의 거리가 너무 ⓔ멀어 이 직선은 최종 후보군에서 제외되는 것이다. 이 과정을 반복하여 최종 후보군을 구하고, 최종 후보군에 포함된 직선 중에서 정상치 집합의 데이터 개수가 최대인 직선을 직선 l로 선택한다. 이 기법은 이상치가 있어도 직선 l을 찾을 가능성이 높다.

WPM
(Word Per Minute)

문단 요약

강의 메모

1

결측치를 처리하는 방법 중 대체 방식에 사용되는 값들은 무엇인지 이 글에서 언급하였나요?

2

중앙값이 어떻게 계산되는지 이 글을 토대로 설명해보세요.

3

이상치가 주로 어떻게 생성되며, 예를 통해 어떤 경우가 이에 해당하는지 이 글에서 설명하였나요?

4

A기법에서의 후보 직선 선택 과정을 이 글을 바탕으로 설명해보세요.