차원의 저주란,

*데이터 학습을 위해 차원이 증가하면서 학습데이터 수가 차원의 수보다 적어져 성능이 저하되는 현상.
*차원이 증가할 수록 개별 차원 내 학습할 데이터 수가 적어지는(sparse) 현상 발생

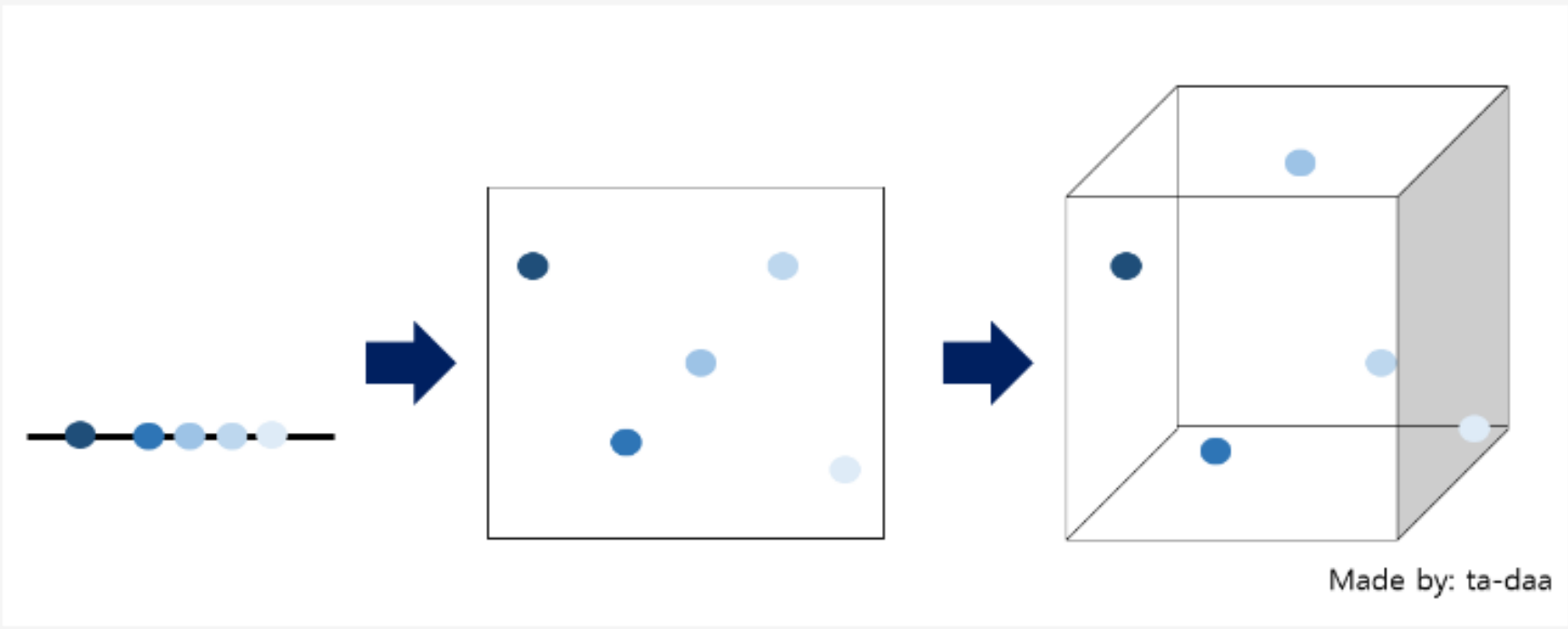
결국, 다중공선성을 띠게 되어, 성능이 저하됨.

*해결책: 차원을 줄이거나(축소시키거나) 데이터를 많이 획득

즉, 간단히 말해서
차원이 증가함에 따라(=변수의 수 증가) 모델의 성능이 안 좋아지는 현상을 의미합니다.
무조건 변수의 수가 증가한다고 해서 차원의 저주 문제가 있는 것이 아니라, 관측치 수보다 변수의 수가 많아지면 발생합니다. (예를들어, 관측치 개수는 200개인데, 변수는 7000개)

왜 이런 현상이 발생할까요?

왜 이런 현상이 발생할까요?



차원의 저주

만약, 변수가 1개인, 1차원을 가정해봅시다.
1차원은 '선'이죠. 선위에 관측치들이 표현될 것 입니다. 제일 왼쪽의 그림처럼요!
선 위에 데이터들이 나란히 있는 것을 볼 수 있습니다.
점들이 빽빽히 들어가 있네요!

그럼 이번에는 같은 데이터를, 차원만 2차원으로 늘려봅시다!
2차원은 '평면'입니다. 가운데 그림처럼 표현이 될 수 있습니다.
점들 사이의 1차원일때보다, 더 벌어져 있음을 알 수 있습니다.

그럼 3차원으로 차원을 늘려보도록 하겠습니다.
3차원은 축이 3개인 차원이겠죠.
제일 오른쪽 그림을 보게되면, 점들 사이에 공간이 많이 비었음을 알 수 있습니다.

이처럼, 차원이 증가함에 따라, 빈 공간이 생기는 것을 차원의 저주라고 합니다.

그럼 빈 공간이 생기는 게 왜 문제가 될까요??

첫줄에 언급했던 것처럼, 빈 공간이 생겼다는 것은, 컴퓨터 상으로 0으로 채워졌다는 뜻입니다.
이는 정보가 없는 셈입니다.
정보가 적으니, 당연히 모델을 돌릴때, 성능이 저하될 수 밖에 없습니다.

차원의 저주 문제에 치명적인 알고리즘이 KNN입니다.

KNN은 K-Nearest Neighborhood의 약자로 우리말로 최근접이웃이라고 합니다.

KNN알고리즘은 자신과 가장 가까운 이웃 K개를 보고 라벨(=결과값)을 정하게 되는데,
위 그림을 보면 알 수 있듯이, 차원이 커질 수록 내 주변의 이웃이 점점.....점.....점 더 멀어져가게 됩니다.

그래서 KNN 알고리즘을 쓸 때, 너무 큰 차원이면 되도록이면 다른 알고리즘을 쓰거나 차원을 줄이는 방법으로 데이터를 한 번 정제해야 합니다.

KNN 알고리즘에 대해서는 다음번 포스팅에 자세히 설명하도록 하겠습니다.

감사합니다!