

063 차원의 저주

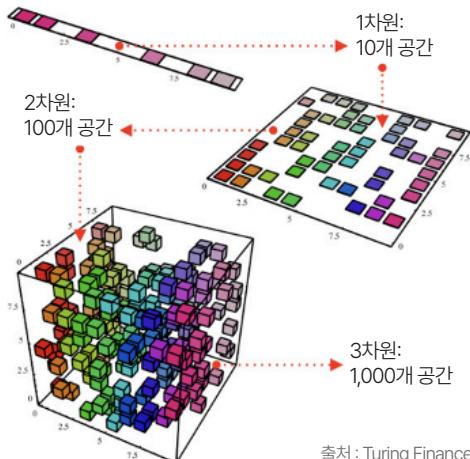
Curse of Dimensionality

데이터의 차원이 높아질수록 분석과 학습이 어려워지는 현상

- 데이터의 특징(차원)이 많아질수록 계산량이 급격히 늘고, 거리 계산이나 분포 추정이 불안정해져 AI 학습 효율이 떨어지는 현상

차원의 저주 발생 원인

차원의 저주는 데이터의 차원, 즉 특징(feature)의 수가 많아질수록 분석과 학습이 어려워지는 현상입니다. 머신러닝에서는 차원이 늘어날 때마다 모델이 학습해야 할 데이터 공간이 기하급수적으로 커지기 때문에, 연산량과 데이터 요구량이 폭발적으로 증가합니다. 예를 들어 1차원 공간을 10개의 구간으로 나눴다면 3차원에서는 100배가 아닌 1,000개의 구간이 필요해집니다. 이렇게 차원이 높아질수록 데이터는 공간상에 희소하게 퍼지고, 모델이 신뢰할 만한 패턴을 학습하기 어려워집니다. 또한 거리 기반 알고리즘의 구분력도 떨어집니다. 고차원에서는 모든 데이터가 서로 비슷한 거리를 가지게 되어, 유사성을 계산하는 K-최근접 이웃(K-Nearest Neighbors, KNN) 알고리즘이나 클러스터링 알고리즘의 성능이 급격히 낮아집니다. 결국 차원이 증가할수록 데이터의 희소성, 거리 왜곡, 계산 복잡도 증가가 동시에 발생하며, 학습 정확도와 일반화 능력이 모두 저하됩니다.



출처 : Turing Finance

차원의 저주에 대한 대응 방법

차원의 저주를 완화하기 위해 차원 축소(Dimensionality Reduction) 기술이 사용됩니다. 대표적으로 주성분 분석(PCA), t-SNE, 오토인코더(Autoencoder) 등이 있으며, 고차원 데이터를 의미를 유지한 채 더 작은 차원으로 압축해 계산 효율을 높이고 과적합을 줄입니다. 또한 불필요한 변수를 제거하거나 핵심 특징만 남기는 특징 선택(Feature Selection) 기법도 차원 축소의 한 방식으로 활용됩니다. 최근에는 데이터의 구조적 특성을 고려한 비선형 차원 축소 방법이 발전하면서, 복잡한 AI 모델에서도 효율적인 학습이 가능해졌습니다. 차원의 저주는 데이터 분석의 근본적 한계로 꼽히지만, 이러한 대응 기법들은 AI가 고차원 환경에서도 안정적이고 효율적으로 학습할 수 있도록 돋는 중요한 연구 분야로 자리 잡고 있습니다.