5월27일~6월4일

∙ 데이터 분석 방법은 끊임없이 증가하는 대량의 고차원 데이터를 분석하는 데 필수적이다. 한쪽 끝에서 클러스터 분석은 데이터 포인트를 분리 된 그룹으로 분할함으로써 데이터를 신속하게 통과시켜 일차 지식을 얻으려고 시도한다. 같은 클러스터에 속하는 데이터 포인트는 유사하지만 다른 클러스터에 속한 데이터 포인트는 서로 다르다고 한다. 가장 널리 사용되고 효율적인 클러스터링 방법 중 하나는 제곱 오차 함수를 최적화하여 클러스터를 나타내기 위해 프로토 타입을 사용하는 K- means 방법이다.

다른 한편으로, 고차원 데이터는 coherent patterns보다 명확하게 검출 될 수 있는 Principal Component Analysis (PCA)릍 통해 lower dimensional data로 변환된다. 이러한 감독되지 않은 차원 감소는 meteorology, image processing, genomic analysis 및 정보 검색과 같은 매우 광범위한 분야에서 사용된다. 또한 PCA가 낮은 차원의 부분 공간에 데이터를 투영하는 데 사용되고 K-means가 부분 공간에 적용되는 것이 일반적이다. 다른 경우, 그래프 Laplacian의 고유 공간과 같은 저 차원 공간에 데이터가 삽입되고 K-means가 적용된다.

PCA 기반 차원 축소의 주된 기반은 PCA가 가장 큰 분산을 갖는 차원을 선택한다는 것이다. 수학적으로 이것은 singular value decomposition (SVD)을 통해 데이터의 최상의 낮은 랭크 근사치를 찾는 것과 같다. 이러한 두 가지 널리 사용되는 방법 간의 연결을 탐색한다.