고소실 7주차 과제

20170101

이은지

PCA에서 압축을 위한 최적 eigenvector 개수 구하기

eigenvector와 eigenvalue의 쌍은 차원의 개수만큼 생성된다. 즉 만약 데이터가 n차원이라면 n개의 쌍이 생성되는 것이다. PCA를 위해서는 그 중 적당한 수인 k 쌍을 골라서 데이터를 재편하여 데이터의 차원을 재구성해야한다. 만약 적당한 eigenvector를 구하지 않는다면 데이터의 손실이 발생하거나 혹은 데이터의 압축 효과를 전혀 얻지 못하기 때문이다. PCA의 eigenvector가 높을수록 데이터의 본래 특징이 잘 드러난다. 따라서 eigenvector를 내림차순으로 정렬한 후 eigenvector을 그에 따라 적절히 선택해야한다.

그렇다면 압축효과를 누리되 데이터의 손실을 방지하기 위해서는 개수를 구하기 이전에

절대적으로 eigenvalue가 0.1 또는 일정 수 또는 양보다 적다면 해당 eigenvector는 사용하

지 않아야한다. 따라서 데이터 쌍이 2개 이상이라면 일정 수 미만의 eigenvalue와 그에

대응하는 eigenvector는 사용하지 않는다.

그리고 이제 몇 개의 쌍을 추출해야하는지를 결정해야한다. 기본적으로 실제값과 투영값

을 최소화하는 것이 데이터의 오차를 최소화하며 가장 정확하므로 아래와 같은 공식을

통해 추출할 쌍을 고를 수 있다.

만약 데이터 쌍이 m개 존재한다면,

그리고 실습과 동일하게 점들의 데이터는 p, 투영값은 Q라고 가정했을 때,

이를 식으로 구체화하면 다음과 같다.

*Then 99% of variance is retained*

그리고 용도에 따라 오차값을 조절할 수 있으므로 만약 오차 범위를 10%로 늘리고 싶다

면 0.01이라는 상수의 범위를 0.1로 수정하면 된다.

따라서 각 eigenvector, eigenvalue 쌍마다 위 식을 계산하여 오차 범위 내에 있는지 판단하

면 된다.