주성분분석-PCA(Principal Component Analysis)

여러 변수들의 변량을 서로 상관성이 높은 여러 변수들의 선형조합으로 만든 새로운 변수들로 축약하는 기법. 이 때 주성분이라 함은 그 방향으로 데이터들의 분산이 가장 큰 방향벡터를 의미한다.

전체적인 과정

1. d차원의 데이터 간의 연관성을 찾기 위해 데이터를 먼저 표준화를 시킨다.
2. 상호 간의 각각의 공분산을 구하기 위해 공분산 행렬을 만든다.
3. 공분산 행렬을 아이겐벨류(고유값)와 아이겐벡터(고유벡터)로 분해한다.
4. 공분산 행렬을 통해 그 두가지를 유도하는 것이 가능한데, 이를 Eigendecomposition이라 한다.
5. 고유값과 고유벡터의 쌍은 최대 d개가 나올 수 있다.
6. 여기서 어떤 특징을 대표한다고 할 만큼의 큰 고유값과 그 쌍이 k개 존재한다면, 데이터의 차원을 k개로 축소하는 것이다. 이 때, k개의 벡터는 새로운 데이터 차원의 basis가 된다.

공분산

일반적인 분산은 모집단에서부터 추출한 표본 데이터들의 편차의 제곱의 산술적 편균을 의미, 즉 평균으로부터 퍼진 정도를 의미한다. 반면 확률론과 통계학에서 공분산은 2개의 확률변수의 상관정도를 나타내는 값이다. X와 y의 공분산은, x, y의 흩어진 정도가 얼마나 서로 상관관계를 가지고 흩어졌는 지를 나타낸다.

공분산 행렬

분산-공분산 행렬은 여러 변수와 관련된 분산과 공분산을 포함하는 정방형 행렬이다.

행렬의 대각선 원소는 각 변수의 분산을 포함하며, 대각선 이외의 원소는 가능한 모든 변수 쌍 간의 공분산을 포함한다.

고유값(eigenvalue)과 고유벡터(eigenvector)

행렬 A를 선형변환으로 봤을 때, 선형변환 A에 의한 변환결과가 자기 자신의 상수배가 되는 0이 아닌 벡터를 고유벡터라 하고 이 상수배 값을 고유값이라 한다. 즉, A에 대해 Av=λv를 만족하는 0이 아닌 열벡터 v를 고유벡터, 상수 λ를 고유값이라 정의한다.

행렬의 mapping의 의미

행렬이란 선형변환이다. 선형 변환은 하나의 벡터 공간을 선형적으로 다른 벡터 공간으로 맵핑하는 기능이 있다.

선형대수학에서의 행렬을 이용한 rotation, shift, scale을 생각해보자.

원래의 성질을 어느정도 유지하면서 새로운 선형 공간으로 벡터들이 이동하거나 변환하여 맵핑된다.

차원 축소는 바로 행렬의 이러한 성질 덕분에 가능한 것이다.

PCA

입력 데이터들의 관계를 보기 위해 만든 공분산 행렬을 분해한 두 가지 성분을 이용한다. 고유벡터는 데이터의 분산이 큰 방향을 나타내는 벡터, 즉 원 데이터의 경향성이 방향 따위를 나타내는 개념이다. 고유값은 그 분산의 크기를 나타내는 것이다. 따라서 고유값이 클수록, 데이터의 경향을 강하게 대표한다고 할 수 있다. 그래서 고유값이 높을수록 좋은 것이다.

예시

1. 빙판길 미끄러짐 사고
2. 수도관 동파
3. 제설 차량 이용으로 인한 소비 금액
4. 폭설로 인한 휴교 횟수
5. 열사병 환자의 수

다음과 같은 5개의 차원의 데이터가 있다고 하자. 하지만 데이터를 잘 살펴보면 모두 온도와 관계된 데이터라는 것을 알 수 있다. 만약 PCA를 통해서 축소를 한다면, 가장 강한 에이겐 벨류를 가지는 벡터를 통해 입력 데이터를 하나의 차원으로 변환시키는 과정을 거치는 것이다. d차원의 데이터가 있을 때, 가장 큰 k개의 고유값에 대한 k개의 벡터를 선택하고, k개의 고유벡터로부터 투영행렬 W를 만든다. 이 W는 d차원의 입력데이터를 k차원의 새로운 피처 X로 변환시킨다. 이것이 PCA의 총체적인 개념이다.