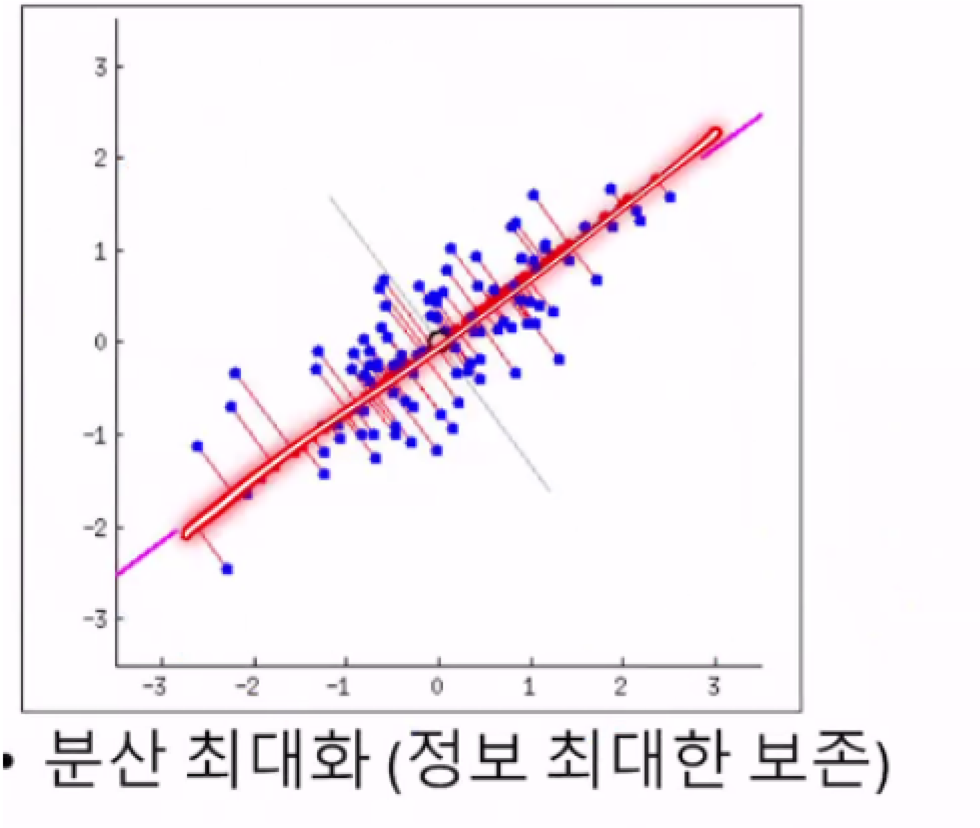
주성분 분석 (Principal Component Analysis): 차원 축소, 설명된 분산

practice_선형대수 · 2024. 9. 30. 15:37

Q. 주성분 분석(Principal Component Analysis, PCA)의 목적과 기본 원리에 대해 설명해 주세요.

•PCA는 데이터의 차원을 축소하여 주요 성분을 추출하는 기법입니다.
•데이터의 분산을 최대화하는 방향으로 새로운 축을 정의하여, 데이터의 주요 변동을 보존하면서 차원을 줄입니다.
아래그림
분산 최대화 - 정보보존 최대화



•집의 특징을 가지고 판매 가격을 예측하는 모델을 만들려고 함

•이를 위해 다양한 집들을 돌아다니며 여러 정보를 수집함

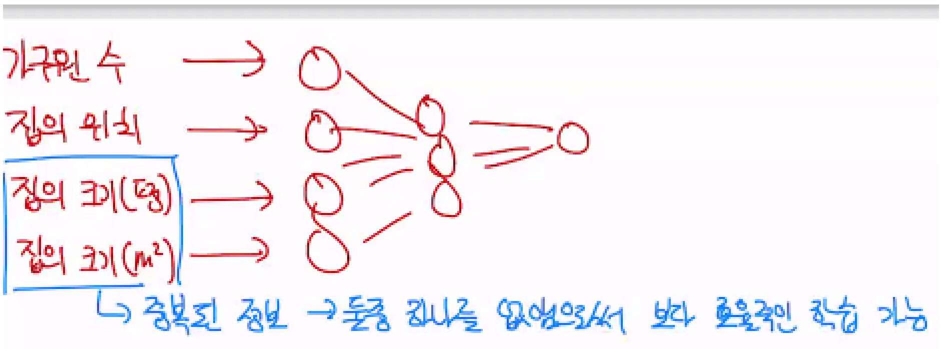
•집의 위치, 가구원 수, 방의 개수, 집의 크기 (평), 집의 크기 (제곱미터), 주변 상권, 지하철역까지의 거리 …

•하지만 이때 몇몇 데이터는 서로 매우 큰 상관관계를 가짐

•집의 크기를 평으로 나타낸 것과, 제곱미터로 나타낸 것은 사실상 같은 정보로 봐도 무방함

•두 정보는 매우 큰 상관계수를 가지고 있으며, 사실상 하나가 없어도 모델이 예측하는 것에는 아무 문제가 없음

이 두 정보(확률변수, 데이터)는 중복되었으며, 하나를 없애는 것이 보다 효율적인 예측 모델을 만드는 것에 도움이 됨



•따라서, 주어진 데이터에서 불필요한 정보는 지울 필요가 있음

•다만 이때 유용한 정보는 최대한 보존해야 함

주성분 분석 (PCA)

•주어진 데이터를 “새로운 축”으로 바라보는 것

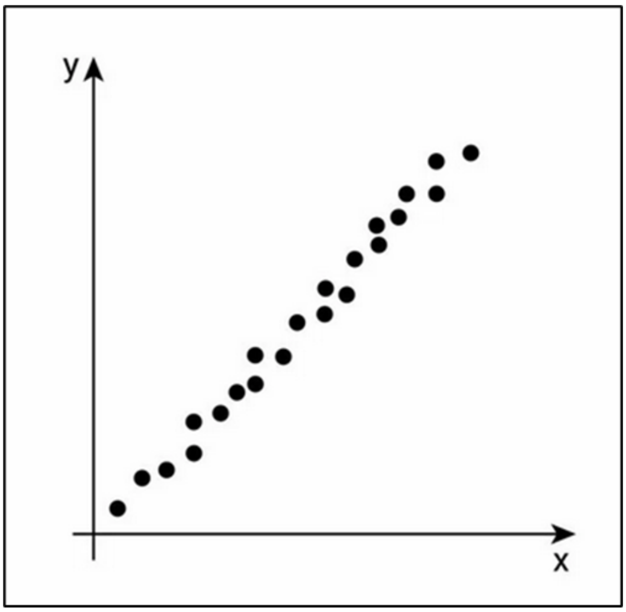
•이때 “새로운 축”은 주어진 데이터의 분산 (정보)를 최대한 보존하는 방향으로 선택

•어떠한 2차원 데이터가 있다고 가정

•X는 집 크기, Y는 집 가격

•그래프에 보이듯, 이 두 정보는 거의 정비례함

중복된 데이터



•(중복된 2차원 데이터를 1차원으로 줄이는 방법) 해결하는 방법은 두가지가 있음

1.단순히 둘중 하나를 삭제

미처 고려하지 못한 정보를 잃을 수 있음

2.PCA를 통해, 두 정보를 하나로 통합하는것 (차원 감소)

(이상적으로) 두 정보를 최대한 보존하는 것이 가능

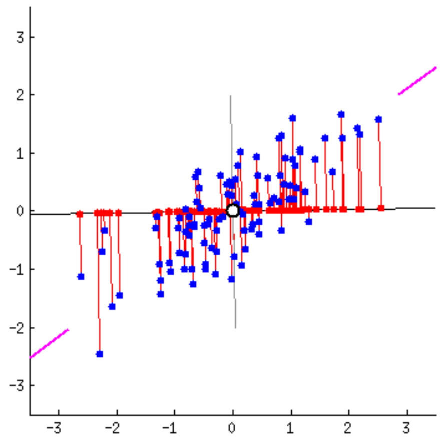
•2차원의 데이터를 1차원으로 나타낼 때, 가능한 많은 정보를 보존해야 함

•정보를 보존한다?

데이터 포인트들의 상대적인 위치를 보존한다

원래부터 가깝던건 차원 축소 후에도 여전히 가깝고, 반대로 원래 멀던건 이후에도 멀어야 한다

출처:
•https://velog.velcdn.com/images%2Fswan9405%2Fpost%2F6b1446e0-fe48-497c-9191-20e3ce447e96%2Fpca02.gif



•이때 우리는 다양한 축을 시도하며, 가장 적절한 (정보를 많이 보존하는) 축을 찾아야 함

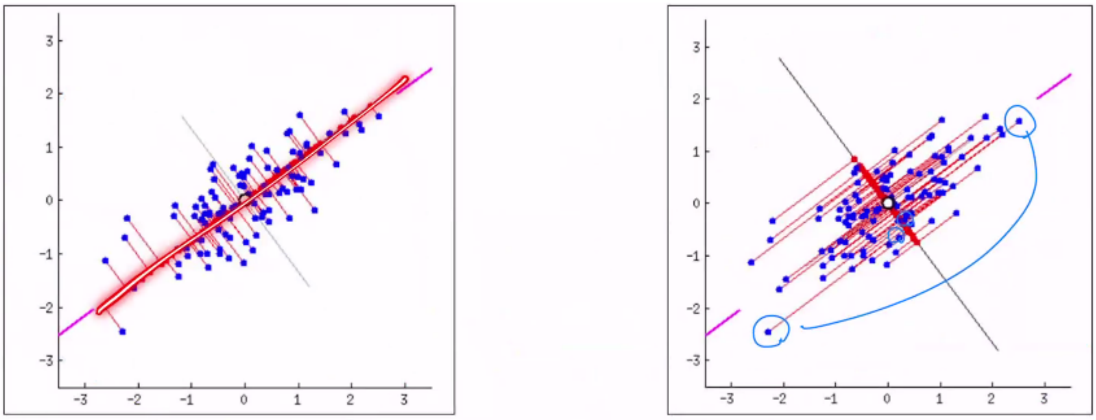
•이때 선택된 축은

1.가능한 많은 정보를 보존 (멀리 있던 포인트들은 여전히 멀리, 가까이 있던 포인트들은 계속 가까이)

2.가능한 많은 원래 데이터의 분산을 보존

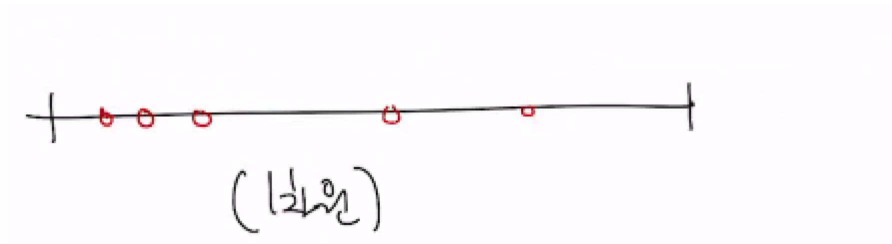
아래그림

분산 최대화 (정보 최대한 보존) 되는 축을 찾음



• 분산 최대화 (정보 최대한 보존)   분산 최소화 (정보 최소한 보존)

아래그림



•위의 예시에서는, 2차원 데이터를 1차원으로 줄였음•

•비슷하게, N차원 데이터를 N보다 작은 차원으로 감소시키는 것도 가능함

•예를 들어, 3차원의 데이터를 2차원으로 축소시킨다면, 정보를 최대한 보존하는 2개의 축을 찾아야 함

•그럼 주어진 데이터에 대해, 분산을 최대화 하는 축은 어떻게 찾을 수 있을까?

수학적으로 바로 계산하는 것 가능 (인공지능 학습 등 불필요)

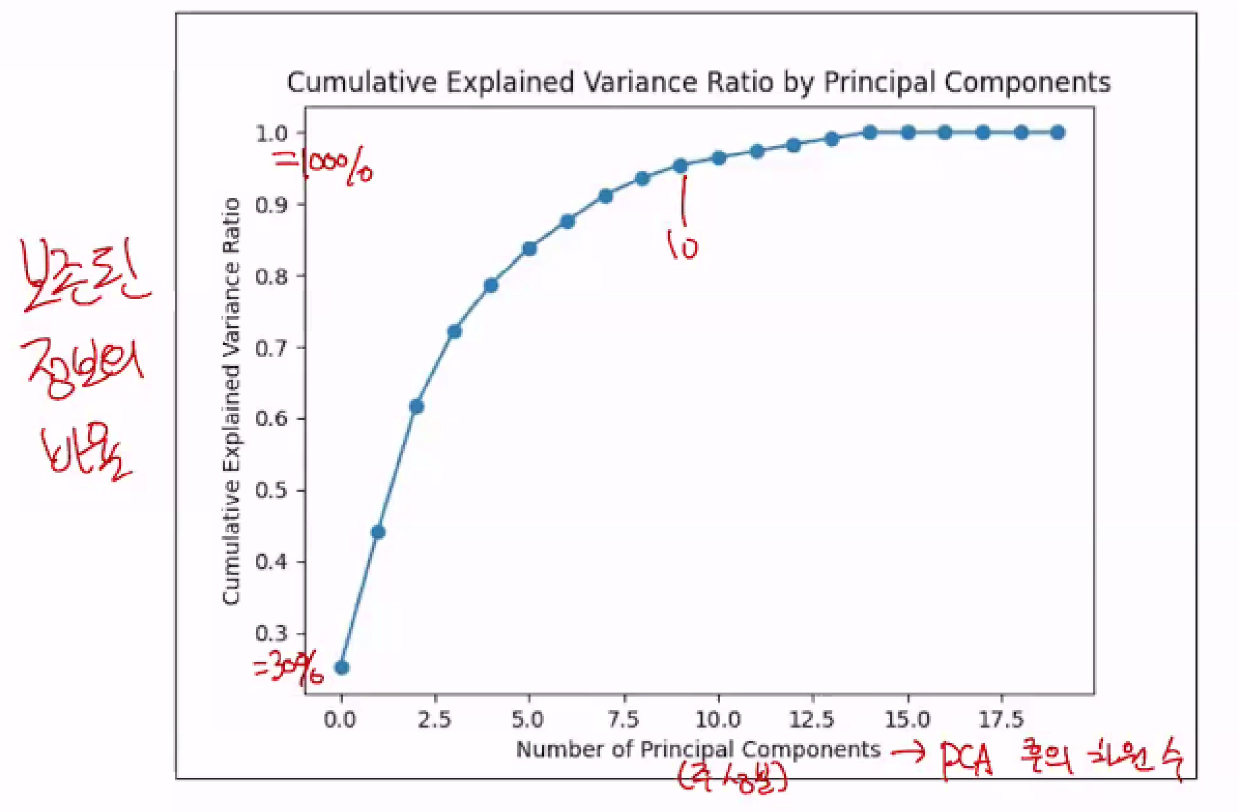
•실제 계산은 고유값과 고유벡터 (Eigenvalue와 Eigenvector)가 사용되므로 단기간에 이해하기 어려움

아래그림

•어떠한 20차원 데이터에 PCA를 적용한 예시 (X는 PCA 적용 후의 차원, Y는 보존된 정보의 비율)

•차원을 감소시킬수록 필연적으로 정보의 손실이 발생

적당한 밸런스를 찾는 것이 필수적



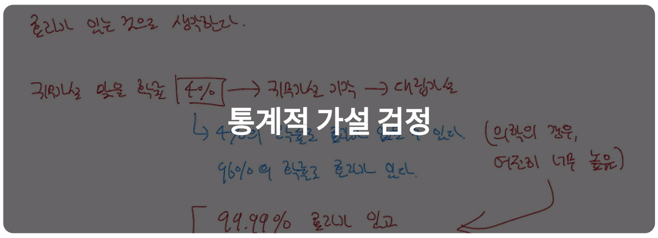
♡ 3 구독하기

'practice_선형대수' 카테고리의 다른 글

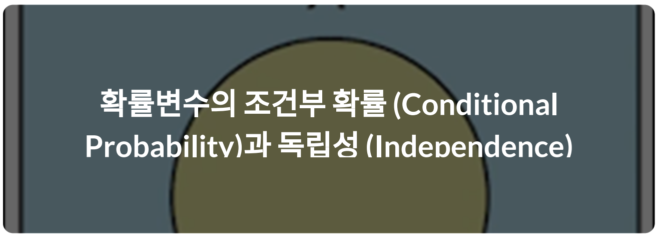
통계적 가설 검정 (0)	2024.10.01
확률변수의 조건부 확률 (Conditional Probability)과 독립성 (Independence) (0)	2024.10.01
확률분포 (0)	2024.09.30
행렬 (1)	2024.09.30
공분산(Covariance)과 상관계수(Correlation Coefficient)의 정의와 차이점 (0)	2024.09.29

관련글

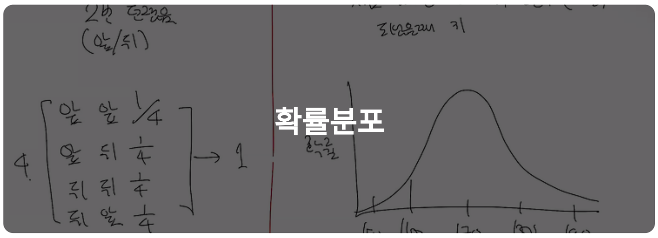
관련글 더보기



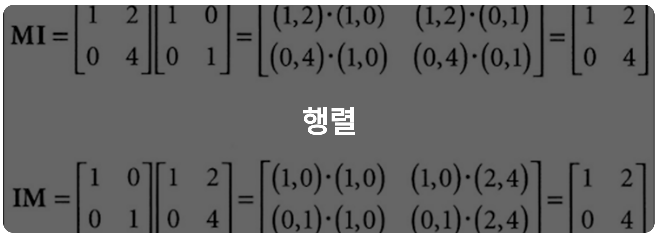
통계적 가설 검정



확률변수의 조건부 확률 (Conditional Probability)과 독립성 (Independence)



확률분포



행렬

자연어(NLP)

네이쳐2024 님의 블로그입니다.

구독하기 +

댓글 2

mwollossna

편안한 밤 되세요~~공감!

2024. 10. 1. 22:09 · 답글

익명

비밀댓글입니다.

2024. 10. 3. 09:29

이름

비밀번호

내용을 입력하세요.

등록