조성진

머신러닝에서 커널은 변환 ϕ(맵핑 함수)를 계산하지 않고
원래 벡터 a와 b에 기반하여 점곱 $\phi(\mathbf{a})^T\phi(\mathbf{b})$를 계산할 수 있는 함수이다.

일반적인 커널

선형: $K(\mathbf{a},\mathbf{b}) = \mathbf{a}^T\mathbf{b}$

다항식: $K(\mathbf{a},\mathbf{b}) = \left(\gamma\mathbf{a}^T\mathbf{b} + r\right)^d$

가우시안 RBF: $K(\mathbf{a},\mathbf{b}) = \exp(-\gamma\|\mathbf{a}-\mathbf{b}\|^2)$

시그모이드: $K(\mathbf{a},\mathbf{b}) = \tanh\left(\gamma\mathbf{a}^T\mathbf{b} + r\right)$

kPCA는 비지도 학습이다

따라서, GridSearchCV를 사용해라
ex) 커널간의 비교
ex) RBF 커널일 때, γ값 찾기

5장 p225, 8장 p290

설명

\# 김유지

Manifold learning

Manifold란

- Manifold란 고차원 데이터가 있을 때 고차원 데이터를 데이터 공간에 뿌리면 sample들을 잘 아우르는 subspace가 있을 것이라는 가정에서 학습 진행
- 이렇게 찾은 manifold는 데이터의 차원을 축소시킬 수 있다

manifold 미국식 [ˈmænɪfoʊld] 영국식 [ˈmænɪfəʊld]

1. 형용사 격식 (수가) 많은, 여러 가지의
2. 명사 전문 용어 (내연 기관의) 매니폴드[다기관]

옥스퍼드 영한사전

김유지

Manifold learning

해결되지않은 질문
d = m 이라면 무엇일까?

Manifold learning

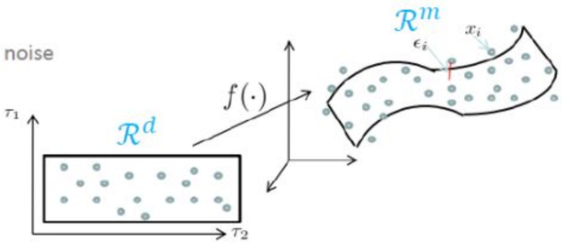
- 고차원 sample -> 저차원 sample로 mapping하는 function f 찾기
- 고차원 sample의 local한 부분만 보면 smooth -> 이를 저차원으로 내리면 해석가능한 point들의 집합을 구할 수 있을 것

Why useful?

- Data compression
- Data visualization
- Curse of dimensionality
- Discovering most important features

- A d dimensional manifold $\mathcal{M}$ is embedded in an m dimensional space, and there is an explicit mapping $f: \mathcal{R}^d \to \mathcal{R}^m$ where $d \le m$
- We are given samples $x_i \in \mathcal{R}^m$ with noise

$$x_i = f(\tau_i) + \epsilon_i$$



https://deepinsight.tistory.com/124

김유지

Manifold learning

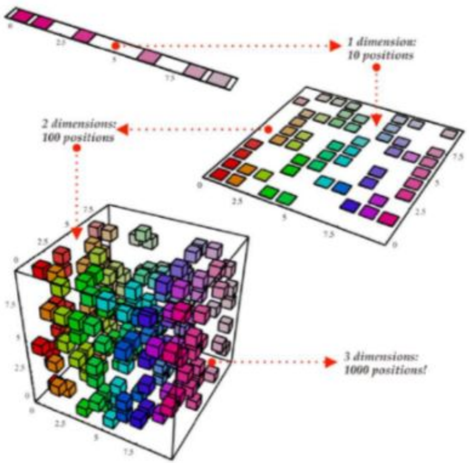
Curse of dimensionality

- 데이터의 차원이 증가할수록 해당 공간의 크기(부피)가 기하급수적으로 증가하기 때문에 동일한 개수의 데이터의 밀도는 차원이 증가할수록 급속도로 희박해진다
- 따라서, 차원이 증가할수록 데이터의 분포 분석 또는 모델추정에 필요한 샘플 데이터의 개수가 기하급수적으로 증가

Manifold Hypothesis

- 고차원의 데이터의 밀도는 낮지만, 이들의 집합을 포함하는 저차원의 매니폴드가 있다
- 이 저차원의 매니폴드를 벗어나는 순간 밀도는 급격히 낮아진다

즉, 고차원의 데이터를 잘 표현하는 manifold를 통해 샘플 데이터의 특징을 파악할 수 있는 것이다.



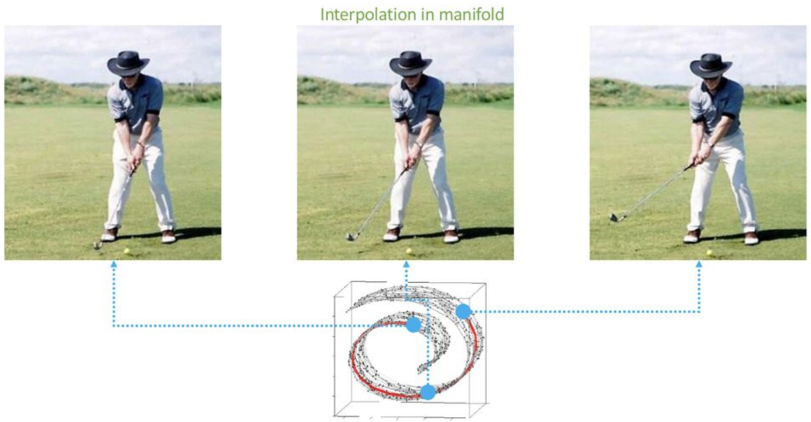
https://deepinsight.tistory.com/124

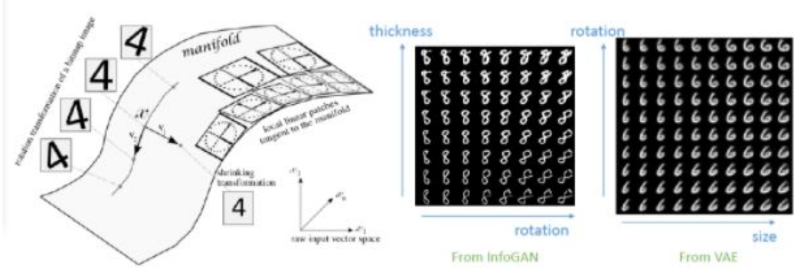
김유지

Manifold learning

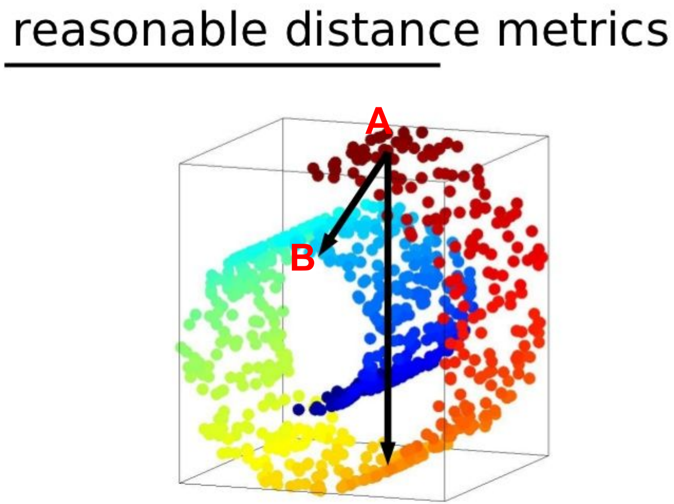
Discovering most important features

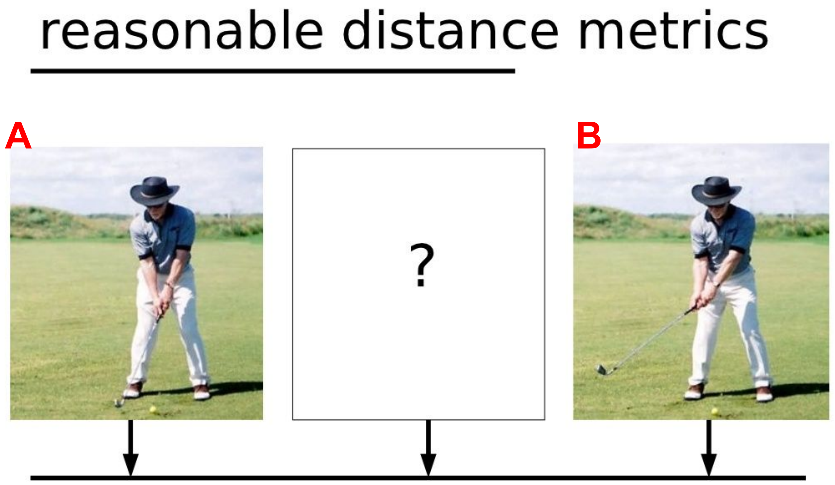
- 고차원의 데이터를 잘 표현한다는 것은 데이터의 중요한 특징을 발견하는 것
- 고차원 데이터의 manifold 좌표들을 조정해보면 manifold 변화에 따라 학습 데이터도 유의미하게 조금씩 변하는 것을 확인할 수 있다





매니폴드(manifold) 예시

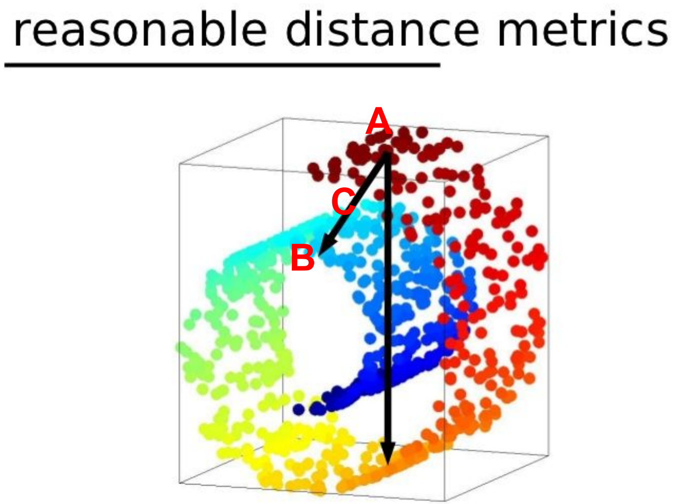


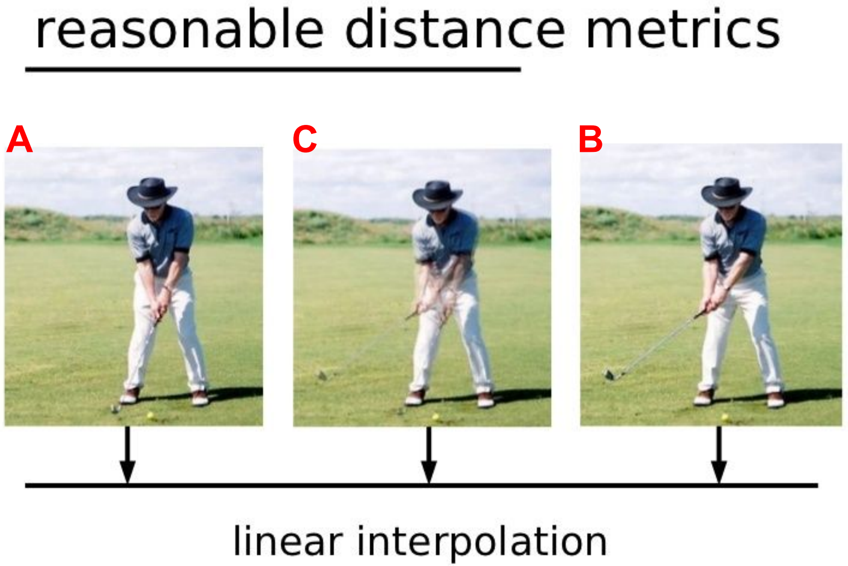


https://www.slideshare.net/zukun/manifold-learning-with-application-to-object-recognition

구건모

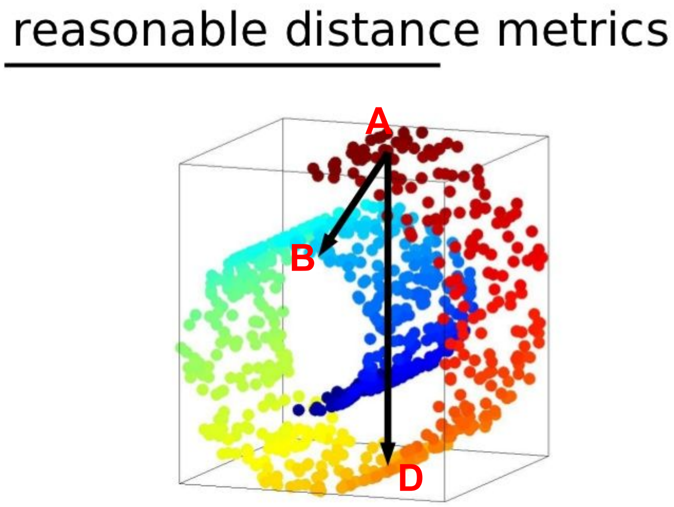
매니폴드(manifold) 예시

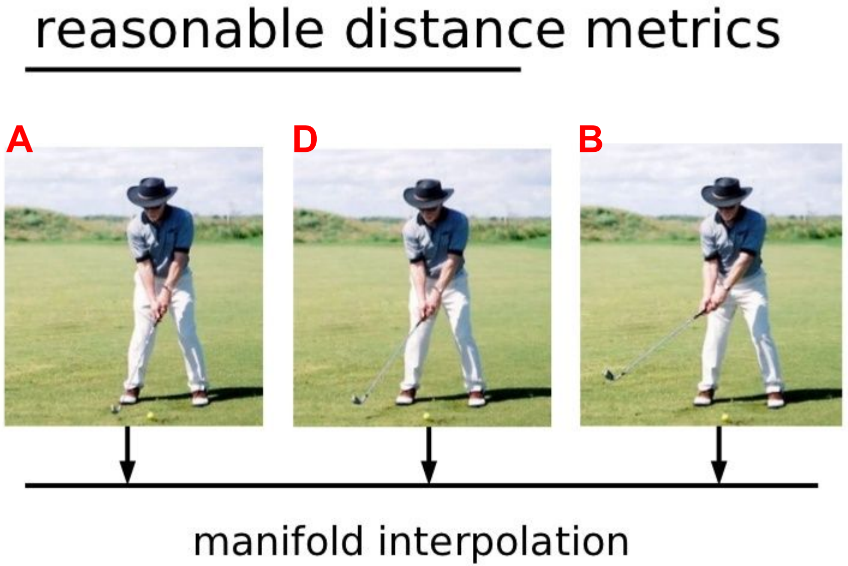




https://www.slideshare.net/zukun/manifold-learning-with-application-to-object-recognition

매니폴드(manifold) 예시





https://www.slideshare.net/zukun/manifold-learning-with-application-to-object-recognition

구건모

외삽을 하면 예측이 더 불안정한 이유와 과대적합의 위험이 커지는 이유

P275

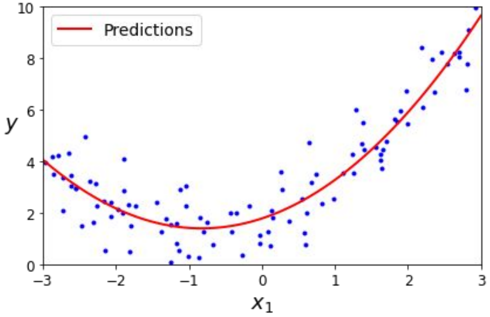
즉, 대부분의 **훈련 데이터가 서로 멀리 떨어져 있습니다**. 이는 **새로운 샘플도 훈련 샘플과 멀리 떨어져 있을 가능성**이 높다는 뜻입니다.

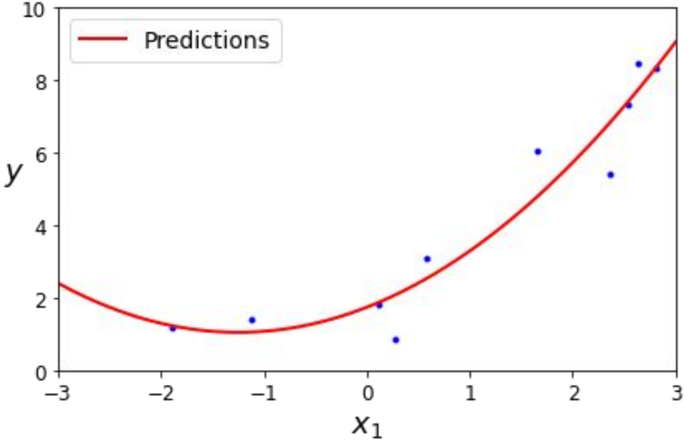
이 경우 예측을 위해 훨씬 많은 외삽을 해야 하기 때문에 저차원일 때보다 예측이 더 불안정합니다. 간단히 말해 훈련 세트의 차원이 클수록 과대적합 위험이 커집니다.

내삽: 주위에 데이터가 많을 때, 결과값을 예측하는 것이고

외삽: 대부분의 데이터와 동떨어진 점에서 결과값을 예측하는 것이라고 생각할 수 있다.

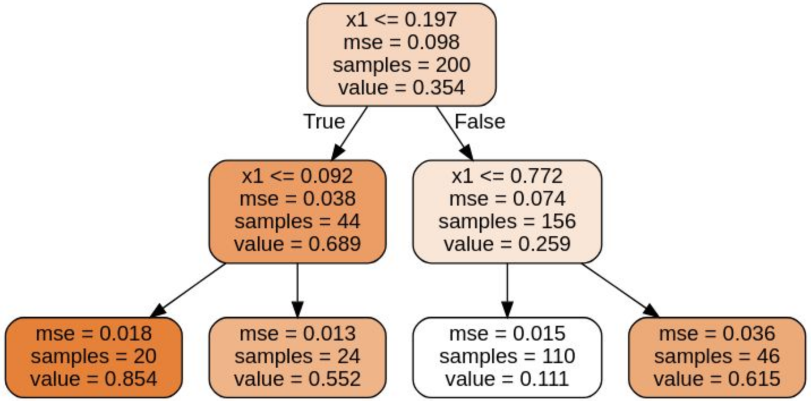
외삽은 새로운 세상?



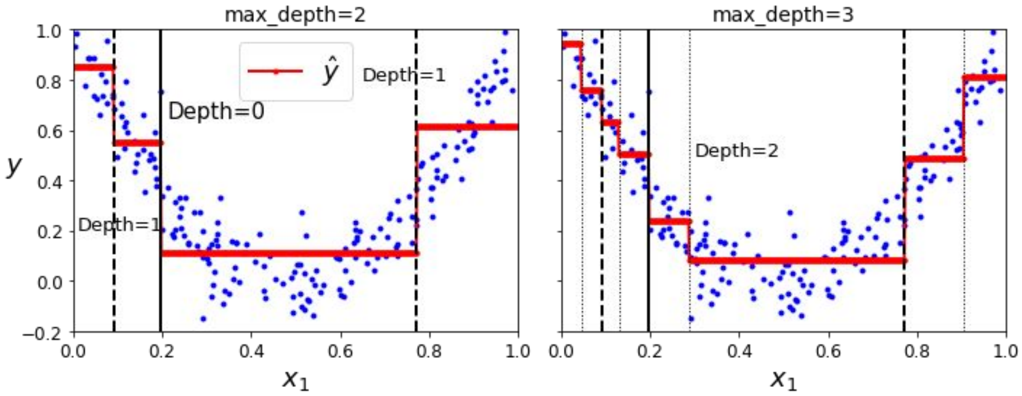


설명 외삽을 하면 예측이 더 불안정한 이유와 과대적합의 위험이 커지는 이유

\# 구건모



결정트리는 외삽을 못합니다



1)

김유지

외삽(Extrapolation)

내삽은 데이터 중에 있고
외삽은 그 밖의 범위에 있다고
이해

p.275

고차원 데이터셋은 매우 희박할 위험이 있습니다. 즉, 대부분의 훈련 데이터가 서로 멀리 떨어져 있습니다. 이는 새로운 샘플도 훈련 샘플과 멀리 떨어져 있을 가능성이 높다는 뜻입니다. 이 경우 예측을 위해 훨씬 많은 **외삽(extrapolation)**을 해야하기 때문에 저차원일 때보다 예측이 더 불안정합니다. 간단히 말해 훈련 세트의 차원이 클수록 과대적합 위험이 커집니다.

<위키백과>

보외법(補外法) 또는 **외삽**(外揷, extrapolation)은 수학에서 원래의 관찰 범위를 넘어서서 다른 변수와의 관계에 기초하여 변수의 값을 추정하는 과정이다. 관찰된 값들 사이의 추정치를 만들어내는 보간법과 비슷하지만 보외법은 더 큰 불확실성과 무의미한 결과 생성에 대한 더 높은 위험에 종속된다.

https://ko.wikipedia.org/wiki/%EB%B3%B4%EC%99%B8%EB%B2%95

김유지

외삽(Extrapolation)

<나무위키>

외삽법 (Extrapolation) 또는 보외법이란 이전의 경험에 비추어, 보다 과학적인 맥락에서는 이전의 실험으로부터 얻은 데이터들에 비추어, 아직 경험/실험하지 못한 경우를 예측해보는 기법이다. 어디까지나 추측이므로 엄밀한 추론이 아니다. 하지만 은유와 실험적인 유비를 통해 새로운 발견을 위한 한 방법으로 유용하다. ⇒ 어느 순간까지의 흐름에 미루어 아직 나타나지 않은, 또는 나타나게 만들 수 없는 부분을 예측하는 기법이다. 외삽 기법은 불완전한 방법이다. 왜냐하면 특이점이 나타날 경우 더 이상 외삽할 수 없기 때문이다. (외삽실패) 그러나 발견의 방법으로서는 매우 유용하다.

※ 특이점 - 수학적으로 특이점이란 그 점의 미분계수가 0이라는 것을 의미한다. 또는 미분계수가 하나가 아니게 되는 점도 특이점이다.

※ 보간(Interpolation 또는 내삽)과 외삽의 차이 - 보간은 특정한 두 점 안쪽에 놓여있는 가능한 값을 구하려는 방법이지만 외삽은 특정한 두 점 바깥에 놓여있는 가능한 값을 구하는 데 있다.

https://namu.wiki/w/%EC%99%B8%EC%82%BD%EB%B2%95

김유지

외삽(Extrapolation)

방법

- 선형 - 기존 데이터의 추세를 활용해 그래프상에서 일직선으로 값을 예측하는 방식이다. 예를 들어 알려진 데이터가 1, 2, 3이고 그 다음에 올 X와 Y를 예측한다 할 때 이 방법을 사용하면 1씩 증가했으므로 각각 4와 5로 예측될 수 있다.
- 다항식 - 기존 데이터와 이들간의 상호작용을 계산해 값을 예측하는 방식이다. 그래프로 그리면 비선형적인 형태가 된다

용도

- 과거나 미래의 예측 - 돌발 변수 출현 가능하기 때문에, 정확한 예측 불가능
- 자료의 범위 밖에 있는 임의의 데이터를 추측하는 목적으로도 이용 가능하다. 반면 자료 가운데 있는 누락된 값을 추측할 때는 내삽법이 사용된다.

https://namu.wiki/w/%EC%99%B8%EC%82%BD%EB%B2%95

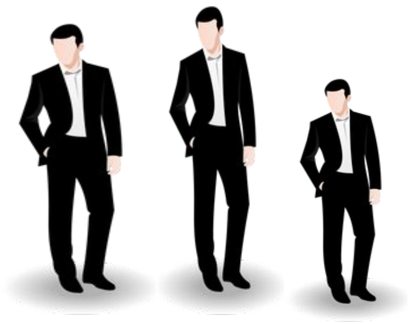
PCA는 스케일 불변성이 아니다.

P-281

다른 방향으로 투영하는 것보다 분산이 최대로 보존되는 축을 선택하는 것이 정보가 가장 적게 손실되므로 합리적으로 보입니다.

Overview and limitations of PCA

- PCA is very widely used. It does not require any distributional assumptions.
- The directions with largest variance are assumed to be of most interest.
- PCA only considers linear combinations of the original variables. (Kernel PCA is an extension of PCA that allows non-linear mappings).
- Dimension reduction can only be achieved if the original variables are correlated. Otherwise, PCA does nothing, except for re-ordering them according to their variance.
- PCA is not scale invariant.



ft : 5.9
lbs: 175

ft : 6.1
lbs: 140

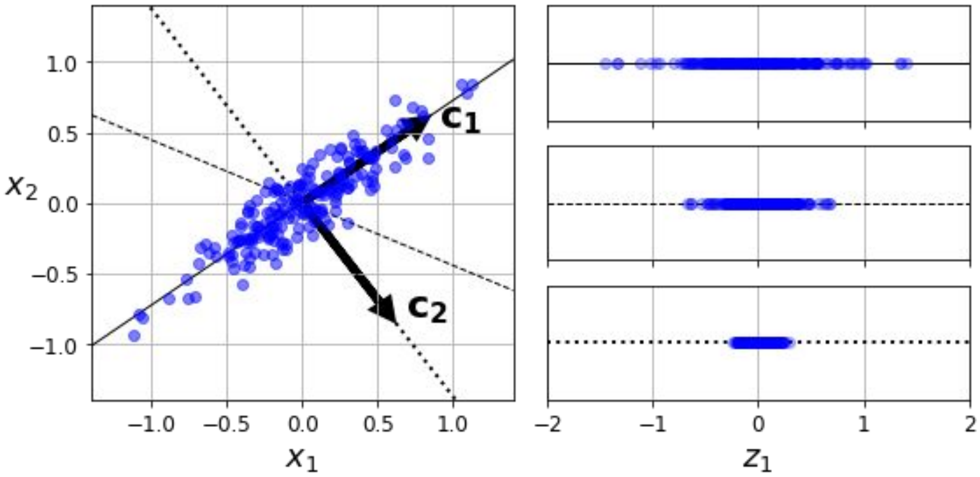
ft : 5.2
lbs: 115

https://dive-into-ds.tistory.com/3

직교를 하고 두번째 축을 찾는 이유가 뭘까요?

P-282

PCA는 훈련 세트에서 분산이 최대인 축을 찾습니다. 또한 **첫 번째 축에 직교**하고 남은 분산을 최대한 보존하는 두 번째 축을 찾습니다.



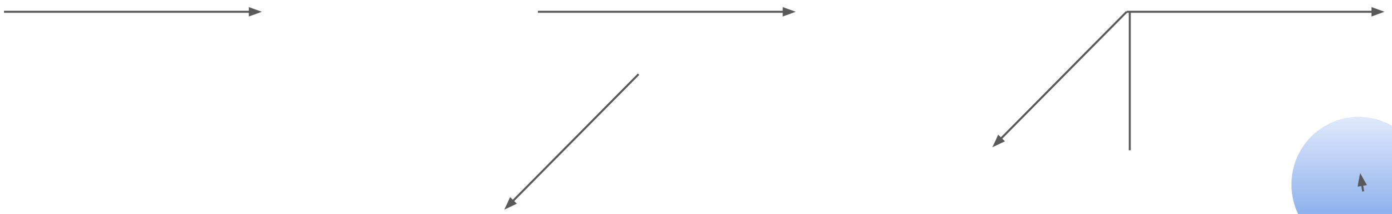
질문

구건모

직교를 하고 두번째 축을 찾는 이유가 뭘까요?

1개의 직선 - 직선
2개의 직선 - 평면
3개의 직선 - 3차원
...

결론: 많은 경우의 수를 줄이기 위해서



1)

질문

 # 백서윤

p 282 책의 오류?! 평면의 수직 축?

[그림 8-2] 에서는 처음 두 개의 PC는 두 화살표가 놓인 **평면의 수직** 축입니다.
그리고 세 번째 PC는 **이 평면에 수직**입니다.

같은 말이 아닌가..?
이해가 잘 되지 않는다

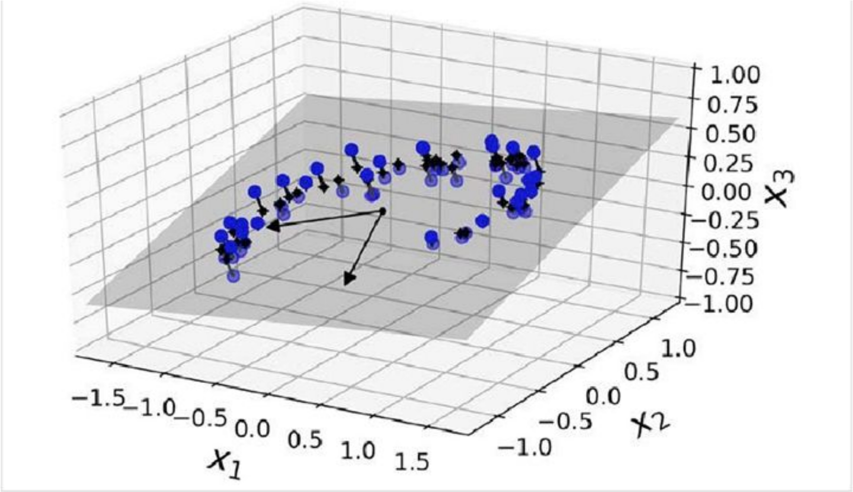


그림 8-2 2차원에 가깝게 배치된 3차원 데이터셋

\# 조성진

차원 축소 vs 특성 제외

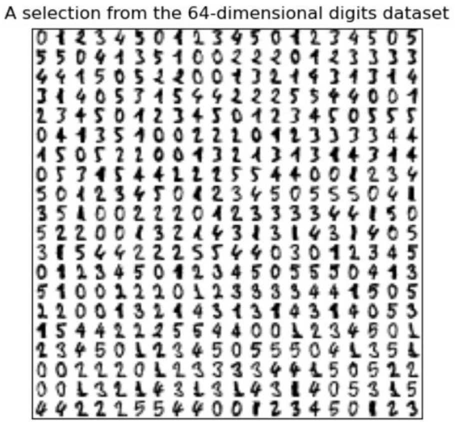
- 무엇이 더 좋을까? ex) 스위스 롤, 타이타닉 객실정보
- 모든 경우를 고려해본 뒤에 결정해야 하는가?

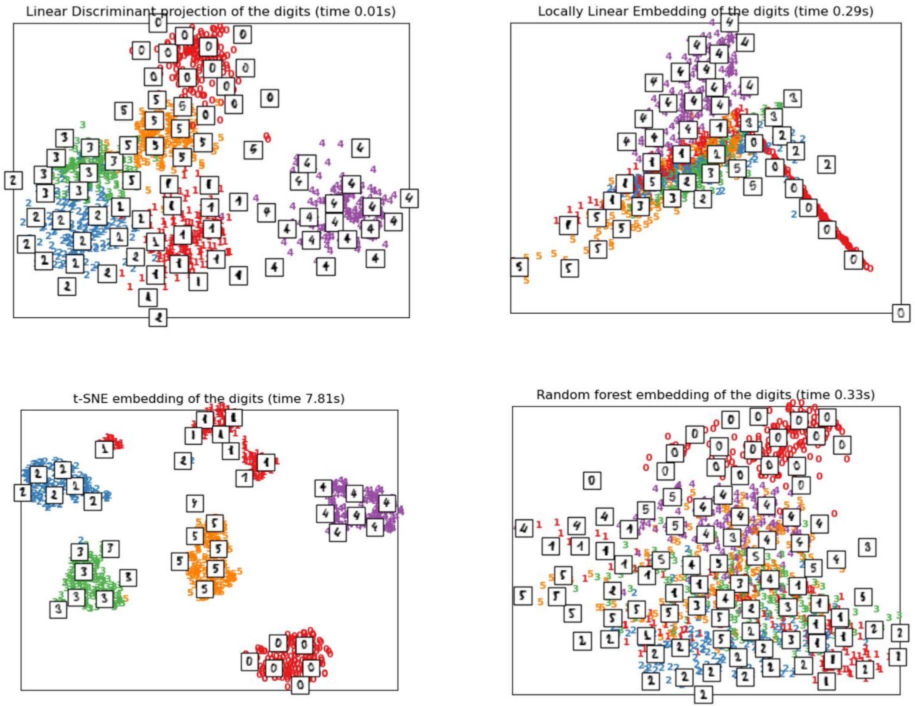
비교대상은 아니라고 본다.
고차원일 때 차원 축소, 상관계수 등으로 사람판단 특성 제외

토의

조성진

매니 폴드 학습 결과에 대한 판단?





https://scikit-learn.org/stable/auto_examples/manifold/plot_lle_digits.html