# 차원 축소와 척도 학습

10.4 커널 선형 차원 축소
10.5 매니폴드 학습
- Isomap
- LLE
10.6 척도 학습

# 10.4.0. 후반부의 내용

Data set 특성의 증가 → 각 특성에 대해 차원 증가

↓

차원의 증가 → 부피가 기하급수적으로 증가, 데이터의 밀도 희소

↓

점 간의 거리가 증가 → 과적합 발생험!

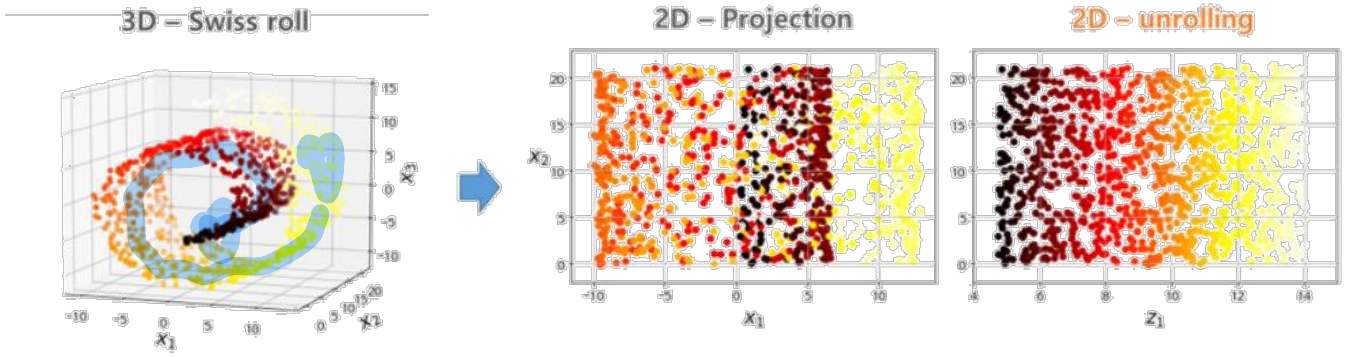
↓

차원 축소 필요

# 10.4.0. 후반부의 내용

## 차원 축소의 두 가지 접근법

- 투영projection : 실제 데이터 셋은 서로 연관된 특성을 가진다. 즉, 특성(차원)이 고르게 분포되어 있지 않다. 데이터 셋은 고차원 공간에서 저차원 부분공간 위에 위치하게 되며, 이를 하나의 특성(차원)으로 표현할 수 있다는 말이 된다. 이를 통해 고차원의 데이터를 저차원으로 투영해 저차원의 데이터 셋으로 만들 수 있다.
- 매니폴드 학습manifold learning: 매니폴드(다양체) 국소적으로 유클리드 공간과 닮은 위상공간이다. 데이터 셋이 위상수학적 구조를 가질 수 있다는 가정에 의해, 고차원인 실제 데이터셋이 더 낮은 저차원 매니폴드에 가깝게 놓여있다고 가정해 차원을 축소시키는 방식이다. 대부분의 차원 축소 알고리즘이 매니폴드를 모델링하는 방식으로 동작된다.



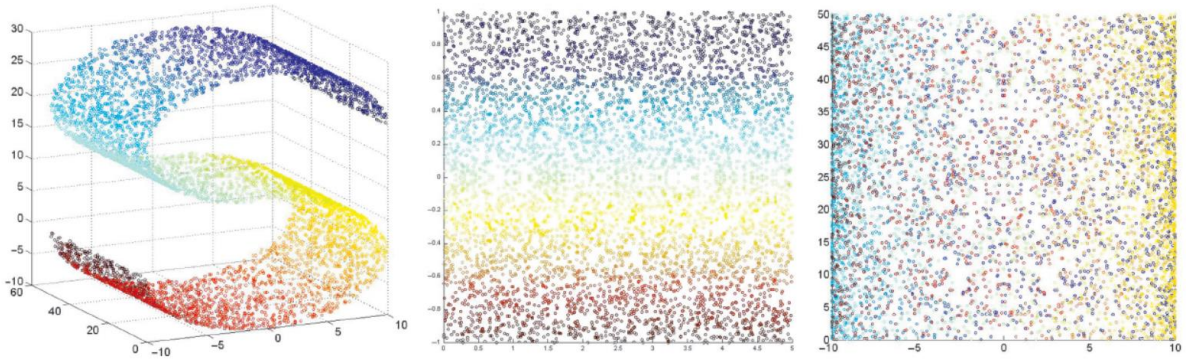
# 10.4.1 커널 선형 차원 축소

- 선형 차원 축소: 고차원 공간에서 저차원 공간으로 매핑하는 함수가 선형이라고 가정.

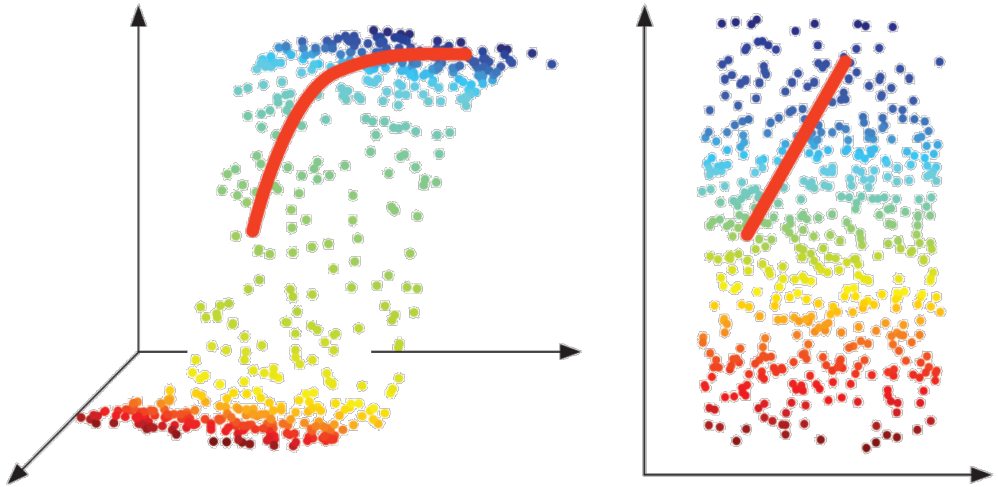
선형 차원 축소의 한계 → 많은 현실 문제는 비선형으로 매핑해야 할 경우가 많음.

- 비선형 차원 축소: 커널 트릭을 통해 선형 차원 축소법에 대해 커널화를 진행. – svm 학습법에서 자세하게 다뤄짐.

  ex) KPCA(커널 주성분 분석) - PCA에 적용 & 비선형 투영으로 차원을 축소



(a) 3차원 공간 (b) 고유 저차원 공간 (c) PCA 차원 축소 결과



(a) 3차원 공간에서 관측한 샘플 포인트 (b) 2차원 공간의 곡면

# 10.4.2 매니폴드 학습

매니폴드 학습: 위상학 개념을 빌린 차원 축소 방법. (매니폴드 - 국소적으로 유클리드 공간과 동형인 공간.)

장점: 차원의 수가 2차원이나 3차원으로 줄어들 때, 데이터에 대한 시각화 가능.

데이터 전처리 때 유용. → SVM, Decision Tree 등 다양한 학습법을 적용하기 전, 데이터를 정제함으로써 학습법의 성능을 올릴 수 있음.

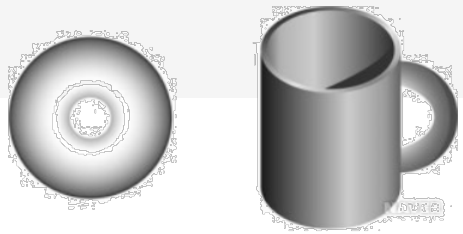
- Isomap – 근접 샘플 사이의 거리를 보존. (고전 알고리즘)
- 국소적 선형 임베딩(Locally Linear Embedding, LEE) – 영역 내 샘플의 선형 관계를 유지 (LLE를 기반으로 한 다양한 알고리즘이 존재)

# 10.4.2 위상학, 동형

- 매니폴드 학습: 위상학 개념을 빌린 차원 축소 방법. (매니폴드 - 국소적으로 유클리드 공간과 동형인 공간.)

**위상학: 연결성이나 연속성 등, 작은 변환에 의존하지 않는 기하학적 성질들을 다루는 수학의 한 분야**

예시는 손잡이가 있는 컵과 도넛, 그리고 안이 꽉 찬 찰흙공과 접시의 같음과 다름을 구분하는 것이다. 어떤 물체를 변형하는데 구부리거나, 늘이거나, 줄일 수 있지만 구멍을 뚫을 수는 없다고 할 때, 찰흙공은 적절한 변형을 통해 접시와 같은 형태로 만들 수 있다. 마찬가지로 컵은 도넛 모양의 찰흙 모형을 가지고 적절한 변형을 통해 손잡이가 있는 컵과 같은 형태로 만들 수 있다.
하지만 찰흙공은 뚫지 않고서는 도넛과 같은 모양이 될 수 없다. 위상수학에서는 손잡이가 있는 컵과 도넛을 같은 형태로, 찰흙공과 접시를 같은 형태로 생각한다. 마찬가지로 도넛과 접시는 다른 형태이다. 이 때, 같은 형태라고 할 수 있는 사물들 사이에 변하지 않는 어떤 공통된 성질을 연구하는 학문으로 위상수학을 소개할 수 있다.
위상수학에서는 위상적 불변성을 공유하는 동형들을 구부리고, 늘이고, 줄이는 것과 같은 변형을 통해 같은 형태로 만들 수 있을 때 같은 도형들로 생각한다. 그리고 위상 수학에서 같은 도형들의 관계는 '동형'이라 일컫는다.
[출처: 네이버 지식백과, 위상수학(Topology)]



손잡이가 있는 컵과 도넛

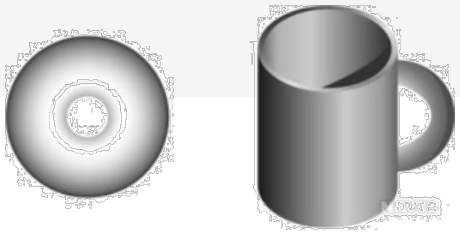
# 10.4.2 위상학, 동형

국소적으로 유클리드 공간의 성질이 있고,
유클리드 거리를 통해 거리를 계산 가능

- 매니폴드 학습: 위상학 개념을 빌린 차원 축소 방법. (매니폴드 - 국소적으로 유클리드 공간과 동형인 공간.)

**위상학**: 연결성이나 연속성 등, 작은 변환에 의존하지 않는 기하학적 성질들을 다루는 수학의 한 분야

예시는 손잡이가 있는 컵과 도넛, 그리고 안이 꽉 찬 찰흙공과 접시의 같음과 다름을 구분하는 것이다. 어떤 물체를 변형하는데 구부리거나, 늘이거나, 줄일 수 있지만 구멍을 뚫을 수는 없다고 할 때, 찰흙공은 적절한 변형을 통해 접시와 같은 형태로 만들 수 있다. 마찬가지로 컵은 도넛 모양의 찰흙 모형을 가지고 적절한 변형을 통해 손잡이가 있는 컵과 같은 형태로 만들 수 있다.
하지만 찰흙공은 뚫지 않고서는 도넛과 같은 모양이 될 수 없다. 위상수학에서는 손잡이가 있는 컵과 도넛을 같은 형태로, 찰흙공과 접시를 같은 형태로 생각한다. 마찬가지로 도넛과 접시는 다른 형태이다. 이 때, 같은 형태라고 할 수 있는 사물들 사이에 변하지 않는 어떤 공통된 성질을 연구하는 학문으로 위상수학을 소개할 수 있다.
위상수학에서는 위상적 불변성을 공유하는 동형들을 구부리고, 늘이고, 줄이는 것과 같은 변형을 통해 같은 형태로 만들 수 있을 때 같은 도형들로 생각한다. 그리고 위상 수학에서 같은 도형들의 관계는 '동형'이라 일컫는다.
[출처: 네이버 지식백과, 위상수학(Topology)]



손잡이가 있는 컵과 도넛

# 10.4.2 매니폴드 학습

매니폴드 학습: 위상학 개념을 빌린 차원 축소 방법. (매니폴드 - 국소적으로 유클리드 공간과 동형인 공간.)

장점: 차원의 수가 2차원이나 3차원으로 줄어들 때, 데이터에 대한 시각화 가능.

데이터 전처리 때 유용. → SVM, Decision Tree 등 다양한 학습법을 적용하기 전,
데이터를 정제함으로써 학습법의 성능을 올릴 수 있음.

- Isomap – 근접 샘플 사이의 거리를 보존. (고전 알고리즘)
- 국소적 선형 임베딩(Locally Linear Embedding, LEE) – 영역 내 샘플의 선형 관계를 유지

( LLE를 기반으로 한 다양한 알고리즘이 존재함. )

# 10.5.1 Isomap

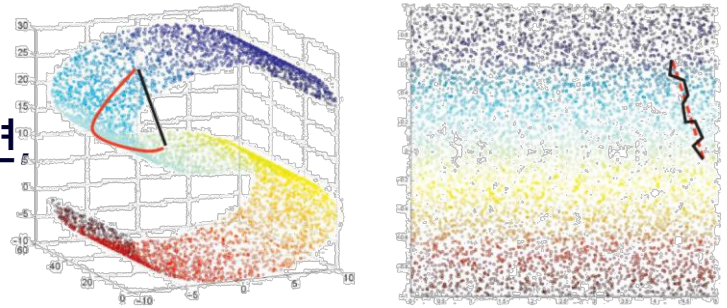
아이디어

→ 저차원 매니폴드를 고차원 공간으로 임베딩한 후 직접 고차원 공간 내에서 직선 거리를 계산하는 것이 적절하지 않다. 매니폴드가 국소적으로 유클리드 공간과 동형인 성질이 있다는 것을 이용해서 각 점에 대해 유클리드 거리에 기반해 최근접 이웃 점들을 찾아 최근접 이웃 그래프를 구축한다. 최근접 이웃 거리에 가까우면 저차원 매니폴드 상의 측지선 거리의 근삿값을 얻을 수 있다.

- 측지선이란? 저차원 매니폴드상에서 두 점 간의 거리

  벌레 한마리가 한 점에서 다른 한 점까지 기어가는 것을 상상해 보면

  곡면 위를 벗어날 수 없는 상황에서는 빨간색 선이 최단 거리가 됨.



(a) 측지선 거리와 고차원 직선 거리 (b) 측지선 거리와 근접 이웃 거리

그림 10.7 ㅣ 저차원 임베딩 매니폴드상의 측지선 거리(빨간색)는 고차원 공간에서 직선 거리 계산을 사용할 수 없다. 하지만 최근접 이웃 거리를 이용해 근삿값을 계산할 수 있다

# 10.5.1 Isomap 알고리즘

**1단계. 인접한 이웃 그래프 구축: 유클리드 거리로 계산.**

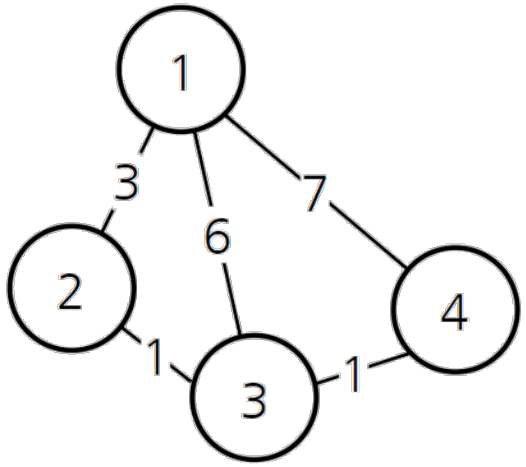
**2단계. 두 점 간의 최단 경로 그래프 계산: 다익스트라 알고리즘, 플로이드 알고리즘 등 이용**

**3단계. MDS 방법론(10.2절)을 사용하여 d차원 임베딩 구축**

최근접 이웃 연결 그래프에서 실제 거리(거리는 최소) 구하는 방법

## 다익스트라 알고리즘

→ 출발-도착 노드의 최단 경로를 찾는 알고리즘




- 각 점의 가장 가까운 이웃과 연결하는 식의 그래프.
- 두 노드 사이의 최단 경로를 구함.
- 간선의 가중치 = 두 연결된 점(노드) 사이의 유클리디안 거리

# 10.5.2 국소적 선형 임베딩 LLE

아이디어

→ 근접 샘플 사이의 거리를 보존하는 Isomap과 달리 영역 내 샘플이 선형 관계를 유지.

국소적 선형 임베딩(Local Linear Embedding, LLE): 고차원의 공간에서 인접해 있는 점(데이터)들 사이의 선형적 구조를 보존하면서 저차원으로 임베딩하는 방법론.

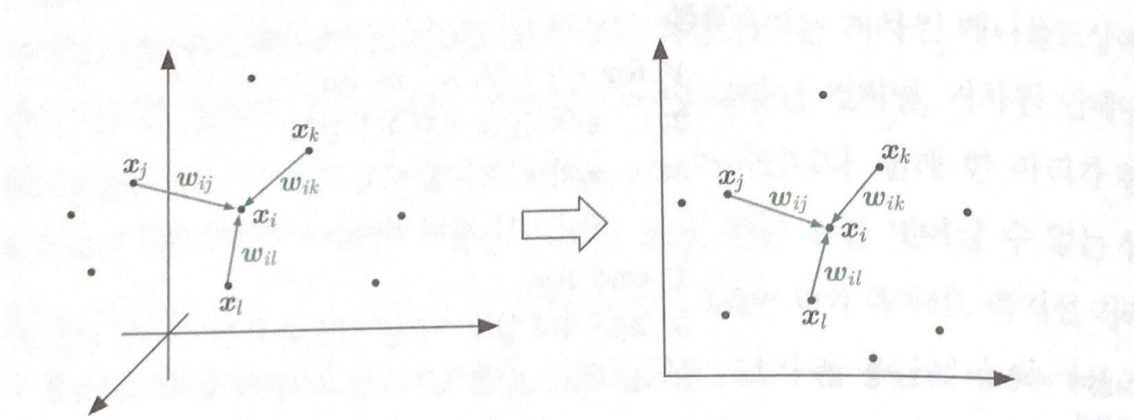



그림 10.9 ＼ 고차원 공간에서의 샘플 구조 관계가 저차원 중에서도 보존되어야 한다

# 10.5.2 LLE 알고리즘: 선형대수의 방법으로 계산

**1단계. 가장 가까운 이웃 검색**

각 점에서 k개의 이웃을 구함.

$$x_i = w_{ij}x_j + w_{ik}x_k + w_{il}x_l \ ,$$

**2단계. 가중치 매트릭스 구성**

현재의 데이터를 나머지 k개의 데이터의 가중치의 합을 뺄 때 최소가 되는 가중치 행렬을 구함.

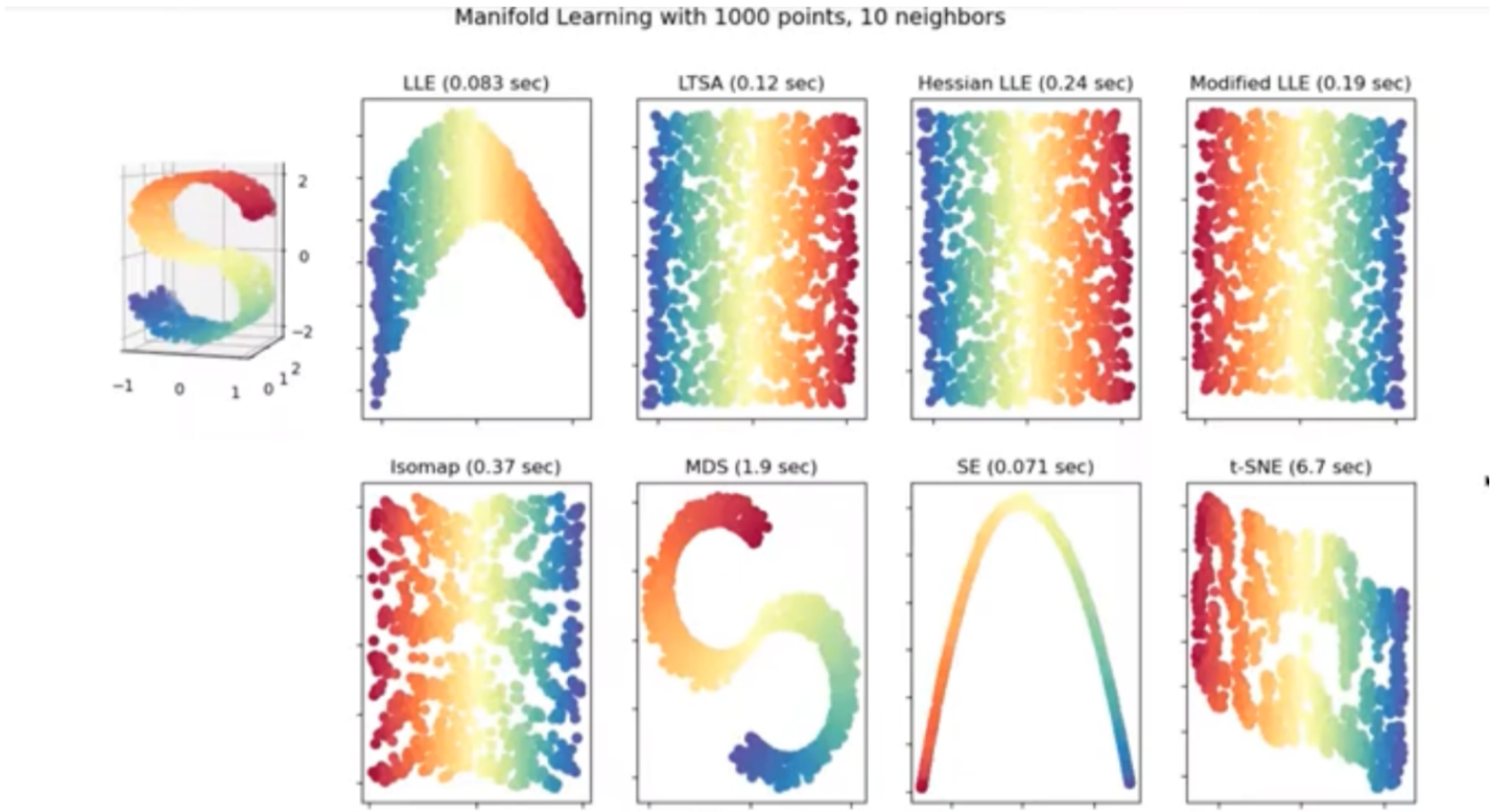
$$E(W) = \sum_i \left| x_i - \sum_j W_{ij}x_j \right|^2$$

**3단계. 부분 고유치 분해**

앞서 구한 가중치를 최대한 보존하며 차원 축소.

이때 차원 축소된 후의 점을 Y로 표현→ 차원 축소된 Yj와의 값 차이를 최소화하는 Y를 구함.

$$\Phi(W) = \sum_i \left| y_i - \sum_j W_i k y_j \right|^2$$




https://youtu.be/jBOVTimnlA8

# 10.6 척도학습

고차원 데이터를 차원 축소하는 목적 → 적합한 저차원 공간을 찾는 것.

이유: 차원 축소된 공간이 원시 공간에서 학습하는 것보다 성능이 좋음.

실제로 각 공간은 샘플속성 위에서 정의한 하나의 거리 척도에 대응함.

적합한 공간을 찾는 것 = 적합한 거리 척도를 찾는 것

척도학습의 동기: 차원 축소 때, 적합한 공간을 찾는 것