M2480.001200 001: 인공지능을 위한 이론과 모델링
2023 가을학기

김지수

2023-11-13

인공지능을 위한 이론과 모델링 소개

비모수적(nonparametric)으로 베이즈(Bayes) 모델링을 하는 방법을 배웁니다.

- 비모수적 베이즈(nonparametric Bayes) 모델링은 다음과 같이 쓸 수 있습니다: 자료 $X_1, \ldots, X_n$이 확률분포 F로부터 iid라 하고, F의 사전분포(prior distribution)을 π라 했을 때, 자료를 관측한 후의 사후분포(posterior distribution)을 알고자 합니다:

$$\pi(F \in A | X_1, \ldots, X_n).$$

- 모수적 베이즈(parametric Bayes)와는 달리, 비모수적 베이즈에서는 사전분포 π, $(X_1, \ldots, X_n)$의 주변분포(marginal distribution), 사후분포 $\pi | X_1, \ldots, X_n$을 수식적으로 닫힌 형태(closed form)으로 구하지 못하는 경우가 많습니다. 그 대신, 각 경우를 계산하는 알고리즘을 배웁니다.

비모수적(nonparametric)으로 베이즈(Bayes) 모델링을 하는 방법을 배웁니다.

- 비모수적 베이즈(nonparametric Bayes) 모델링은 다음과 같이 쓸 수 있습니다: 자료 $X_1, \ldots, X_n$이 확률분포 F로부터 iid라 하고, F의 사전분포(prior distribution)을 π라 했을 때, 자료를 관측한 후의 사후분포(posterior distribution)을 알고자 합니다:

$$\pi(F \in A | X_1, \ldots, X_n).$$

- 다음과 같은 통계적인 문제를 어떻게 계산하는지 배웁니다: 누적분포함수(cumulative distribution function) 추정, 밀도(density function) 추정

Concentration inequality는 확률변수의 행태를 확률적으로 통제합니다.

- 약한 대수의 법칙(law of large number)은, $X_1, \ldots, X_n$이 iid 자료이고 모평균 $\mathbb{E}[X_1] = \mu < \infty$일 때, 표본평균이 모평균으로 확률수렴함을 뜻합니다:

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right) = 0 \text{ for all } \epsilon > 0.$$

- 표본평균을 더 일반적인 함수로 확장시켰을 때 확률변수의 행태를 다음과 같이 확률적으로 부등호로 통제하는 것 (및 이에 필요한 기술들)을 Concentration inequality라 합니다:

$$P\left(\sup_{f\in\mathcal{F}} |f(X_1, \ldots, X_n) - \mathbb{E}\left[f(X_1, \ldots, X_n)\right]| > \epsilon\right) < \delta.$$

Concentration inequality는 확률변수의 행태를 확률적으로 통제합니다.

- 표본평균을 더 일반적인 함수로 확장시켰을 때 확률변수의 행태를 다음과 같이 확률적으로 부등호로 통제하는 것 (및 이에 필요한 기술들)을 Concentration inequality라 합니다:

$$P\left(\sup_{f\in\mathcal{F}}|f(X_1,\ldots,X_n)-\mathbb{E}\left[f(X_1,\ldots,X_n)\right]|>\epsilon\right)<\delta.$$

- 다음과 같은 concentration inequality를 배웁니다: Hoeffding's inequality, McDiarmid's inequality, Bernstein's inequality, Rademacher Complexity, VC dimension, uniform bound 등

추정량(estimator)의 최대위험(maximum risk)은 추정량의 최악의 경우에 오차의 기대값(expected error)입니다.

- 추정량(estimator) $\hat{\theta}_n$ 의 최대위험(maximum risk)은 최악의 경우에 추정량 $\hat{\theta}_n$ 이 만들어낼 수 있는 오차의 기대값(expected error)입니다.
- $$\sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\ell \left(\hat{\theta}_n(X),\ \theta(P) \right) \right]$$
  - $X = (X_1, \cdots, X_n)$는 고정된 분포 P에서 추출하고, P는 확률분포의 집합 $\mathcal{P}$에 속합니다.
  - 추정량 $\hat{\theta}_n$ 은 자료 X의 임의의 함수입니다.
  - 손실함수 $\ell(\cdot, \cdot)$는 추정량 $\hat{\theta}_n$의 오차를 잽니다.

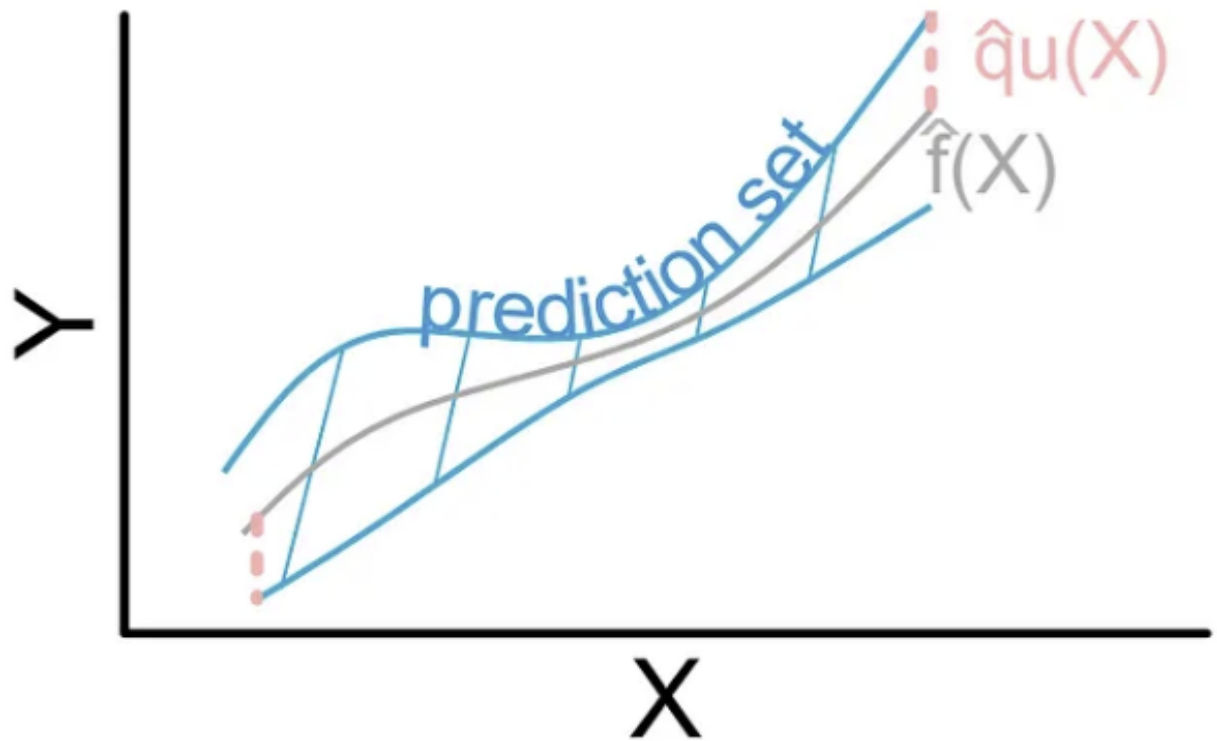
미니맥스 위험(minimax risk)은 모수(parameter) 추정의 통계적 어려움을 묘사합니다.

- 미니맥스 위험(minimax risk) R_n은 최악의 경우에도 잘 작동하는 추정량(estimator)의 위험(risk)입니다. 이를 표본크기(sample size)의 함수로 봅니다.
- $$R_n = \inf_{\hat{\theta}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\ell \left(\hat{\theta}_n(X),\ \theta(P) \right) \right]$$
  - $X = (X_1, \cdots, X_n)$는 고정된 분포 P에서 추출하고, P는 확률분포의 집합 $\mathcal{P}$에 속합니다.
  - 추정량 $\hat{\theta}_n$ 은 자료 X의 임의의 함수입니다.
  - 손실함수 $\ell(\cdot, \cdot)$는 추정량 $\hat{\theta}_n$의 오차를 잽니다.

Conformal Prediction은 예측에 대한 신뢰구간을 제공합니다.

- 자료 $(X_1, Y_1), \ldots, (X_n, Y_n)$이 주어졌을 때, 함수값이 랜덤한 집합이면서 다음을 만족하는 C_n을 찾습니다:

$$P\left(Y_{n+1} \in C_n(X_{n+1})\right) \geq 1 - \alpha.$$

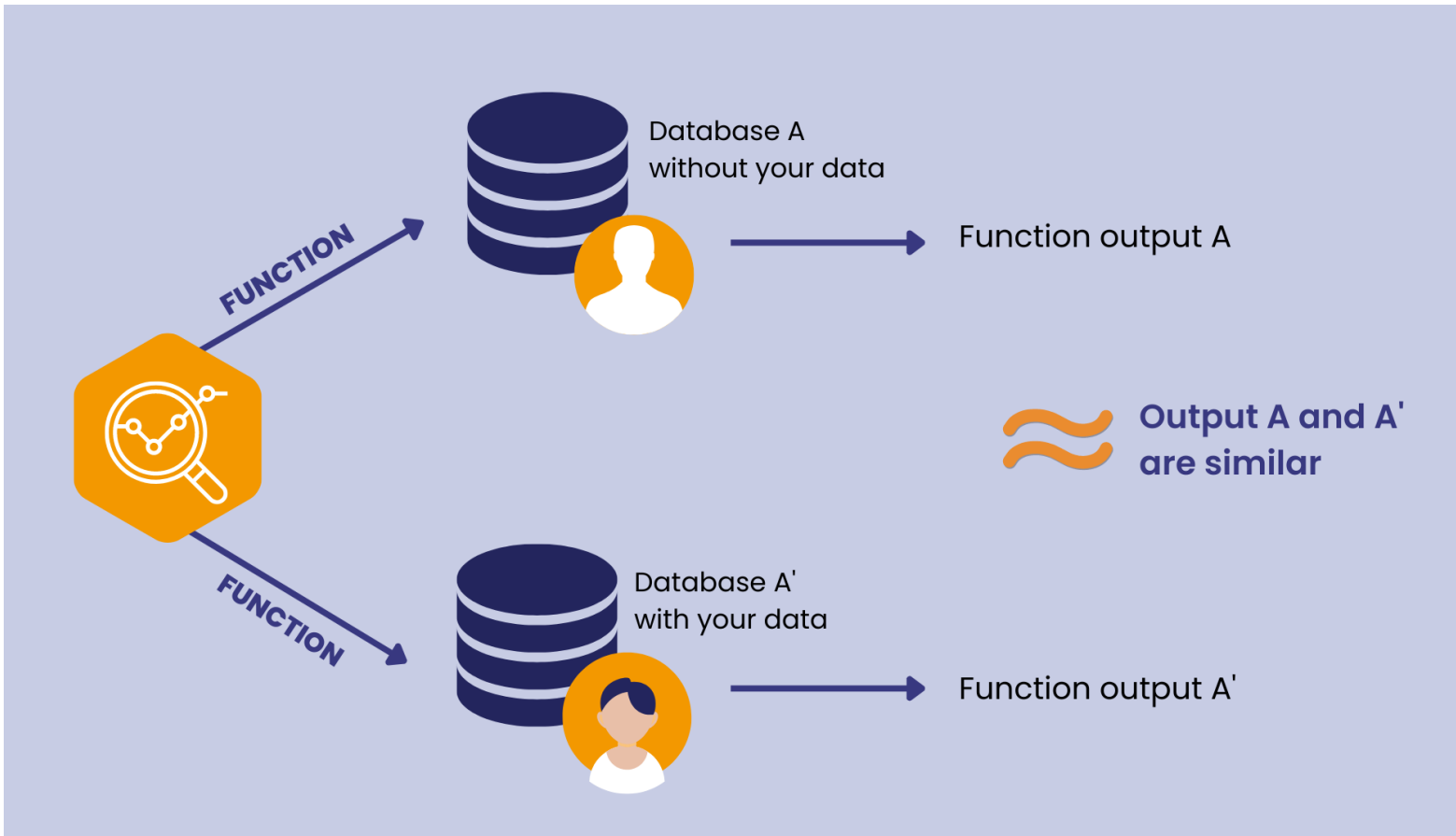


[1]

[1] https://towardsdatascience.com/conformal-prediction-4775e78b47b6

Differential Privacy는 개개인의 정보를 숨기면서 집단에 대한 자료 분석을 가능하게 하는 통계적인 방법론입니다.

- 자료 $D = \{X_1, \ldots, X_n\}$으로부터 자료 분석의 결과물 Z가 나온다고 할 때, Differential Privacy는 $X_1, \ldots, X_n$을 정확히 알지 않고서도 자료 분석의 결과물이 비슷하게 나오도록 합니다.



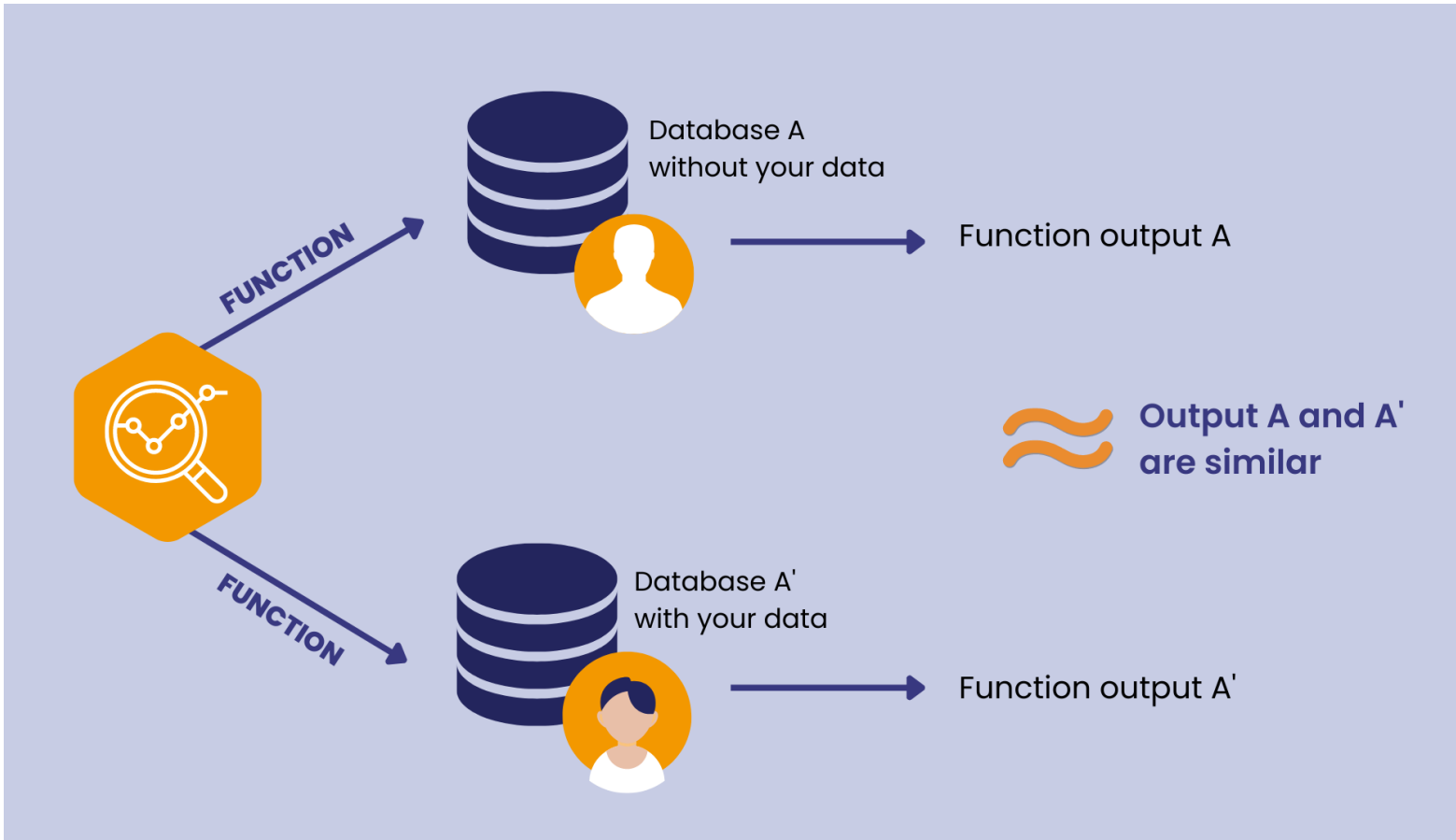
[2]

[2] https://www.statice.ai/post/what-is-differential-privacy-definition-mechanisms-examples

Differential Privacy는 개개인의 정보를 숨기면서 집단에 대한 자료 분석을 가능하게 하는 통계적인 방법론입니다.

- $D = \{X_1, \ldots, X_{i-1}, X_i, X_{i+1}, \ldots X_n\}$로부터 표본 하나를 바꾼 자료 $D' = \{X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots X_n\}$가 있을 때, 다음을 만족하면 ϵ-differential privacy가 있다고 합니다:

$$P(Z \in A|D) \leq e^{\epsilon} P(Z \in A|D')$$



[3]

[3]

Wasserstein distance는 하나의 확률분포를 다른 확률분포로 옮기는 데에 비용이 얼마나 드는지 계산합니다.

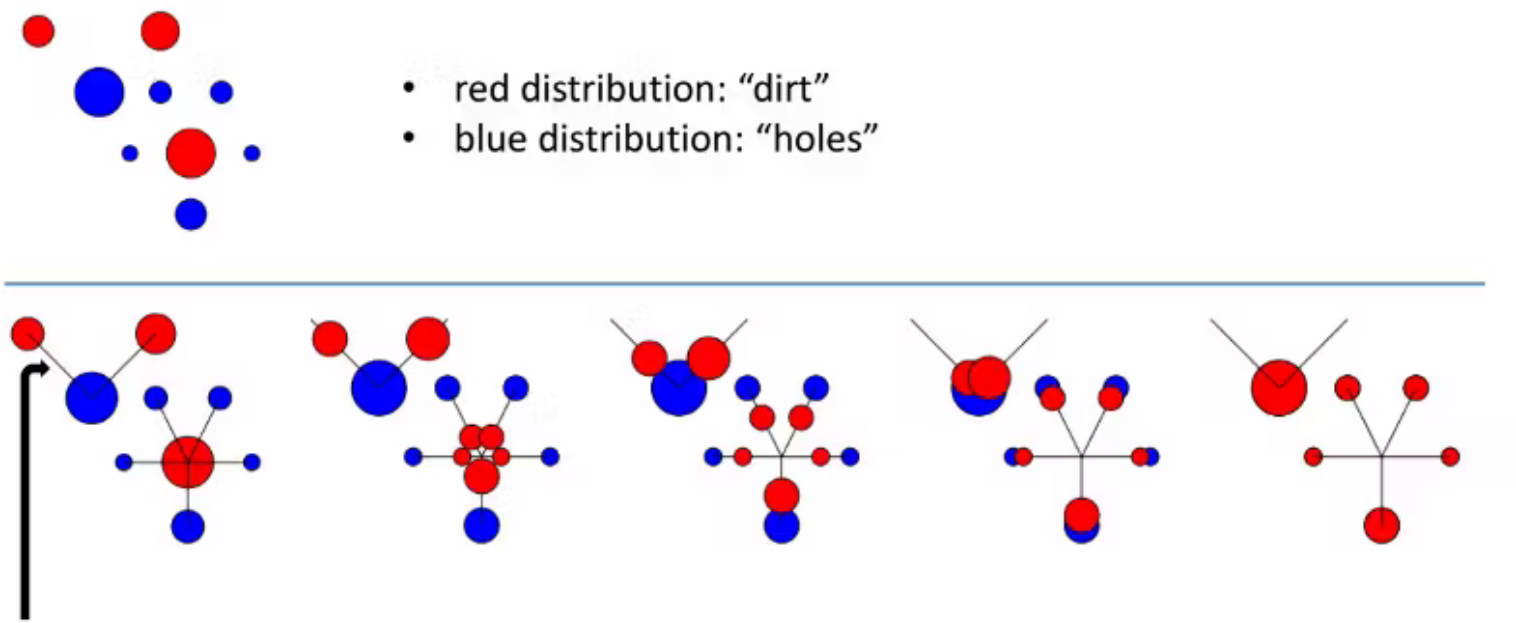
- 두 확률분포 P, Q가 있을 때, $\mathcal{J}(P, Q)$가 주변분포(marginal distribution)가 P와 Q인 결합분포(joint distribution)를 모아놓은 집합이라고 하면, p-Wasserstein distance는 다음과 같이 정의됩니다:

$$W_p(P, Q) = \left(\inf_{J \in \mathcal{J}(P,Q)} \int \|x - y\|^p \, dJ(x, y) \right)^{1/p}.$$

- 최소값을 주는 J를 optimal transport plan 내지는 optimal coupling 이라고 부릅니다.
- $p = 1$일 때 Earth Mover distance라고도 부릅니다: 흙을 옮겨서 구멍을 메우는 데에 필요한 노동력을 뜻합니다.

Wasserstein distance는 하나의 확률분포를 다른 확률분포로 옮기는 데에 비용이 얼마나 드는지 계산합니다.

- $$W_p(P,Q) = \left(\inf_{J \in \mathcal{J}(P,Q)} \int \|x-y\|^p \, dJ(x,y) \right)^{1/p}.$$
- $p = 1$일 때 Earth Mover distance라고도 부릅니다: 흙을 옮겨서 구멍을 메우는 데에 필요한 노동력을 뜻합니다.



4

[4] https://anebz.eu/earth-mover-distance

고차원에서 이표본검정(two sample testing)을 합니다.

- 두 iid 자료

$$X_1, \ldots, X_n \sim P, \qquad Y_1, \ldots, Y_m \sim Q$$

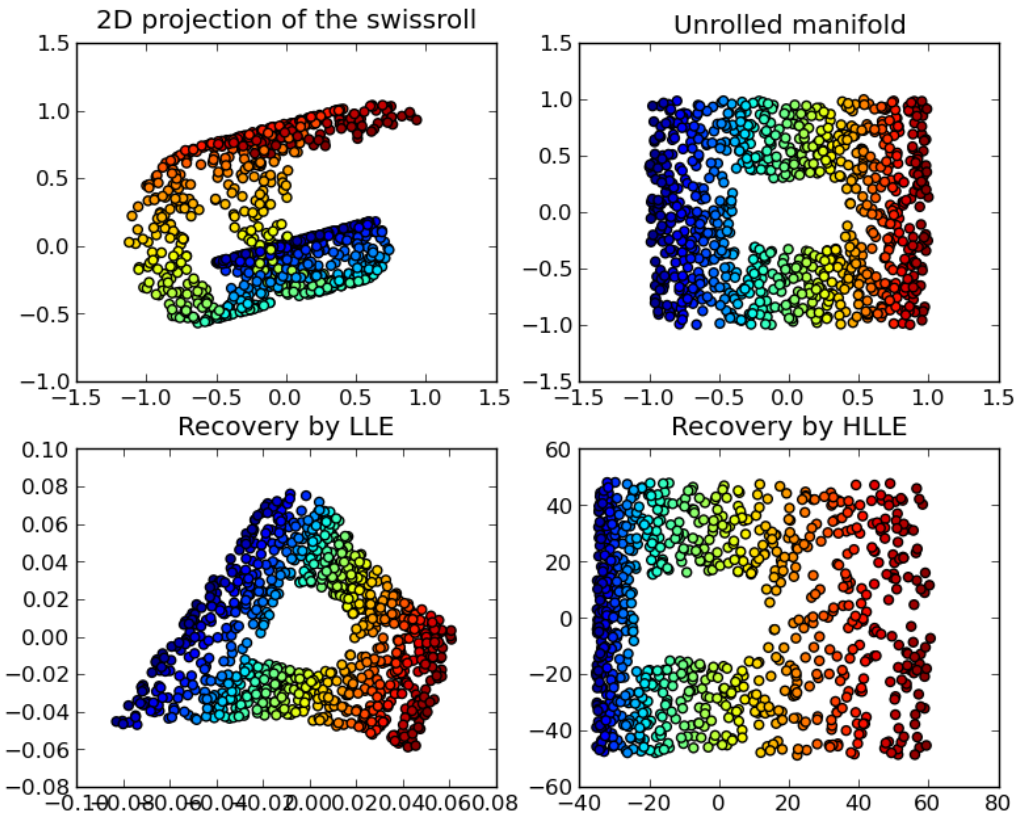
를 관측했을 때, 다음을 검정하고자 합니다:

$$H_0 : P = Q \qquad \text{versus} \qquad H_1 : P \neq Q.$$

- 특히 자료가 고차원에 있을 때, 저차원의 이표본검정(two sample testing)이 어떤 문제가 생기고 고차원에선 어떤 검정을 쓸 수 있는지 알아봅니다.

차원축소(dimension reduction)의 다양한 방법을 알아봅니다.

- 고차원의 자료를 저차원으로 표현하는 것을 차원축소(dimension reduction)이라고 합니다.

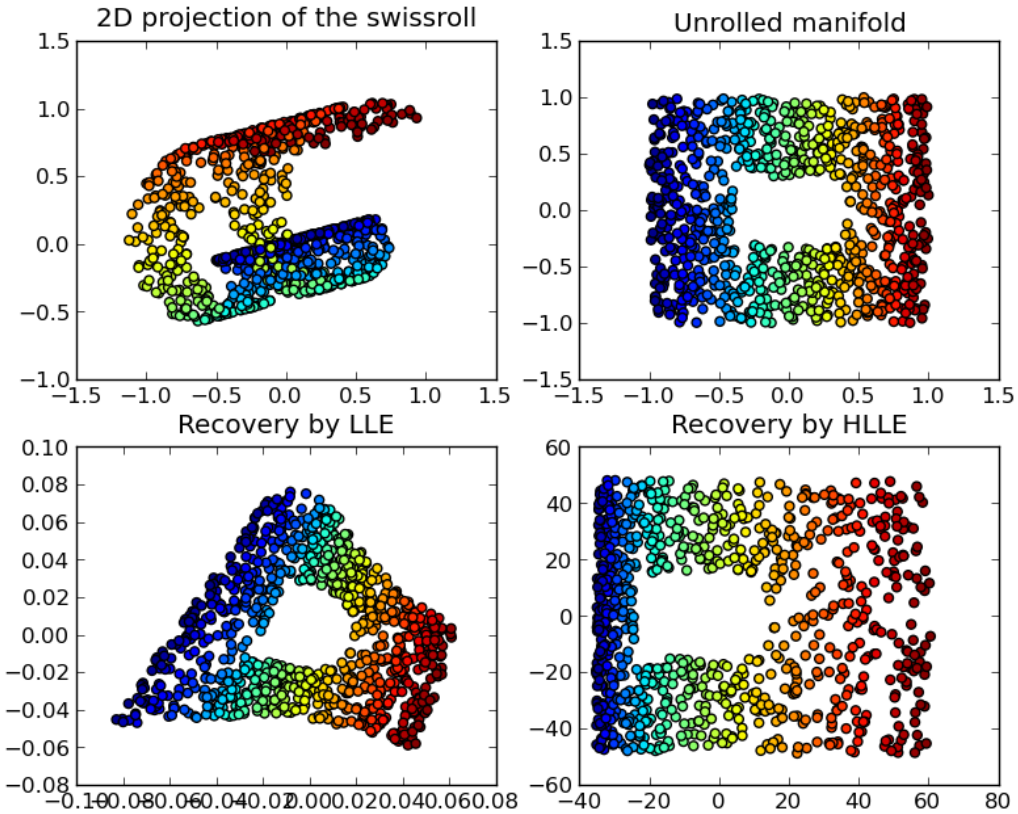


[5]

[5] By Olivier Grisel - Generated using the Modular Data Processing toolkit and matplotlib., CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=10661091

차원축소(dimension reduction)의 다양한 방법을 알아봅니다.

- 차원축소의 다양한 방법을 알아봅니다: Principal Component Analysis(PCA), Multidimensional Scaling, Local Linear Embedding, Isomap, Laplacian Eigenmaps, Manifold Learning, Random Projection 등



6

[6]