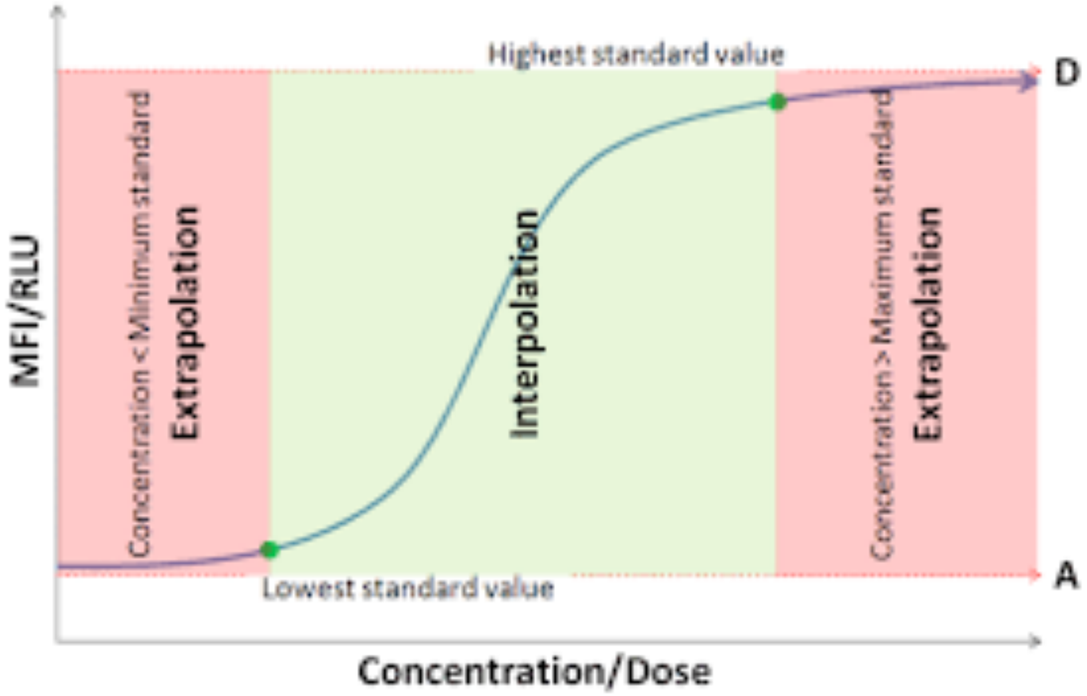
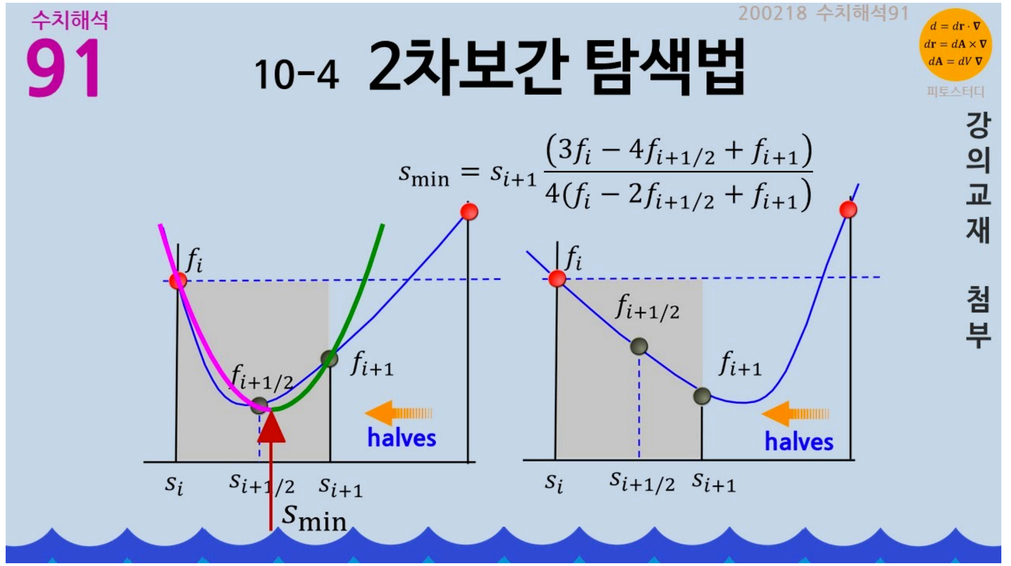
230924_7기_데이터분석기초

결측치 & 누락값 대체방법

보간법과 보외법 (수치해석)

- 보간법 (Interpolation)
- 내삽이라 하며 특정한 두 점 안쪽에 놓여있는 가능한 값을 구하려는 방법
- 보외법 (Extrapolation)
- 외삽 관찰 범위를 넘어서 다른 변수와의 관계에 기초하여 변수의 값을 추정하는 과정





E.g)라그랑주 보간법

Lagrange interpolation

수치해석에서 **라그랑주 다항식**은 **라그랑주 형식**에서 데이터 포인트의 주어진 집합으로부터 다항식을 보간하는 방법으로, 조제프루이 라그랑주의 이름에서 왔다. 이것은 1779년 에드워드 웨어링에 의해 처음으로 발견되었고, 1783년에 레온하르트 오일러에 의해 마지막으로 재발견되었다.

정의 [편집]

$k+1$ 데이터 포인트의 주어진 집합

$$(x_0, y_0), \ldots, (x_j, y_j), \ldots, (x_k, y_k)$$

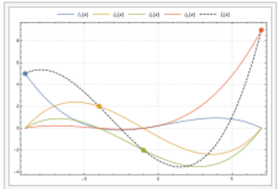
여기서 x_j는 두 개의 같은 값이 존재하지 않고, **라그랑주 형식의 보간 다항식**은 선형 결합

$$L(x) := \sum_{j=0}^{k} y_j \ell_j(x)$$

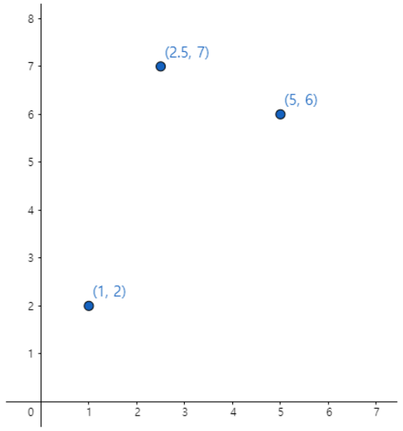
이다. 이것의 라그랑주 기초 다항식은 다음과 같다.

$$\ell_j(x) := \prod_{f=0,\, f\neq j}^{k} \frac{x - x_f}{x_j - x_f} = \frac{(x-x_0)}{(x_j-x_0)} \cdots \frac{(x-x_{j-1})}{(x_j-x_{j-1})} \frac{(x-x_{j+1})}{(x_j-x_{j+1})} \cdots \frac{(x-x_k)}{(x_j-x_k)}.$$

x_i는 두 개의 같은 값이 존재하지 않기 때문에(그리고 존재할 수도 없다, 그렇지 않으면 데이터 집합의 의미가 모순된다), $x_j - x_f \neq 0$이 표현이 잘 정의 된다.



이 그림은 네개의 점((−9, 5), (−4, 2), (−1, −2), (7, 9))에 대하여 (입방체의) 보간 다항식 $L(x)$ (검정)을 보여준다. 이것은 *크기가 변형된* 기초 다항식 ($y_0L_0(x)$, $y_1L_1(x)$, $y_2L_2(x)$ 그리고 $y_3L_3(x)$)의 합이다. 보간 다항식은 4개의 모든 컨트롤 포인트를 지나고, 각각 *크기가 변형된* 기초 다항식은 그것들 각각의 컨트롤 포인트를 지나고 x가 다른 세개의 컨트롤 포인트에 부합되는 곳에서 0이다.

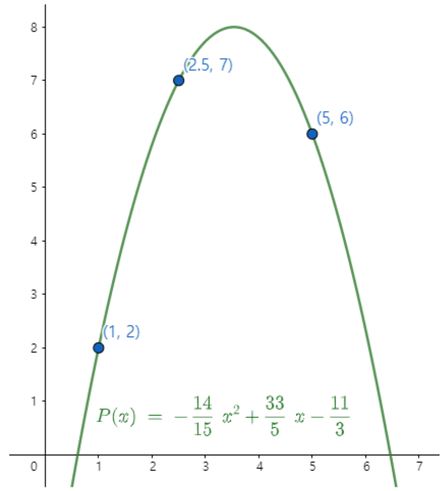


(1,2), (2.5,7), (5,6) 을 가지고 Ln,k 를 세우면

$$L_{2,0} = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} = \frac{(x-2.5)(x-5)}{(1-2.5)(1-5)}$$

$$L_{2,1} = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} = \frac{(x-1)(x-5)}{(2.5-1)(2.5-5)}$$

$$L_{2,2} = \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} = \frac{(x-1)(x-2.5)}{(5-1)(5-2.5)}$$



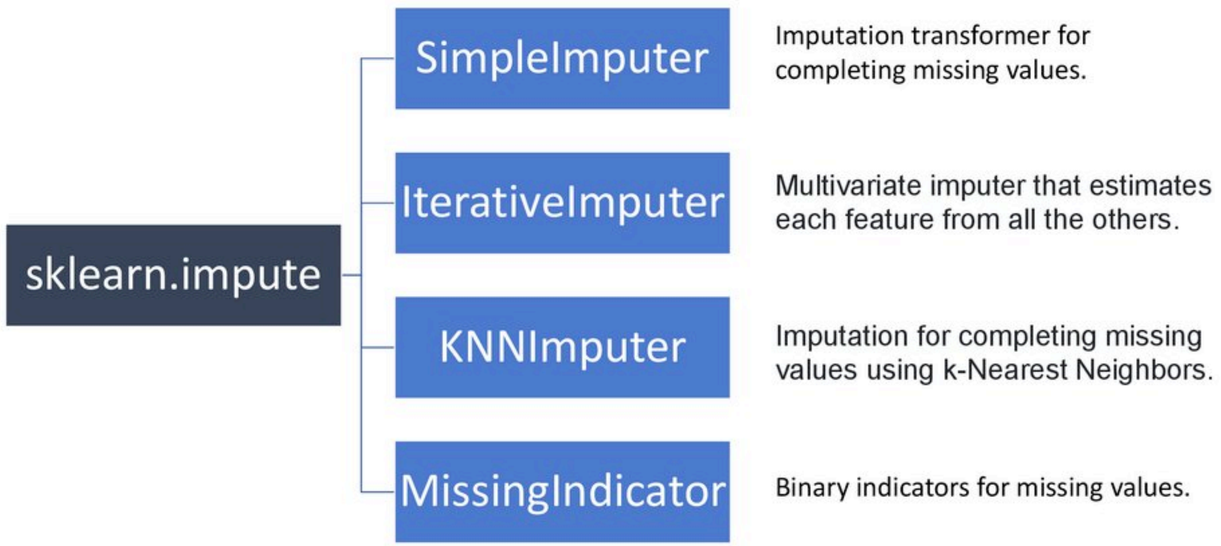
주어진 데이터와 L2,k 를 사용해 라그랑주 다항식을 세웁니다.

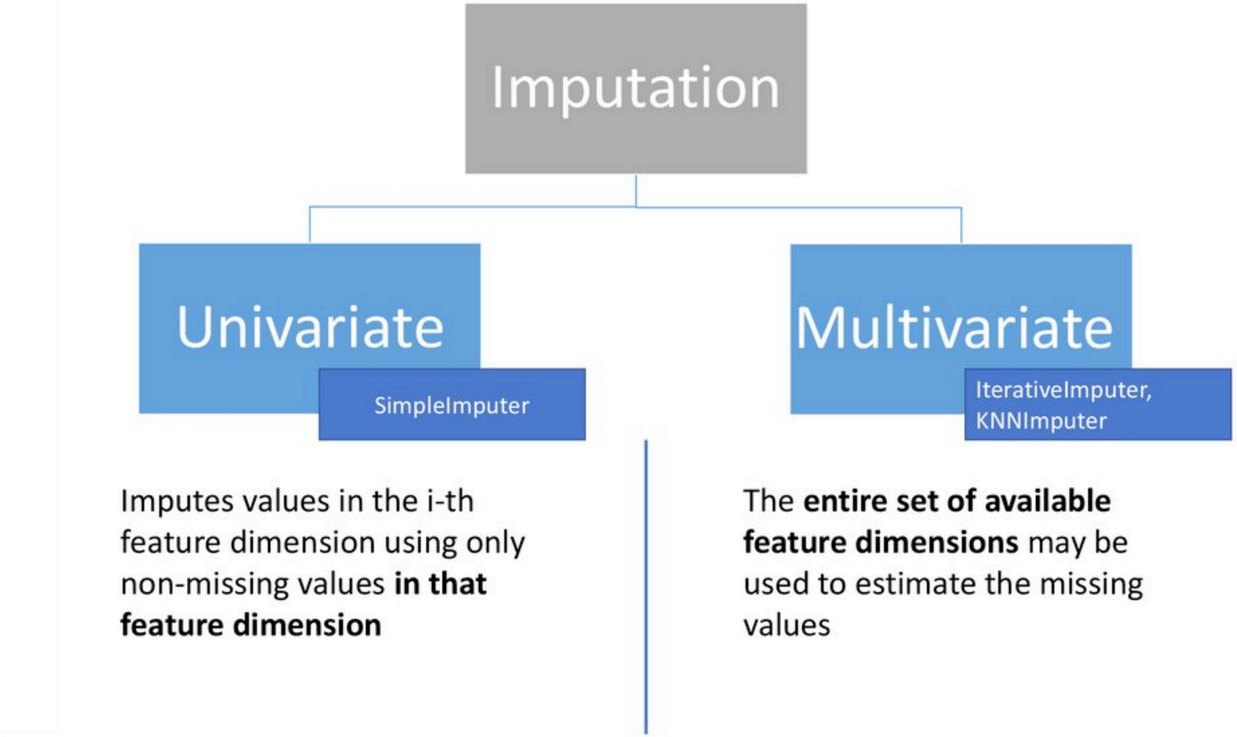
$$P(x) = \frac{2}{6}(x^2 - 7.5x + 12.5) - \frac{28}{15}(x^2 - 6x + 5) + \frac{6}{10}(x^2 - 3.5x + 2.5)$$

$$\Rightarrow P(x) = -\frac{14}{15}x^2 + \frac{33}{5}x - \frac{11}{3}$$

Sklearn impute

What's inside this module?



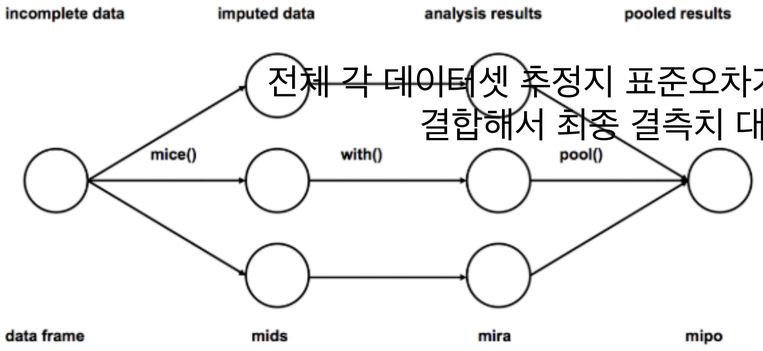


Iterative Imputer

결측치를 대체할 때
예를 들어 단순선형회귀 대치 -> 연령과 소득
소득에 대한 데이터가 결측치가 있다. 연령이란 데이터는 결측치가 없다.
단순하게 연령과 소득에 대한 관계로 둘의 선형관계 상관성을 보고
둘의 회귀식의 관계로 단순하게 결칙치를 대체하는 것
연령 어느정 높아지면 소득도 올라간다는 패턴

1. **Imputation**: Impute the missing entries of the incomplete data sets m times (m=3 in the figure). Note that imputed values are drawn from a distribution. Simulating random draws doesn't include uncertainty in model parameters. Better approach is to use Markov Chain Monte Carl (MCMC) simulation. This step results in m complete data sets.
2. **Analysis**: Analyze each of the m completed data sets.
3. **Pooling**: Integrate the m analysis results into a final result

서로다른 결측치를 대체할 수 있는 다른 변수들과 함께 데이터 표본을 추출하면서 진행하고



전체 각 데이터셋 추정지 표준오차가 계산된 n개의 데이터 셋에서
결합해서 최종 결측치 대치값을 산출하는 조건

생성된 데이터들 -> 통계적으로 모수를 추정치 그리고 표준오차등을 계산해서

단순 선형 회귀 대치법

단순하게 접근하는 것이 아니라 회귀식에 확률 오차항 추가해서
확률적으로 회귀 대치를 진행
~~표본오차 문제~~ 여러가지 문제들이 발생해서 이 부분을 해결하기 위해서 다중대치법을 사용한다.

확률적 회귀 대치법

위의 단순, 확률적 회귀대치법을 보완하기 위해서 단순 대치를 여러번 수행하는 것 가상 데이터를 샘플 뽑아서 대치값들을 분석하고 결국 최종 결측치를 대체할 것을 선정

다중 대치법

MICE 설명 :https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/

몬테카를로 MCMC

A Markov Chain is a mathematical process that undergoes transitions from one state to another. Key properties of a Markov process are that it is random and that each step in the process is "memoryless;" in other words, the future state depends only on the current state of the process and not the past

https://www.publichealth.columbia.edu/research/population-health-methods/markov-chain-monte-carlo