4.5 변분법

데이터 분석에서는 함수와 더불어 다양한 범함수가 사용된다. 이 절에서는 범함수의 개념과 범함수의 미분에 해당하는 변분법을 공부한다.

범함수

함수(function)는 실수(real number)를 입력받아 실수를 출력한다.

$$\text{real number } x \rightarrow \boxed{\text{function} \quad f} \rightarrow \text{real number} \tag{4.5.1}$$

그런데 앞으로 공부하게 될 기댓값, 엔트로피 등을 계산할 때는 함수를 입력받고 이를 기반으로 실수를 출력한다. 이렇게 **함수를 입력으로 받아 실수를 출력하는 것을 범함수(functional)**라고 한다.

$$\text{function } y(x) \rightarrow \boxed{\text{functional} \quad F} \rightarrow \text{real number} \tag{4.5.2}$$

범함수는 보통 알파벳 대문자로 표기한다. 함수에서는 입력변수를 실수를 소괄호(parenthesis)로 감싸지만 범함수는 입력변수인 함수를 대괄호(square bracket)로 감싼다.

$$F[y(x)] \tag{4.5.3}$$

범함수 값은 정적분으로 계산하는 경우가 많다. 예를 들어 확률변수 X의 기댓값과 엔트로피는 확률밀도함수 $p(x)$를 다음처럼 적분한 값이다. 구체적인 정의는 확률론에서 공부하게 된다. 여기에서는 함수 $p(x)$를 입력으로 받아서 스칼라 실수를 출력하는 범함수 E와 H라는 것이 있다는 것만 알면 된다.

$$\text{E}[p(x)] = \int_{-\infty}^{\infty} xp(x)dx \tag{4.5.4}$$

$$\text{H}[p(x)] = -\int_{-\infty}^{\infty} p(x)\log p(x)dx \tag{4.5.5}$$

입력인 함수가 변할 때 범함수의 출력이 어떻게 달라지는지를 계산하는 학문을 **변분법(functional calculus, calculus of variations)**이라고 한다.

범함수의 테일러 전개

함수 $f(x)$의 도함수 $\frac{df}{dx}$를 알면 다음처럼 함수의 근사값을 구할 수 있다. 이를 함수의 테일러 전개(Taylor expansion)라고 한다. 이 식에서 ϵ은 아주 작은 실수를 의미한다.

$$f(x+\epsilon) \approx f(x) + \frac{df}{dx}\epsilon \tag{4.5.6}$$

범함수에 대해서도 마찬가지로 테일러 전개를 할 수 있다. 범함수의 테일러 전개를 이해하기 위해 우선 범함수가 아닌 단변수 함수의 테일러 전개부터 해보자. 설명의 편의를 위해 함수이름을 F, 독립변수를 y라고 하자. $F(y)$는 실수를 입력으로 받는 함수이지만 나중에 범함수와 비교하기 위해 편의상 대문자로 표시하였다. 입력 변수도 보통 사용하는 x 대신 y라는 알파벳을 사용하였다.

함수 F의 테일러 전개식은 다음과 같다.

$$F(y+\epsilon) \approx F(y) + \frac{dF}{dy}\epsilon \tag{4.5.7}$$

F가 아직은 범함수가 아니라 함수이고 y도 아직은 함수가 아니라 변수이므로 이 식은 위 테일러 전개식을 함수와 변수 이름만 바꿔 쓴 것이다.

만약 F가 단변수 함수가 아니라 $y_1, y_2, \ldots, y_N$ 이라는 N개의 실수 입력을 받는 다변수 함수라면 테일러 전개식은 다음처럼 쓸 수 있다.

$$\begin{aligned} F(y_1+\epsilon_1, y_2+\epsilon_2, \ldots, y_N+\epsilon_N) &\approx F(y_1, y_2, \ldots, y_N) + \frac{\partial F}{\partial y_1}\epsilon_1 + \frac{\partial F}{\partial y_2}\epsilon_2 + \cdots \frac{\partial F}{\partial y_N}\epsilon_N \\ &= F(y_1, y_2, \ldots, y_N) + \sum_{i=1}^{N} \frac{\partial F}{\partial y_i}\epsilon_i \end{aligned}$$

위 식에서 y_i 이 x_i를 입력으로 받아서 계산된 함수의 값 $y(x_i)$ 이라고 가정한다.

$$y_i = y(x_i) \tag{4.5.9}$$

그리고 ϵ_i는 x_i를 입력으로 받는 임의의 함수 $\eta(x)$의 값 $\eta(x_i)$ 에 아주 작은 공통 상수 ϵ을 곱한 값이라고 가정한다.

$$\epsilon_i = \epsilon\eta(x_i) \tag{4.5.10}$$

그러면 위 테일러 전개식은 다음처럼 쓸 수 있다.

$$F\big(y(x_1)+\epsilon\eta(x_1), y(x_2)+\epsilon\eta(x_2), \ldots, y(x_N)+\epsilon\eta(x_N)\big) \approx F\big(y(x_1), y(x_2), \ldots, y(x_N)\big)$$

위 식은 실수 벡터 또는 수열(sequence)를 입력으로 받고 실수를 출력하는 함수에 대한 테일러 전개이다.

$$\text{sequence } \{y(x_1), \cdots, y(x_N)\} \quad \rightarrow \quad \text{function } F\big(\{y(x_1), \cdots, y(x_N)\}\big) \tag{4.5.12}$$

여기에서 수열의 크기 N을 무한대로 늘리면 수열 $\{y(x_1), \cdots, y(x_N)\}$는 $y(x)$라는 함수를 나타낸다고 볼 수 있다. 그러면 함수 F도 이제는 함수가 아니라 함수 y를 입력으로 받는 범함수가 된다.

$$\text{function } y(x) \quad \rightarrow \quad \text{functional } F[y(x)] \tag{4.5.13}$$

위 식에서 수열 $\{y(x_1), \cdots, y(x_N)\}$를 함수 y로 바꿔 쓰면 다음과 같다.

$$F[y(x)+\epsilon\eta(x)] \approx F[y(x)] + \epsilon\int \frac{\delta F}{\delta y(x)}\eta(x)dx \tag{4.5.14}$$

이 식이 바로 범함수에 대한 테일러 전개이다.

범함수의 도함수

위 식에서 ϵ의 변화에 의한 범함수 값의 변화는

$$\frac{F[y(x)+\epsilon\eta(x)]-F[y(x)]}{\epsilon}=\int\frac{\delta F}{\delta y(x)}\eta(x)dx \tag{4.5.15}$$

이다. 어떤 $\eta(x)$ 함수에 대해서도 이 값이 0이 되려면

$$\frac{\delta F}{\delta y(x)}=0 \tag{4.5.16}$$

이 되는 수 밖에 없다. 여기에 나온

$$\frac{\delta F}{\delta y(x)} \tag{4.5.17}$$

를 범함수의 도함수(functional derivative)라고 하며 함수의 도함수와 같은 역할을 한다. y를 변수, F를 함수라고 놓고 미분하여 구한 도함수와 같다. 하지만 범함수를 함수로 미분하였다는 것을 강조하기 위해 d 기호가 아니라 δ 기호를 사용하였다.

적분형 범함수의 도함수

대부분의 범함수는 x에 대한 적분으로 정의되며 적분 기호안의 연산은 $y(x)$와 x 를 입력 변수로 받는 함수 $G\big(y,x\big)$라고 할 수 있다.

$$F[y(x)]=\int G(y,x)dx \tag{4.5.18}$$

이러한 범함수 F의 도함수는 다음처럼 계산할 수 있다.

$$\frac{\delta F}{\delta y}=\frac{\partial G}{\partial y} \tag{4.5.19}$$

y가 원래는 함수이지만 마치 변수처럼 생각하고 G의 편미분을 구했기 때문에 $\dfrac{\delta}{\delta y}$ 기호가 아닌 $\dfrac{\partial}{\partial y}$ 기호를 사용하였다.

예제

기댓값 E는 다음과 같이 정의되는 범함수이다.

$$\text{E}[p(x)]=\int_{-\infty}^{\infty}xp(x)dx \tag{4.5.20}$$

위 식에 대응시키면

$$G(y,x)=xy \tag{4.5.21}$$

가 된다. 이 때 y에 대한 F의 도함수는

$$\frac{\delta F}{\delta y}=\frac{\partial G}{\partial y}=x \tag{4.5.22}$$

이다.

예제

그레디언트 부스팅(gradient boosting) 방법에서는 주어진 목표함수 $y(x)$와 가장 비슷한 모형함수 $\hat{y}(x)$를 구하기 위해 다음과 같은 범함수인 손실함수를 사용한다.

$$L = \int \frac{1}{2}(\hat{y}(x) - y(x))^2 dx \tag{4.5.23}$$

모형함수 $\hat{y}(x)$에 대한 범함수 손실함수 L의 그레디언트를 구하면

$$G(\hat{y}) = \frac{1}{2}(\hat{y}(x) - y(x))^2 \tag{4.5.24}$$

에서

$$\frac{\delta L}{\delta \hat{y}} = \frac{\partial G}{\partial y} = \hat{y}(x) - y(x) \tag{4.5.25}$$

이 된다.

오일러-라그랑주 공식

가끔씩 x, $y(x)$ 이외에 추가로 $y(x)$의 x에 대한 도함수인 $y'(x) = \dfrac{dy}{dx}$도 입력 변수로 받는 함수 $G\big(y(x), y'(x), x\big)$로 정의된 도함수도 있을 수 있다.

$$F[y(x)] = \int G(y, y', x) dx \tag{4.5.26}$$

이 때는 범함수F의 함수 y에 대한 도함수를 다음처럼 구한다.

$$\frac{\partial G}{\partial y} - \frac{d}{dx}\left(\frac{\partial G}{\partial y'}\right) \tag{4.5.27}$$

위 식에서 $\dfrac{\partial G}{\partial y}$와 $\dfrac{\partial G}{\partial y'}$는 함수 y와 y'을 마치 별개의 변수인 것처럼 생각하고 편미분한 도함수를 뜻한다. $\dfrac{d}{dx}\left(\dfrac{\partial G}{\partial y'}\right)$는 그렇게 구해진 함수를 다시 변수 x로 미분한 결과를 말한다.

이를 **오일러-라그랑주(Euler-Lagrange) 공식**이라고 한다.

최적 제어

함수 $f(x)$가 있을 때 이 값이 가장 커지도록 혹은 가장 작아지도록 하는 독립변수 x의 값을 찾아내는 것을 최적화(optimization)이라고 한다. 이와 비슷하게 **범함수 $F[y(x)]$가 있을 때 이 값이 가장 커지도록 혹은 가장 작아지도록 하는 독립함수 $y(x)$를 찾는 것을 최적 제어(optimal control)**라고 한다.

최적화를 위한 필요조건은 최적의 독립변수의 값 x^*을 입력하면 함수의 도함수의 값이 0이 되어야 한다는 것이었다.

$$\frac{df}{dx}(x^*) = 0 \tag{4.5.28}$$

최적제어에서도 최적의 함수 $y^*(x)$를 입력하면 범함수의 도함수의 값이 0이 되어야 한다는 것이 최적 조건이다.

$$\frac{\delta f}{\delta y}[y^*(x)] = 0 \tag{4.5.29}$$