딥러닝을 위한 수학 I

김영록

한국외대
교육대학원

2024년 5월 2일

- 다음으로 지수함수의 미분에 대해 알아봅시다. 로그함수에서는 밑을 e로 해서 깔끔하게 미분된 식을 얻을수 있었습니다. 그래서 지수함수의 밑으로 우선 e를 써서 생각해 보겠습니다.
- $y = e^x$라는 지수함수가 있다고 가정합시다. 지수함수와 로그함수는 서로 역함수의 관계이기 때문에 다음 식이 성립합니다.

$$x = \log y$$

- 이 식을 미분하면 다음과 같이 표현할 수 있습니다.

$$\frac{dx}{dy} = (\log y)' = \frac{1}{y}$$

- 따라서 2.7절에서 설명한 역함수의 미분에 따라 다음과 같은 식을 얻을 수 있습니다.

$$\frac{dy}{dx} = \frac{1}{\frac{dx}{dy}} = \frac{1}{\frac{1}{y}} = y$$

- 놀랍게도 y의 미분은 y 자신이 되어 버렸습니다. y를 원래의 e^x 형태로 다시 쓰면 다음과 같이 표현할수 있습니다.

$$(e^x)' = e^x \tag{1}$$

이것이 네이피어 상수 e를 밑으로 하는 지수함수의 미분 공식입니다.

- e를 밑으로 하지 않는 지수함수를 미분할 때는 양변을 자연로그 형태로 변형한 다음, 미분해주면 됩니다. 이러한 계산 방법을 '로그 미분법 (logarithmic differentiation)'이라고 합니다.

$$\log y = \log a^x = x \log a$$

- 양변을 x로 미분하면 다음 식과 같습니다.

$$\frac{d \log y}{dx} = \frac{d(x \log a)}{dx} = \log a \tag{2}$$

- 이 식은 합성함수의 미분 공식을 따라 다음과 같이 쓸 수 있습니다.

$$\frac{d \log y}{dx} = \frac{d \log y}{dy}\frac{dy}{dx} = \frac{1}{y}\frac{dy}{dx} \tag{3}$$

- 그리고 수식 (2)와 (3)에 의해 다음 식이 성립합니다.

$$\log a = \frac{1}{y}\frac{dy}{dx}$$

- 이 식은 다음과 같이 정리할 수 있습니다.

$$y' = \frac{dy}{dx} = (\log a)y = (\log a)a^x \tag{4}$$

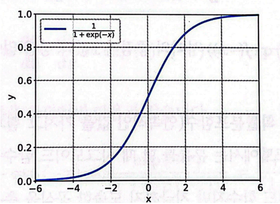
이것이 자연로그 이외의 수를 밑으로 하는 지수함수의 미분 공식입니다.

- 앞서 살펴본 것처럼 네이피어 상수 e를 밑으로 하는 지수함수는 미분 결과가 자신이 되는 수학적으로 아름다운 특징이 있다 보니 이 책의 이후 내용에도 자주 사용됩니다.
- 실제로 이러한 형태의 지수함수를 사용할 때는 인수 부분에 복잡한 식이 들어가고 합성함수의 모양인 경우가 많습니다. 전형적인 예로는 6.2절에 나오는 정규분포함수가 있습니다.
- 지수함수의 우측 상단에 복잡한 수식을 위첨자로 쓰게 되면 수식 자체가 복잡해서 알아보기 힘듭니다. 그래서 'e^x'라는 표기 대신 '$\exp(x)$라는 표기법을 많이 씁니다. 이 책에서도 이후 내용에서는 지수함수를 표기할 때 이 같은 방식으로 표기할 것입니다.

- 다음 함수를 살펴봅시다. 이 함수는 '시그모이드 함수(sigmoid function)'[1]라고 합니다.

$$y = \frac{1}{1 + \exp(-x)}$$

- 그림 5-10은 이 함수를 그래프로 표현한 것입니다.



- 그래프를 보면 다음과 같은 특징이 있다는 것을 알 수 있습니다.
  - 값이 항상 증가하는 함수다.[2]
  - x값이 음의 무한대로 갈 때 함숫값은 0에 가까워진다.

[1]정확하게는 시그모이드 함수란 매개변수 a를 포함한 $y = \frac{1}{1+\exp(-ax)}$과 같은 형태의 함수를 말합니다. 이에 반해 매개변수 a가 없는 함수를 '표준 시그모이드 함수라고 합니다. 머신러닝에서는 편의상 '표준'이라는 말을 생략하고 그냥 시그모이드 함수라고 줄여 부르는 경향이 있습니다. 이러한 관례에 따라 시그모이드 함수라고 표기합니다.

[2]이런 특징을 가진 함수를 단조증가함수(monotone increasing function)라고 합니다.

-
  - x값이 양의 무한대로 갈 때 함숫값은 1에 가까워진다.
  - $x = 0$일 때 함숫값은 0.5다.
  - 그래프 모양은 점 $(0, 0.5)$에 대해 점대칭이다.
- 마지막 특징은 다음의 계산으로 확인할 수 있습니다.

$$f(x) = \frac{1}{1 + \exp(-x)}$$

- $f(x)$가 위와 같을 때 다음 식이 성립합니다.

$$\begin{aligned} f(x) + f(-x) &= \frac{1}{1 + \exp(-x)} + \frac{1}{1 + \exp(x)} \\ &= \frac{1}{1 + \exp(-x)} + \frac{\exp(-x)}{1 + \exp(-x)} = 1 \end{aligned}$$

- 위 식을 정리하면 다음과 같이 쓸 수 있습니다.

$$\frac{1}{2}(f(x) + f(-x)) = \frac{1}{2}$$

- 이것은 두 개의 점 $(x, f(x))$와 $(-x, f(-x))$ 사이의 중점이 x값과 상관 없이 항상 $\left(0, \frac{1}{2}\right)$에 있다는 것을 의미합니다.
- 이러한 성질은 6장에서 설명할 확률분포함수(연속적인 값을 가지고 결과가 확률값인 함수)의 특징으로 적합합니다. 그래서 머신러닝 모델에서는 분류를 할 때 시그모이드 함수를 자주 사용합니다.
- 조금은 복잡해 보이는 시그모이드 함수지만 지금까지 도출한 공식을 총동원하면 어렵지 않게 미분할 수 있습니다. 실제로 계산해 봅시다.
- 우선 다음과 같은 시그모이드 함수가 있다고 할 때

$$y = \frac{1}{1 + \exp(-x)}$$

- 분모 부분을 다음과 같은 함수로 표현합시다.

$$u(x) = 1 + \exp(-x)$$

- 그러면 다음과 같이 간단한 형태로 바꿔 쓸 수 있습니다.

$$y(u) = \frac{1}{u}$$

- 여기에 합성함수의 미분 공식을 적용합니다. 구체적으로는 다음과 같은 식이 만들어집니다.
$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$$
- 여기서 우변의 왼쪽 부분을 미분한 결과는 다음과 같습니다.
$$\frac{dy}{dx} = \left(\frac{1}{u}\right)' = (u^{-1})' = (-1) \cdot u^{-2} = -\frac{1}{u^2}$$
- 이번에는 우변의 오른쪽 부분인 합성함수의 미분 공식을 써 봅시다. $v = -x$ 라고 할 때 u와 v는 다음과 같은 관계가 됩니다.
$$u = 1 + \exp(-x) = 1 + \exp(v)$$
- 따라서 우변의 오른쪽 부분은 다음과 같이 풀어 쓸 수 있습니다.
$$\frac{du}{dx} = \frac{du}{dv} \cdot \frac{dv}{dx} = \exp(v) \cdot (-1) = -\exp(-x)$$

- 우변의 왼쪽 부분과 오른쪽 부분을 조합하면 다음과 같습니다.

$$\frac{dy}{dx} = -\frac{1}{u^2} \cdot -\exp(-x) = \frac{\exp(-x)}{(1+\exp(-x))^2} = \frac{1+\exp(-x)-1}{(1+\exp(-x))^2}$$
$$= \frac{1}{1+\exp(-x)} - \frac{1}{(1+\exp(-x))^2}$$

- 결론적으로 수식을 정리하면 다음과 같이 쓸 수 있습니다.

$$f'(x) = y(1-y) \tag{5}$$

- 수식 (5)이 시그모이드 함수의 미분 결과입니다. 자세히 보면 원래의 함숫값만 사용해 미분값을 계산할 수 있다는 것을 알 수 있습니다. 시그모이드 함수의 이러한 특징은 뒤에 나올 머신러닝 모델에서 학습을 진행할 때 활용하게 됩니다.

- 앞에서 소개한 시그모이드 함수는 실수를 입력하면 (확률값으로 해석할 수 있는) 0에서 1까지의 값을 출력하는 함수였습니다.
- 이제부터 소개할 '소프트맥스(softmax) 함수'는 벡터를 입력하면 (확률값으로 해석할 수 있는) 0에서 1 까지의 값을 가진 같은 차수의 벡터를 출력하는 함수입니다. 기능도 시그모이드 함수와 비슷하고 출력결과도 확률값으로 쓸 수 있는 함수입니다. 4장에서 설명한 다변수함수가 n개의 입력에 1개의 출력이었다면 이번에는 n개의 입력에 n개의 출력이므로 다변수함수를 더 확장한 함수라고 볼 수 있습니다. 이런 함수를 '벡터함수 (vector function)'라고도 합니다.
- 그림 5-11은 $n = 3$일 때 소프트맥스 함수의 개념도입니다.

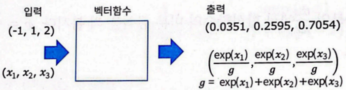


그림 5–11 소프트맥스 함수 (n=3)

- 입력과 출력이 다음과 같을 때
  - 입력 벡터: (x_1, x_2, x_3)
  - 출력 벡터: (y_1, y_2, y_3)

- 결과를 표현하는 식은 다음과 같습니다.

$$\begin{cases} y_1 = \frac{\exp(x_1)}{g(x_1,x_2,x_3)} \\ y_2 = \frac{\exp(x_2)}{g(x_1,x_2,x_3)} \\ y_3 = \frac{\exp(x_3)}{g(x_1,x_2,x_3)} \end{cases}$$

- 이때 $g(x_1, x_2, x_3)$은 다음과 같습니다.

$$g(x_1, x_2, x_3) = \exp(x_1) + \exp(x_2) + \exp(x_3)$$

- 소프트맥스 함수의 정의에 의해 다음 식이 성립합니다.

$$y_1 + y_2 + y_3 = 1,\ 0 \leq y_i \leq 1\ (i = 1, 2, 3)$$

- 이러한 특징을 살펴보면 세 개의 출력값을 확률값으로도 쓸 수 있다는 것을 알 수 있습니다.
- 다음으로 소프트맥스 함수의 미분을 계산해 봅시다. 이 함수는 다변수함수이므로 미분을 할 때 4.2절에 설명한 편미분으로 계산해야 합니다.

- 우선 x와 y의 첨자가 같은 경우로 편미분해 봅시다. 수식이 간결해지도록 $\exp(x_1)$을 $h(x_1)$로 표기하겠습니다.

$$y_1 = \frac{h(x_1)}{g(x_1, x_2, x_3)} = \frac{h}{g}$$

- 2.8절에서 설명한 몫의 미분 공식 (2.8.1)에 의해 위의 식을 다음과 같이 쓸 수 있습니다.

$$\frac{\partial y_1}{\partial x_1} = \frac{g \cdot h_{x_1} - h \cdot g_{x_1}}{g^2}$$

- 위 식에서 h_{x_1}과 g_{x_1}을 따로 풀어보면 다음과 같습니다.

$$h_{x_1} = \exp(x_1)' = \exp(x_1) = h$$
$$g_{x_1} = \frac{\partial g}{\partial x_1} = \exp(x_1) = h$$

- 이들을 조합하면 다음과 같은 결과가 나옵니다.

$$\frac{\partial y_1}{\partial x_1} = \frac{g \cdot h - h \cdot h}{g^2} = \frac{h}{g} \cdot \frac{g - h}{g} = \frac{h}{g} \cdot \left(1 - \frac{h}{g}\right) = y_1(1 - y_1)$$

- 편미분한 결과는 원래의 함수 값 y_1만으로도 표현할 수 있고 앞 절에서 본 시그모이드 함수의 미분 결과인 수식 (5)과 모양이 똑같은 것을 알 수 있습니다.
- 지금까지 x와 y의 첨자가 같을 때의 편미분 결과를 봤습니다. 그러면 x와 y의 첨자가 같지 않은 경우는 어떻게 될까요? 이해를 돕기 위해 y_2를 x_1로 편미분하는 경우를 예로 들어 보겠습니다.

$$y_2 = \frac{\exp(x_2)}{g(x_1, x_2, x_3)} = \frac{h(x_2)}{g}$$

- 이때 분자 부분은 x_1의 관점에서 상수($h' = 0$)로 볼 수 있고 몫의 미분 공식을 사용하면 다음과 같이 식을 쓸 수 있습니다.

$$\frac{\partial y_2}{\partial x_1} = \frac{g \cdot h(x_2)_{x_1} - h(x_2) \cdot g_{x_1}}{g^2} = \frac{g \cdot 0 - h(x_2) \cdot g_{x_1}}{g^2} = -\frac{h(x_2) \cdot g_{x_1}}{g^2}$$

- g_{x_1}은 g를 x_1로 편미분한 결과이므로 앞의 계산 결과에 의해 $h(x_1)$이 됩니다.

$$\frac{\partial y_2}{\partial x_1} = -\frac{h(x_2) \cdot h(x_1)}{g^2} = -\frac{h(x_2)}{g} \cdot \frac{h(x_1)}{g} = -y_2 \cdot y_1$$

- 이제까지의 내용을 정리하면 다음과 같습니다.

$$\frac{\partial y_j}{\partial x_i} = \begin{cases} y_i(1 - y_i) & (i = j) \\ -y_i y_j & (i \neq j) \end{cases} \tag{6}$$

시그모이드 함수와 소프트맥스 함수의 관계

- 지금까지의 계산 결과를 보면 시그모이드 함수와 소프트맥스 함수 사이에 어떤 관계가 있는 것으로 보여집니다. 이러한 생각은 $n = 2$일 때 소프트맥스 함수에 다음 계산을 해 보면 사실이라는 것을 알 수 있습니다. 참고로 마지막 수식은 분자와 분모를 $exp(x_i)$로 나누고 5.1절에서 도출한 지수함수의 공식 (**??**)를 적용했습니다.

$$y_1 = \frac{\exp(x_1)}{\exp(x_1) + \exp(x_2)} = \frac{1}{1 + \exp(-(x_1 - x_2))}$$

- 이때 $x_1 - x_2$를 x로 대체하면 시그모이드 함수와 같은 식이 되는 것을 알 수 있습니다. 즉, $n = 2$일 때의 소프트맥스 함수는 사실상 시그모이드 함수와 동일하며, 반대로 시그모이드 함수를 $n = 3$ 이상으로 확장한 것이 소프트맥스 함수라고 볼 수 있습니다. 이러한 시그모이드 함수와 소프트맥스 함수 간의 관계는 뒤에 나올 실습편에서 8장의 이진 분류와 9장의 다중 클래스 분류의 관계로 연결되니 참고하기 바랍니다.

- Gradient Descent based Optimization Algorithms for Deep Learning Models Training
  https://arxiv.org/pdf/1903.03614.pdf
  위의 논문을 번역을 하여서 중간고사 숙제로 제출하세요.