

2022학년도 1학기 컴퓨터언어학

제10강 신경망 언어 모형 (4)

박수지

서울대학교 인문대학 언어학과

2022년 4월 6일 수요일

오늘의 목표

- 1 순방향신경망 학습에 필요한 손실함수를 도출할 수 있다.
- 2 합성함수의 계산 그래프를 그리고 순전파를 수행할 수 있다.
- 3 순방향신경망을 이루는 각 계층의 미분값을 역방향으로 구하는 오차역전파법을 설명할 수 있다.

손실함수

로지스틱 회귀분석(이항 분류)의 교차엔트로피 손실함수

$$L_{CE}(\hat{y}, y) = -\log p(y|\vec{x}) \quad \cdots \text{교차엔트로피 손실함수의 정의.}$$

$$= -[y \log \hat{y} + (1 - y) \log (1 - \hat{y})] \quad \cdots \text{3강 슬라이드 22-23면 참조.}$$

일반화(\rightarrow 다항 분류)

스칼라 $y \in \{0, 1\}$ 를 원-핫 인코딩 벡터 $\vec{y} = [y_1, y_2] \in \{0, 1\}^2$ 로 표현할 수 있다.

정답 표상 1 $y = 1 \Rightarrow y_1 = 1 \Rightarrow \vec{y} = [1, 0]$

2 $y = 0 \Rightarrow y_2 = 1 \Rightarrow \vec{y} = [0, 1]$

추정치 1 $\hat{y} = P(y = 1|\vec{x}) \Rightarrow \hat{y}_1 = P(y_1 = 1|\vec{x})$

2 $1 - \hat{y} = P(y = 0|\vec{x}) \Rightarrow \hat{y}_2 = P(y_2 = 1|\vec{x})$

손실함수

로지스틱 회귀분석(이항 분류)의 교차엔트로피 손실함수

$$L_{CE}(\hat{y}, y) = - [y \log \hat{y} + (1 - y) \log (1 - \hat{y})]$$

소프트맥스 회귀분석(다항 분류)의 교차엔트로피 손실함수

부류가 K개 있을 때

$$K = 2 \quad L_{CE}(\hat{\vec{y}}, \vec{y}) = - [y_1 \log \hat{y}_1 + y_2 \log \hat{y}_2]$$

$$K = 3 \quad L_{CE}(\hat{\vec{y}}, \vec{y}) = - [y_1 \log \hat{y}_1 + y_2 \log \hat{y}_2 + y_3 \log \hat{y}_3]$$

$$\cdots \quad L_{CE}(\hat{\vec{y}}, \vec{y}) = - \sum_{k=1}^K y_k \log \hat{y}_k$$

손실함수

소프트맥스 회귀분석(다항 분류)의 교차엔트로피 손실함수

$$\begin{aligned} L_{\text{CE}}(\hat{\vec{y}}, \vec{y}) &= - \sum_{k=1}^K y_k \log \hat{y}_k \\ &= - \log \hat{y}_c \end{aligned} \quad (\text{c가 정답인 경우})$$

관찰

두 번째 등식이 성립하는 이유

- \vec{y} 가 원-핫 인코딩 벡터이므로, \vec{x} 의 정답이 c 번째 부류일 때 $y_c = 1$ 이다.
- $k \neq c$ 인 k 의 경우 $y_k = 0$ 이므로 $y_k \log \hat{y}_k = 0$ 이 된다.

손실함수

소프트맥스 회귀분석(다항 분류)의 매개변수

n 차원 입력 벡터 $\vec{x} = [x_1, x_2, \dots, x_n]^T$ 를 K 개 중 하나의 부류로 분류할 때

$$\text{가중치 행렬 } \mathbf{W} = \begin{bmatrix} \vec{w}_1 \\ \vec{w}_2 \\ \vdots \\ \vec{w}_K \end{bmatrix} = \begin{bmatrix} w_{1,1} & w_{1,2} & \dots & w_{1,n} \\ w_{2,1} & w_{2,2} & \dots & w_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{K,1} & w_{K,2} & \dots & w_{K,n} \end{bmatrix}$$

$$\text{편향 벡터 } \vec{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_K \end{bmatrix}$$

손실함수

예시: 소프트맥스 회귀분석의 매개변수

4차원 입력 벡터 $\vec{x} = [x_1, x_2, x_3, x_4]^T$ 를
3개 중 하나의 부류로 분류할 때

$$\mathbf{W} = \begin{bmatrix} \vec{w}_1 \\ \vec{w}_2 \\ \vec{w}_3 \end{bmatrix} = \begin{bmatrix} w_{1,1} & w_{1,2} & w_{1,3} & w_{1,4} \\ w_{2,1} & w_{2,2} & w_{2,3} & w_{2,4} \\ w_{3,1} & w_{3,2} & w_{3,3} & w_{3,4} \end{bmatrix}$$

$$\vec{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

가중합과 확률 추정치

$$\vec{z} = \mathbf{W}\vec{x} + \vec{b} = \begin{bmatrix} \vec{w}_1 \cdot \vec{x} + b_1 \\ \vec{w}_2 \cdot \vec{x} + b_2 \\ \vec{w}_3 \cdot \vec{x} + b_3 \end{bmatrix}$$

$$\hat{y} = \begin{bmatrix} \text{softmax}(z_1) \\ \text{softmax}(z_2) \\ \text{softmax}(z_3) \end{bmatrix} = \begin{bmatrix} \frac{\exp(z_1)}{\sum_{k=1}^3 \exp(z_k)} \\ \frac{\exp(z_2)}{\sum_{k=1}^3 \exp(z_k)} \\ \frac{\exp(z_3)}{\sum_{k=1}^3 \exp(z_k)} \end{bmatrix}$$

손실함수

소프트맥스 회귀분석(다항 분류)의 교차엔트로피 손실함수

$$\begin{aligned} L_{CE}(\hat{\vec{y}}, \vec{y}) &= -\log \hat{y}_c \quad (c \text{가 정답인 경우}) \\ &= -\log \text{softmax}(z_c) \\ &= -\log \frac{\exp(z_c)}{\sum_{k=1}^K \exp(z_k)} \\ &= -\log \frac{\exp(\vec{w}_c \cdot \vec{x} + b_c)}{\sum_{j=1}^K \exp(\vec{w}_j \cdot \vec{x} + b_j)} \end{aligned}$$

기울기 계산하기

목표

손실함수 L_{CE} 의 값을 최소로 하는 매개변수 \mathbf{W}, \vec{b} 의 값을 찾는다.

필요한 것

각 매개변수에 대한 편도함숫값 $\frac{\partial L_{CE}}{\partial w_{k,i}}$

로지스틱 회귀분석에서 손실함수의 편도함수

$$\frac{\partial L_{CE}}{\partial w_j} = (\hat{y} - y)x_j \quad (j = 1, 2, \dots, n)$$

기울기 계산하기

신경망의 마지막 계층(출력층)...

입력층 $\vec{x} = [x_1, x_2, \dots, x_n] \in \mathbb{R}^n$

가중합 $\vec{z} = \mathbf{W}\vec{x} + \vec{b} \in \mathbb{R}^K$

출력층 $\hat{y} = \text{softmax}(\vec{z}) \in \mathbb{R}^K$

$$\mathbf{W} = \begin{bmatrix} \vec{w}_1 \\ \vec{w}_2 \\ \vdots \\ \vec{w}_K \end{bmatrix} = \begin{bmatrix} w_{1,1} & w_{1,2} & \dots & w_{1,n} \\ w_{2,1} & w_{2,2} & \dots & w_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{K,1} & w_{K,2} & \dots & w_{K,n} \end{bmatrix}$$

...에서 손실함수의 편도함수

$$\begin{aligned} \frac{\partial L_{CE}}{\partial w_{k,i}} &= (\hat{y}_k - y_k) x_i \\ &= -(y_k - \hat{y}_k) x_i \\ &= - \left(y_k - \frac{\exp(\vec{w}_k \cdot \vec{x} + b_k)}{\sum_{j=1}^K \exp(\vec{w}_j \cdot \vec{x} + b_j)} \right) x_i \end{aligned}$$

문제

출력층 이전의 은닉층에는 적용할 수 없다!

계산 그래프

계산 그래프

계산을 여러 연산으로 쪼개어 각 연산을 노드로 표현한 그래프

예시: 순전파

$L(a, b, c) = c(a + 2b)$
이 함수를 쪼개어 보자.

$$d = 2b$$

$$e = a + d$$

$$L = ce$$

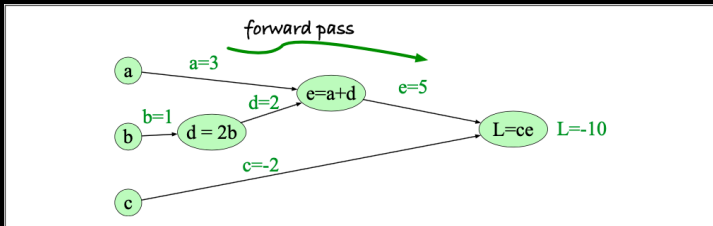


Figure 7.14 Computation graph for the function $L(a, b, c) = c(a + 2b)$, with values for input nodes $a = 3$, $b = 1$, $c = -2$, showing the forward pass computation of L .

계산 그래프에서의 역방향 미분

미분의 연쇄법칙

$$\text{합성함수의 미분 } f(x) = u(v(x)) \quad \Rightarrow \quad \frac{df}{dx} = \frac{du}{dv} \times \frac{dv}{dx}$$

예시: 역전파

$$L(a, b, c) = c(a + 2b) = c(a + d) = ce$$

$$\frac{\partial L}{\partial d} = \frac{\partial L}{\partial e} \frac{\partial e}{\partial d} = \frac{\partial}{\partial e}(ce) \times \frac{\partial}{\partial d}(a + d)$$

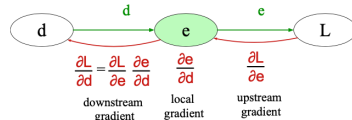


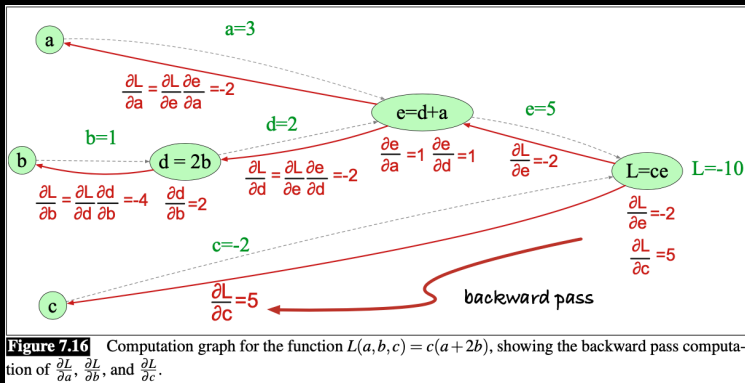
Figure 7.15 Each node (like e here) takes an upstream gradient, multiplies it by the local gradient (the gradient of its output with respect to its input), and uses the chain rule to compute a downstream gradient to be passed on to a prior node. A node may have multiple local gradients if it has multiple inputs.

계산 그래프에서의 역방향 미분

예시: 역전파

$$\begin{aligned} L(a, b, c) &= c(a + 2b) \\ &= c(a + d) \\ &= ce \end{aligned}$$

$$\begin{aligned} \frac{\partial L}{\partial a} &= \frac{\partial L}{\partial e} \frac{\partial e}{\partial a} \\ \frac{\partial L}{\partial b} &= \frac{\partial L}{\partial e} \frac{\partial e}{\partial d} \frac{\partial d}{\partial b} \\ \frac{\partial L}{\partial c} & \end{aligned}$$



계산 그래프에서의 역방향 미분

로지스틱 함수의 계산 그래프 그려 보기

로지스틱 함수

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

$$a = -z$$

$$b = \exp(a)$$

$$c = 1 + b$$

$$\sigma = \frac{1}{c}$$

$$\frac{d\sigma}{dz} = \frac{d\sigma}{dc} \frac{dc}{db} \frac{db}{da} \frac{da}{dz}$$

$$\frac{da}{dz} = -1$$

$$\frac{db}{da} = \exp(a) = b$$

$$\frac{dc}{db} = 1$$

$$\frac{d\sigma}{dc} = -\frac{1}{c^2}$$

계산 그래프에서의 역방향 미분

로지스틱 함수의 미분

로지스틱 함수...

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

$$a = -z$$

$$b = \exp(a)$$

$$c = 1 + b$$

$$\sigma = \frac{1}{c}$$

$$\frac{da}{dz} = -1$$

$$\frac{db}{da} = \exp(a) = b$$

$$\frac{dc}{db} = 1$$

$$\frac{d\sigma}{dc} = -\frac{1}{c^2}$$

...의 도함수 구하기

$$\begin{aligned} \frac{d\sigma}{dz} &= \frac{d\sigma}{dc} \times \frac{dc}{db} \times \frac{db}{da} \times \frac{da}{dz} \\ &= \left(-\frac{1}{c^2}\right) \times 1 \times b \times (-1) \\ &= \frac{b}{c^2} \\ &= \frac{\exp(-z)}{[1 + \exp(-z)]^2} \end{aligned}$$

계산 그래프에서의 역방향 미분

로지스틱 함수의 미분

로지스틱 함수의 도함수 정리하기

$$\begin{aligned}\frac{d\sigma}{dz} &= \frac{d}{dz} \left[\frac{1}{1 + \exp(-z)} \right] = \frac{\exp(-z)}{[1 + \exp(-z)]^2} \\ &= \frac{1}{1 + \exp(-z)} \times \frac{\exp(-z)}{1 + \exp(-z)} \\ &= \frac{1}{1 + \exp(-z)} \times \frac{1 + \exp(-z) - 1}{1 + \exp(-z)} \\ &= \frac{1}{1 + \exp(-z)} \times \left[\frac{1 + \exp(-z)}{1 + \exp(-z)} - \frac{1}{1 + \exp(-z)} \right] \\ &= \sigma(1 - \sigma)\end{aligned}$$

계산 그래프에서의 역방향 미분

2층 신경망의 손실함수를 1층 매개변수로 편미분하기

입력값 $\vec{x} = [x_1, x_2] \in \mathbb{R}^2$

가중합1 $\vec{z}^{[1]} = \mathbf{W}^{[1]}\vec{x} + \vec{b}^{[1]} \in \mathbb{R}^2$

$$\text{1 } z_1^{[1]} = w_{1,1}^{[1]}x_1 + w_{1,2}^{[1]}x_2 + b_1^{[1]} \in \mathbb{R}$$

$$\text{2 } z_2^{[1]} = w_{2,1}^{[1]}x_1 + w_{2,2}^{[1]}x_2 + b_2^{[1]} \in \mathbb{R}$$

활성화값1 $\vec{a}^{[1]} = \text{ReLU}(\vec{z}^{[1]}) = \vec{h} \in \mathbb{R}^2$

$$\text{1 } a_1^{[1]} = \text{ReLU}(z_1^{[1]}) = h_1 \in \mathbb{R}$$

$$\text{2 } a_2^{[1]} = \text{ReLU}(z_2^{[1]}) = h_2 \in \mathbb{R}$$

가중합2 $z^{[2]} = \vec{w}^{[2]}\vec{a}^{[1]} + b^{[2]} = w_1^{[2]}a_1^{[1]} + w_2^{[2]}a_2^{[1]} + b^{[2]} \in \mathbb{R}$

활성화값2 $a^{[2]} = \sigma(z^{[2]}) = \hat{y} \in \mathbb{R}$

손실함숫값 $L(a^{[2]}, y) = -[y \log a^{[2]} + (1 - y) \log (1 - a^{[2]})]$

계산 그래프에서의 역방향 미분

2층 신경망의 손실함수를 1층 매개변수로 편미분하기

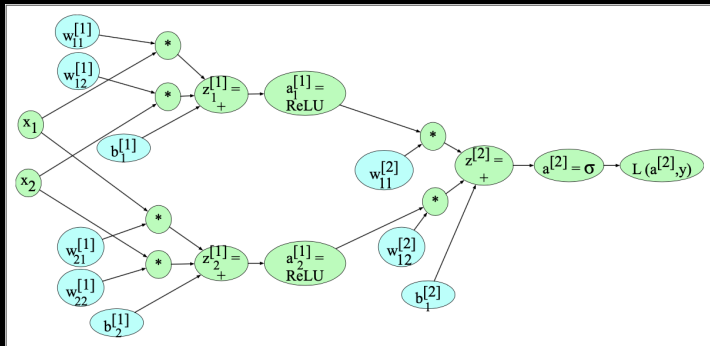


Figure 7.17 Sample computation graph for a simple 2-layer neural net (= 1 hidden layer) with two input units and 2 hidden units. We've adjusted the notation a bit to avoid long equations in the nodes by just mentioning the function that is being computed, and the resulting variable name. Thus the $*$ to the right of node $w_{11}^{[1]}$ means that $w_{11}^{[1]}$ is to be multiplied by x_1 , and the node $z_1^{[1]} = +$ means that the value of $z_1^{[1]}$ is computed by summing the three nodes that feed into it (the two products, and the bias term $b_1^{[1]}$).

계산 그래프에서의 역방향 미분

2층 신경망의 손실함수를 1층 매개변수로 편미분하기

손실함수의 편도함수: 연쇄법칙

$$\frac{\partial L}{\partial w_{1,1}^{[1]}} = \frac{\partial L}{\partial a^{[2]}} \frac{\partial a^{[2]}}{\partial z^{[2]}} \frac{\partial z^{[2]}}{\partial a_1^{[1]}} \frac{\partial a_1^{[1]}}{\partial z_1^{[1]}} \frac{\partial z_1^{[1]}}{\partial w_{1,1}^{[1]}}$$

1. 손실함수 → 활성화값2

$$L = - \left[y \log a^{[2]} + (1 - y) \log (1 - a^{[2]}) \right]$$

$$\begin{aligned} \frac{\partial L}{\partial a^{[2]}} &= - \left[\frac{y}{a^{[2]}} - \frac{1 - y}{1 - a^{[2]}} \right] \\ &= - \left[\frac{y(1 - a^{[2]}) - (1 - y)a^{[2]}}{a^{[2]}(1 - a^{[2]})} \right] \\ &= \frac{a^{[2]} - y}{a^{[2]}(1 - a^{[2]})} \end{aligned}$$

계산 그래프에서의 역방향 미분

2층 신경망의 손실함수를 1층 매개변수로 편미분하기

손실함수의 편도함수: 연쇄법칙

$$\frac{\partial L}{\partial w_{1,1}^{[1]}} = \frac{\partial L}{\partial a^{[2]}} \frac{\partial a^{[2]}}{\partial z^{[2]}} \frac{\partial z^{[2]}}{\partial a_1^{[1]}} \frac{\partial a_1^{[1]}}{\partial z_1^{[1]}} \frac{\partial z_1^{[1]}}{\partial w_{1,1}^{[1]}}$$

$$\frac{\partial L}{\partial a^{[2]}} = \frac{a^{[2]} - y}{a^{[2]} (1 - a^{[2]})}$$

2. 활성화값2 → 가중합2(로지스틱 계층)

$$a^{[2]} = \sigma(z^{[2]})$$

$$\begin{aligned} \frac{\partial a^{[2]}}{\partial z^{[2]}} &= \sigma(z^{[2]}) [1 - \sigma(z^{[2]})] \\ &= a^{[2]} (1 - a^{[2]}) \end{aligned}$$

계산 그래프에서의 역방향 미분

2층 신경망의 손실함수를 1층 매개변수로 편미분하기

손실함수의 편도함수: 연쇄법칙

$$\frac{\partial L}{\partial w_{1,1}^{[1]}} = \frac{\partial L}{\partial a^{[2]}} \frac{\partial a^{[2]}}{\partial z^{[2]}} \frac{\partial z^{[2]}}{\partial a_1^{[1]}} \frac{\partial a_1^{[1]}}{\partial z_1^{[1]}} \frac{\partial z_1^{[1]}}{\partial w_{1,1}^{[1]}}$$

$$\frac{\partial L}{\partial a^{[2]}} = \frac{a^{[2]} - y}{a^{[2]} (1 - a^{[2]})}$$

$$\frac{\partial a^{[2]}}{\partial z^{[2]}} = a^{[2]} (1 - a^{[2]})$$

3. 가중합2 → 활성화값1 (Affine 계층)

$$z^{[2]} = w_1^{[2]} a_1^{[1]} + w_2^{[2]} a_2^{[1]} + b^{[2]}$$

$$\frac{\partial z^{[2]}}{\partial a_1^{[1]}} = w_1^{[2]}$$

계산 그래프에서의 역방향 미분

2층 신경망의 손실함수를 1층 매개변수로 편미분하기

손실함수의 편도함수: 연쇄법칙

$$\frac{\partial L}{\partial w_{1,1}^{[1]}} = \frac{\partial L}{\partial a^{[2]}} \frac{\partial a^{[2]}}{\partial z^{[2]}} \frac{\partial z^{[2]}}{\partial a_1^{[1]}} \frac{\partial a_1^{[1]}}{\partial z_1^{[1]}} \frac{\partial z_1^{[1]}}{\partial w_{1,1}^{[1]}}$$

$$\frac{\partial L}{\partial a^{[2]}} = \frac{a^{[2]} - y}{a^{[2]} (1 - a^{[2]})}$$

$$\frac{\partial a^{[2]}}{\partial z^{[2]}} = a^{[2]} (1 - a^{[2]})$$

$$\frac{\partial z^{[2]}}{\partial a_1^{[1]}} = w_1^{[2]}$$

4. 활성화값1 → 가중합1 (ReLU 계층)

$$a_1^{[1]} = \text{ReLU}(z_1^{[1]}) = \begin{cases} 0, & \text{if } z_1^{[1]} \leq 0 \\ z_1^{[1]}, & \text{if } z_1^{[1]} > 0 \end{cases}$$

$$\begin{aligned} \frac{\partial a_1^{[1]}}{\partial z_1^{[1]}} &= \begin{cases} 0, & \text{if } z_1^{[1]} \leq 0 \\ 1, & \text{if } z_1^{[1]} > 0 \end{cases} \\ &= \mathbb{1} \{z_1^{[1]} > 0\} \end{aligned}$$

계산 그래프에서의 역방향 미분

2층 신경망의 손실함수를 1층 매개변수로 편미분하기

손실함수의 편도함수: 연쇄법칙

$$\frac{\partial L}{\partial w_{1,1}^{[1]}} = \frac{\partial L}{\partial a^{[2]}} \frac{\partial a^{[2]}}{\partial z^{[2]}} \frac{\partial z^{[2]}}{\partial a_1^{[1]}} \frac{\partial a_1^{[1]}}{\partial z_1^{[1]}} \frac{\partial z_1^{[1]}}{\partial w_{1,1}^{[1]}}$$

$$\frac{\partial L}{\partial a^{[2]}} = \frac{a^{[2]} - y}{a^{[2]} (1 - a^{[2]})}$$

$$\frac{\partial a^{[2]}}{\partial z^{[2]}} = a^{[2]} (1 - a^{[2]})$$

$$\frac{\partial z^{[2]}}{\partial a_1^{[1]}} = w_1^{[2]}$$

5. 가중합1 → 가중치1 (Affine 계층)

$$z_1^{[1]} = w_{1,1}^{[1]} x_1 + w_{1,2}^{[1]} x_2 + b_1^{[1]}$$

$$\frac{\partial z_1^{[1]}}{\partial w_{1,1}^{[1]}} = x_1$$

계산 그래프에서의 역방향 미분

2층 신경망의 손실함수를 1층 매개변수로 편미분하기

손실함수의 편도함수: 연쇄법칙

$$\begin{aligned}\frac{\partial L}{\partial w_{1,1}^{[1]}} &= \frac{\partial L}{\partial a^{[2]}} \frac{\partial a^{[2]}}{\partial z^{[2]}} \frac{\partial z^{[2]}}{\partial a_1^{[1]}} \frac{\partial a_1^{[1]}}{\partial z_1^{[1]}} \frac{\partial z_1^{[1]}}{\partial w_{1,1}^{[1]}} \\&= \frac{a^{[2]} - y}{a^{[2]} (1 - a^{[2]})} \times a^{[2]} (1 - a^{[2]}) \times w_1^{[2]} \times \mathbb{1} \{z_1^{[1]} > 0\} \times x_1 \\&= (a^{[2]} - y) \times w_1^{[2]} \times \mathbb{1} \{z_1^{[1]} > 0\} \times x_1 \\&= \begin{cases} (\hat{y} - y) w_1^{[2]} x_1, & \text{if } z_1^{[1]} > 0 \\ 0, & \text{if } z_1^{[1]} \leq 0 \end{cases}\end{aligned}$$

계산 그래프에서의 역방향 미분

지금까지 살펴본 계층

- 1 손실함수
- 2 로지스틱 계층 (\rightarrow 소프트맥스 계층으로 일반화 가능)
- 3 ReLU 계층
- 4 Affine 계층(가중합)

심화: tanh 계층의 역전파를 계산해 보자. (시험에 안 나옴)

관찰

위의 계층들을 조합하여 3층 이상의 신경망으로 확장할 수 있다.

남은 문제

《밑바닥부터 시작하는 딥러닝 1》6장 〈학습 관련 기술들〉 참조 — 심화 과제

- 매개변수 갱신(최적화) 방법
- 가중치의 초기값 설정
- 배치 정규화
- 과적합(Overfitting) 방지

다음 주에 배울 것

합성곱신경망(CNN: Convolutional Neural Networks)

- 《밑바닥부터 시작하는 딥러닝 1》7장
- Kim (2014) “Convolutional Neural Networks for Sentence Classification”
<https://arxiv.org/abs/1408.5882>