인공지능을 위한 머신러닝 알고리즘

7. 역전파

Tacademy

CONTENTS

학습 목표

- 역전파 알고리즘을 이해하고 다층 퍼셉트론의 파라미터 값을 계산할 수 있다.
- 활성함수 미분값의 특징을 이해할 수 있다.

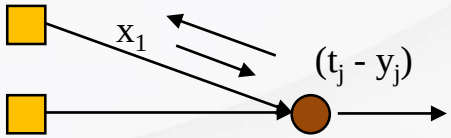
1. 역전파 학습 방법

다층 퍼셉트론

다층 퍼셉트론의 적절한 가중치를 어떻게 찾을 수 있을까?

- 단층 퍼셉트론 모델에서 적절한 가중치를 찾기 위해 경사 하강법 (**Gradient Descent**)을 사용

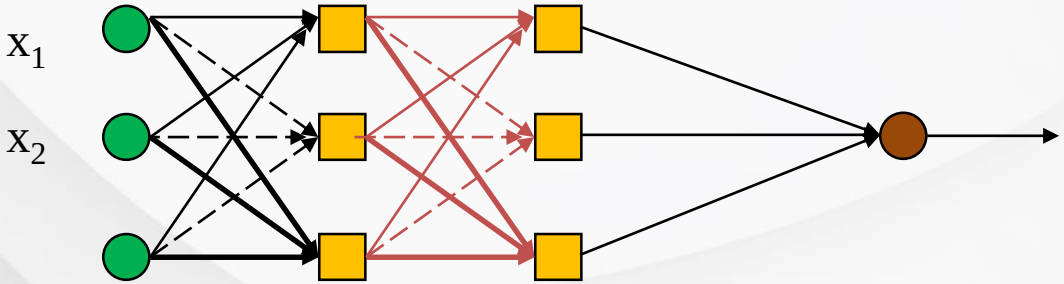
$$\Delta w_{ji} = (t_j - y_j)\, x_i$$



- 노드 i와 출력 j를 연결하는 가중치 w_{ji}는
  출력 j로부터 받은 에러 신호$(t_j - y_j)$와 노드의 입력(x_i)에 의해서만 영향을 받음

다층 퍼셉트론

다층 퍼셉트론의 적절한 가중치를 어떻게 찾을 수 있을까?



- 위와 같이 여러 층을 갖는 다층 퍼셉트론에서 에러가 세번째 층에서만 계산된다면 처음 두 개의 층은 어떻게 가중치를 학습할까?
- 입력층에서는 직접적인 에러 신호 $(t_j - y_j)$ 가 존재하지 않음

기여도 할당 문제 (Credit Assignment Problem)

- 전체 학습 모델을 구성하는데 관여하고 있는 모든 개별 요소들
  예> 은닉 유닛들)에 '기여도' 또는 '책임'을 할당하는 문제
- 다층 신경망에서는 어떤 가중치들을 얼마큼, 어떤 방향으로 학습시켜야 하는지 관련됨
- 앞부분 층의 가중치들이 최종 출력(또는 에러)에 얼마큼 영향을 미치는 결정하는 문제와 비슷
- 가중치 w_{ji}가 에러에 미치는 영향을 계산해야 함

$$\frac{\partial E(t)}{\partial w_{ij}(t)}$$

역전파 (Backpropagation)

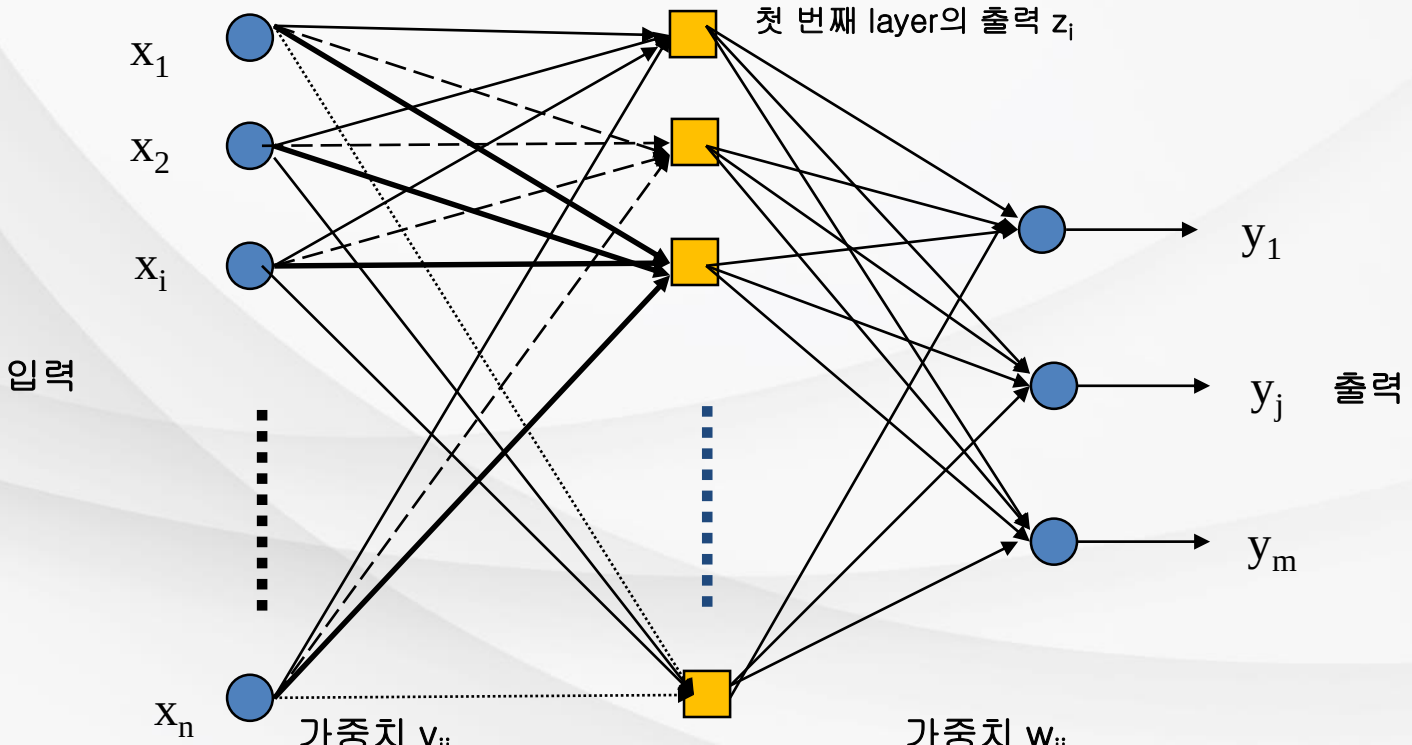
- 다층 퍼셉트론에서 기여도 할당 문제에 대한 해결책

  Rumelhart, Hinton and Williams (1986)

- 역전파는 두 단계로 나뉘어짐

1. 앞먹임 단계	입력 값들을 사용하여 다층 퍼셉트론의 최종 출력을 계산
2. 오류 후방 전파 단계	에러 값을 계산한 뒤, 최종 출력 유닛들부터 시작하여 네트워크의 후방으로 에러 값을 전파

입력층 - 은닉층 – 출력층 구조를 갖는 다층 퍼셉트론의 모습



가중치 v_{ij}
– 입력층의 j번째 노드와 은닉층의 i번째 노드의 연결 가중치

가중치 w_{ij}
– 은닉층의 j번째 노드와 출력층의 i번째 노드의 연결 가중치

다층 퍼셉트론의 노드 활성 값 계산

$$z_i(t) = g(\Sigma_j \; v_{ij}(t)\, x_j(t)) \quad \text{at time t}$$
$$= g(\; u_i(t))$$

$$y_i(t) = g(\Sigma_j \; w_{ij}(t)\, z_j(t)) \quad \text{at time t}$$
$$= g(\; a_i(t))$$

- g는 활성함수 예> 시그모이드 함수
- **Bias**는 추가 가중치로 여겨짐

역전파 – (1) 앞먹임 단계

1. 은닉 유닛들의 값 계산

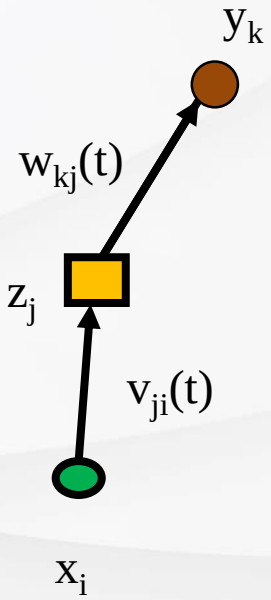
$$u_j(t) = \sum_i v_{ji}(t) x_i(t)$$

$$z_j = g(u_j(t))$$

2. 출력 유닛들의 값 계산

$$a_k(t) = \sum_j w_{kj}(t) z_j$$

$$y_k = g(a_k(t))$$



역전파 – (2) 오류 후방 전파 단계

- 에러의 제곱의 합을 사용할 경우, 다음과 같은 손실 함수 식을 얻음

$$E(t) = \frac{1}{2}\sum_{k=1}(d_k(t) - y_k(t))^2$$

- d_k는 타겟 벡터의 k차원의 값
- E를 줄이기 위해 경사 하강법을 사용하여 가중치를 변경함

$$w_{ij}(t+1) - w_{ij}(t) \propto -\frac{\partial E(t)}{\partial w_{ij}(t)}$$

- 출력 유닛과 은닉 유닛 모두 적용

역전파 – (2) 오류 후방 전파 단계

편미분 방정식은 체인룰을 사용하여 두 개 항의 곱으로 나타낼 수 있음

$$\frac{\partial E(t)}{\partial w_{ij}(t)} = \frac{\partial E(t)}{\partial a_i(t)} \bullet \frac{\partial a_i(t)}{\partial w_{ij}(t)}$$

출력 유닛과 은닉 유닛 모두 적용

은닉 유닛:

$$u_j(t) = \sum_i v_{ji}(t) x_i(t)$$

$$z_j = g(u_j(t))$$

출력 유닛:

$$a_k(t) = \sum_j w_{kj}(t) z_j$$

$$y_k = g(a_k(t))$$

Term A i 번째 출력 유닛의 $a_i(t)$ 값에 대한 에러 값의 변화량

Term B i 번째 출력 유닛에 연결되어있는 j 번째 가중치에 대한 $a_i(t)$의 변화량

역전파 – (3) 손실 함수의 미분

B항

$$\frac{\partial u_i(t)}{\partial v_{ij}(t)} = x_j(t) \qquad \frac{\partial a_i(t)}{\partial w_{ij}(t)} = z_j(t)$$

은닉 유닛의 경우 | 출력 유닛의 경우

은닉 유닛:

$$u_j(t) = \sum_i v_{ji}(t) x_i(t)$$

$$z_j = g(u_j(t))$$

출력 유닛:

$$a_k(t) = \sum_j w_{kj}(t) z_j$$

$$y_k = g(a_k(t))$$

A항

$$\frac{\partial E(t)}{\partial u_j(t)} \qquad \frac{\partial E(t)}{\partial a_j(t)}$$

은닉 유닛의 경우 | 출력 유닛의 경우

체인룰에 의해서 계산 가능

역전파 – (3) 손실 함수의 미분

각 출력 유닛에 대하여 아래의 식을 계산

$$\Delta_j(t) = \frac{\partial E(t)}{\partial a_j(t)} = g'(a_j(t)) \frac{\partial E(t)}{\partial y_j(t)}$$

$$\Delta_j(t) = -g'(a_j(t))(d_j(t) - y_j(t))$$

은닉 유닛:

$$u_j(t) = \sum_i v_{ji}(t) x_i(t)$$

$$z_j = g(u_j(t))$$

출력 유닛:

$$a_k(t) = \sum_j w_{kj}(t) z_j$$

$$y_k = g(a_k(t))$$

손실 함수:

$$E(t) = \frac{1}{2} \sum_{k=1} (d_k(t) - y_k(t))^2$$

역전파 – (3) 손실 함수의 미분

각 은닉 유닛에 대하여 체인룰 사용

$$a_j(t) = \sum_m w_{jm}(t) g(u_m(t))$$

$$\delta_i(t) = \frac{\partial E(t)}{\partial u_i(t)} = \sum_j \frac{\partial E(t)}{\partial a_j(t)} \frac{\partial a_j(t)}{\partial u_i(t)}$$

$$\delta_i(t) = g'(u_i(t)) \sum_j w_{ji} \Delta_j$$

은닉 유닛:

$$u_j(t) = \sum_i v_{ji}(t) x_i(t)$$

$$z_j = g(u_j(t))$$

출력 유닛:

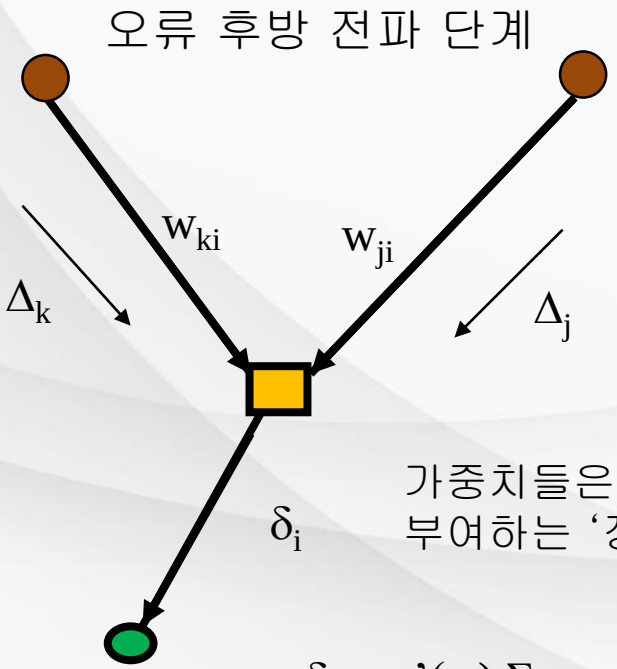
$$a_k(t) = \sum_j w_{kj}(t) z_j$$

$$y_k = g(a_k(t))$$

손실 함수:

$$E(t) = \frac{1}{2} \sum_{k=1} (d_k(t) - y_k(t))^2$$

역전파 – (4) 기여도 할당 문제의 해결



가중치들은 은닉 유닛들에게 '기여도' 또는 '책임'을 부여하는 '정도'의 값으로 해석할 수 있음

$$\delta_i = g'(u_i) \Sigma_j w_{ji} \Delta_j$$

역전파 – (5) 가중치 업데이트

A와 B를 결합

$$\frac{\partial E(t)}{\partial v_{ij}(t)} = \delta_i(t)x_j(t)$$

$$\frac{\partial E(t)}{\partial w_{ij}(t)} = \Delta_i(t)z_j(t)$$

A B

$$\frac{\partial E(t)}{\partial v_{ij}(t)} = \frac{\partial E(t)}{\partial u_i(t)} \bullet \frac{\partial u_i(t)}{\partial v_{ij}(t)}$$

$$\frac{\partial E(t)}{\partial w_{ij}(t)} = \frac{\partial E(t)}{\partial a_i(t)} \bullet \frac{\partial a_i(t)}{\partial w_{ij}(t)}$$

E에 대한 경사 하강법을 하기 위해서 가중치를 다음과 같이 변경해야 함

$$v_{ij}(t+1)\text{-}v_{ij}(t) = \eta\,\delta_i(t)\,x_j(n)$$
$$w_{ij}(t+1)\text{-}w_{ij}(t) = \eta\,\Delta_i(t)\,z_j(t)$$

η은 학습률을 나타내는 파라미터 (0 < η <=1)

역전파 – (5) 가중치 업데이트

❖ 가중치 학습식

$$v_{ij}(t+1) - v_{ij}(t) = \eta \delta_i(t) x_j(t)$$

$$w_{ij}(t+1) - w_{ij}(t) = \eta \Delta_i(t) z_j(t)$$

역전파 – (5) 가중치 업데이트

❖ 출력 유닛

$$w_{ij}(t+1) - w_{ij}(t) = \eta \Delta_i(t) z_j(t)$$
$$= \eta (d_i(t) - y_i(t)) g'(a_i(t)) z_j(t)$$

에러의 크기 / 활성함수의 미분값 / 입력의 크기

❖ 은닉 유닛

$$v_{ij}(t+1) - v_{ij}(t) = \eta \delta_i(t) x_j(t)$$
$$= \eta g'(u_i(t)) x_j(t) \sum_k \Delta_k(t) w_{ki}$$

활성함수의 미분값 / 입력의 크기 / 상위층 유닛들의 기여도를 가중치에 따라 평균 낸 값

2. 활성함수의 미분값

2. 활성함수의 미분값

활성함수에 대한 에러의 변화량

활성함수가 에러의 변화량에 얼만큼 영향을 미칠까?

$$\Delta_i(t) = (d_i(t) - y_i(t))g'(a_i(t))$$

$$\delta_i(t) = g'(u_i(t))\sum_k \Delta_k(t) w_{ki}$$

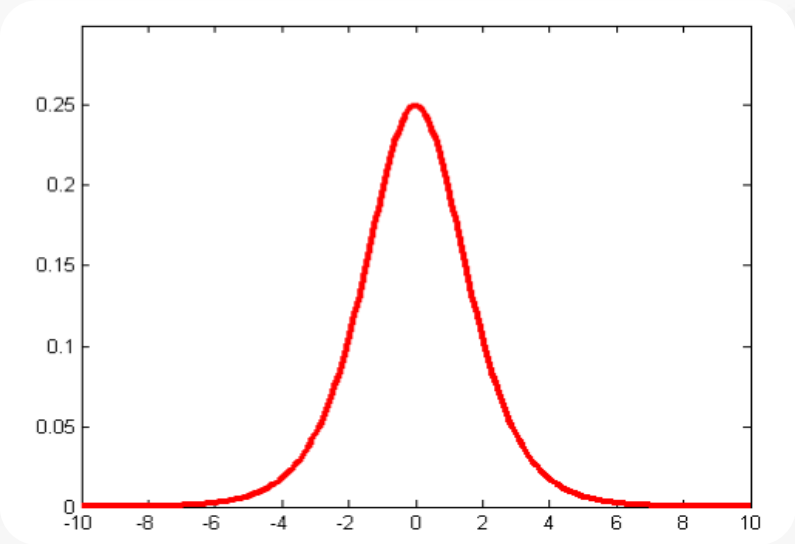
Where : $g'(a_i(t)) = \frac{dg(a)}{da}$

- 활성함수 g에 대한 미분을 계산해야 함
- 미분이 가능하기 위해서 활성함수는 '**smooth**'해야 함

시그모이드 함수

$$g'(a_i(t)) = \frac{k\exp(-k\,a_i(t))}{[1+\exp(-k\,a_i(t))]^2} = kg(a_i(t))[1-g(a_i(t))]$$

since: $y_i(t) = g(a_i(t))$ we have: $g'(a_i(t)) = ky_i(t)(1-y_i(t))$



활성함수 미분값과 가중치의 관계

❖ 가중치의 변화량은 활성함수 미분값의 비례

$$\Delta_i(t) = (d_i(t) - y_i(t))g'(a_i(t))$$

$$\delta_i(t) = g'(u_i(t))\sum_k \Delta_k(t)w_{ki}$$

- 가중치의 학습은 유닛의 $a_i(t)$ 또는 $u_i(t)$ 값이 너무 크거나 작지 않을 경우 잘 됨
- 값이 너무 크거나 작을 때 미분값은 **0**에 가까워짐
- 딥신경망이 학습이 잘 안되었던 이유

학습정리

지금까지 [역전파]에 대해서 살펴보았습니다.

역전파 학습 방법

앞먹임 단계: 입력 값들을 사용하여 다층 퍼셉트론의 최종 출력을 계산
오류 후방 전파 단계: 에러값을 계산한 뒤, 최종 출력 유닛들부터 시작하여 네트워크의 후방으로 에러값을 전파

역전파 학습 방법

은닉 유닛의 가중치를 학습하기 위해서 체인룰을 사용,
은닉 유닛의 출력 유닛들에 대한가중치들은 해당 은닉 유닛에 대한 '기여도' 또는 '책임'을 부여하는 '정도'의 값으로 해석할 수 있음

활성함수의 미분값

시그모이드 함수의 미분 형태 $g'(a_i(t)) = k(1 - \frac{1}{1+e^{-k\,a_i(t)}})\frac{1}{1+e^{-k\,a_i(t)}}$