2021 Spring

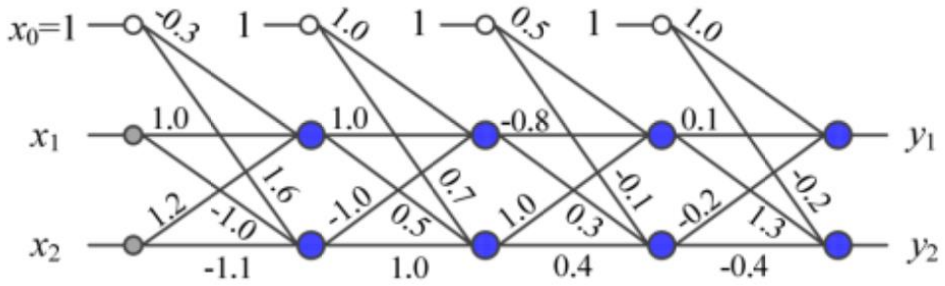
Artificial Intelligence & Deep Learning

Prof. Minsuk Koo

Department of Computer Science & Engineering
Incheon National University

HW Chapter 4.

- #5



- #7

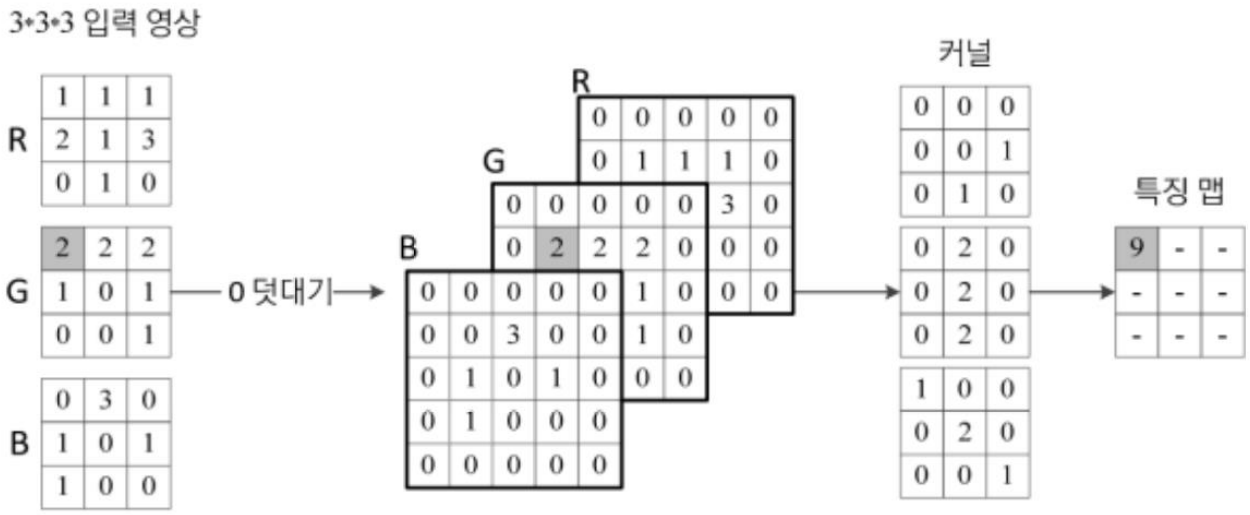
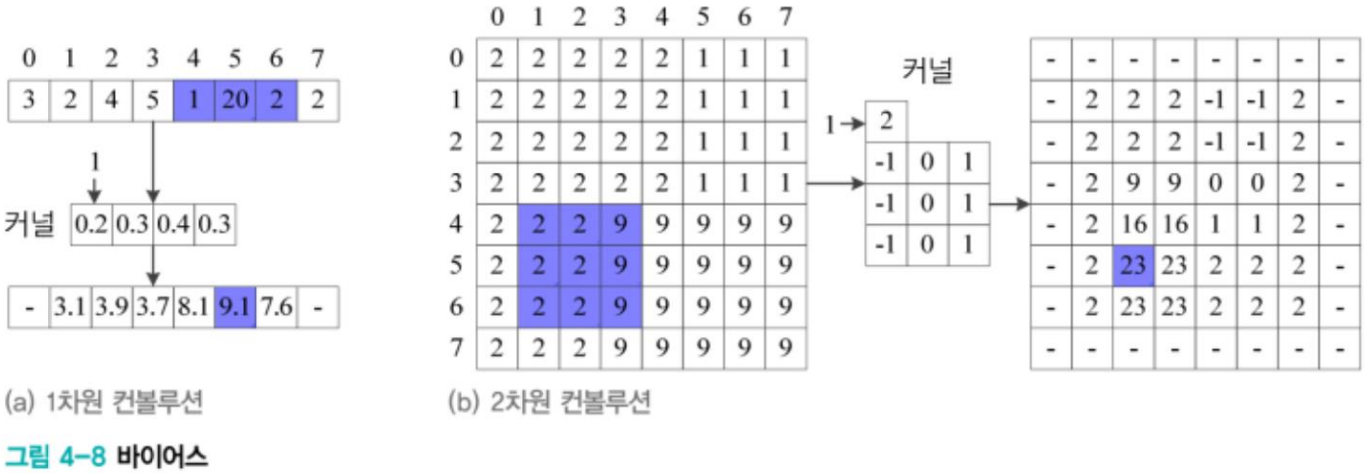


그림 4-14 텐서의 컨볼루션 연산(0 덧대기 적용)

HW Chapter 4.

- #8



(a) 1차원 컨볼루션 (b) 2차원 컨볼루션

그림 4-8 바이어스

- #9
- #10

Chapter 5. 딥러닝 최적화

- 최적화
  - MSE의 단점 및 교차 엔트로피, 로그우도 소개
  - SGD 성능 향상에 효과적인 전처리, 가중치 초기화, 모멘텀, 적응적 학습률, 활성함수, 배치 정규화 설명
  - 규제의 필요성과 원리
  - 규제기법: 가중치 벌칙, 조기 멈춤, 데이터 확대, 드롭아웃, 앙상블
  - 하이퍼 매개변수의 중요성 및 최적화
  - 2차 미분 정보의 활용 및 SGD 보완

Recap: Gradient Descent

current W:

[0.34,
-1.11,
0.78,
0.12,
0.55,
2.81,
-3.1,
-1.5,
0.33,…]

loss 1.25347

W + h (first dim):

[0.34 + **0.0001**,
-1.11, + **0.0001**,
0.78, + **0.0001**,
0.12,
0.55,
2.81,
-3.1,
-1.5,
0.33,…]

loss 1.25347

gradient dW:

[-2.5,
0.6,
0,
?,
?,
?,…]

(1.25347 - 1.25347)/0.0001
= 0

$$\frac{df(x)}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

From CS231n Stanford Univ.

2.2.6 정보이론

- 메시지가 지닌 정보를 수량화할 수 있나?
  - "고비 사막에 눈이 왔다"와 "대관령에 눈이 왔다"라는 두 메시지 중 어느 것이 더 많은 정보를 가지나?
  - 정보이론의 기본 원리 → 확률이 작을수록 많은 정보
- 자기 정보self information
  - 사건(메시지) e_i의 정보량 (단위: 비트 또는 나츠)

$$h(e_i) = -\log_2 P(e_i) \quad \text{또는} \quad h(e_i) = -\log_e P(e_i) \tag{2.44}$$

- 엔트로피
  - 확률변수 x의 불확실성을 나타내는 엔트로피

$$\text{이산 확률분포} \quad H(x) = -\sum_{i=1,k} P(e_i)\log_2 P(e_i) \quad \text{또는} \quad H(x) = -\sum_{i=1,k} P(e_i)\log_e P(e_i) \tag{2.45}$$

$$\text{연속 확률분포} \quad H(x) = -\int_{\mathbb{R}} P(x)\log_2 P(x) \quad \text{또는} \quad H(x) = -\int_{\mathbb{R}} P(x)\log_e P(x) \tag{2.46}$$

2.2.6 정보이론

자기 정보와 엔트로피 예제

예제 2-8

윷을 나타내는 확률변수를 *x*라 할 때 *x*의 엔트로피는 다음과 같다.

$$H(x) = -\left(\frac{4}{16}\log_2\frac{4}{16} + \frac{6}{16}\log_2\frac{6}{16} + \frac{4}{16}\log_2\frac{4}{16} + \frac{1}{16}\log_2\frac{1}{16} + \frac{1}{16}\log_2\frac{1}{16}\right) = 2.0306\text{비트}$$

주사위는 눈이 6개인데 모두 1/6이라는 균일한 확률을 가진다. 이 경우 엔트로피를 계산하면 다음과 같다.

$$H(x) = -\left(\frac{1}{6}\log_2\frac{1}{6} + \frac{1}{6}\log_2\frac{1}{6} + \frac{1}{6}\log_2\frac{1}{6} + \frac{1}{6}\log_2\frac{1}{6} + \frac{1}{6}\log_2\frac{1}{6} + \frac{1}{6}\log_2\frac{1}{6}\right) = 2.585 \text{ 비트}$$

- 주사위가 윷보다 엔트로피가 높은 이유는?

2.2.6 정보이론

- 교차 엔트로피cross entropy
  - 두 확률분포 P와 Q 사이의 교차 엔트로피

$$H(P,Q) = -\sum_{x} P(x)\log_2 Q(x) = -\sum_{i=1,k} P(e_i)\log_2 Q(e_i) \qquad (2.47)$$

  - 식을 전개하면,

$$\begin{aligned} H(P,Q) &= -\sum_{x} P(x)\log_2 Q(x) \\ &= -\sum_{x} P(x)\log_2 P(x) + \sum_{x} P(x)\log_2 P(x) - \sum_{x} P(x)\log_2 Q(x) \\ &= H(P) + \underbrace{\sum_{x} P(x)\log_2 \frac{P(x)}{Q(x)}}_{\text{KL 다이버전스}} \end{aligned}$$

2.2.6 정보이론

- KL 다이버전스
  - 식 (2.48)은 P와 Q 사이의 KL 다이버전스
  - 두 확률분포 사이의 거리를 계산할 때 주로 사용

$$KL(P \parallel Q) = \sum_{x} P(x)\log_2 \frac{P(x)}{Q(x)} \qquad (2.48)$$

- 교차 엔트로피와 KL 다이버전스의 관계

$$P와\ Q의\ 교차\ 엔트로피\ H(P,Q) = H(P) + \sum_{x} P(x)\log_2 \frac{P(x)}{Q(x)} \qquad (2.49)$$

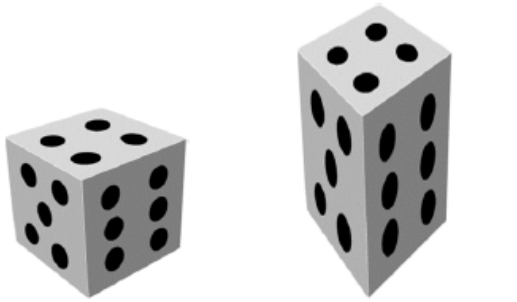
$$= P의\ 엔트로피 + P와\ Q\ 간의\ KL\ 다이버전스$$

2.2.6 정보이론

예제 2-9

[그림 2-21]과 같이 정상적인 주사위와 찌그러진 주사위가 있는데, 정상적인 주사위의 확률분포는 P, 찌그러진 주사위의 확률분포는 Q를 따르며, P와 Q가 다음과 같이 분포한다고 가정하자.

$$P(1)=\frac{1}{6},\ P(2)=\frac{1}{6},\ P(3)=\frac{1}{6},\ P(4)=\frac{1}{6},\ P(5)=\frac{1}{6},\ P(6)=\frac{1}{6}$$
$$Q(1)=\frac{3}{12},\ Q(2)=\frac{1}{12},\ Q(3)=\frac{1}{12},\ Q(4)=\frac{1}{12},\ Q(5)=\frac{3}{12},\ Q(6)=\frac{3}{12}$$



(a) 정상 주사위 (b) 찌그러진 주사위

그림 2-21 확률분포가 다른 두 주사위

확률분포 P와 Q 사이의 교차 엔트로피와 KL 다이버전스는 다음과 같다.

$$H(P,Q)=-\left(\frac{1}{6}\log_2\frac{3}{12}+\frac{1}{6}\log_2\frac{1}{12}+\frac{1}{6}\log_2\frac{1}{12}+\frac{1}{6}\log_2\frac{1}{12}+\frac{1}{6}\log_2\frac{3}{12}+\frac{1}{6}\log_2\frac{3}{12}\right)=2.7925$$
$$KL(P\parallel Q)=\frac{1}{6}\log_2\frac{2}{3}+\frac{1}{6}\log_2 2+\frac{1}{6}\log_2 2+\frac{1}{6}\log_2\frac{2}{3}+\frac{1}{6}\log_2\frac{2}{3}=0.2075$$

[예제 2-8]에서 P의 엔트로피 $H(P)$는 2.585이었다. 따라서 식 (2.49)가 성립함을 알 수 있다.

5.1 목적함수: 교차 엔트로피와 로그우도

- 5.1.1 평균제곱 오차를 다시 생각하기
- 5.1.2 교차 엔트로피 목적함수
- 5.1.3 softmax 활성함수와 로그우도 목적함수

시험에서는 틀린 만큼 합당한 벌점을 받는 것이 중요하다. 그래야 다음 시험에서 심기일전으로 공부하여 틀리는 개수를 줄일 가능성이 크기 때문이다. 틀린 개수에 상관없이 비슷한 벌점을 받는다면 나태해져 성적을 올리는 데 지연이 발생할 것이다. 이러한 원리가 기계 학습에도 적용될까?

5.1.1 평균제곱 오차 다시 생각하기

- 평균제곱 오차(MSE) 목적함수

$$e = \frac{1}{2}\|\mathbf{y} - \mathbf{o}\|_2^2 \qquad (5.1)$$

  - 오차가 클수록 e값이 크므로 벌점으로 훌륭함

- 하지만 큰 허점
  - 왼쪽 상황은 $e = 0.2815$, 오른쪽 상황은 $e = 0.4971$이므로 오른쪽이 더 큰 벌점을 받아야 마땅함

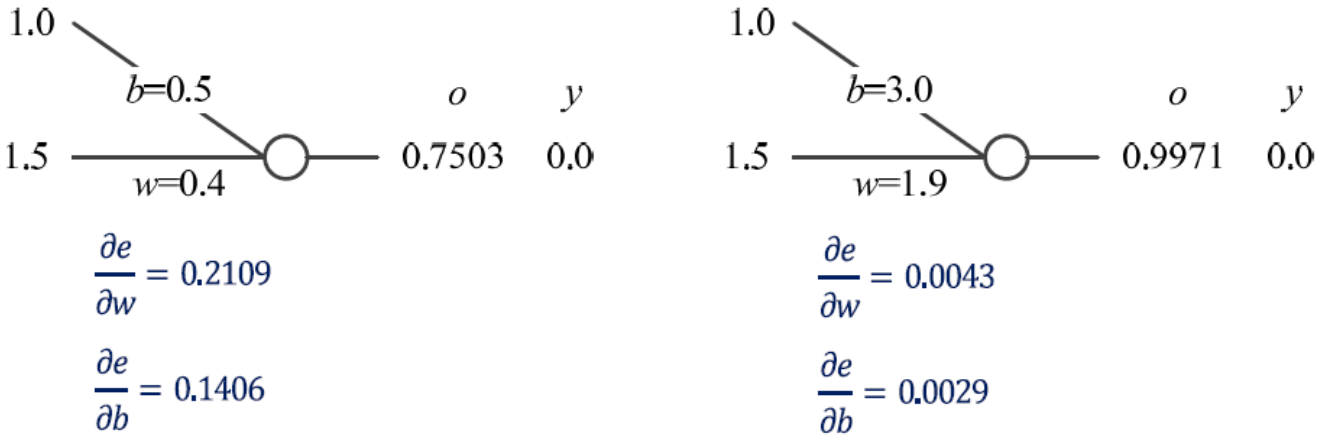


그림 5-1 MSE가 목적함수로서 부적절한 상황

5.1.1 평균제곱 오차 다시 생각하기

- 큰 허점
  - 식 (5.3)의 그레이디언트가 벌점에 해당

$$e = \frac{1}{2}(y - o)^2 = \frac{1}{2}(y - \sigma(wx + b))^2 \tag{5.2}$$

$$\left.\begin{aligned} \frac{\partial e}{\partial w} &= -(y - o)x\sigma'(wx + b) \\ \frac{\partial e}{\partial b} &= -(y - o)\sigma'(wx + b) \end{aligned}\right\} \tag{5.3}$$

  - 그레이디언트를 계산해보면 왼쪽 상황의 그레이디언트가 더 큼 → 더 많은 오류를 범한 상황이 더 낮은 벌점을 받은 꼴 → 학습이 더딘 부정적 효과

- 이유
  - $wx + b$(아래 그래프의 가로축에 해당)가 커지면 그레이디언트가 작아짐

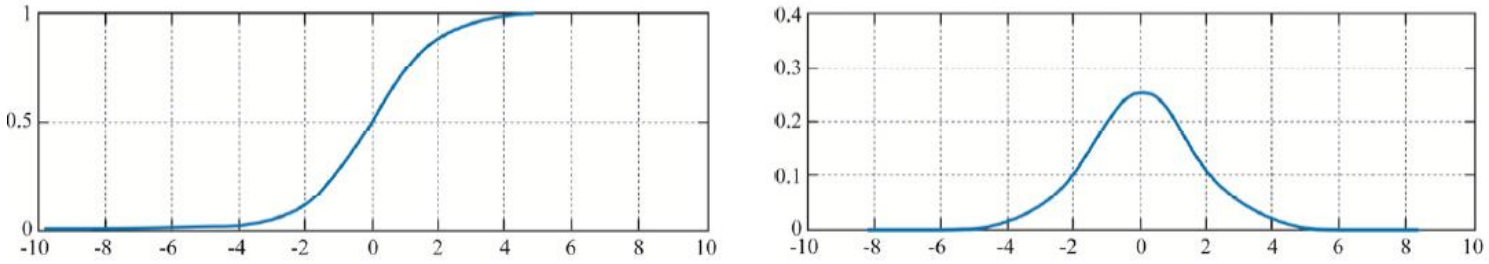


그림 5-2 로지스틱 시그모이드함수와 도함수

5.1.2 교차 엔트로피 목적함수

교차 엔트로피

- 레이블에 해당하는 y가 확률변수 (부류가 2개라고 가정하면 $y \in \{0,1\}$)
- 확률 분포: P는 정답 레이블, Q는 신경망 출력

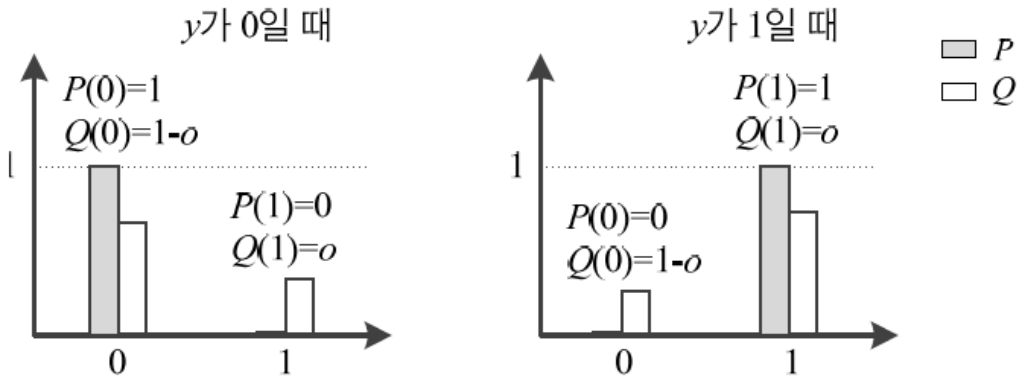


그림 5-3 레이블 y가 0일 때와 1일 때의 P와 Q의 확률분포

- 확률분포를 통일된 수식으로 쓰면,

$$P(0) = 1 - y \quad Q(0) = 1 - o$$
$$P(1) = y \quad Q(1) = o$$

- 교차 엔트로피 식은 $H(P,Q) = -\sum_{y\in\{0,1\}} P(y)\log_2 Q(y)$

5.1.2 교차 엔트로피 목적함수

교차 엔트로피 목적함수

$$e = -(y \log_2 o + (1-y)\log_2(1-o)), \quad \text{이때, } o = \sigma(z) \text{이고} z = wx + b \qquad (5.4)$$

- 제구실 하는지 확인
  - y가 1, o가 0.98일 때 (예측이 잘된 경우)
    - 오류 $e = -(1 \log_2 0.98 + (1-1)\log_2(1-0.98)) = 0.0291$로서 낮은 값
  - y가 1, o가 0.0001일 때 (예측이 엉터리인 경우)
    - 오류 $e = -(1 \log_2 0.0001 + (1-1)\log_2(1-0.0001)) = 13.2877$로서 높은 값

5.1.2 교차 엔트로피 목적함수

- 공정한 벌점을 부여하는지 확인 (MSE의 느린 학습 문제를 해결하나?)
  - 도함수를 구하면,

$$\begin{aligned}\frac{\partial e}{\partial w} &= -\left(\frac{y}{o}-\frac{1-y}{1-o}\right)\frac{\partial o}{\partial w}\\ &= -\left(\frac{y}{o}-\frac{1-y}{1-o}\right)x\sigma'(z)\\ &= -x\left(\frac{y}{o}-\frac{1-y}{1-o}\right)o(1-o)\\ &= x(o-y)\end{aligned}$$

$$\left.\begin{aligned}\frac{\partial e}{\partial w} &= x(o-y)\\ \frac{\partial e}{\partial b} &= (o-y)\end{aligned}\right\} \quad (5.5)$$

  - 그레이디언트를 계산해 보면, 오류가 더 큰 오른쪽에 더 큰 벌점(그레이디언트) 부과

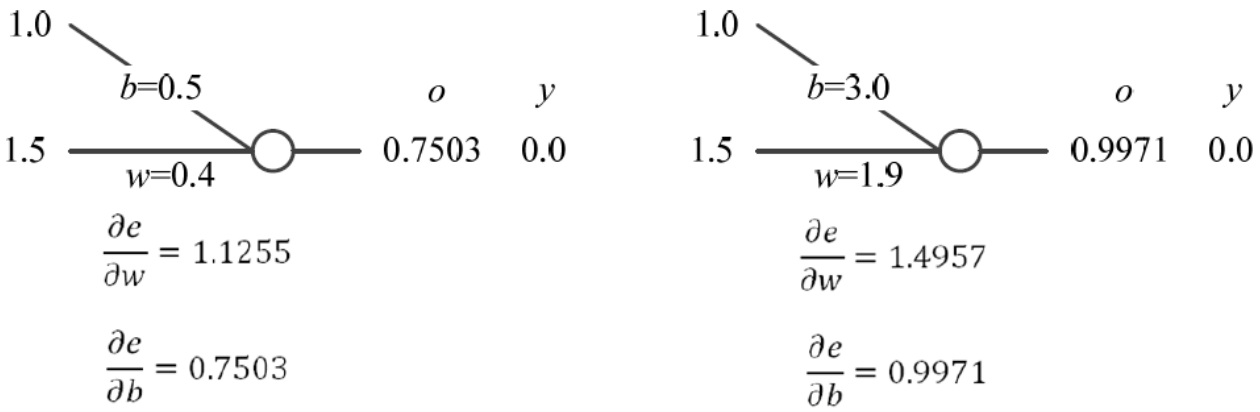


그림 5-4 교차 엔트로피를 목적함수로 사용하여 느린 학습 문제를 해결

5.1.2 교차 엔트로피 목적함수

- 식 (5.4)를 c개의 출력 노드를 가진 경우로 확장
  - 출력 벡터 $\mathbf{o} = (o_1, o_2, \cdots, o_c)^{\mathrm{T}}$인 상황으로 확장 ([그림 4-3]의 DMLP)

$$e = -\sum_{i=1,c} (y_i \log_2 o_i + (1 - y_i) \log_2 (1 - o_i)) \qquad (5.6)$$