인 공 지 능

[심층학습 최적화 I]

jaekoo@kookmin.ac.kr

KMU 국민대학교 KOOKMIN UNIVERSITY

미리보기

- 과학 혹은 공학에서 최적화
  - 예) 우주선의 최적궤도, 운영체제의 작업 할당 계획 등
- 기계학습의 최적화도 매우 복잡함
  - 훈련집합으로 학습을 마친 후, 현장에서 발생하는 새로운unknown 샘플을 잘 예측해야 함

    → 즉, 일반화generalization 능력이 좋아야 함
    - 훈련집합은 전체 데이터 (실제, 알 수 없음) 대리자 역할
    - 검증집합은 테스트집합 대리자 역할
    - MSE, log-likelihood 등의 손실함수는 주어진 과업의 학습 성능(=판단 기준) 대리자 역할
- 기계 학습의 최적화가 어려운 이유
  - 대리자 관계
  - 매개탐색 공간에서 목적함수의 비볼록non-covex 성질, 고차원 특징 공간, 데이터의 희소성 등
  - 긴 훈련 시간
- 해당 장은 깊은 신경망 최적화에 효과적인 방안을 제시함

5.1 목적함수: 교차 엔트로피와 로그우도

- 5.1.1 평균제곱 오차를 다시 생각하기
- 5.1.2 교차 엔트로피 목적함수
- 5.1.3 소프트맥스softmax 함수와 로그우도 목적함수

5.1 목적함수: 교차 엔트로피와 로그우도

- 학습 과정의 성능 판정의 중요성

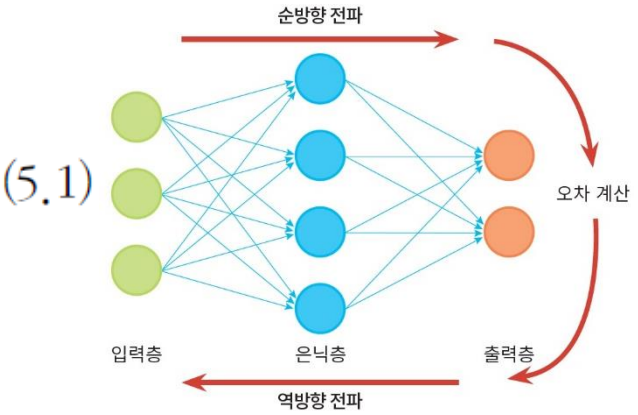
시험에서는 틀린 만큼 합당한 벌점을 받는 것이 중요하다. 그래야 다음 시험에서 심기일전으로 공부하여 틀리는 개수를 줄일 가능성이 크기 때문이다. 틀린 개수에 상관없이 비슷한 벌점을 받는다면 나태해져 성적을 올리는 데 지연이 발생할 것이다. 이러한 원리가 기계 학습에도 적용될까?

5.1.1 평균제곱 오차 다시 생각하기

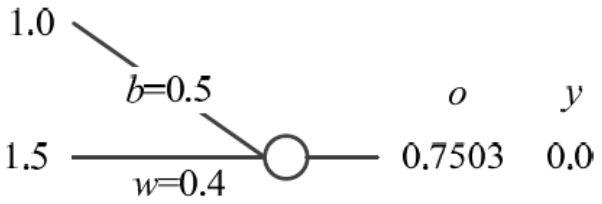
- 평균제곱 오차(MSE) 목적함수

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

  - 오차가 클수록 e값이 크므로 벌점(정량적 성능)으로 활용됨

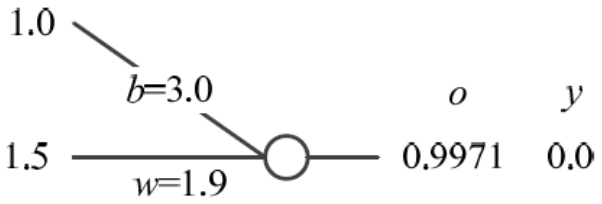


- 하지만, 큰 허점이 존재
  - 왼쪽 상황은 $e = 0.2815$,
  - 오른쪽 상황은 $e = 0.4971$이므로 오른쪽이 더 큰 벌점을 받아야 마땅함



$$\frac{\partial e}{\partial w} = 0.2109$$

$$\frac{\partial e}{\partial b} = 0.1406$$



$$\frac{\partial e}{\partial w} = 0.0043$$

$$\frac{\partial e}{\partial b} = 0.0029$$

그림 5-1 MSE가 목적함수로서 부적절한 상황

신경망 학습 과정에서
학습은 오류를 줄이는 방향으로 가중치와 편향을 교정
← 큰 교정이 필요함에도 작은 경사도로 작게 갱신됨

5.1.1 평균제곱 오차 다시 생각하기

큰 허점

- 식 (5.3)의 경사도gradient가 벌점에 해당

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

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

- 경사도를 계산해보면 왼쪽 상황의 경사도가 더 큼

  → 더 많은 오류를 범한 상황이 더 낮은 벌점을 받은 꼴

  → 학습이 더딘 부정적 효과

이유

- $wx+b$ (그림의 가로축에 해당)가 커지면 경사도가 작아짐

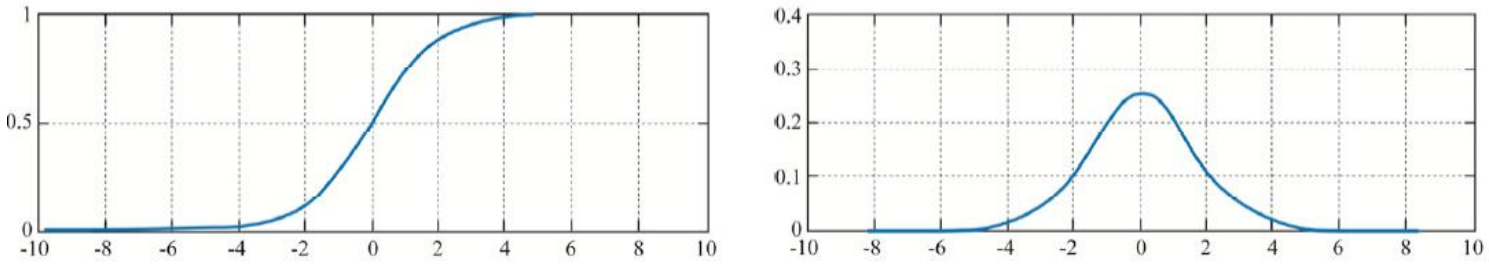


그림 5-2 로지스틱 시그모이드함수와 도함수

5.1.2 교차 엔트로피 목적함수

교차 엔트로피cross entropy

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

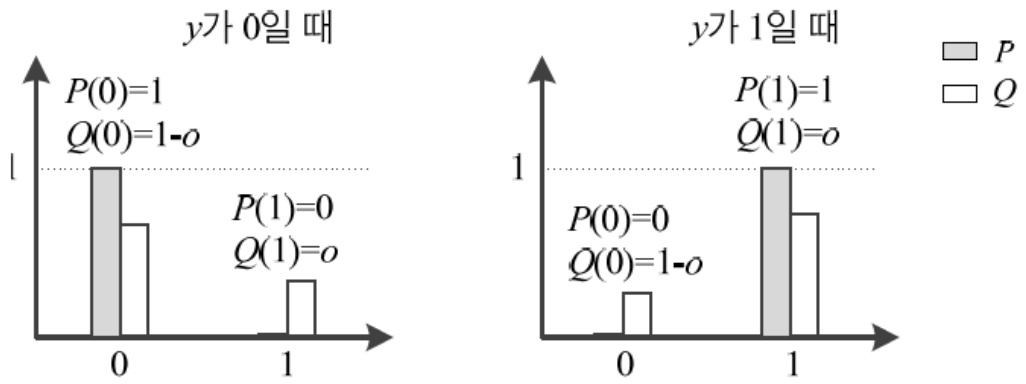


그림 5-3 레이블 y가 0일 때와 1일 때의 P와 Q의 확률분포

- 확률분포를 통일된 수식으로 쓰면,

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

신경망 출력으로 표기하면,
$o = \sigma(z)$이고 $z = wx + b$

- 식 (2.47) 교차 엔트로피 $H(P,Q) = -\sum_x P(x)\log_2 Q(x) = -\sum_{i=1,\ldots,k} P(e_i)\log_2 Q(e_i)$을 적용

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

5.1.2 교차 엔트로피 목적함수

간단한 예

- $P(x)$와 $Q(x)$의 값을 극단적인 경우로 가정 1 혹은 0으로 단순화 가정
- 교차 엔트로피 $H(P,Q) = -\sum P(x) log Q(x)$를 구하면,

잘못된 학습으로 오분류된 손실은

$$-[1\ 0]\begin{bmatrix}\log 0\\ \log 1\end{bmatrix}=-(-\infty+0)=\infty$$

잘된 학습으로 제대로 분류된 손실은

$$-[1\ 0]\begin{bmatrix}\log 1\\ \log 0\end{bmatrix}=-(0+0)=0$$

5.1.2 교차 엔트로피 목적함수

교차 엔트로피 목적함수

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

- 역할을 잘 수행하는지 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 잘못된 경우, 혹은 오분류된 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 해결 확인)
  - 도함수를 구하면,

식 (5.4) 대입, 연쇄법칙 적용
(log f(x))′=f′(x)/f(x)

$e = -(y \log_2 o + (1-y)\log_2(1-o))$

$o = \sigma(z)$이고 $z = wx + b$

$$\frac{\partial e}{\partial w} = -\left(\frac{y}{o} - \frac{1-y}{1-o}\right)\frac{\partial o}{\partial w}$$

$$= -\left(\frac{y}{o} - \frac{1-y}{1-o}\right) x\sigma'(z)$$

$\sigma'(z) = \sigma(z)(1-\sigma(z)) = o(1-o)$

$$= -x\left(\frac{y}{o} - \frac{1-y}{1-o}\right) o(1-o)$$

$$= x(o-y)$$

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

  - 경사도를 계산해 보면, 오류가 더 큰 오른쪽에 더 큰 벌점 (경사도) 부과

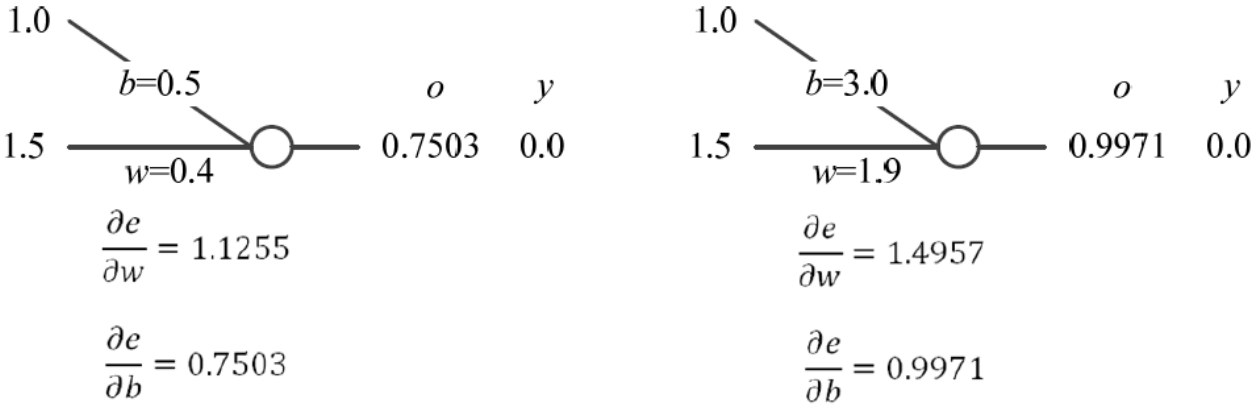


그림 5-4 교차 엔트로피를 목적함수로 사용하여 느린 학습 문제를 해결

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)

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

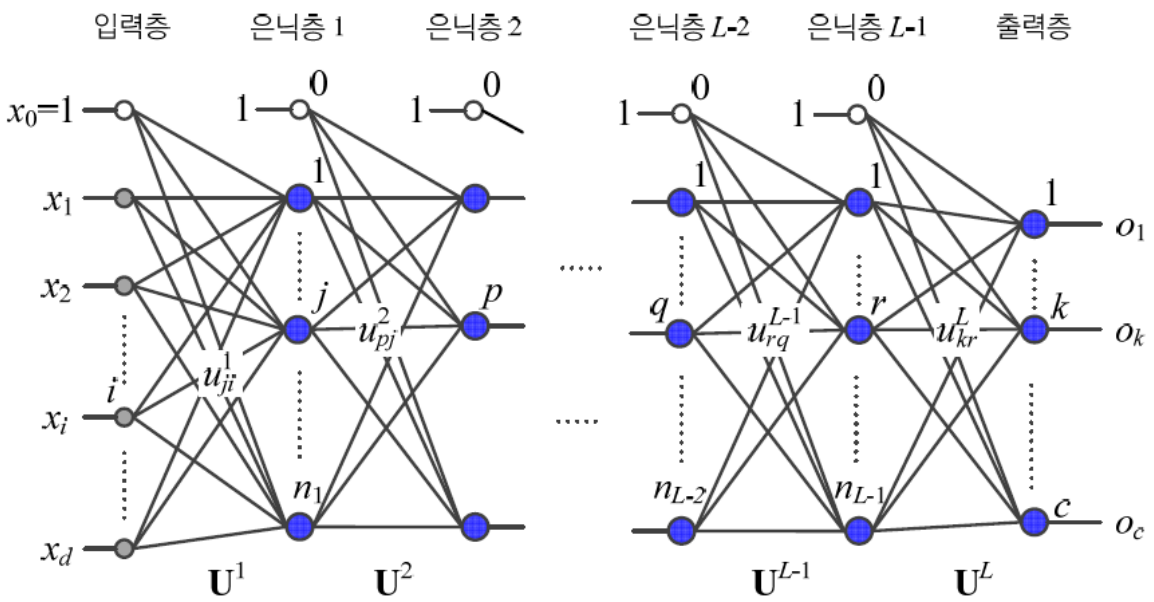


그림 4-3 깊은 MLP(DMLP)의 구조

  - 예) $y \in \{$개, 고양이, 사람$\}$ 경우, $P(y) = [0, 0, 1]$와 $Q(y) = [0.2, 0.3, 0.5]$일 때, 교차 엔트로피는 $-\log(0.5)$

5.1.3 소프트맥스 활성함수와 로그우도 목적함수

소프트맥스softmax 함수

$$o_j = \frac{e^{s_j}}{\sum_{i=1,c} e^{s_i}} \qquad (5.7)$$

- 동작 예
  - 소프트맥스는 최대max를 모방

← 출력 노드의 중간 계산 결과 s_i^L의 최댓값을 더욱 활성화하고 다른 작은 값들은 억제

  - 모두 더하면 1이 되어 확률 모방

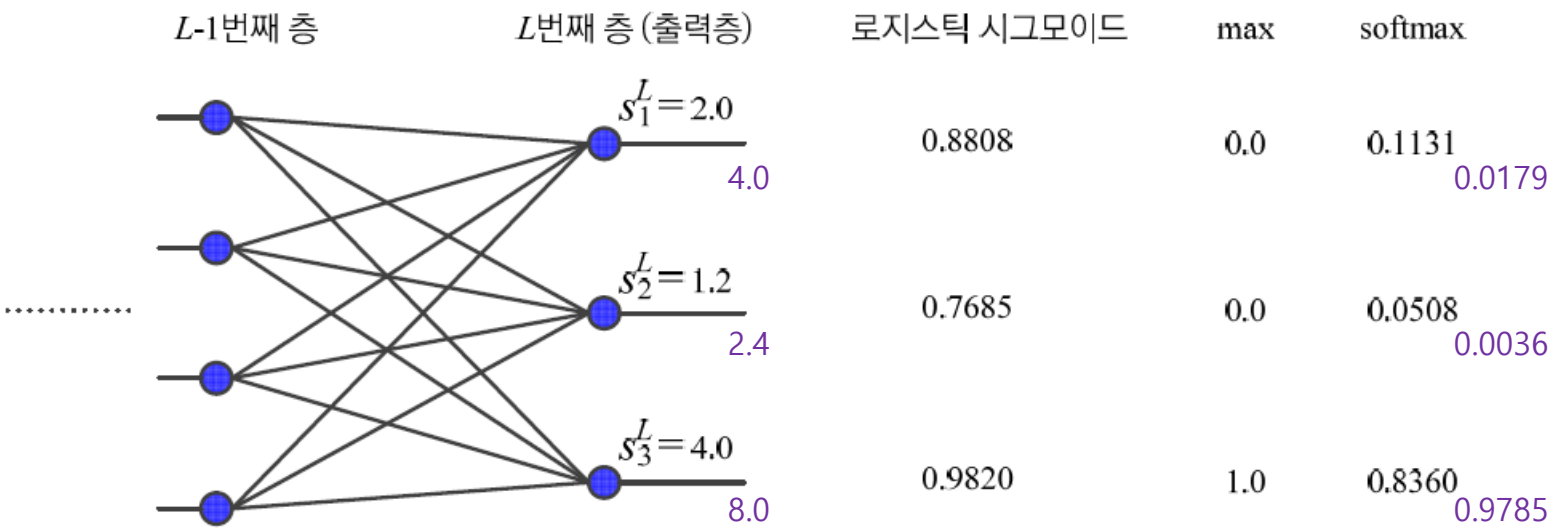


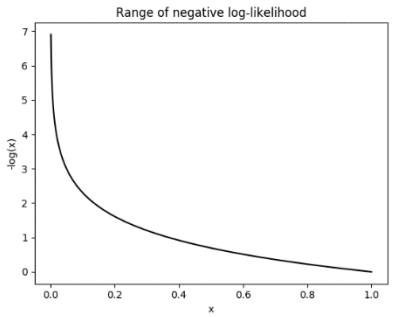
그림 5-5 출력층의 활성함수로 로지스틱 시그모이드와 softmax 비교

출력층의 변화에 따라
소프트맥스의 결과도 최대 출력에 더 가까워짐
따라서 부드러운 최대 함수로 불림

5.1.3 소프트맥스 활성함수와 로그우도 목적함수

음의 로그우도negative log-likelihood 목적함수

$$e = -\log_2 o_y \qquad (5.8)$$



- 모든 출력 노드값을 사용하는 MSE나 교차 엔트로피와 달리 o_y라는 하나의 노드만 적용
- o_y는 샘플의 정답에 해당하는 노드의 출력값
  - 동작 예시1) [그림 5-5]에서 현재 샘플이 두 번째 부류라면 o_y는 o_2
    $e = -\log_2 0.0508 = 4.2990$. 잘못 분류한 셈이므로 목적함수값이 큼
  - 동작 예시2) [그림 5-5]에서 현재 샘플이 세 번째 부류라면 o_y는 o_3
    $e = -\log_2 0.8360 = 0.2584$. 제대로 분류한 셈이므로 목적함수값이 작음

소프트맥스와 로그우도

- 소프트맥스는 최댓값이 아닌 값을 억제하여 0에 가깝게 만든다는 의도 내포
- 신경망에 의한 샘플의 정답에 해당하는 노드만 보겠다는 로그우도와 잘 어울림
- 따라서 둘을 결합하여 사용하는 경우가 많음

5.1.3 소프트맥스 활성함수와 로그우도 목적함수

- 소프트맥스와 교차 엔트로피 목적함수
  - 로그우도 손실함수 ~ 교차엔트로피 최소화

$$s = f(x_i; W)$$

$$P(Y = k|X = x_i) = \frac{e^{s_k}}{\sum_j e^{s_j}}$$

Softmax Function

Probabilities must be >= 0

Probabilities must sum to 1

$$L_i = -\log P(Y = y_i|X = x_i)$$

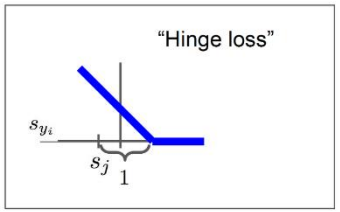
	Unnormalized log-probabilities / logits	exp	unnormalized probabilities	normalize	probabilities	compare	Correct probs
cat	3.2	→	24.5	→	0.13	→ ←	1.00
car	5.1	→	164.0	→	0.87		0.00
frog	-1.7	→	0.18	→	0.00		0.00

Cross Entropy

$$H(P, Q) = H(p) + D_{KL}(P\|Q)$$

5.1.3 소프트맥스 활성함수와 로그우도 목적함수

- 소프트맥스와 힌지로스hinge loss 출력 비교



$$L_i = \sum_{j \neq y_i} \begin{cases} 0 & \text{if } s_{y_i} \geq s_j + 1 \\ s_j - s_{y_i} + 1 & \text{otherwise} \end{cases}$$
$$= \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

matrix multiply + bias offset

0.01	-0.05	0.1	0.05
0.7	0.2	0.05	0.16
0.0	-0.45	-0.2	0.03

W

-15
22
-44
56

x_i

+

0.0
0.2
-0.3

b

분류 정답 y_i 2

hinge loss (SVM)

-2.85
0.86
0.28

max(0, -2.85 - 0.28 + 1) +
max(0, 0.86 - 0.28 + 1)
=
1.58

cross-entropy loss (Softmax)

-2.85	exp	0.058	normalize (to sum to one)	0.016	- log(0.353) = **0.452**
0.86		2.36		0.631	
0.28		1.32		0.353	

5.1.3 소프트맥스 활성함수와 로그우도 목적함수

- 소프트맥스 분류기softmax classifier
  - 다항 로지스틱 회귀 분석multinomial logistic regression의 예
  - 분류기의 최종 값을 확률로 표현
  - 소프트맥스와 로그우도 목적함수

$$s = f(x_i; W)$$

$$P(Y = k|X = x_i) = \frac{e^{s_k}}{\sum_j e^{s_j}}$$

Softmax Function

$$L_i = -\log P(Y = y_i|X = x_i)$$

Probabilities must be >= 0

Probabilities must sum to 1

cat	3.2	24.5	0.13
car	5.1	164.0	0.87
frog	-1.7	0.18	0.00

exp

normalize

Unnormalized log-probabilities / logits

unnormalized probabilities

probabilities

→ L_i = -log(0.13) = **2.04**

Maximum Likelihood Estimation
Choose probabilities to maximize the likelihood of the observed data

5.1.3 소프트맥스 활성함수와 로그우도 목적함수

- 미니배치 단위 예