

CSCI E-89B Lecture 1

아무것도 몰라도 이해되는 신경망 기초

정리: InJoo 학습노트

Fall 2025

읽기 가이드

이 문서는 전혀 모르는 사람을 위해 작성되었다. 수학식은 꼭 필요한 만큼만, 대신 **비유** + **손계산 예시** + **체크리스트**로 이해를 도와준다.

- **핵심 비유:** 신경망은 “입력 재료 → 여러 단계의 조리(층) → 결과 요리”를 만드는 레시피이다.
- **핵심 목표:** 예측이 실제와 얼마나 다른지(오차)를 숫자로 재고(손실 함수), 그걸 줄이는 방향으로 레시피(가중치)를 조금씩 조정(경사 하강법)한다.
- **읽는 순서:** 1) 신경망이 뭔지 직관, 2) 층/뉴런/활성화함수, 3) 손실/비용 함수, 4) 순전파/역전파 계산, 5) 경사하강법(배치/미니배치/SGD), 6) 실전 팁 & FAQ.

1 신경망을 아주 직관적으로 이해하기

한 줄 요약

신경망(Neural Network)은 입력 x 를 받아 여러 층(layer)을 거치며 출력 \hat{y} 를 만드는 함수들의 합성이다:

$$\hat{y} = f^{(L)}(f^{(L-1)}(\dots f^{(1)}(x) \dots)).$$

여기서 각 $f^{(\ell)}$ 은 선형변환(가중치/편향)과 비선형 활성화 함수로 이루어진다.

왜 굳이 “깊게(Deep)” 써야 할까?

- 얇은(한두 단계) 모델은 직선/완만한 곡선 같은 단순한 경계만 만든다.
- 복잡한 문제(예: 이미지/언어)는 비선형 변환을 여러 번 적용해 복잡한 모양의 결정 경계가 필요하다.
- 층을 쌓으면, 간단한 조각 특징 → 중간 특징 → 고수준 의미 식으로 표현이 점점 “추상화”된다.

2 뉴런, 층, 활성화 함수

2.1 두 입력, 은닉 2, 출력 1인 2-2-1 미니 네트워크

구성: 입력 $x = (x_1, x_2)$, 은닉층 뉴런 u_1, u_2 , 출력 \hat{y} . 은닉층은 ReLU를, 출력층은 회귀면 선형, 이진분류면 시그모이드, 다중분류면 소프트맥스를 쓴다고 생각하면 된다.

$$\begin{aligned}u_1 &= f\left(w_{01}^{(1)} + w_{11}^{(1)}x_1 + w_{21}^{(1)}x_2\right), \\u_2 &= f\left(w_{02}^{(1)} + w_{12}^{(1)}x_1 + w_{22}^{(1)}x_2\right), \\ \hat{y} &= g\left(w_0^{(2)} + w_1^{(2)}u_1 + w_2^{(2)}u_2\right).\end{aligned}$$

여기서 f 는 은닉층 활성화, g 는 출력층 활성화다.

2.2 활성화 함수(왜 비선형이 필요?)

- **ReLU** ($f(x) = \max(0, x)$): 빠르고 간단, 깊은 네트워크에서도 학습 잘 됨. 기본값으로 생각해도 좋다.
- **Sigmoid** ($\sigma(x) = 1/(1 + e^{-x})$): 출력이 (0,1) 이라 확률처럼 해석 쉬움(이진분류 출력층에 주로 사용).
- **Softmax**: $\text{softmax}_j(z) = \frac{e^{z_j}}{\sum_k e^{z_k}}$ (다중분류 출력층).
- **Tanh**: $(-1, 1)$ 범위. 옛날엔 자주 썼지만 ReLU에 밀림.

핵심: 활성화가 비선형이어야 층을 쌓을 의미가 생긴다. 선형만 쌓으면 전체가 결국 또 선형이 된다.

3 손실(LOSS)과 비용(COST) 정확히 구분하기

3.1 손실 함수 $L^{(i)}$ (한 샘플의 틀림 정도)

- 회귀(실수 예측): $L^{(i)} = (\hat{y}^{(i)} - y^{(i)})^2$ (MSE 단일항)
- 이진분류: $L^{(i)} = -(y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}))$
- 다중분류: $L^{(i)} = -\sum_{c=1}^M y_c^{(i)} \log \hat{y}_c^{(i)}$ (원-핫 y 가정)

3.2 비용 함수 $J(\mathbf{w})$ (데이터 전체 평균 오차)

$$J(\mathbf{w}) = \frac{1}{m} \sum_{i=1}^m L^{(i)}(\mathbf{w}).$$

요약: L = 개별샘플 오류, J = 전체 평균 오류(우리가 최소화하려는 목표).

마무리 요약(핵심만 쪽)

- 신경망은 선형 + 비선형 변환을 층층이 쌓아 복잡한 함수를 근사한다.
- 손실은 한 샘플, 비용은 전체 평균. 우리는 비용을 내리도록 가중치를 업데이트한다.
- 순전파로 예측, 역전파로 기울기, 경사하강법으로 업데이트. 학습률/미니배치가 실전 핵심 노브.