Data Science

데이터 모델링

분석 기법의 적용

강사 양석환

분석 기법의 개요

분석 기법

- **학습 유형에 따른 분석 모델**
  - **지도학습(Supervised Learning)**
    - **정답(Label)이 있는 데이터를 활용해 데이터를 학습시키는 방법**
    - **입력 값이 주어질 때 정답이 무엇인지 알려주면서 모델을 학습시키는 방법**
    - **대표적으로 분류(Classification), 회귀(Regression)로 구분됨**
      - **분류: 주어진 데이터를 여러 가지 카테고리 중 하나로 분류하는 것**
        - 이진분류: 주어진 데이터에 대하여 두 가지 중 하나로 분류
        - 다중분류: 주어진 데이터에 대하여 여러 가지 중 하나로 분류
      - **회귀: 주어진 데이터의 특징을 기반으로 새로운 데이터 값을 예측하는 것**

분석 기법

- **학습 유형에 따른 분석 모델**
  - **비지도학습(Unsupervised Learning)**
    - 정답(Label)이 없는 데이터를 스스로 학습하여 숨겨진 의미, 패턴을 찾아내고 구조화 하는 방법
    - 입력 값은 있으나 정답이 없기 때문에 기준이 되는 출력 값이 존재하지 않으므로 모델의 성능 평가가 어려움
    - 군집분석(Clustering), 연관성분석(Association Analysis), 인공신경망(Neural Networks), 오토 인코더(Auto-Encoder) 등의 모델이 있음
  - **준지도학습(Semi-Supervised Learning)**
    - 정답이 있는 데이터와 정답이 없는 데이터를 동시에 학습하는 방법
    - 정답이 있는 소수의 데이터만으로 부분학습 모델을 만들고, 이 모델을 사용해서 나머지 정답이 없는 데이터에 정답을 생성한 후 지도학습을 수행함

분석 기법

- 학습 유형에 따른 분석 모델
  - 강화학습(Reinforcement Learning)
    - 주어진 환경에서 보상을 최대화하도록 에이전트를 학습시키는 기법
    - 에이전트와 환경의 상태 등이 인공신경망으로 입력
    - 에이전트가 행동을 결정하고 환경을 통해 보상(또는 벌칙)이 있으면 이전의 입력 값과 행동 등을 긍정적(또는 부정적)으로 평가하여 학습을 수행함

분석 기법

- **데이터 분석 알고리즘과 분야**

알고리즘	주 활용 분야
업리프트 모델링	단계적 추정, 예측 분석
생존 분석	의료 통계, 설비 분야 사건 예측
회귀 분석	예측, 추정 분석
시각화	원인과 관계 분석
기초 통계	기초 통계현황 파악
부스팅, 배깅	분류 분석
시계열 분석	시간 상의 예측(이자율)
요인 분석	차원 축소
텍스트 마이닝	감성 분석

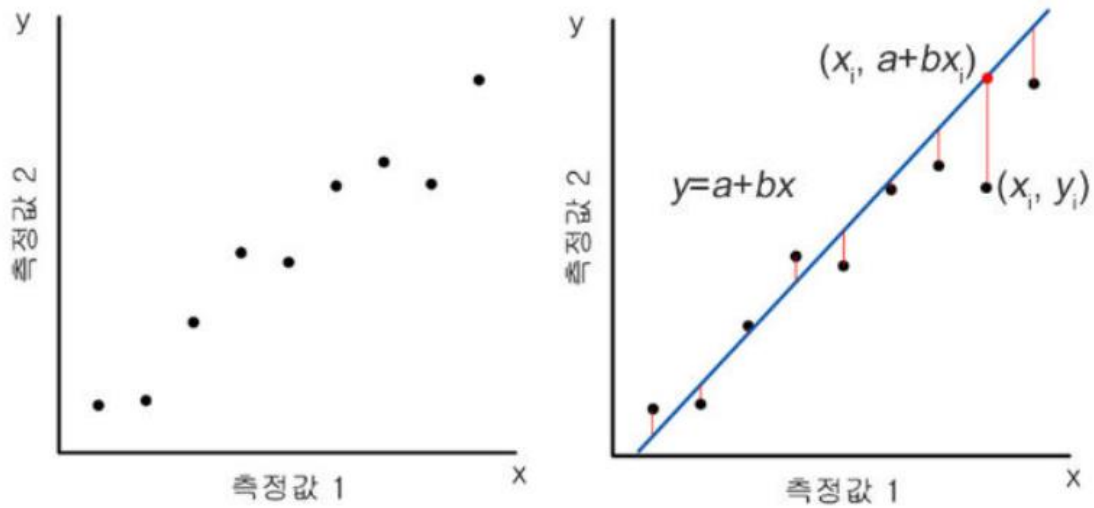
알고리즘	주 활용 분야
의사결정나무, 랜덤포레스트	분류
신경회로망	예측 분석
군집분석	독립 변수들만의 분류, 그룹화
추천-협업 필터링	아이템과 이용자 간의 상호 분석을 통한 추천
앙상블 기법	추정, 예측, 규범 등의 결합 분석
소셜 네트워크 분석	관계망 분석
서포트 벡터 머신	분류 분석
주성분 분석	원인 분석, 차원 축소

다양한 분석 기법 (AI/통계기반)

AiDA Lab.

회귀분석

- 회귀분석이란?
  - 특정 변수가 다른 변수에 어떤 영향을 미치는지를 수학적 모형으로 설명, 예측하는 기법
  - 독립변수로 종속변수를 예측하는 기법으로도 사용됨
    - 독립변수: 입력값 또는 원인을 설명하는 변수
    - 종속변수: 결과값 또는 효과를 설명하는 변수
    - 회귀계수(회귀선)
      - 독립변수가 주어질 때의 종속변수의 기대값
      - 일반적으로 최소제곱법을 이용함
    - 최소제곱법(Method of Least Squares, MLS)
      - 잔차(Residual, 관측값 y와 예측값 $\hat{y}$ 간 차이)의 제곱의 합이 최소가 되게 하는 직선을 찾는 방법



최소제곱법

회귀분석

- 회귀분석의 종류
  - 선형회귀분석 (Linear Regression)
    - 종속변수 y와 한 개 이상의 독립변수 x와의 선형 상관성을 파악하는 회귀분석 기법
    - 종속변수와 독립변수 모두 연속형 변수여야 함
    - 단순 선형회귀분석
      - 한 개의 종속변수 y와 한 개의 독립변수 x로 두 개의 변수 사이의 관계를 분석함
      - $y = ax + b$ (a: 회귀계수, b: y절편)
    - 다중 선형회귀분석
      - 독립변수가 두 개 이상이고 종속변수가 y 하나인 선형회귀분석
      - $y = ax_1 + bx_2 + \cdots + c$ (a, b, …: 회귀계수, c: y절편)

- 회귀분석의 종류
  - 선형회귀분석 (Linear Regression)
    - 선형회귀분석의 기본적인 가정
      - 선형성: 독립변수와 종속변수가 선형적이어야 한다.
      - 잔차 정규성: 잔차의 기댓값은 0 이며 정규분포를 이루어야 한다.
      - 잔차 독립성: 잔차들은 서로 독립적이어야 한다.
      - 잔차 등분산성: 잔차들의 분산이 일정해야 한다.
      - 다중 공선성: 다중 회귀분석을 수행할 경우, 3개 이상의 독립변수 간에 상관관계로 인한 문제가 없어야 한다.

회귀분석

- **회귀분석의 종류**
  - **로지스틱 회귀분석 (Logistic Regression)**
    - 종속변수가 연속형이 아닌 범주형으로 입력 데이터가 주어졌을 때 특정 분류로 결과가 나타나는 회귀분석
    - 종속변수와 독립변수 간의 관계를 함수를 통해서 예측하는 것은 선형회귀분석과 유사함
    - 단순 로지스틱 회귀분석
      - 종속변수가 이항형 문제(범주의 개수가 두 개인 경우)인 회귀분석
    - 다중 로지스틱 회귀분석
      - 종속변수가 이항형 문제가 아닌 두 개 이상의 범주를 가지는 문제인 경우의 회귀분석
      - 각 모수에 대해 비선형식이며 승산(odds)으로 로짓변환(0과 1로 조정하는 과정)을 통해 선형함수로 치환 가능
        - 승산: 임의의 사건 A가 발생하지 않을 확률 대비 일어날 확률의 비율

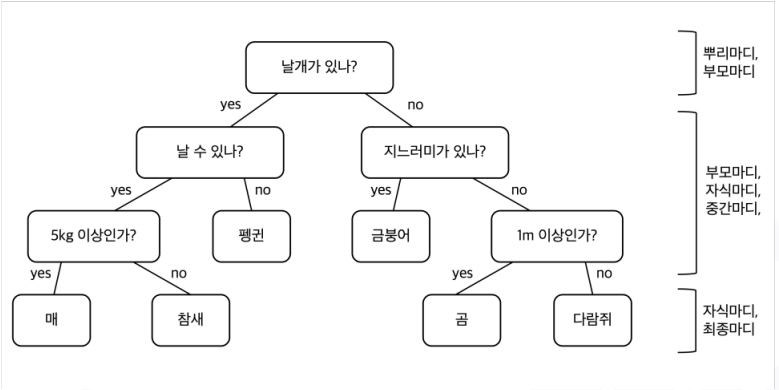
이외에도 다양한 종류의 로지스틱 회귀분석이 있음

회귀분석

- 회귀분석의 장단점
  - 장점
    - 크기와 관계없이 계수들에 대한 명료한 해석과 손쉬운 통계적 유의성 검증이 가능함
  - 단점
    - 선형적인 관계로 데이터가 구성되어 있어야 적용할 수 있음

의사결정나무 (Decision Tree)

- **의사결정나무란?**
  - **의사결정 규칙을 나무 모양으로 나타내어 전체 자료를 몇 개의 소집단으로 분류(Classification)하거나 예측(Prediction)을 수행하는 기법**
    - **상위노드로부터 하위노드로 트리구조를 형성하는 매 단계마다 분류 변수와 분류 기준값의 선택이 중요함**
    - **상위노드에서 분류된 각각의 하위노드는 노드 내 동질성이 커지고, 노드 간에는 이질성이 커지는 방향으로 분류 변수와 기준값을 선택함**
    - **모형의 크기는 과대적합(또는 과소적합) 되지 않도록 조절하여야 함**
    - **시장조사, 광고조사, 품질관리 등 다양한 분야에서 활용되고 있으며, 타겟 고객 분류, 고객 신용분류, 행동 예측 등에 사용됨**



의사결정나무 (Decision Tree)

- **의사결정나무의 구성**
  - **뿌리 마디(Root Node, 뿌리 노드):** 나무가 시작되는 마디. 부모가 없는 마디. 대상이 되는 모든 자료집합을 포함함
  - **중간 마디(Internal Node):** 뿌리 마디에서 나온 각 나무줄기의 중간에 있는 마디
  - **끝 마디(Terminal Node, 잎 노드):** 각 나무줄기의 끝에 있는 마디. 자식이 없는 마디
  - **자식 마디(Child Node):** 하나의 마디로부터 분리된 2개 이상의 마디
  - **부모 마디(Parent Node):** 자식 마디의 상위 마디
  - **가지(Branch):** 하나의 마디로부터 끝 마디까지 연결된 마디들
  - **깊이(Depth):** 가장 긴 가지의 크기(마디의 개수)

의사결정나무 (Decision Tree)

- **의사결정나무의 분석 과정**

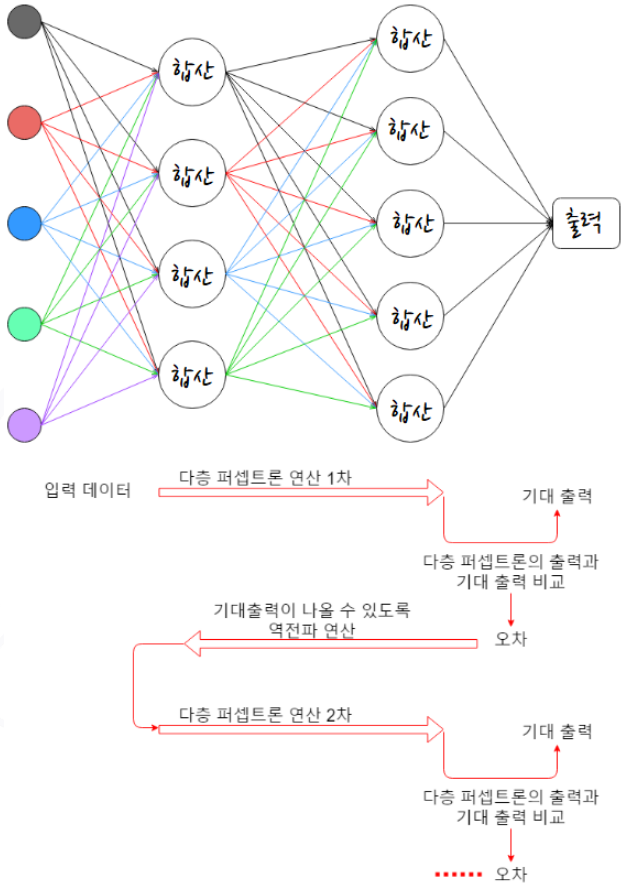
1. **변수 선택**
   - 목표변수와 관련된 설명(독립)변수들을 선택
2. **의사결정나무 형성**
   - 분석 목적에 따른 적절한 분리 기준과 정지 규칙, 평가 기준에 따라 의사결정나무를 만듦
3. **가지치기**
   - 부적절한 나뭇가지 제거
   - 과적합을 막고 일반화 성능을 높여줌
4. **모형 평가 및 예측**
   - 이익(Gain), 위험(Risk), 비용(Cost) 등을 고려하여 모형을 평가하며 분류 및 예측을 수행함

의사결정나무 (Decision Tree)

- 의사결정나무의 장단점
  - 장점
    - 연속형, 범주형 변수 모두 적용 가능
    - 변수의 비교가 가능하고 규칙에 대해 이해하기 쉬움
    - 데이터로부터 규칙을 도출하는 데에 유용함
      - DB 마케팅, CRM, 시장조사, 기업 부도/환율예측 등 다양한 분야에서 활용
  - 단점
    - 트리구조가 복잡할 시, 예측/해석력이 떨어짐
    - 데이터 변형에 민감함

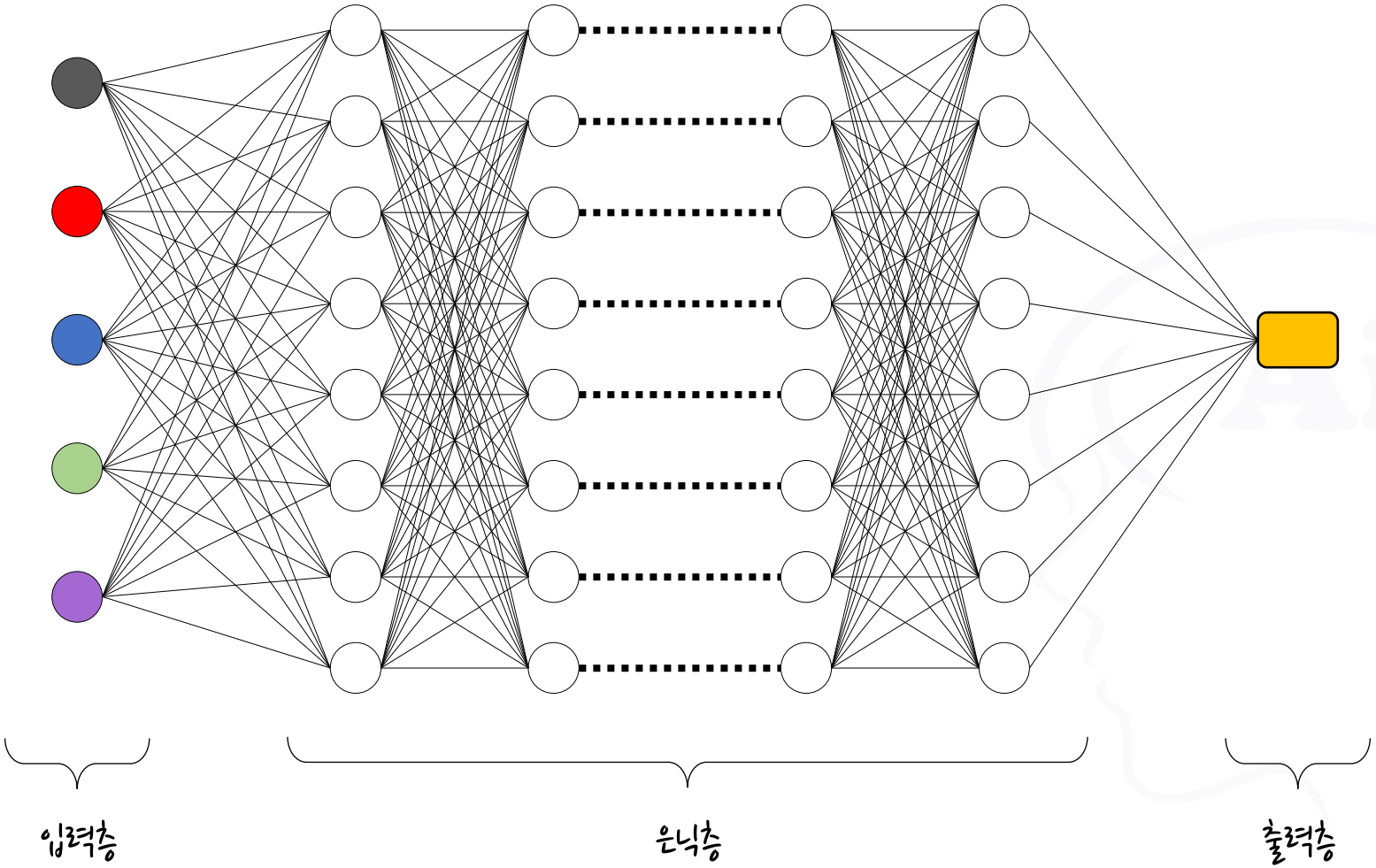
인공신경망 / 딥러닝

- 인공신경망의 구성과 구조
  - 노드(뉴런에 해당)
    - 입력값에 가중치를 곱하고 합산 후, 활성함수를 통해 다음 노드로 전달
  - 가중치(시냅스에 해당)
    - 노드와의 연결 계수
  - 활성함수
    - 임계값을 이용하여 노드의 활성화 여부를 결정함
  - 입력층, 출력층
  - 은닉층
    - 다층 네트워크에서 입력층과 출력층의 사이. 데이터를 전파 학습



인공신경망 / 딥러닝

- 딥러닝 모델의 구조



인공신경망 / 딥러닝

- **입력층**

  어떤 데이터들이 입력되는가?

  - 이미지(사진) 데이터
  - 동영상 데이터
  - 센서 데이터
  - 주식 데이터
  - 기상관측 데이터
  - 등…

  수치 데이터

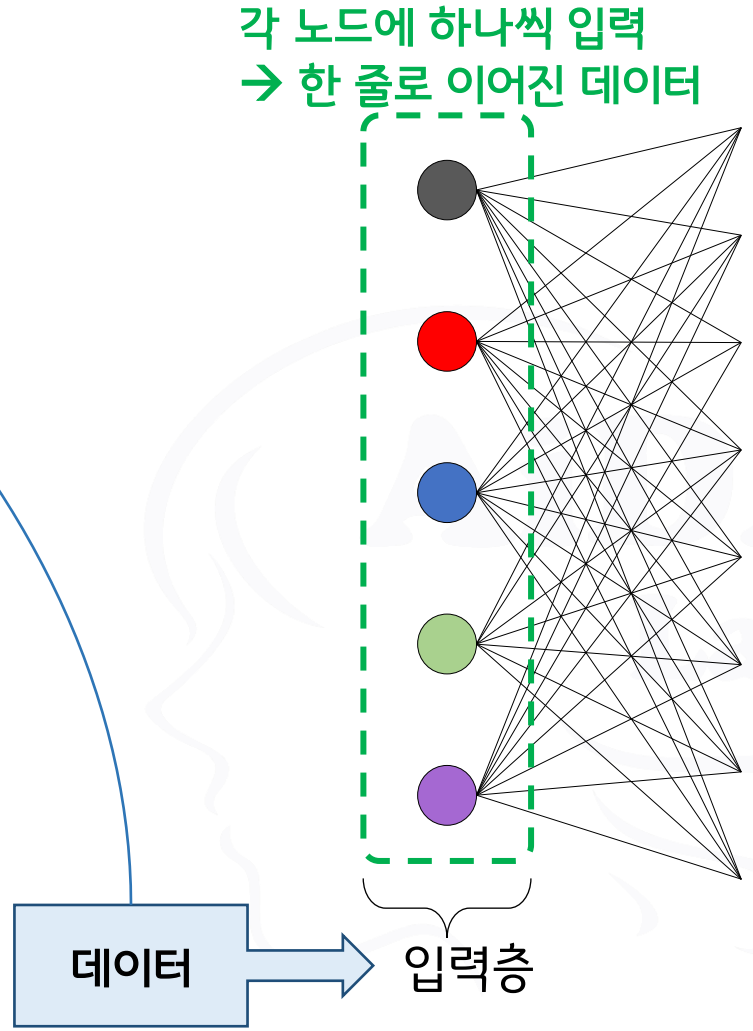
  - SNS 데이터
  - Web Scraping 데이터
  - 등…

  문자(열) 데이터 → 수치화

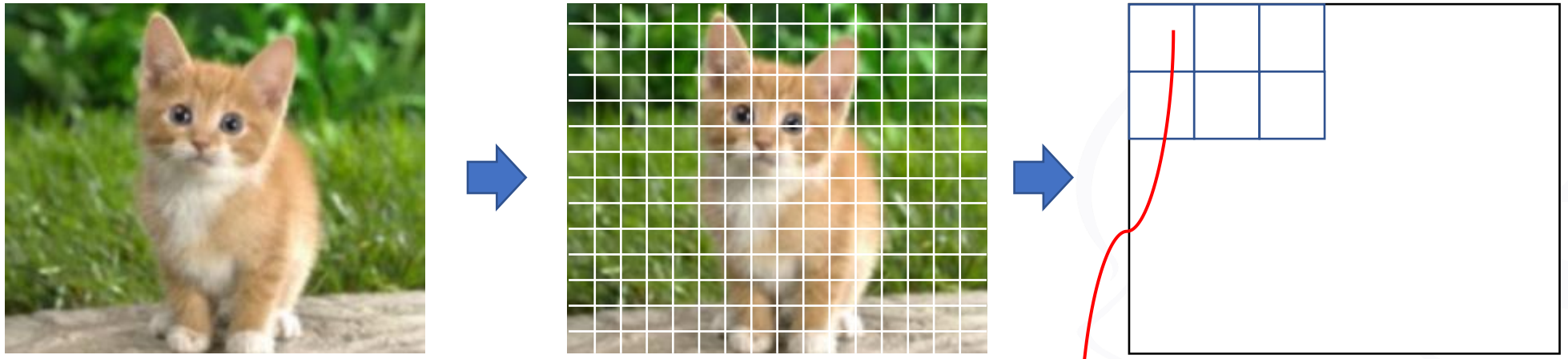
  딥러닝에서 사용되는 데이터는

  기본적으로 수치 데이터

  1차원 배열로 이루어진 수치 데이터



- **입력층**
  - **이미지 / 영상 데이터가 왜 수치 데이터인가?**
    - 이미지 데이터는 색깔을 가진 수많은 점이 가로x세로 크기의 2차원 배열 속에 모인 데이터



배열 데이터를 1차원으로 변환하여 입력 데이터로 사용

32832 = #008040 = #00 #80 #40 =
R G B

동영상 데이터는 다수의 이미지가 순서대로 연결된 것

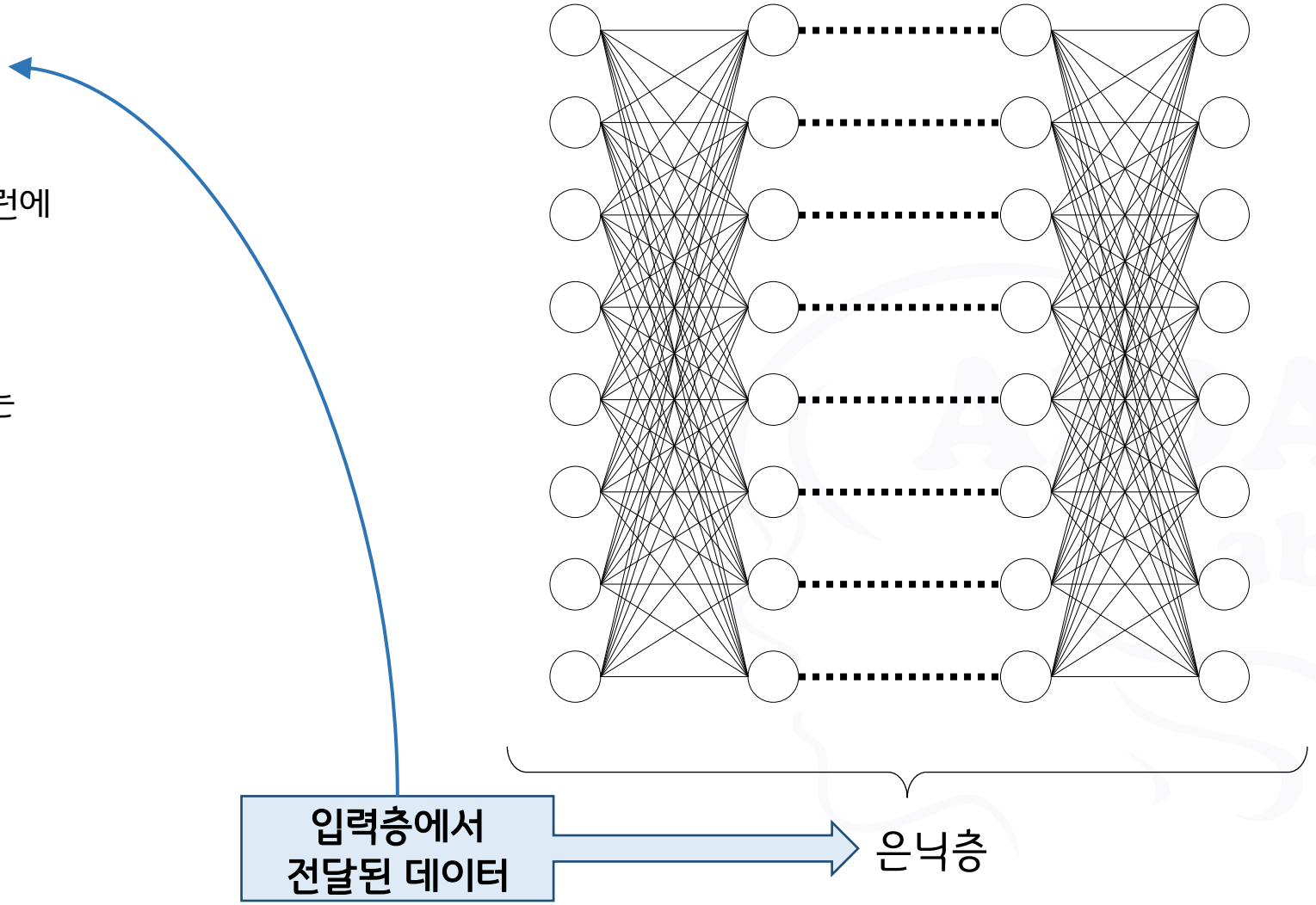
인공신경망 / 딥러닝

- **은닉층**

  데이터는 어떻게 전달되는가?

  - 입력층에서 입력된 데이터는
  - 각 데이터가 첫 번째층의 모든 뉴런에
  - 동일하게 전달됨

  - 각 층의 모든 뉴런이 가진 데이터는
  - 다음 층의 모든 뉴런에
  - 동일하게 전달됨

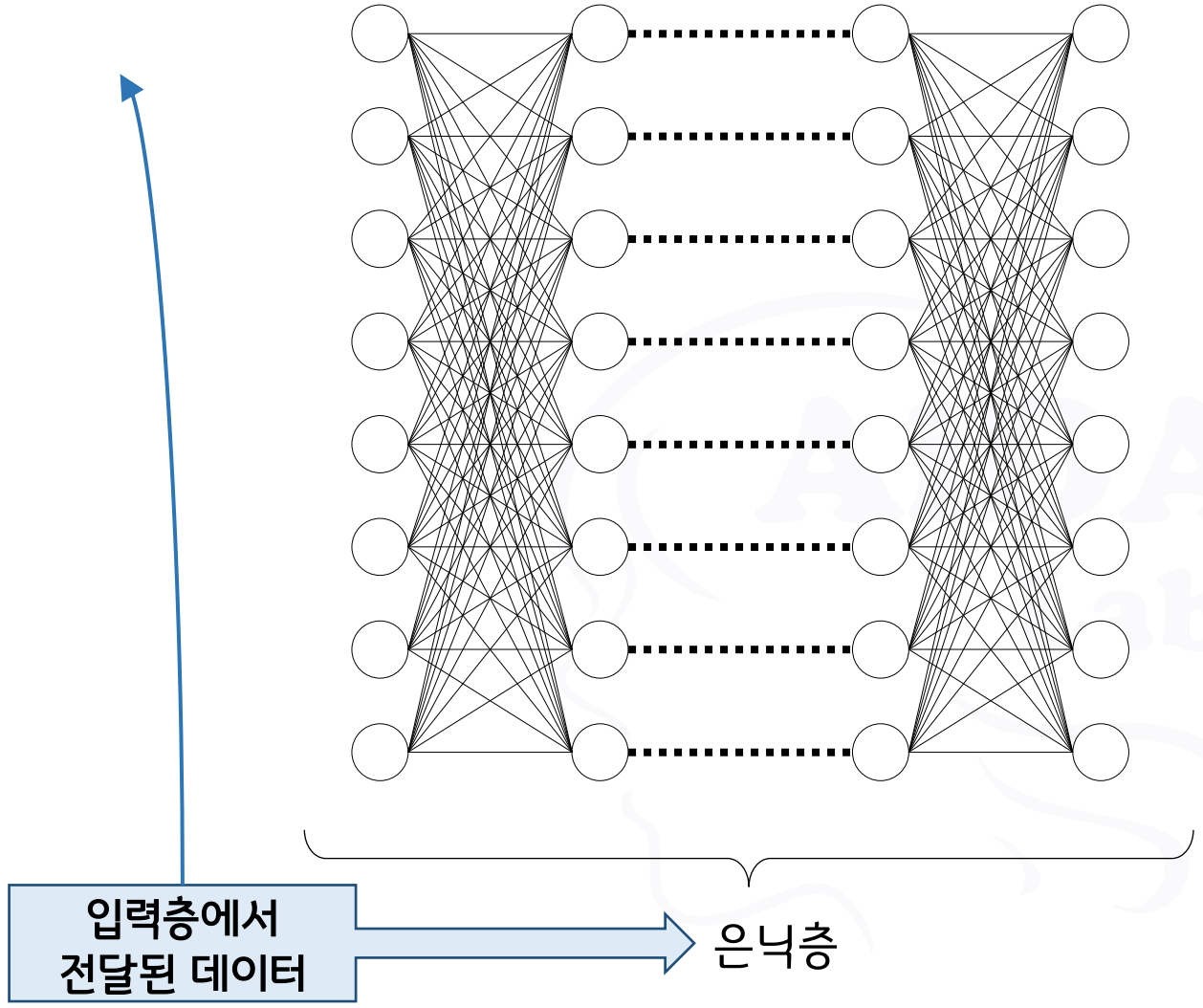


- **은닉층**

  각 뉴런에 전달된 데이터는 어떻게 가공되는가?

  - 각 층의 뉴런은
  - 입력되는 모든 데이터를
  - 다 더한 후(합산)

  - 합산 결과를 활성화 함수에 적용하고
  - 활성화 함수(F1) 적용 결과를
  - 다음 층의 뉴런으로 전달

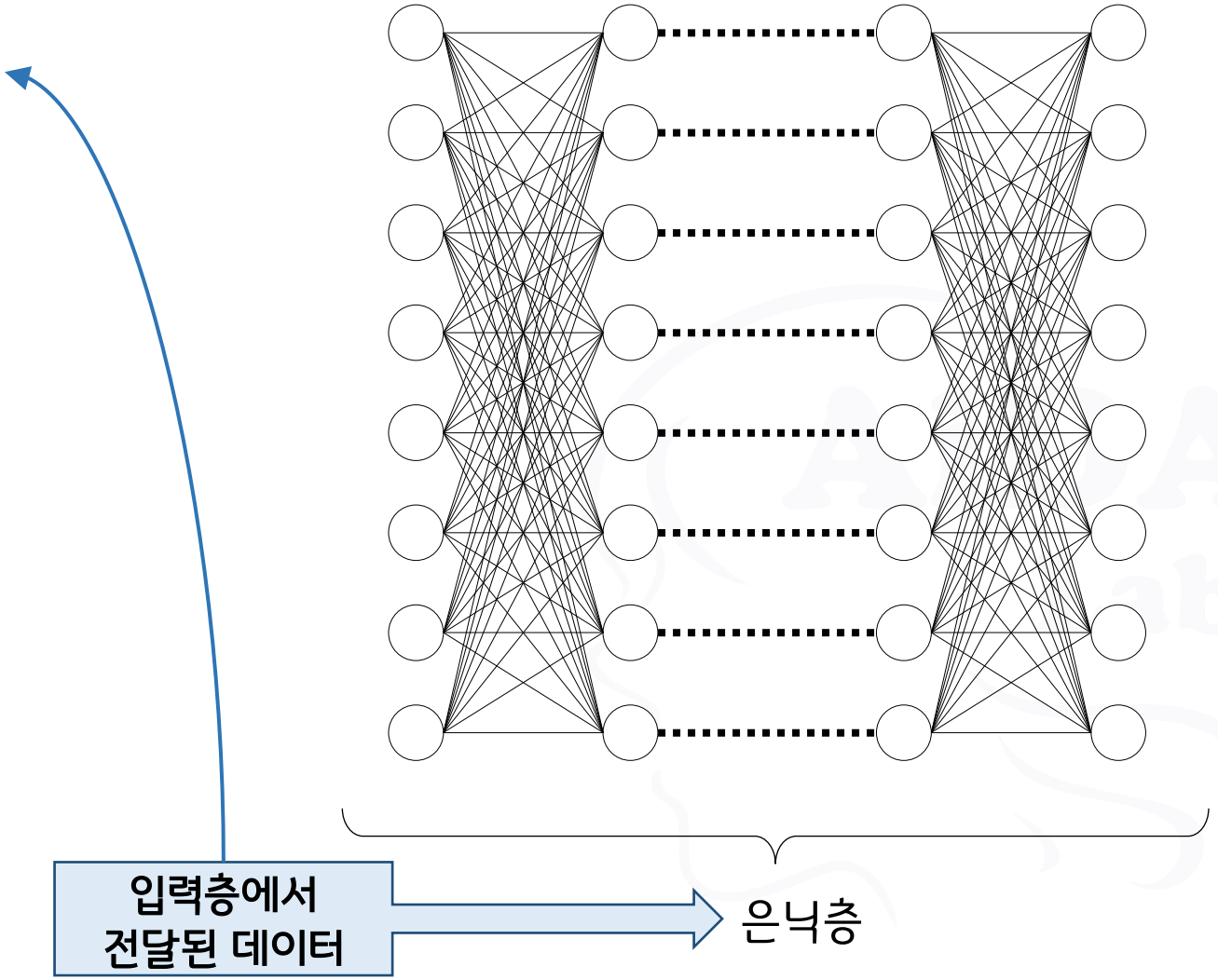


- **은닉층**

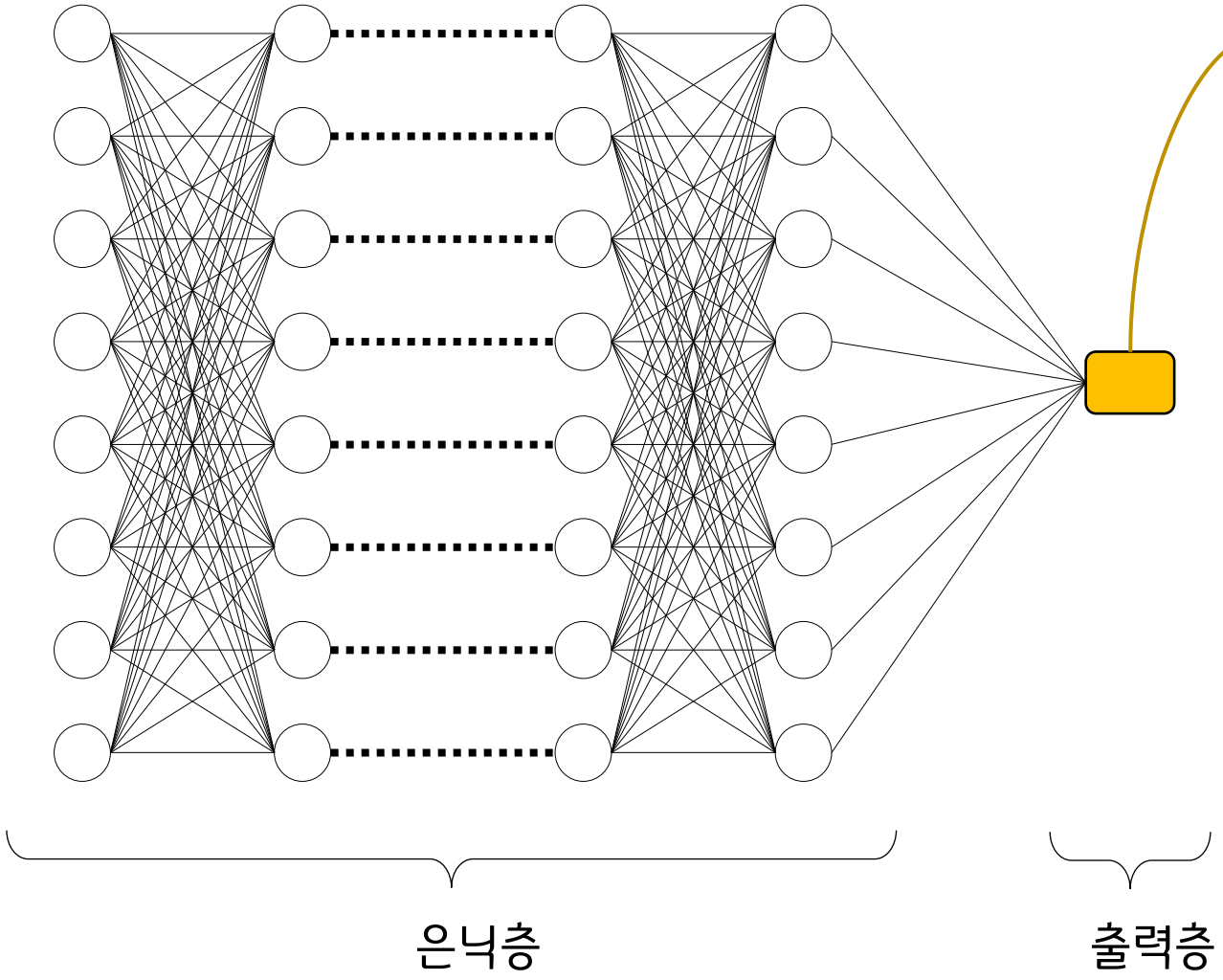
  은닉층은 데이터에 어떤 작업을 하는가?

  - 각 층에 입력된 데이터는
  - 다음 층의 뉴런으로 가는 경로에 설정된 가중치를
  - 현재의 뉴런이 가진 데이터에 곱하여
  - 그 결과를 다음 층의 뉴런으로 전달

  - 모든 층의 데이터 이동에서
  - 가중치의 곱셈과 입력 값의 합산은
  - 동일하게 적용됨



- **은닉층-출력층**



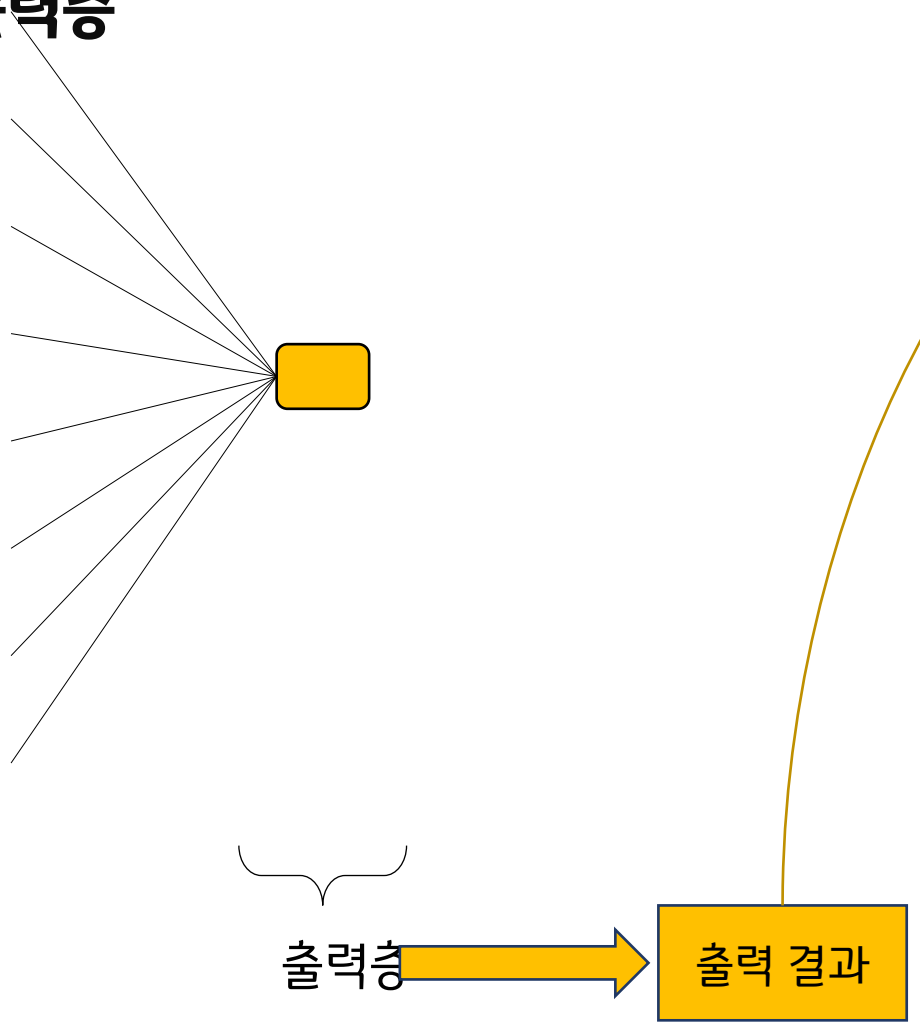
은닉층과 출력층은 함께 어떤 작업을 하는가?

- 은닉층의 마지막 층의 뉴런들은
- 자기에게 입력된 데이터를
- 합산 후
- 활성화 함수(F1)를 적용하고
- 적용 결과를
- 출력층의 출력 뉴런으로 전달

- 출력 뉴런은
- 전달된 모든 데이터를
- 순서대로 나열하여(벡터화)
- 활성화 함수(F2)를 적용하고
- 적용 결과(벡터값)를
- 결과로서 출력함

인공신경망 / 딥러닝

- **출력층**



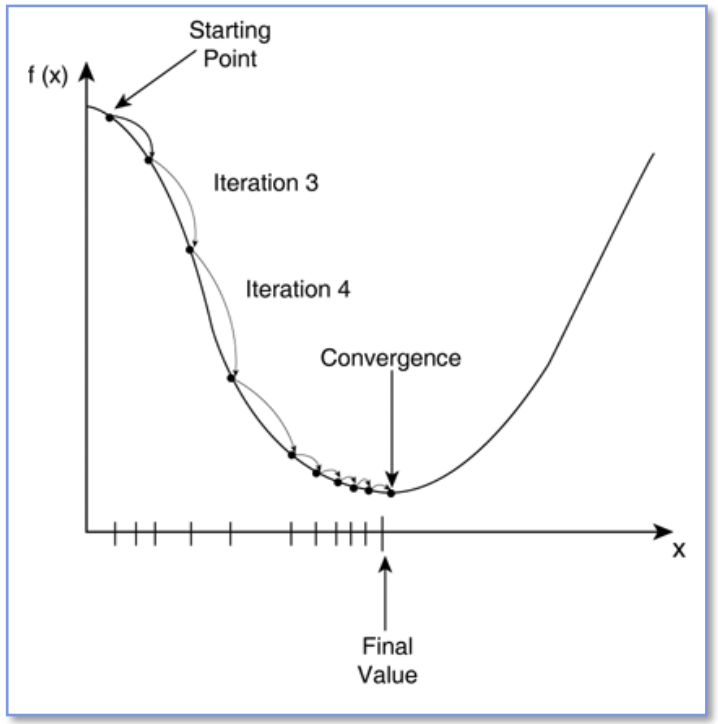
출력층에서는 어떤 결과가 나오는가?

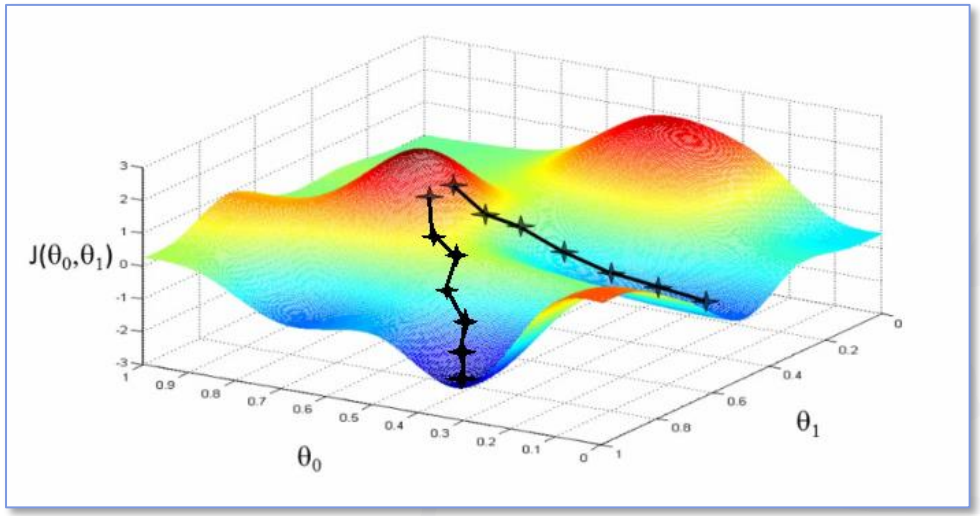
- 출력 뉴런은
- 전달된 모든 데이터를
- 순서대로 나열하여(벡터화)
- 활성화 함수(F2)를 적용하고
- 적용 결과(벡터값)를
- 결과로서 출력함

- 출력층의 결과는
- 미리 입력된 기대 결과값과 비교하여(지도학습)
- 일치하면 → 학습 완료
- 일치하지 않으면 → 진행의 반대 방향으로 다시 전달

- 가중치의 조정
  - 가장 많이, 기본적으로 사용되는 가중치 조정 방법
    - 경사 하강법 (Gradient Descent)

초기값에 따라
최저점이
달라질 수 있다



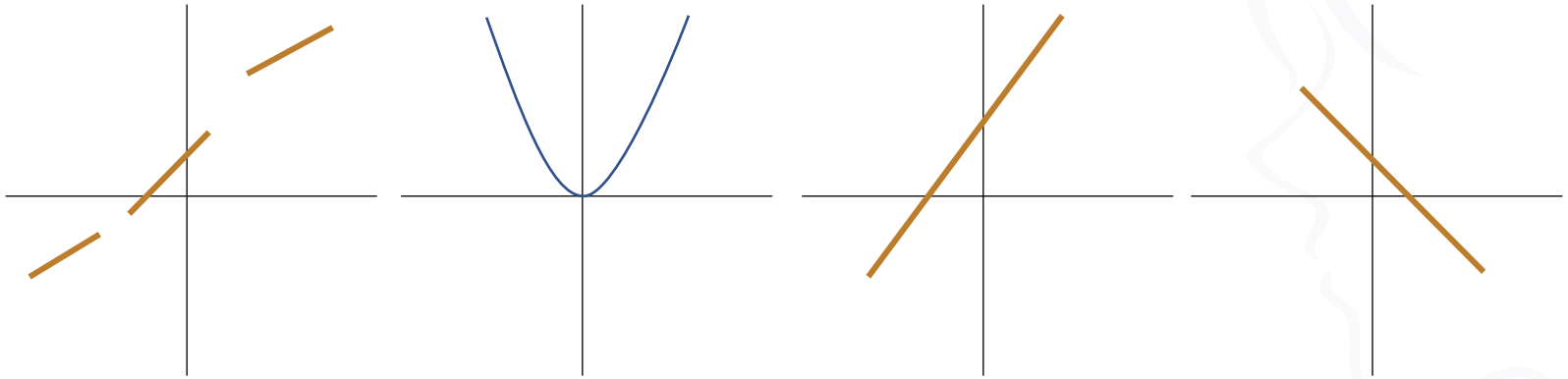


최저점은 하나가 아니다

이런 이유로...
같은 목적과 같은 데이터로 학습하더라도
모든 학습된 모델의 내부(가중치 집합)는 모두 다름

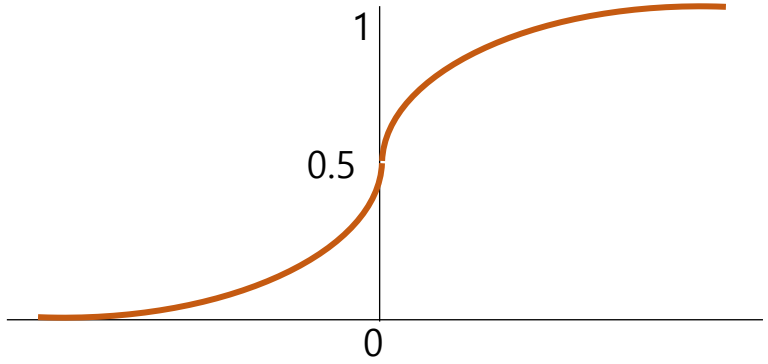
- **활성화 함수 (Activation Function)**
  - 신경망을 구성하는 각 퍼셉트론에서 임계 값을 넘었을 때 출력을 처리하는 함수
  - 함수의 정의
    - 입력: 이전 층의 디바이스 또는 퍼셉트론들로부터 전달되는 데이터
    - 함수의 동작: 입력 값의 합산 + 합산결과와 임계 값의 비교 + 출력 결정
      (활성화 조건)
    - 출력: 퍼셉트론 층의 연산 결과값. 다음 층의 뉴런에 대한 입력 또는 최종 층의 출력

- 활성화 함수의 조건
  - 정의역(함수에 입력 가능한 값의 범위, 집합) 안에서 연속이며 무한해야 한다
  - 단조 함수여야 한다 (방향을 바꾸지 않아야 한다)
  - 비선형 함수여야 한다.
  - 계산 효율이 좋아야 한다.



이런 함수는 부적합

- 표준 은닉 활성화 함수
  - Sigmoid

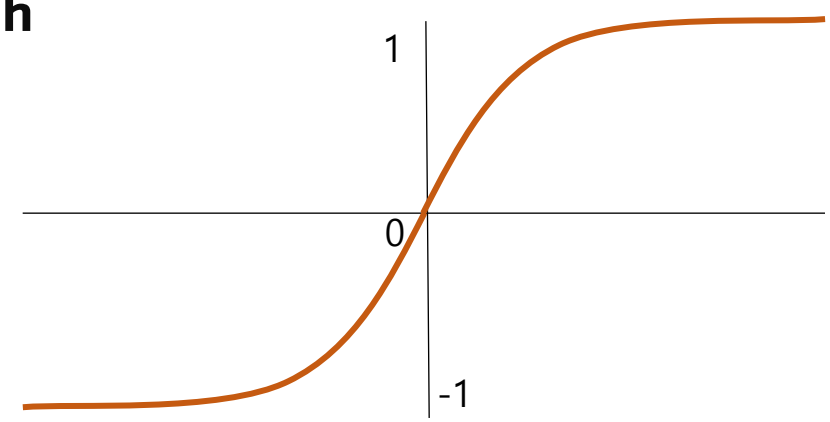


가장 많이 사용되어 왔고

가장 중요한 활성화 함수

(활성화 함수의 기본 형태)

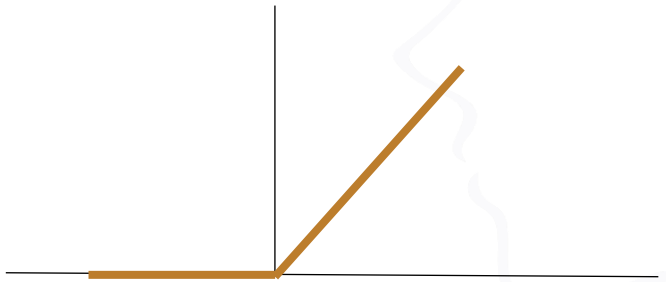
  - tanh



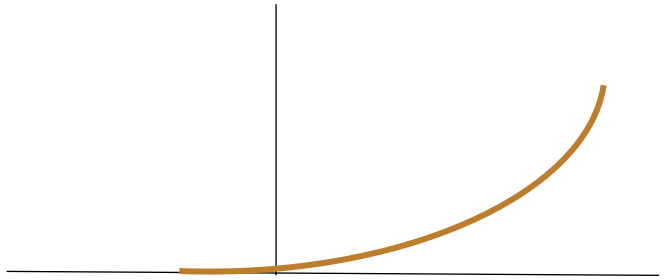
은닉층에서는 sigmoid 보다 tanh 함수가 더 좋음

음의 상관관계도 지원

- 표준 출력 계층 활성화 함수
  - 신경망의 목적에 따라 최선의 선택이 달라진다
    - 일반 데이터값 예측 → 활성화 함수 미적용
    - 서로 무관한 항목에 대한 예/아니오 확률 예측 → sigmoid
    - 여러 가능성 중 하나의 확률 예측 → softmax
  - 최근 가장 많이 쓰이는 활성화 함수
    - Relu (Rectified Linear Unit, 정류된 선형 유닛)



- **표준 출력 계층 활성화 함수**
  - **Softmax**



- k개의 값이 존재할 때 각각의 값의 편차를 확대시켜 큰 값은 상대적으로 더 크게, 작은 값은 상대적으로 더 작게 만든 후, 정규화 시키는 함수
- Softmax 함수를 거친 k개의 값의 총합은 1이 됨
- 지수증가를 기반으로 하는 함수

- 과적합(Over Fitting)
  - 주어진 데이터로 학습을 너무 많이 하면 오히려 역효과!!
    - 학습에 입력된 데이터는 완벽에 가깝게 처리함
    - 학습에 입력되지 않은 데이터는 제대로 처리되지 않음
  - 원인: 잡음 데이터 (대부분)
    - 불필요한 정보가 많이 포함된 데이터로 학습이 반복됨에 따라 불필요한 정보가 분류의 기준에 포함되어 버리는 것이 원인

- **과적합의 해결방안**
  - **조기 종료:** 적당한 선에서 학습을 종료시킴 → 데이터의 정규화와 관련
  - **정규화 (데이터를 일반화 시키기)**
    - 필요한 신호는 학습하고 잡음은 제거하는 효과
    - 모델의 학습 난이도를 높임으로써 학습 데이터의 세부 사항(잡음 포함)에 대한 일반화를 활용하도록 하는 기법의 일부로 사용됨
  - **Drop Out**
    - 학습 중에 무작위로 선택한 뉴런을 0으로 설정
      → 군데군데 망의 연결고리를 잘라 내어 대형 신경망이 소형 신경망처럼 동작하게 만듦
    - 소형 신경망에서는 과적합이 거의 발생하지 않음 (표현능력이 협소하기 때문)
    - 대형 신경망(딥러닝 모델)을 Drop Out을 통해 소형 신경망처럼 동작하게 하여 과적합 발생률을 떨어뜨리는 방법