머신러닝 기반 데이터 분석

부산IT교육센터 박경미

과목 전체 목제

1. **머신러닝과 딥러닝 소개**
    1. 인공지능, 머신러닝, 딥러닝
    2. 머신러닝이란?
    3. 딥러닝이란?
2. **PyTorch Basic**
    1. 파이토치 개요
    2. 환경설정
    3. 파이토치 기초 문법
3. **선형회귀분석(Linner Regression)**
    1. 선형회귀
    2. 자동 미분
    3. 다중선형 회귀
    4. nn.Model로 선형 회귀 구현
    5. 클래스로 선형회귀 모델 구현하기
4. **로지스틱 회귀(Logistic Regression)**
    1. 로지스틱 회귀?
    2. 로지스틱 회귀 손실함수
    3. 로지스틱 회귀 수식
5. **인공신경망(Aritficial Neural Network)**
    1. 딥러닝 이해와 퍼셉트론(Perceptron)
    2. XOR 문제
    3. 역전파 알고리즘
    4. 그래디언트 손실 문제와 렐루(ReLU)
6. **CNN(Convolution and Pooling)**
    1. 합섭곱과 풀링
    2. 패턴 추출의 원리
7. **최적화와 오버피팅 방지, 정규화**
    1. 최적화를 위한 하이퍼 파라메타, 학습률 사용
    2. 오버피팅 방지를 위한 검증 데이터 셋, 테스트 데이터 셋 사용
    3. 가중치 감쇠, 데이터 증강, 드롭 아웃, 배치 정규화
8. **RNN(Recurrent Neural Network)**
    1. 순환신경망 사례
    2. LSTM
9. **PyTorch 모델 앱 배포**

1장. 머시러닝과 딥러닝 개요

1. 인공지능, 머신러닝, 딥러닝
2. 머신러닝이란?
3. 딥러닝이란?

1. 인공지능, 머신러닝, 딥러닝

❖ 인공지능, 머신러닝, 딥러닝

- 인공지능 (Artificial Intelligence)

  인간의 학습능력, 추론능력 등을 컴퓨터를 통해 구현하는 포괄적인 개념

- 머신러닝 (Machine Learning)

  데이터를 이용하여 명시적으로 정의되지 않은 패턴을 학습하여 미래 결과(값, 분포)를 예측

  ※ 데이터마이닝(Data Mining): 데이터간의 상관관계나 속성을 찾는 것이 주목적

- 딥러닝 (Deep Learning)

  머신러닝의 한 분야로서 신경망(Neural Network)을 통하여 학습하는 알고리즘의 집합

1. 인공지능, 머신러닝, 딥러닝

❖ 인공지능 (Artificial Intelligence, AI)

- **인공지능 :** 인간의 지능을 모방하여 사람이 하는 일을 컴퓨터(기계)가 할 수 있도록 하는 기술
- 인공지능을 구현하는 방법 : 머신 러닝(machine learning)과 딥러닝(deep learning)이 있음
- 인공지능과 머신 러닝, 딥러닝의 관계는 다음과 같음

인공지능 > 머신 러닝 > 딥러닝

- 목적과 주어진 환경에 맞게 데이터를 분석하려면 머신 러닝과 딥러닝 차이를 명확하게 이해해야 함
- 머신 러닝과 딥러닝 모두 학습 모델을 제공하여 데이터를 분류할 수 있는 기술
- 둘은 접근 방식에 차이가 있음

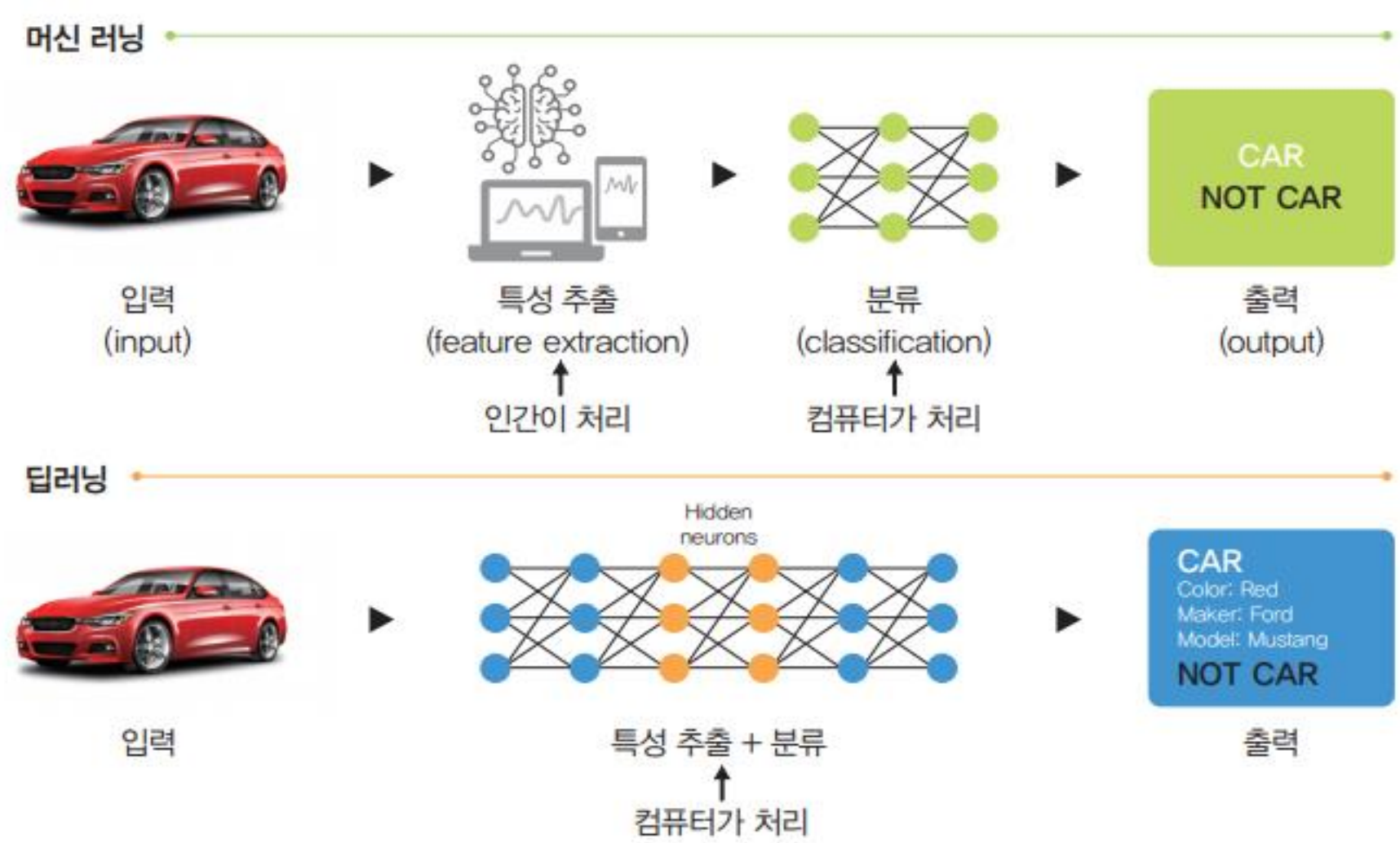
1. 인공지능, 머신러닝과 딥러닝

❖ 머신러닝

- **머신 러닝** : 주어진 데이터를 인간이 먼저 처리(전처리)
- 이미지 데이터라면 사람이 학습(train) 데이터를 컴퓨터가 인식할 수 있도록 준비해 두어야 함
- 머신 러닝은 범용적인 목적을 위해 제작된 것으로 데이터의 특징을 스스로 추출하지 못함, 이 과정을 인간이 처리해 주어야 하는 것이 머신 러닝
- 즉, 머신 러닝의 학습 과정은 각 데이터(혹은 이미지) 특성을 컴퓨터(기계)에 인식시키고 학습시켜 문제를 해결
- 딥러닝은 인간이 하던 작업을 생략, 대량의 데이터를 신경망에 적용하면 컴퓨터가 스스로 분석한 후 답을 찾음

1. 인공지능, 머신러닝, 딥러닝

▼ 그림 1-2 머신 러닝과 딥러닝 차이



1. 인공지능, 머신러닝, 딥러닝

▼ 표 1-1 머신 러닝과 딥러닝

구분	머신 러닝	딥러닝
동작 원리	입력 데이터에 알고리즘을 적용하여 예측을 수행한다.	정보를 전달하는 신경망을 사용하여 데이터 특징 및 관계를 해석한다.
재사용	입력 데이터를 분석하기 위해 다양한 알고리즘을 사용하며, 동일한 유형의 데이터 분석을 위한 재사용은 불가능하다.	구현된 알고리즘은 동일한 유형의 데이터를 분석하는 데 재사용된다.
데이터	일반적으로 수천 개의 데이터가 필요하다.	수백만 개 이상의 데이터가 필요하다.
훈련 시간	단시간	장시간
결과	일반적으로 점수 또는 분류 등 숫자 값	출력은 점수, 텍스트, 소리 등 어떤 것이든 가능

2. 머신 러닝이란?

❖ 머신 러닝이란?

- 머신 러닝은 인공지능의 한 분야로, 컴퓨터 스스로 대용량 데이터에서 지식이나 패턴을 찾아 학습하고 예측을 수행하는 것
- 즉, 컴퓨터가 학습할 수 있게 하는 알고리즘과 기술을 개발하는 분야라고 할 수 있음

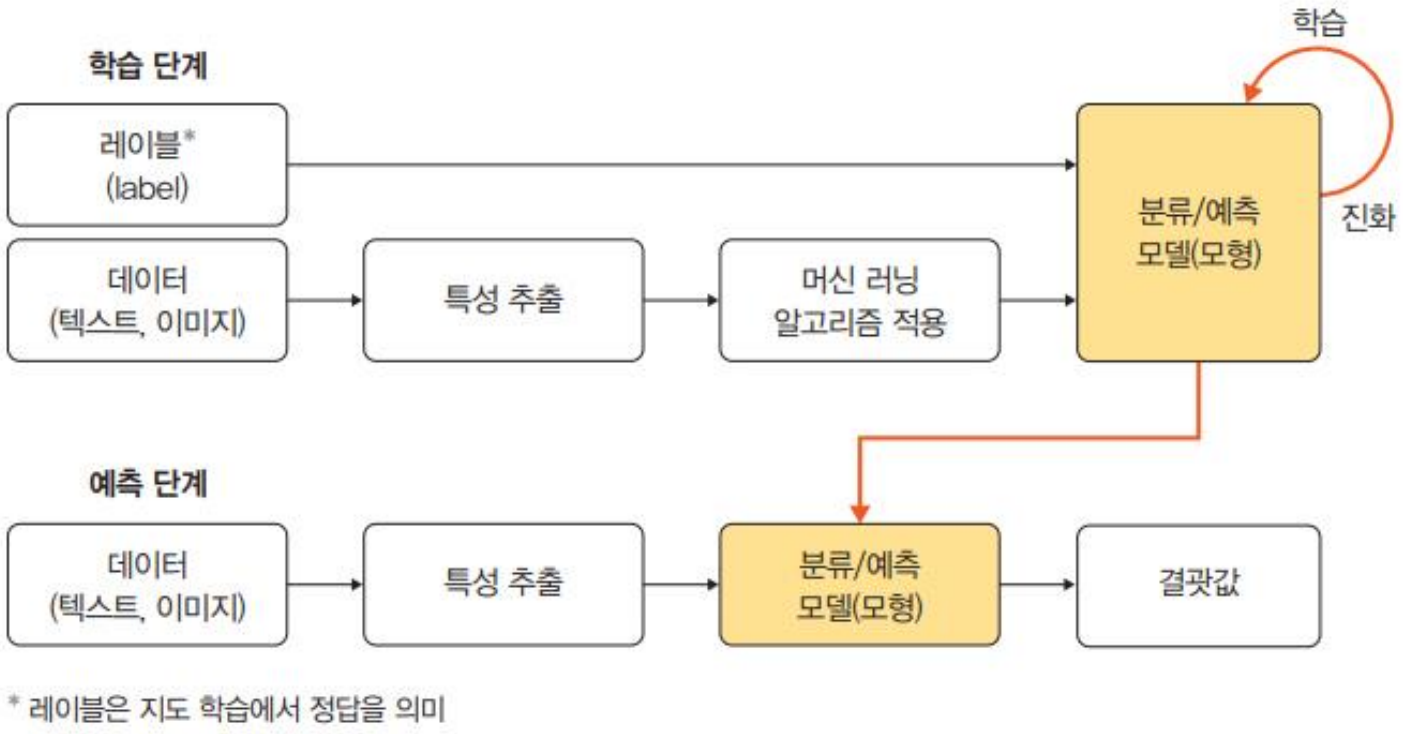
❖ 머신러닝의 정의

- 머신러닝은 데이터에서부터 학습하도록 컴퓨터 프로그래밍하는 과학(또는 예술)
- "머신러닝은 명시적인 프로그래밍 없이 컴퓨터가 학습하는 능력을 갖추게 하는 연구 분야" - 아서 새뮤얼(Arthur Samuel), 1959
- "어떤 작업 T에 대한 컴퓨터 프로그램의 성능을 P로 측정했을 때 경험 E로 인해 성능이 향상됐다면, 이 컴퓨터 프로그램은 작업 T와 성능 측정 P에 대해 경험 E로 학습한 것" - 톰 미첼(Tom Mitchell), 1997

2. 머신 러닝이란?

❖ 머신 러닝 학습 과정

- 학습 단계(learning)와 예측 단계(prediction)로 구성
- **학습 단계** : 훈련 데이터를 머신 러닝 알고리즘에 적용하여 학습시키고, 이 학습 결과로 모델이 생성
- **예측 단계** : 학습 단계에서 생성된 모형에 새로운 데이터를 적용하여 결과를 예측



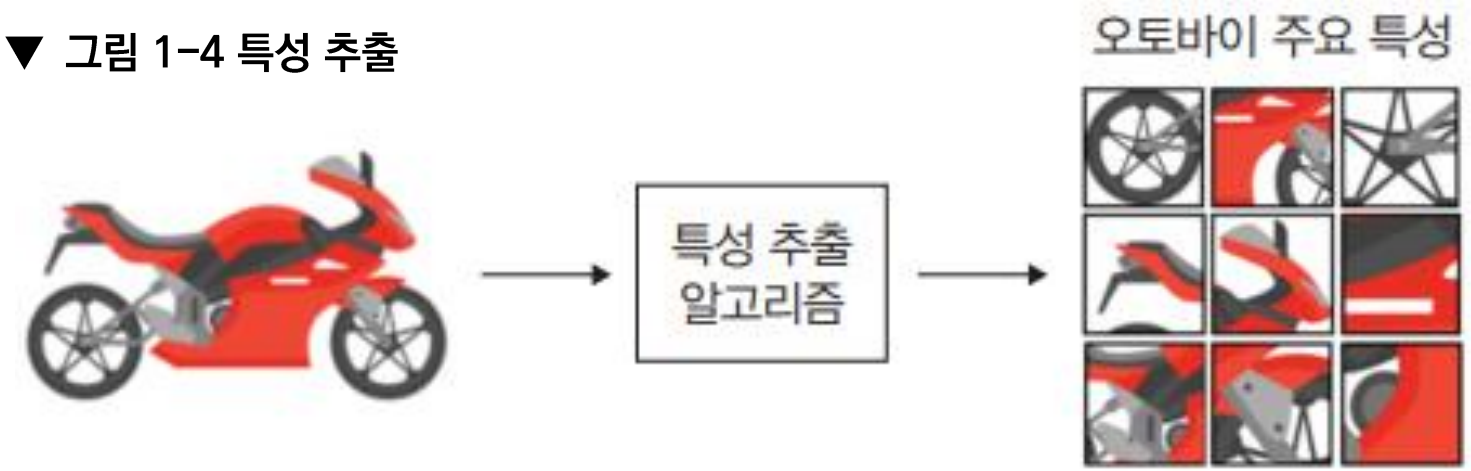
◀ 그림 1-3 머신 러닝 학습 과정

2. 머신 러닝이란?

❖ 머신 러닝 학습 과정

- 특성 추출
  - 머신 러닝에서 컴퓨터가 스스로 학습하려면, 즉 컴퓨터가 입력받은 데이터를 분석하여 일정한 패턴이나 규칙을 찾아내려면 사람이 인지하는 데이터를 컴퓨터가 인지할 수 있는 데이터로 변환해 주어야 함
  - 이때 데이터별로 어떤 특징을 가지고 있는지 찾아내고, 그것을 토대로 데이터를 벡터로 변환하는 작업을 특성 추출(feature extraction)이라고 함

▼ 그림 1-4 특성 추출



2. 머신 러닝이란?

❖ 머신 러닝 학습 과정

- 머신 러닝의 주요 구성 요소 : **데이터**와 모델(모형)
- **데이터** : 머신 러닝이 학습 모델을 만드는 데 사용하는 것
- 훈련 데이터가 나쁘다면 실제 현상의 특성을 제대로 반영할 수 없으므로 실제 데이터의 특징이 잘 반영되고 편향되지 않는 훈련 데이터를 확보하는 것이 중요
- 학습에 필요한 데이터가 수집되었다면 '**훈련 데이터셋**' 과 '**테스트 데이터셋**' 용도로 분리해서 사용
- '**훈련 데이터셋**'을 또 다시 '**훈련 데이터셋(train dataset)**' 과 '**검증 데이터셋(validation dataset)**'으로 분리해서 사용
- 보통 데이터의 70~80%는 훈련용으로, 20~30%는 테스트용으로 분리해서 사용

2. 머신 러닝이란?

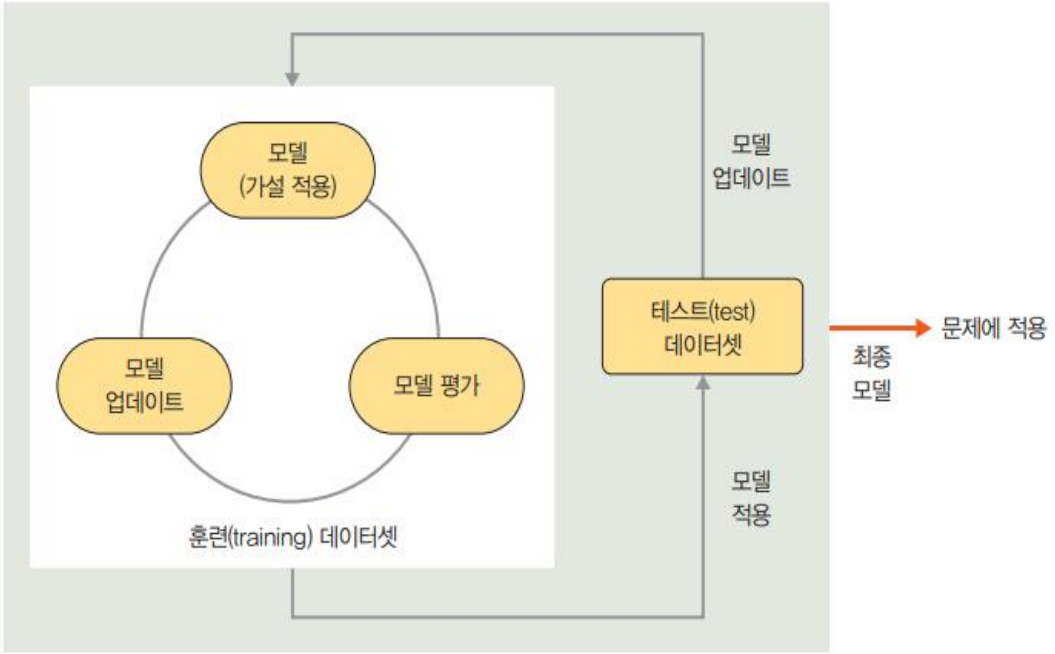
❖ 머신 러닝 학습 과정

- 모델 : 머신 러닝의 학습 단계에서 얻은 최종 결과물로 가설이라고도 함
- 예를 들어 “입력 데이터의 패턴은 A와 같다.”라는 가정을 머신 러닝에서는 모델이라고 함
- 모델의 학습 절차는 다음과 같음

1. 모델(또는 가설) 선택
2. 모델 학습 및 평가
3. 평가를 바탕으로 모델 업데이트

- 이 세 단계를 반복하면서 주어진 문제를 가장 잘 풀 수 있는 모델을 찾음
- 최종적으로 완성된 모델(모형)을 해결하고자 하는 문제에 적용해서 분류 및 예측 결과를 도출

▼ 그림 1-5 머신 러닝의 문제 풀이 과정



2. 머신 러닝이란?

❖ 머신 러닝 학습 과정

- 훈련, 검증, 테스트 데이터 셋
  - 수집된 데이터 셋은 크게 **훈련(training), 테스트(test) 데이터셋**으로 분리하여 사용
  - **훈련 데이터 셋**을 다시 **훈련과 검증(validation)** 용도로 분리해서 사용하는 경우를 볼 수 있는데 이들 간의 차이를 알아보자

▼ 그림 1-6 훈련과 검증, 테스트 데이터셋

2. 머신 러닝이란?

❖ 머신 러닝 학습 과정

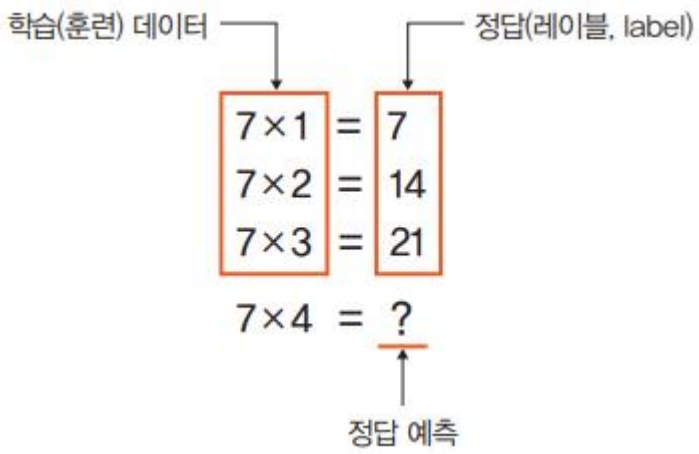
- 모델 성능의 평가는 왜 필요할까?
- 첫 번째는 테스트 데이터 셋에 대한 성능을 가늠해 볼 수 있기 때문임
- 딥러닝의 목적은 새롭게 수집될 데이터에 대해 정확한 예측을 하는 데 있음
- 이때 검증 데이터 셋을 사용해서 새롭게 수집될 데이터에 대해 예측을 평가해 볼 수 있음
- 두 번째는 모델 성능을 높이는 데 도움을 줌
- **예** : 훈련 데이터 셋에 대한 정확도는 높은데 검증 데이터 셋에 대한 정확도가 낮다면 훈련 데이터 셋에 과적합이 일어났을 가능성을 생각해 볼 수 있음
- 이 경우 정규화(regularization)를 하거나 에포크(epoch)를 줄이는 방식으로 과적합을 막을 수 있음

2. 머신 러닝이란?

❖ 머신 러닝 학습 알고리즘

- 머신 러닝의 학습 알고리즘의 종류 : 지도 학습, 비지도 학습, 강화 학습이 있음
- 지도 학습 : 이름에서 알 수 있듯이 정답이 무엇인지 컴퓨터에 알려 주고 학습시키는 방법

▼ 그림 1-7 지도 학습

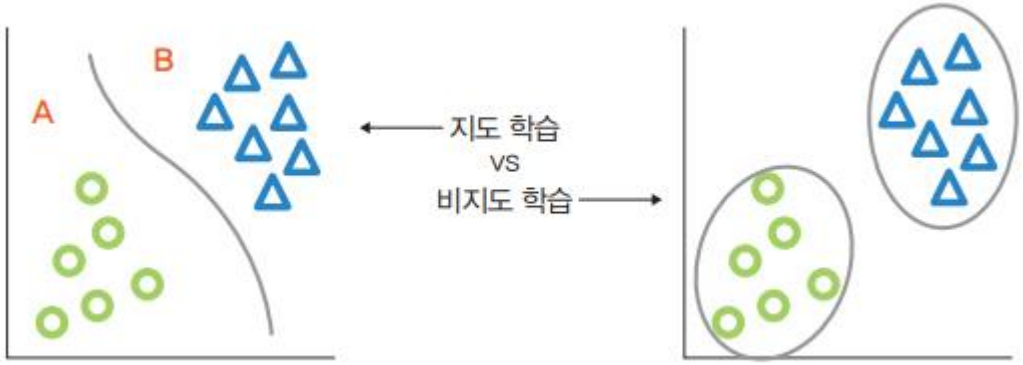


2. 머신 러닝이란?

❖ 머신 러닝 학습 알고리즘

- 비지도 학습 : 정답을 알려 주지 않고 특징(다리 길이가 짧은 초식 동물)이 비슷한 데이터(토끼, 다람쥐)를 클러스터링(범주화)하여 예측하는 학습 방법
- 즉, 다음 그림과 같이 지도 학습은 주어진 데이터에 대해 A 혹은 B로 명확한 분류가 가능
- 비지도 학습은 유사도 기반(데이터 간 거리 측정)으로 특징이 유사한 데이터끼리 클러스터링으로 묶어서 분류

▼ 그림 1-8 지도 학습과 비지도 학습



2. 머신 러닝이란?

❖ 머신 러닝 학습 알고리즘

- **강화 학습** : 머신 러닝의 꽃이라고 부를 만큼 어렵고 복잡함
- 분류할 수 있는 데이터가 있는 것도 아니고 데이터가 있다고 해도 정답이 없기 때문임
- 강화 학습은 **자신의 행동에 대한 보상을 받으며 학습을 진행**
- 게임이 대표적인 사례
- 예: <쿠키런>국내 게임
- 쿠키가 에이전트(agent)이며(즉, 게이머가 에이전트가 되겠죠?) 게임 배경이 환경(environment)
- 이때 에이전트가 변화하는 환경에 따라 다른 행동(action)을 취하게 됨
- 동전이나 젤리를 취득하는 등 행동에 따라 보상(몸집이 커짐)을 얻음
- **강화 학습**은 이러한 **보상이 커지는 행동은 자주** 하도록 하고, **줄어드는 행동은 덜** 하도록 하여 **학습을 진행**

2. 머신 러닝이란?

▼ 그림 1-9 강화 학습(〈쿠키런〉 게임)
(출처: https://www.devsisters.com/ko/product/games/)

2. 머신 러닝이란?

▼ 표 1-2 지도 학습, 비지도 학습, 강화 학습

구분	유형	알고리즘
지도 학습 (supervised learning)	분류(classification)	• K-최근접 이웃(K-Nearest Neighbor, KNN) • 서포트 벡터 머신(Support Vector Machine, SVM) • 결정 트리(decision tree) • 로지스틱 회귀(logistic regression)
	회귀(regression)	선형 회귀(linear regression)
비지도 학습 (unsupervised learning)	군집(clustering)	• K-평균 군집화(K-means clustering) • 밀도 기반 군집 분석(DBSCAN)
	차원 축소 (dimensionality reduction)	주성분 분석 (Principal Component Analysis, PCA)
강화 학습 (reinforcement learning)	-	마르코프 결정 과정 (Markov Decision Process, MDP)

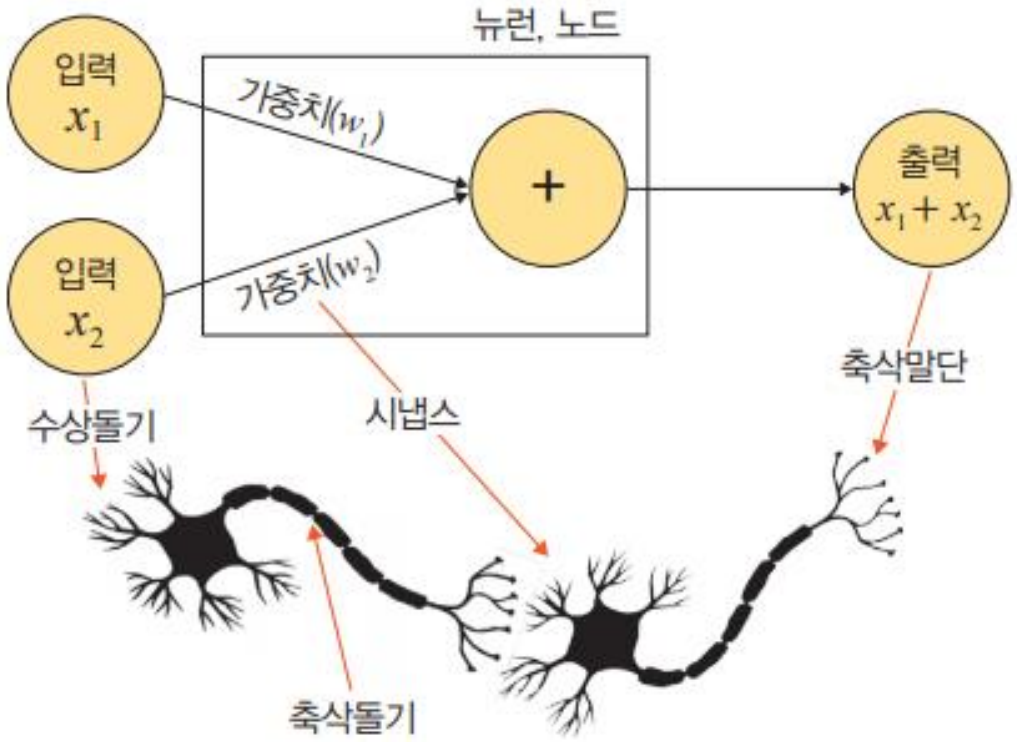
3. 딥러닝이란?

❖ 딥러닝

- **딥러닝** : 인간의 신경망 원리를 모방한 **심층 신경망 이론**을 기반으로 고안된 머신 러닝 방법의 일종
- 즉, 딥러닝이 머신 러닝과 다른 큰 차이점은 인간의 뇌를 기초로 하여 설계했다는 것
- 인간의 뇌가 엄청난 수의 뉴런(neuron)과 시냅스(synapse)로 구성되어 있는 것에 착안하여 컴퓨터에 뉴런과 시냅스 개념을 적용
- 각각의 뉴런은 복잡하게 연결된 수많은 뉴런을 병렬 연산하여 기존에 컴퓨터가 수행하지 못했던 음성.영상 인식 등 처리를 가능하게 함

3. 딥러닝이란?

▼ 그림 1-10 인간의 신경망 원리를 모방한 심층 신경망



❖ 딥러닝이란

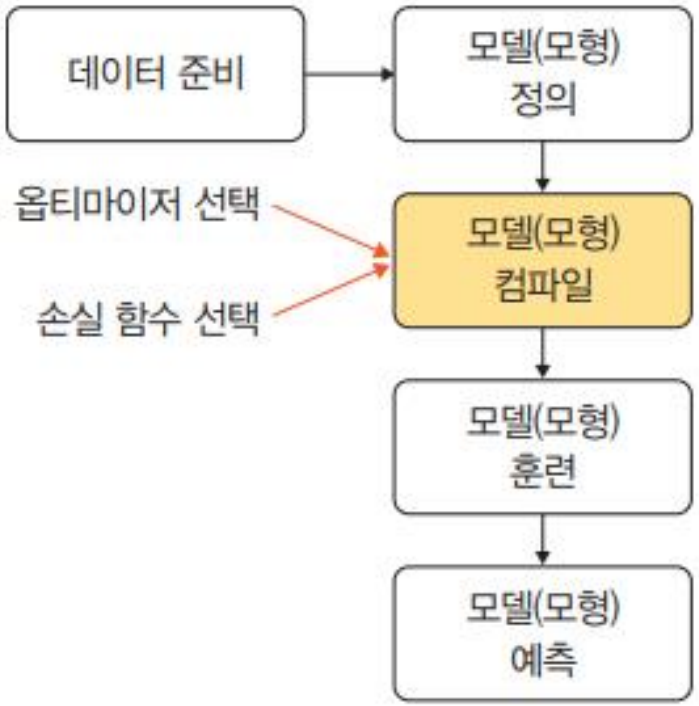
- **수상돌기** : 주변이나 다른 뉴런에서 자극을 받아들이고, 이 자극들을 전기적 신호 형태로 세포체와 축삭돌기로 보내는 역할
- **시냅스** : 신경 세포들이 이루는 연결 부위로, 한 뉴런의 축삭돌기와 다음 뉴런의 수상돌기가 만나는 부분
- **축삭돌기** : 다른 뉴런(수상돌기)에 신호를 전달하는 기능을 하는 뉴런의 한 부분 뉴런에서 뻗어 있는 돌기 중 가장 길며, 한 개만 있음
- **축삭말단** : 전달된 전기 신호를 받아 신경 전달 물질을 시냅스 틈새로 방출

3. 딥러닝이란?

❖ 딥러닝 학습 과정

- 딥러닝의 학습 과정도 머신 러닝과 크게 다르지 않음
- 물론 자세히 다룬다면 데이터를 구하고 전처리하는 방법부터 튜닝하는 방법까지 포함되겠지만, 세세한 부분까지 작성하고 다루기에는 딥러닝 분야가 너무 넓음
- 데이터 준비부터 모델(모형)을 정의하고 사용하는 상위 레벨에서 짚고 넘어감

▼ 그림 1-11 딥러닝 모델의 학습 과정



3. 딥러닝이란?

❖ 데이터 준비 :

- 초보자가 데이터를 쉽게 구할 수 있는 방법은 두 가지
- 첫째, **파이토치(https://tutorials.pytorch.kr/)**나 케라스(https://keras.io/)에서 제공하는 데이터 셋을 사용하는 것
- 제공되는 데이터들은 이미 전처리를 했기 때문에 바로 사용할 수 있으며, 수많은 예제 코드를 쉽게 구할 수 있는 장점이 있음
- 둘째, 캐글(Kaggle) 같은 곳에 공개된 데이터를 사용하는 것
- 물론 국내의 공개 데이터들도 사용할 수 있으나 상당히 많은 전처리를 해야 하기에
- 가능하면 캐글 같은 플랫폼에 제공된 데이터를 활용하길 권장

3. 딥러닝이란?

❖ 모델(모형) 컴파일 :

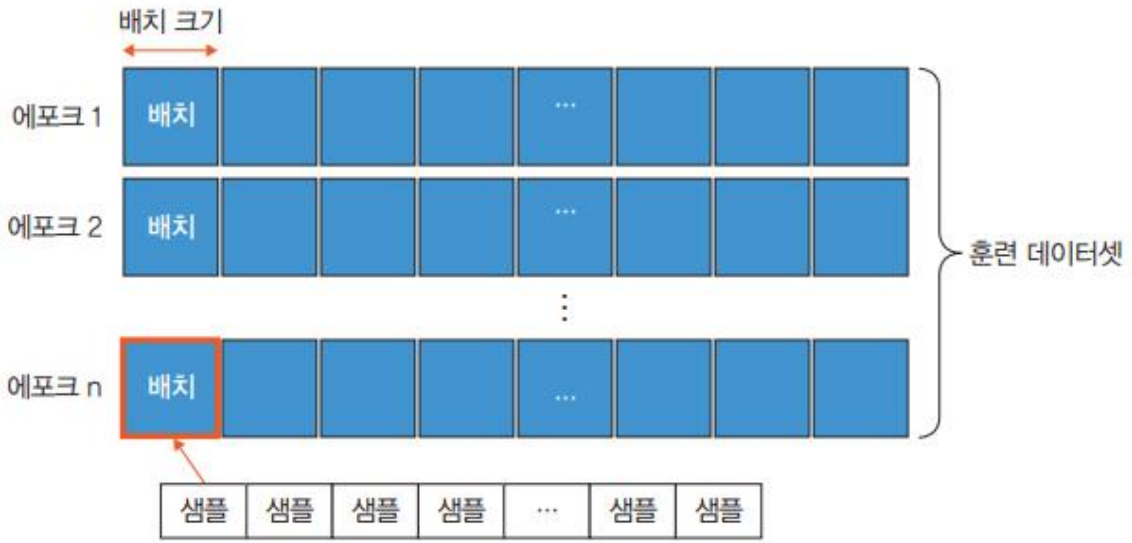
- 컴파일 단계에서 **활성화 함수, 손실 함수, 옵티마이저**를 선택
- 이때 데이터 형태에 따라 다양한 옵션이 가능
- 훈련 데이터 셋 형태가 연속형이라면 **평균 제곱 오차**(Mean Squared Error, MSE)를 사용할 수 있으며, **이진 분류**(binary classification)라면 **크로스 엔트로피**(crossentropy)를 선택
- 또한, **과적합**을 피할 수 있는 **활성화 함수** 및 **옵티마이저** 선택이 중요

❖ 모델(모형) 훈련

- 훈련 단계에서는 한 번에 처리할 데이터 양을 지정
- 이때 한 번에 처리해야 할 데이터 양이 많아지면 학습 속도가 느려지고 메모리 부족 문제를 야기할 수 있기 때문에 적당한 데이터 양을 선택하는 것이 중요
- 전체 훈련 데이터셋에서 일정한 묶음으로 나누어 처리할 수 있는 **배치**와 훈련의 횟수인 **에포크** 선택이 중요
- 이때 훈련 과정에서 값의 변화를 시각적으로 표현하여 눈으로 확인하면서 파라미터와 하이퍼파라미터에 대한 최적의 값을 찾을 수 있어야 함

3. 딥러닝이란?

▼ 그림 1-12 모델 훈련에 필요한 하이퍼파라미터



❖ 배치사이즈와 에포크

- '훈련 데이터셋 1000개에 대한 배치 크기가 20'이라면 샘플 단위 20개마다 모델 가중치를 한 번씩 업데이트시킨다는 의미
- 즉, 총 50번(=1000/20)의 가중치가 업데이트
- 이때 에포크가 10이고 배치 크기가 20이라면, 가중치를 50번 업데이트하는 것을 총 열 번 반복한다는 의미
- 각 데이터 샘플이 총 열 번씩 사용되는 것이므로 결과적으로 가중치가 총 500번 업데이트

3. 딥러닝이란?

❖ 딥러닝 학습 과정

- 성능이 좋다는 의미는?
  - 머신 러닝/딥러닝에서 '성능(performance)'에 대한 공식적인 정의는 없음
  - 궁극적으로 모델 성능은 데이터가 수집된 산업 분야와 모델이 생성된 목적에 의존한다고 볼 수 있음
  - 즉, 모델 성능이 좋다는 의미는 다음과 같은 다양한 의미로 사용할 수 있음
    - 예측을 잘함(정확도가 높음)
    - 훈련 속도가 빠름
- 모델(모형) 예측: 검증 데이터셋을 생성한 모델(모형)에 적용하여 실제로 예측을 진행해보는 단계
- 이때 예측력이 낮다면 파라미터를 튜닝하거나 신경망 자체를 재설계해야 할 수도 있음

3. 딥러닝이란?

❖ 딥러닝 학습 알고리즘

- 딥러닝에서 지도 학습, 비지도 학습, 강화 학습을 정리하면 다음 표와 같음
- 단순한 알고리즘만 고려했을 때의 구분이며, 서로 혼합하여 사용하거나 분석 환경에 제약을 둘 경우 구분이 달라질 수 있음에 주의

▼ 표 1-3 지도 학습, 비지도 학습, 강화 학습

구분	유형	알고리즘
지도 학습(supervised learning)	이미지 분류	• CNN • AlexNet • ResNet
	시계열 데이터 분석	• RNN • LSTM
비지도 학습 (unsupervised learning)	군집 (clustering)	• 가우시안 혼합 모델(Gaussian Mixture Model, GMM) • 자기 조직화 지도(Self-Organizing Map, SOM)
	차원 축소	• 오토인코더(AutoEncoder) • 주성분 분석(PCA)
전이 학습(transfer learning)	전이 학습	• 버트(BERT) • MobileNetV2
강화 학습(reinforcement learning)	–	마르코프 결정 과정(MDP)