

# 인공지능 FAQ 1~13, 28~32

YEAR MONTH DAY

## Q1. 인공지능에서 지능에 해당하는 기능은 무엇인가?

- 인공지능은 인간의 학습능력, 추론 능력, 언어 이해 능력을 컴퓨터 프로그램으로 실현하는 학문 또는 기술  
↳ 따라서 인공지능에서의 지능은 사람의 지능과 유사하지만 더 방대한 데이터를 다룬다.
- 인공지능에서 지능에 해당하는 기능은 학습, 추론, 문제 해결, 지각, 언어 이해 등이 있다.

1. 학습 (Learning) [데이터를 분석하여 패턴을 인식하고, 경험을 바탕으로 새로운 정보를 습득하는 능력]  
    ↳ 머신러닝 (기계학습), 딥러닝 (심층학습) 해당

2. 추론 및 문제 해결 (Reasoning & Problem - Solving)

    ↳ 주어진 정보를 바탕으로 논리적인 결론을 도출 + 최적의 해결책을 찾는 능력

3. 지각 (Perception) - 주변 환경을 인식하고, 데이터를 처리하는 능력

4. 자연어 처리 (Natural Language Processing, NLP)

    ↳ 인간이 사용하는 언어를 이해하고 생성하는 능력

## Q2. 인공지능의 종류 3가지에 대하여 설명하시오. (지도학습, 반지도학습, 강화학습)

- 지도학습, 반지도학습, 강화학습은 모두 머신러닝의 학습 방법이다.

\* 머신러닝: 빅데이터를 스스로 분석하고, 그 내용을 바탕으로 결론을 도출하는 기술

### 1. 지도학습 (Supervised Learning)

- 입력값과 결과 값을 같이 주고 학습을 시키는 방법

    ↳ 정답기반으로 오류를 줄여서 학습하는 방법

    = 반복 학습을 통해 오류를 줄여 가면서 점차 정답에 가까워지는 방법

- 분류 (Classification)

- 데이터가 범주형 변수를 예측하기 위해 사용될 때

- 이진 분류 (Binary Classification) : 레이블이 2개인 경우

- 다중클래스 분류 (multi-class classification) : 범주가 2개 이상인 경우

- 회귀 (Regression)

- 연속된 값을 예측 할 때

- 트레이닝 데이터를 이용하여 연속적인 값을 예측하는 것

## 2. 반지도 학습 (Semi-Supervised Learning)

- 군집을 학습한 후에, 군집의 일부 데이터만 사람이 정답을 매겨주면, 군집 전체를 사람이 매긴 정답으로 볼수 있다는 원리
- 소량의 정답 데이터와 대량의 비정답 데이터를 함께 사용하여 학습 성능 향상

## 3. 강화 학습 (Reinforcement Learning)

- 에이전트 (Agent)가 환경과 상호작용하여 보상을 최대화하는 방향으로 학습하는 방식
  - ↳ 정답을 직접 주지 않고, 행동에 대한 보상을 통해 최적의 전략을 학습
- 환경에 대한 사전지식이 없는 상태로 학습 진행
- 결정을 순차적으로 내려야 하는 문제에 적용 (MDP 사용)

↳ Markov Decision Process

정리

학습 방식	정답 여부	주요 특징
지도 학습	있음	정답 데이터 기반으로 학습
반지도 학습	일부 있음	작은 양의 정답 데이터로 대량의 비정답 데이터 활용
강화 학습	없음 (보상 기반)	보상을 최대화하는 방향으로 행동 학습

## Q3. 전통적인 프로그래밍 방법과 인공지능 프로그램의 차이점은 무엇인가?

- 전통적인 프로그래밍: 개발자가 명확한 규칙과 로직을 직접 코딩 (명시적인 알고리즘)
- 인공지능 프로그램: AI가 입력된 데이터로 규칙을 학습 (파편, 확률적 예측)

전통적 " : 입력  $\rightarrow$  규칙(알고리즘)  $\rightarrow$  출력

인공적인 " : 입력 + 성답  $\rightarrow$  학습  $\rightarrow$  모델 생성  $\rightarrow$  예측

Q4. 딥러닝과 머신러닝의 차이점은 무엇인가?

• 머신러닝 - 데이터를 기반으로 패턴을 학습하는 알고리즘으로 사람이 직접 특징추출을 위해 설계

1. 학습방식: 지도학습, 비지도학습, 강화학습

2. 모델구조: DT, RF, SVM, KNN, LR 등

3. 응용분야: 간단한 데이터 분석, 추천 시스템 등

• 딥러닝: - 인공신경망을 이용한 머신러닝의 하위분야

1. 학습방식: 지도학습, 비지도학습

2. 모델구조: 신경망(DNN) 사용

3. 응용분야: 이미지 인식, 음성 인식, 자연어 처리 등

### 차이점

구조적 차이: 머신러닝은 특징을 사람이 직접 설계한 후 이를 기반으로 학습하지만,

딥러닝은 특징을 사람이 직접 정의하지 않고, 인공 신경망이 자동으로 학습

데이터 필요량: 머신러닝은 비교적 적은 데이터로도 학습이 가능하지만, 딥러닝은 대량의 데이터 필요

응용분야: 머신러닝은 구조화된 데이터를 사용한 예측 및 분류 작업에 사용되지만,

딥러닝은 이미지, 음성, 자연어 등을 처리하는 것과 같이 인간의 뇌가 처리하는 복잡한 작업에 사용

Q5. Classification과 Regression의 주된 차이점은?

- 분류와 회귀는 출력값의 유형에 따라 구분

• Classification - 출력값은 이산적(Discrete)인 형태

ex) email 스팸 필터링, 질병 진단, 이미지 분류 → 범주형 데이터

• Regression - 출력값은 연속적(continuous)인 형태

ex) 주식 시장 예측, 기온 예측, 주택 가격 예측 → 연속적인 숫자 데이터

Q6. 머신러닝에서 차원의 저주(curse of dimensionality)란?

- 차원의 저주는 데이터의 차원이 증가할수록 데이터가 희소해지고, 계산량이 증가하며,

모델의 성능이 저하되는 문제가 발생하는 것.

\* 차원 : 머신러닝에서 차원은 데이터의 특징 또는 속성을 의미

Q7. Dimensionality Reduction는 왜 필요한가?

• 차원 축소(Dimensionality Reduction) - 고차원 데이터에서 불필요한 특성을 제거하여 모델의 성능을 향상시키는 기법.

< 필요한 이유 >

1. 차원의 저주 해결: 차원이 증가할수록 데이터가 희소해지고 성능이 저하되는 것을 차원 축소를 통해 불필요한 특성을 제거하여 모델이 효과적으로 학습 할 수 있게 함.

2. 과적합(Overfitting) 방지: 특성이 많아지면 모델이 불필요한 부분까지 학습하여 과적합이 발생할 수 있어 차원 축소로 중요한 정보만 유지

## Q8. Ridge 와 Lasso의 공통점과 차이점? (Regularization, 규제, Scaling)

공통점: Ridge, Lasso 둘다 규제 기법을 사용하여 과적합을 방지 + 특징들의 크기를 맞추기 위해 스케일링 필요.

\* 규제 기법: 회귀 계수의 크기를 제한하여 모델을 단순화하는 기법

\* 가중치 규제: 모델의 손실함수값이 너무 작아지지 않도록 특정한 값(함수)를 추가

차이점: Lasso 방식은 L1 규제를, Ridge는 L2 규제를 적용한다.

L1의 경우에는 특정한 변수를 삭제할 수 있지만 L2는 모든 변수를 유지한다.

### 수학적 차이점

Lasso : L1 규제  $\rightarrow$  회귀 계수의 절대값의 합  $\sum |w|$

Ridge : L2 규제  $\rightarrow$  회귀 계수의 제곱합  $\sum w^2$

## Q9. Overfitting VS Underfitting

### • 과적합(Overfitting)

- 필요 이상으로 훈련 데이터에 맞춰져 테스트 데이터에서 성능이 낮음. (훈련 때는 높음)

- 해결방안: 규제 적용, 데이터 증가, 특성 줄이기

### • 과소적합(Underfitting)

- 모델이 너무 단순하여 훈련 데이터를 제대로 학습하지 못하여 훈련 데이터와 테스트 데이터 모두 성능 낮음.

- 해결 방안: 더 복잡한 모델 사용, 더 많은 특성 추가

## Q10. Feature Engineering과 Feature Selection의 차이점은?

특성 공학 [데이터의 기본 패턴을 더 잘 찾아내기 위해서 새로운 특성을 생성]

주요기법: 특성 변환, 범주형 데이터 인코딩 등

특성 선택 [특성 중에 중요한 특성을 식별하고 선택 한다. (필요없는 특성 제거)]

주요 기법: 필터 방법, 랩퍼 방법, 임베디드 방법 등

Q11. 전처리(Preprocessing)의 목적과 방법? (노이즈, 이상치, 결측치)

- 전처리 목적: 데이터의 품질 향상시켜 모델의 성능을 최적화하고, 불필요한 데이터를 정리하여 훈련 속도를 개선한다.

#### • 전처리 방법

1. 결측치 처리: 결측치를 제거하거나 결측값을 평균, 중앙값, 최빈값으로 대체

↳ 데이터가 비어있는 경우

2. 이상치 처리: 박스플롯(boxplot) 사용하거나 IQR 방법을 이용하거나 Z-score를 활용한다.

↳ 데이터가 정상 범위를 극단적으로 벗어난 값

3. 노이즈 처리: 이동 평균을 이용하거나 로버스트 스케일링(Robust Scaling) 사용한다.

↳ 데이터에 포함된 필요하지 않은 변동성

#### 4. 데이터 정규화 & 데이터 표준화

• 정규화: 데이터 범위를 [0, 1]로 변환

• 표준화: 데이터의 평균을 0, 표준편차를 1로 변환

Q12. EDA(Exploratory Data Analysis)란? (데이터의 특성 파악: 분포, 상관관계)

- EDA [데이터를 시각화하고 요약 통계를 분석하여 데이터의 특성을 파악하고 문제 해결 방향을 설정하는 과정.  
마신러닝 모델을 학습시키기 전에 준비 과정]

- 주요 과정 -

- 데이터 구조 이해: 데이터셋의 크기, 변수 유형, 컬럼 정보 확인
- 데이터 분포 분석: 컬럼이 정규분포를 따른지 아니면 치우친 분포를 보이는지 파악
- 결측치 및 이상치 탐색: 결측치 및 이상치 확인 후 처리
- 변수 간 상관관계 분석: 범주형 데이터인지, 수치형 데이터인지 확인 후 데이터 조합에 따라 통계 방법과 시각화 방법을 선택한다.

데이터 조합	요약 통계 방법	시각화 방법
범주형 - 범주형	교차 테이블	모자이크 플롯
수치형 - 범주형	카테고리별 통계 값	박스 플롯
수치형 - 수치형	상관계수	산점도

Q13. 회귀에서 절편과 기울기가 의미하는 바는? 딥러닝과 어떻게 연관되는가?

- 선형회귀에서 절편은 기준값이고, 기울기는 변수 간 관계의 강도를 나타낸다.

$$y = mx + b$$

$x$ : 입력 데이터

$y$ : 입력  $x$ 에 대한 예측값

$m$ :  $x$ 가 증가할 때  $y$ 가 얼마나 증가(감소) 하는지 나타내는 변화량

$b$ : 절편으로  $x=0$  일 때,  $y$  값

딥러닝 연관성

- 선형회귀에서의 기울기와 절편의 개념이 딥러닝에서 뉴런이 학습하는 가중치와 편향과 동일한 역할을 한다.

Q28. 결정트리에서 불순도(impurity) - 지니 계수(Gini Index)란 무엇인가?

- 불순도(impurity): 특정 노드에 있는 데이터들이 어느 정도 혼합되어 있는지를 나타내는 척도.  
↳ 클래스가 섞여 있을 수록 불순도↑

- 지니 계수(gini Index): 불순도를 계산하는 방법

$$Gini(t) = 1 - \sum_{i=1}^m p_i^2 \quad p_i: 현재 노드에서 i에 속할 확률$$

→ 지니계수가 높을수록 불순도가 높다 = 데이터가 더 분산되었다.  
= 데이터가 혼잡하고 분류하기 어렵다.

Q29. 앙상블(Ensemble)이란?

- 앙상블(Ensemble): 여러개의 개별 모델을 조합하여 최적의 모델로 일반화 하는 방법 (약한 모델 → 강한 모델)  
↳ why 과적합을 줄이기 위해서

- 유형 -

1. 보팅(Voting): 여러개의 모델들이 예측한 결과를 투표로 최종 예측하는 방법

↳ 하드 보팅: 가장 많이 나온 클래스 (다수결)

↳ 소프트 보팅: 확률 평균을 내서 확률이 가장 높은 클래스 (신뢰도)

2. 배깅(bagging): 전체 데이터에서 랜덤으로 샘플링을 추출하여 생성 이후에 병렬로 학습 후 보팅 or 평균으로 결과를 합치는 방식. ex. 랜덤 포레스트

3. 부스팅(Boosting): 약한 모델들을 순차적으로 학습하고, 이전 모델이 예측이 틀린 데이터는 가중치를 부여해 오류를 보완.

Q. 30. 부트스트랩핑(Bootstrapping)이란?

- 부트스트랩핑은 원본 데이터를 여러 번 샘플링하여 새로운 데이터셋을 만드는 방법이다.
    - 중복이 허용되는 복원 추출 방식으로 진행된다.
    - 하나의 데이터셋에서 많은 훈련 샘플을 만들 수 있다.
- ↳ 데이터의 다양성 확보 + 안정적이고 일반화된 모델 생성을 위해서

Q. 31. 배깅(Bagging)이란 무엇인가?

• 배깅(Bagging): Bootstrap + Aggregating

- 부트스트랩 샘플링을 이용해 데이터셋을 만들고, 이를 독립적인 모델로 학습한 뒤 합계(Aggregating)
  - 분산을 줄이고 과적합 문제 완화
  - 합계 방식: 대수결(분류), 평균값(회귀)

Q. 32. 주성분 분석(PCA)이란 무엇인가?

- 주성분 분석: 차원축소 기법 중 하나로 고차원 데이터셋의 차원을 축소하여 불필요한 특징 제거하는 방법.

↳ 원리  
[자포에 존재하는 직선 중에 데이터의 분산이 가장 큰 방향을 찾음 (데이터가 가장 넓게 퍼져있는)]  
= 이 직선축이 주성분(PC)  
주성분인 축으로 사영한다.

※ 사영: 고차원 데이터를 특성축에 그림자처럼 투영하여 낮은 차원으로 표현.

과정 - 1. 데이터를 표준화: 스케일 조정

2. 공분산 행렬 계산: 데이터 분포와 구조 파악

3. 고유값 및 백터 계산

4. 주성분 선택

5. 데이터 사영.