2025학년도 1학기

# 인공지능 기초를 위한 FAQ


| | |
|---|---|
| **과목명** | 인공지능개론 |
| **담당** | 이명규 교수님 |
| **제출일** | 2025년 3월 7일 |



| | |
|---|---|
| **학과** | 컴퓨터공학과 |
| **학번** | 202136017 |
| **성명** | 이준상 |

1. 인공지능에서 지능에 해당하는 기능은 무엇인가

- 인간의 인지 능력을 모방하거나 능가하는 다양한 기능을 포괄한다. 학습능력, 추론능력, 인식능력, 창의성으로 크게 분류할 수 있다.

2. 인공지능의 종류 3가지에 대해서 설명하시오 (지도학습, 반지도학습, 강화학습)

- 지도학습은 정답이 표시된 데이터를 사용하여 모델을 학습시키는 방식이다.
지도학습은 데이터 준비, 모델 선택, 모델 학습, 모델 평가, 모델 활용 순으로 작동한다.
미리 정의된 종류 중 하나로 분류하거나 입력 데이터에 대해서 연속되는 값을 예측하는 회귀 유형이 있다.

- 비지도 학습은 정답이 표시되지 않은 데이터를 사용하여 모델을 학습시키는 방식이다.
데이터 준비, 모델 선택, 모델 학습, 결과 분석 순으로 이루어지며 비슷한 특성을 가진 데이터들을 묶는 작업인 군집화, 많은 특성을 가진 데이터로 핵심 정보를 유지하면서 가진 특성은 줄이는 차원 축소 등의 유형이 있다.

- 강화학습은 반복적인 시도를 통해서 최적의 결과를 얻기 위한 방식이다.
개체(에이전트)는 설정된 환경과 상호작용하며 행동을 수행한다. 현재 환경은 인식하고 가능한 행동 중 하나를 선택한다. 선택한 행동에 따라 환경은 피드백(보상)을 준다. 받은 피드백을 바탕으로 자신의 전략을 개선한다.

3. 전통적인 프로그래밍 방법과 인공지능 프로그램의 차이점은 무엇인가

- 전통적인 프로그래밍은 입력에 대해서 개발자가 정의해둔 규칙과 논리에 기반으로 결과를 도출하는 방식이지만
인공지능 프로그래밍은 데이터를 기반으로 인공지능이 스스로 학습하고 학습한 규칙을 기반으로 결과를 예측하는 방식이다.

4. 딥러닝과 머신러닝의 차이점은 무엇인가

- 딥러닝은 데이터의 특성을 스스로 학습하고 추출한다. 학습된 데이터의 정확성을 개선하기 위해 여러 층(단계)을 거치기에
머신러닝보다 복잡한 구조를 가지고 있다. 머신러닝은 사전에 특성을 추출한 데이터를 가지고 학습하기 때문에 비교적
단순한 구조를 가진다.

5. Classification과 Regression의 각각 차이점은

- 모두 지도 학습의 한 종류 이지만 Classification(분류)는 미리 정의된 클래스들 중 하나로 분류하는 작업이다. 예측 결과는 범주형 또는 이산형 값이다. Regression(회귀)는 입력 데이터에 대한 연속적인 숫자 값을 예측하는 작업이다. 입력 데이터와 출력값 사이의 관계를 파악하고 데이터에 대한 정확한 값을 예측하는 것이 목표이다.

6. 머신러닝에서 차원의 저주(Curse of dimensionality)란

- 데이터의 특성(차원)이 많아짐에 따라 발생하는 학습에 필요한 계산 복잡성의 증가, 새로운 데이터에 대한 예측 성능 저하를 야기하는 과적합, 일반적인 패턴은 학습하기 어렵게 만드는 데이터 희소성 등을 총칭하는 용어이다.
차원 축소나 모델의 복잡도를 제한하는 규제 등을 통해 해결할 수 있다.

7. Dimensionality Reduction은 왜 필요한가

- Dimensionality Reduction은 데이터의 중요한 정보를 최대한 유지하면서 불필요한 정보를 제거하여 데이터의 복잡성을 줄이고 효율성을 높이기 위해 필요하다.

8. Ridge와 Lasso의 공통점과 차이점? (Regularization, 규제, Scaling)

- 공통점은 모델의 가중치의 페널티를 부과하여 가중치가 너무 커지는 것을 방지하고 선형 회귀 모델을 기반으로 한다.
특성 스케일링을 통해 모델의 성능을 향상시킬 수 있다. 차이점은 Ridge는 가중치 제곱의 합에 페널티를 부여하는 L2 규제를 사용하고 Lasso는 가중치 절대값의 합에 페널티를 부여하는 L1 규제를 사용한다.
Ridge는 가중치를 0에 가깝게 만들어 특성 선택 기능은 없지만 Lasso는 0으로 만들어 특성 선택 효과를 가지고 중요하지 않은 특성을 자동으로 제거한다. 특성의 스케일이 다르면 규제 효과가 달라질 수 있기에 특성 스케일링을 통해 모든 특성의 스케일을 동등하게 맞추는 것이 중요하다.

9. Overfitting Vs Underfitting

- Overfitting은 (과적합) 은 모델이 학습 데이터에 과하게 맞춰져서 실제 데이터의 일반적인 패턴뿐만 아니라 노이즈나 이상치까지 학습하여 새로운 데이터에 대한 성능이 떨어지는 현상이다. Underfitting은 모델이 너무 단순하여 데이터의 기본적인 패턴조차 학습하지 못하여 학습된 데이터에 대해서 성능이 떨어지는 현상이다. 최적의 모델은 학습 데이터의 패턴을 잘 학습하고 새로운 데이터에 대해서도 일반화가 가능해야 한다. Overfitting과 Underfitting의 균형을 찾는 것이 중요하다.

10. Feature Engineering과 Feature selection의 차이점은

- Feature Engineering은 기존 특성을 변환하거나 새로운 특성을 생성하여 모델의 성능을 향상시키는 과정이다. 모델의 예측 능력 향상 및 데이터의 숨겨진 패턴발견, 모델의 해석력 향상 등의 효과를 기대할 수 있다.
Feature Selection은 모델 학습에 불필요하거나 성능을 저하시키는 특성을 제거하고 중요특성만 선택하는 과정이다. 모델의 복잡도를 줄이고 일반화 성능을 향상시키는 데 목적이 있다.

11. 전처리(preprocessing)의 목적과 방법은 (노이즈, 이상치, 결측치)

- 전처리는 노이즈 등이 불필요한 데이터를 제거하거나 수정하여 데이터의 품질 및 이해도를 향상시키고 학습하기 쉬운 형태로 데이터를 변환하여 모델 예측 성능을 향상시키는 등의 목적이 있다. 노이즈 전처리 방법은 통계적 방법, 시각화 방법 등을 사용하여 노이즈를 식별하고 필터링, Smoothing 등의 기법을 통해 노이즈를 제거한다.
이상치 전처리는 통계적, 시각화 방법등을 사용하여 식별하고 이상치를 제거하거나 다른 값으로 대체 혹은 유지한다.
결측치 전처리 방법은 데이터에서 결측치를 식별하고 존재하는 행 또는 열을 삭제하거나 다른 값(평균, 중앙값 등)으로 대체하거나 예측 모델을 사용하여 결측치를 예측한다.

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

- EDA는 데이터 분석 초기 단계에서 데이터를 이해하고 탐색하는 과정이다. 주요 목표는 데이터의 특성을 파악하고 숨겨진 패턴이나 관계를 발견하며, 데이터 전처리 및 모델링 방향을 설정하는 것이다.
EDA에서 데이터의 분포를 파악하는 방법에는 평균, 중앙값 등 통계량을 계산하여 데이터의 중심 경향과 분산 정도를 파악하는 기술 통계 방법이나 밀도 그래프 등을 사용하여 분포의 형태를 시각화하는 등의 방법이 있다.
상관관계를 파악하는 방법에는 상관계수 등을 계산하여 특성간의 선형 또는 비선형 상관관계를 파악하거나 산점도, 히트맵 등을 사용하여 특성간의 상관관계를 시각화 하는 등의 방법이 있다.

13. 회귀에서 결정과 기울기가 의미하는 바는 딥러닝과 어떻게 연관되는가

- 회귀 분석에서 결정은 변수 x가 0 일때의 변수 y의 예측값을 뜻하며 모델의 기본값을 설정하는 역할이다.
기울기는 변수 사이의 선형적인 관계의 강도와 방향을 나타내며 x에 따른 y의 값을 예측하는 데에 사용된다.
딥러닝 모델, 특히 인공 신경망은 여러 개의 뉴런들이 복잡하게 연결된 구조를 가지고 있는데 각 뉴런은 입력값에 가중치를 곱하고 편향(bias)을 더하여 출력값을 생성한다. 이때 가중치는 회귀의 기울기와 비슷한 역할을 하며 편향은 절편과 유사한 역할을 한다. 딥러닝 모델은 데이터를 가장 잘 설명하는 최적의 기울기와 절편을 찾는 과정이라고 할 수 있다.

28. 결정 트리에서 불순도(Impurity) - 지니 계수(Gini Index)란 무엇인가

- 결정 트리에서 불순도는 데이터 집합 내에서 서로 다른 클래스가 혼합되는 정도를 나타내는 지표이다. 불순도가 높을수록 데이터 집합이 혼란스럽고 불순도가 낮을수록 데이터 집합이 순수하다고 판단한다. 지니 계수는 불순도를 측정하는 대표적인 방법이다. 순수하다고 판단하는 0부터 불순하다고 판단하는 1 사이의 값으로 불순도를 측정한다.
지니 계수를 통해 데이터의 특성을 파악하고 모델의 성능을 향상시킬 수 있다.

29. 앙상블이란 무엇인가

- 머신러닝에서 여러 개의 모델을 결합하여 단일 모델보다 더 나은 예측 성능을 얻는 기법이다. 각 모델의 예측 결과를 결합하는 방식에 따라 다양한 앙상블 기법이 존재한다.

30. 부트 스트래핑(bootstrapping)이란 무엇인가

- 부트스트래핑은 통계적 방법론으로서는 주어진 표본 데이터에서 복원 추출을 통해 여러 개의 가상 표본을 생성하고 이를 통해 모집단의 통계적 특성을 추정하는 방법으로 모델의 안정성과 일반화 능력을 평가하는 데 사용될 수 있다. 앙상블 학습 기법 중 배깅(Bagging)에서는 모델의 다양성을 확보하고 개별 모델의 예측 오류를 상쇄하고 분산을 줄여 안정성을 높이며 과적합을 방지하는 데에도 핵심적인 역할을 수 행한다.

31. 배깅(Bagging)이란 무엇인가

- 배깅은 머신러닝에서 앙상블 기법 중 하나로, 여러 개의 모델을 결합하여 단일 모델보다 더 나은 예측 성능을 얻는 방법이다. 부트스트랩 샘플링을 통해 얻은 여러 개의 표본을 사용하여 개별 모델을 학습시킨다. 서로 다른 데이터셋으로 학습되므로 약간의 차이가 있는 예측 결과를 가지고 분류 문제인 경우에는 다수결, 회귀 문제일 경우에는 결과의 평균을 계산하여 예측 결과를 결정한다.

32. 주성분 분석(PCA)이란 무엇인가

- 주성분 분석은 고차원 데이터의 복잡성을 줄이고 중요한 정보를 유지하면서 데이터를 더 낮은 차원으로 변환하는 차원 축소 기법이다. 데이터의 분산을 최대한 보존하는 새로운 축(주성분)을 찾아 데이터를 변환하는 알고리즘이다. 데이터의 특성과 분석 목적에 따라 적절한 주성분 분석 활용을 통해 데이터 분석의 효율성과 정확성을 높일 수 있다.