인공지능 기초를 위한 FAQ

1. 인공지능에서 지능에 해당하는 기능은 무엇인가?

인공지능에서 '지능'이란 인간의 사고 능력을 모방하여 문제를 해결하는 다양한 기능을 의미한다.
대표적으로 학습, 추론, 문제 해결, 지각, 자연어 처리, 계획 및 의사 결정과 같은 기능으로 구현된다.
학습 (Learning)은 데이터를 분석하고 패턴을 찾아 새로운 지식을 습득하는 과정이다.
추론 (Reasoning)은 기존 정보를 바탕으로 논리적으로 결론을 도출하는 능력이다.
문제 해결 (Problem Solving)은 최적의 해결방법을 찾는 과정이며, 지각 (Perception)은 이미지 인식, 음성 인식 등 외부 환경을 감지하고 이해하는 기능이다. 계획 및 의사결정 (Planning & Decision Making)은 목표를 달성하기 위해 전략을 수립하는 과정이다.

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

지도학습 : 입력값과 결과값을 함께 주고 학습을 시키는 방법으로, 과거의 데이터를 기반으로 미래를 예측할 때 유용하게 사용된다.

반지도학습 : 입력 데이터의 일부에만 정답이 있을 경우 이를 이용하여 모델을 학습하는 방법이다. 즉 일부만 정답이 있는 데이터셋과, 나머지 정답이 없는 데이터셋을 이용하여 학습한다.

강화학습 : 결과값이 아닌, 어떤 일을 잘했을 때 보상을 주는 방식으로 어떠한 액션이 최선인지를 학습시키는 방법이다. 게임, 네비게이션 등에서 활용되며 보상을 극대화시킬 수 있는 동작을 선택할 수 있도록 학습시킨다.

3. 전통적인 프로그래밍 방법과 인공지능 프로그램의 차이점은 무엇인가?

전통적인 프로그램은 데이터를 입력하고, 프로그래머가 규칙을 설정해서 결과를 도출하는 식이다.
반면 인공지능 프로그램은 데이터와 결과를 같이 입력하고, 그것을 AI가 학습하여 규칙을 도출하는 형식이다.

4. 딥러닝과 머신러닝의 차이점은 무엇인가?

머신러닝은 전통적인 알고리즘을 사용하여 수동으로 각 데이터의 특징을 알아내고, 입력과 출력 데이터 사이의 관계를 학습하는 것에 초점을 맞추어 반면 딥러닝은 인공 신경망 기반의 모델을 사용해서 데이터의 특징을 자동으로 알아내고, 머신러닝보다 더 복잡하고 다양한 데이터를 처리할 수 있기 때문에 더 좋은 성능을 보인다.

5. Classification과 Regression의 주된 차이점은?

Classification (분류) 모델은 예측값으로 이산적인 값을 출력한다. 예를 들어 사진을 개와 고양이로 분류하는 것은 Classification이다. 반면 Regression (회귀) 모델은 예측값으로 연속적인 값을 출력한다. 어떤 사람의 키와 몸무게를 데이터로 얻어 허리둘레를 예측하는 모델이 이에 속한다.

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

공간의 차원이 증가함에 따라 데이터의 밀도가 급격히 감소하고, 이로 인해 데이터 분석이나 머신러닝 모델의 성능에 부정적인 영향을 미치는 현상을 말한다. 차원이 증가할수록 필요한 데이터는 기하급수적으로 증가하는데, 이는 연산 비용 증가, 과적합, 예측 성능 저하를 가져올 수 있다.

7. Dimensionality Reduction은 왜 필요한가?

차원이 증가하면 비용, 시간, 자원, 용량, 과적합 등의 문제가 발생할 수 있고, 군집화 분석 결과도 좋지 않다. 따라서 차원 축소를 함으로써 설명력이 높은 feature만 사용하여 효과적인 모델을 생성한다.

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

Ridge와 Lasso는 선형 회귀 모델의 과적합을 방지하기 위한 규제 (Regularization) 기법으로, 손실 함수에 패널티를 추가하여 모델의 일반화 성능을 높인다. 두 방법 모두 스케일링 (Standardization)이 필요하며, 다중공선성 (독립 변수들끼리 너무 유사해서 회귀 분석이 불안정해지는 문제)을 완화하는 효과가 있다. Ridge는 L2 정규화를 사용하여 회귀 계수를 작게 만들지만 0으로 만들지는 않으며, Lasso는 L1 정규화를 적용하여 일부 계수를 0으로 만들어 변수 선택 기능을 수행한다. Ridge는 모든 변수를 유지하면서 모델을 안정적으로 만들고, Lasso는 불필요한 변수를 제거하여 희소 모델 (Sparse Model)을 생성한다.

9. Overfitting vs. Underfitting

Overfitting은 학습 데이터에 대해 과하게 학습된 상황이다. 분산이 높게 학습되어 학습 데이터 이외의 데이터에 대해서 모델이 잘 동작하지 못한다. 반면 Underfitting은 학습 데이터조차 제대로 학습하지 못하고 편향이 높게 학습되어 새로운 데이터를 예측하지 못하는 상태다.

10. Feature Engineering과 Feature Selection의 차이점은?

Feature Engineering은 기존 데이터를 변형하거나 새로운 feature를 만들어 모델 성능을 향상시키는 과정이다 예를 들어 날짜 데이터를 요일, 계절 등으로 변환하는 경우가 있다
반면 Feature Selection은 주어진 feature 중 중요한 feature만 선택하여 모델 성능을 최적화하는 과정으로, 차원의 저주를 줄이고 과적합을 방지한다

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

전처리의 목적은 데이터의 노이즈 (측정된 변수에 무작위의 오류 또는 분산이 존재하는 것) 제거, 이상치 (관측된 데이터의 범위에서 많이 벗어난 값) 처리, 결측치 (수집 과정에서 누락된 데이터) 보완 등을 통해 모델 성능을 최적화하는 것이다 노이즈 처리 방법으로는 이동 평균, 이상치 제거 (IQR, Z-score), 데이터 변환 (로그, 스케일링), 군집화 (K-Means), Autoencoder 등을 사용하는 방법이 있다. 이상치 처리 방법으로는 IQR (사분위 범위) 또는 Z-score을 사용하여 이상치를 탐지하고 제거 또는 대체한다 결측치는 삭제하거나 평균, 중앙값, 최빈값으로 대체하며, KNN, 선형 회귀, 딥러닝 기반 예측 모델을 사용하기도 한다

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

EDA란 탐색적 데이터 분석으로, 분석 초기 단계에서 데이터의 분포와 패턴을 파악하여 인사이트를 얻는 과정이다 EDA는 이상치 탐지, 데이터 품질 확인, feature selection 등에 중요한 역할을 한다 히스토그램, 박스 플롯, KDE 그래프를 통해 데이터 분포 확인을 할 수 있고, 피어슨 상관계수, 히트맵 등의 방법을 통해 상관관계를 분석할 수 있다 산점도, 쌍플롯 등을 통해 변수 간 관계를 분석할 수 있다

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

절편 (Intercept, b)는 독립 변수가 0일 때 종속 변수 (y)의 예상값이다 기울기 (Slope, w)는 독립 변수 x가 변할 때 종속 변수 y가 얼마나 변하는지 나타낸다. 선형 회귀의 가중치 w와 절편 b 개념이 딥러닝의 뉴런에서 가중치 (weight)와 바이어스 (Bias) 역할을 한다.
각 뉴런은 $y = w_1x_1 + w_2x_2 + \cdots + w_nx_n + b$ 와 같은 수식을 이용해 데이터를 처리하는데, 이때 뉴런이 데이터를 학습하며 w와 b를 최적화한다

28. 결정트리에서 불순도 (Impurity) - 지니 계수 (Gini Index)란 무엇인가?

노드에 다양한 클래스의 데이터가 포함될수록 불순도가 높다. 지니 계수는 불순도를 측정하는 지표로, 값이 작을수록 노드가 더 순수하다. $Gini = 1 - \sum p_i^2$ 공식으로 나타낼 수 있으며, p_i는 클래스 i가 해당 노드에 속할 확률을 의미한다. 예를 들어 클래스 A가 노드에 속할 확률이 80%, B가 속할 확률이 20%라면, $1 - (0.8^2 + 0.2^2) = 0.32$ 처럼 구할 수 있다.

29. 앙상블이란 무엇인가?

앙상블이란 여러 개의 모델을 조합하여 개별 모델보다 더 강력한 예측 성능을 내는 기법이다. 과적합 감소 효과가 있으며, 개별 모델 성능이 잘 안 나올 때 앙상블 학습을 이용하면 성능이 향상될 수 있다. 배깅 (Bagging), 부스팅 (Boosting), 스태킹 (Stacking) 등의 유형이 있다.

30. 부트 스트래핑 (bootstraping)이란 무엇인가?

데이터에서 중복을 허용하여 여러 개의 샘플을 랜덤하게 뽑는 과정이다. 작은 데이터셋으로도 여러 개의 모델을 만들 수 있으며, 원본 데이터와 비슷한 분포를 가진 데이터를 만들 수 있다. 배깅 (Bagging)과 랜덤 포레스트 (Random Forest)에서 사용된다.

31. 배깅 (Bagging)이란 무엇인가?

Bootstrap Aggregating의 약자다. 원본 데이터에서 부트스트래핑으로 여러 개의 데이터셋을 생성한 후, 각 데이터셋으로 개별 모델 (약한 학습기)을 학습한다. 이후 평균 (회귀) 또는 투표 (분류)를 통해 최종 예측한다. 모델 간 상관관계를 낮춰 과적합을 방지하고, 모델이 독립적으로 학습하므로 병렬 처리가 가능하다. 대표적인 예로 랜덤 포레스트가 있다.

32. 주성분 분석 (PCA)이란 무엇인가?

PCA는 차원 축소 기법으로, 고차원 데이터를 저차원으로 변환하여 데이터의 중요한 정보만 유지한다. 이때 주성분이란 전체 데이터 (독립변수)의 분산을 가장 잘 설명하는 성분을 말한다. PCA의 과정으로는 우선 데이터를 표준화하고, 공분산 행렬 계산을 수행한다. 이후 고유값과 고유벡터를 계산하고 고유값이 큰 순서대로 주성분을 선택한다. 데이터의 분산을 최대한 보존하면서 차원을 줄이지만, feature의 해석력이 감소할 수 있다. 데이터 시각화, 노이즈 제거, 연산 속도 향상에 사용된다.