학과: 응용통계학과 이름: 백형빈

인공지능 기초를 위한 FAQ

1. 인공지능에서 지능에 해당하는 기능은 무엇인가?

인공지능에서 지능에 해당하는 기능은 인간의 인지 능력을 모방하여 구현하는 다양한 기능을 의미한다.

<대표적인 기능>

1. 학습(Learning)
   - 데이터에서 패턴을 찾아내고, 경험을 통해 개선하는 기능 (지도학습, 비지도학습, 강화학습)
2. 추론 (Reasoning)
   - 기존의 지식을 활용하여 새로운 사실을 도출하는 기능 (연역적 추론, 귀납적 추론)
3. 문제 해결 (Problem Solving)
   - 주어진 목표를 달성하기 위해 최적의 방법을 찾는 기능. (탐색 알고리즘, 최적화 개념)

2. 인공지능의 종류 3가지에 대해서 설명하세요 (지도학습, 반지도 학습, 강화학습)

1. 지도학습(Supervised Learning)
   - 정의 : 입력 데이터(Input)와 정답(Label)이 주어진 상태에서, 모델이 입력과 출력의 관계를 학습하는 방식
   - 특징 : 데이터에 대한 명확한 정답이 있어야 하며, 학습이 끝난 후 새로운 데이터에 대해 정확한 예측을 수행하는 것이 목표
2. 반지도 학습(Semi-Supervised Learning)
   - 정의 : 일부 데이터에만 정답(Label)이 있고, 대부분의 데이터에는 정답이 없는 상태에서 학습하는 방식
   - 특징 : 지도학습과 비지도 학습의 중간 형태로, 소량의 라벨링된 데이터와 대량의 라벨링 되지 않은 데이터를 함께 활용
3. 강화학습(Reinforcement Learning, RL)
   - 정의 : 환경과 상호작용하며 시행 착오를 통해 보상을 최적화하는 방식
   - 특징 : 지도학습처럼 정답이 주어지지 않고, 행동에 대한 보상과 벌점을 통해 최적의 정책을 학습

3. 전통적인 프로그래밍 방법과 인공지능 프로그램의 차이점은 무엇인가?

	전통적인 프로그램	인공지능 프로그래밍
동작방식	: 개발자가 직접 규칙을 정의	데이터에서 패턴을 학습
입력과 출력	: 입력 → 명시적인 규칙 적용 → 출력	입력 → 학습된 모델 적용 → 출력
문제해결 접근법	: 논리적 규칙과 알고리즘을 직접 설계	데이터에서 패턴을 찾아 자동으로 규칙을 학습
변화 대응 력	: 새로운 조건이 생기면 개발자가 직접 수정	학습을 통해 새로운 패턴 적용 가능

4. 딥러닝과 머신러닝의 차이점은 무엇인가?
- 머신러닝 : 데이터에서 패턴을 학습하여 예측하거나 분류하는 알고리즘을 포함하는 인공지능의 하위 분야
- 딥러닝 : 머신러닝의 한 종류로, 인공신경망(ANN)을 기반으로 여러 층의 뉴런을 활용해 복잡한 패턴을 학습하는 방법
- 머신러닝은 사람이 직접 특성을 선택, 딥러닝은 모델이 자동으로 중요한 특징을 학습
- 머신러닝은 적은 데이터로도 학습이 가능하고 상대적으로 연산량이 적다, 딥러닝은 대량의 데이터가 필요하고 높은 연산량이 요구된다.

5. Classification과 Regression의 주된 차이점은?
- 분류(classification) : 데이터를 특정 범주(category) 또는 클래스(class)로 구분하는 문제
- 회귀(Regression) : 연속적인 숫자 값(Numeric Value)을 예측하는 문제
→ 출력 값이 이산적인지(분류) 또는 연속적인지(회귀)에 따라 적절한 알고리즘을 선택하는 차이 존재

6. 머신러닝에서 차원의 저주(curse of dimensionality)란?
차원의 저주는 데이터의 차원이 증가할수록 모델성능이 저하되는 현상을 의미한다. 즉, 특성(Feature)의 개수가 많아질수록 데이터가 희소해지고, 계산 복잡도가 증가하며, 일반화 성능이 떨어지는 문제 발생

7. Dimensionality Reduction은 왜 필요한가?
차원축소(Dimensionality Reduction)은 데이터의 특성(Feature) 개수를 줄여 모델의 성능을 향상시키는 기법이다. 차원이 증가할수록 데이터가 희소해지고 연산 비용이 증가하며, 과적합 위험이 커지기 때문에 차원 축소가 필요하다.

8. Ridge와 Lasso의 공통점과 차이점은? (Regularization, 규제, Scaling)
- 공통점
- 둘다 회귀 모델의 과적합(Overfitting)을 방지하기 위한 정규화(Regularization) 기법으로 사용된다.
- 손실하는 손실 함수(Loss Function)에 패널티(Term)를 추가하여 모델이 과도하게 복잡해지는 것을 방지한다.
- 특성(Feature) 스케일링이 필요 (StandardScaler 나 MinMaxScaler 사용)
- 차이점
- Ridge는 L2 정규화, Lasso는 L1 정규화 방식 사용
- Ridge는 가중치를 0에 가깝게 만든다, Lasso는 불필요한 계수를 완전히 0으로 만들어 특성 선택 효과
- Ridge는 모든 특성이 골고루 중요할 때 사용, Lasso는 일부 특성만 중요한 경우 사용

9. Overfitting Vs Underfitting

1. overfitting (과적합)

정의: 모델이 훈련 데이터에 너무 최적화 되어 테스트 데이터에서 성능이 낮아지는 현상

원인: 모델이 너무 복잡함, 특성이 너무 많음, 학습 데이터에 과하게 적합

해결방법: Regularization (L1/L2, Dropout), 불필요한 특성 제거, 데이터 증가, 훈련 에포크 줄이기

2. Underfitting (과소적합)

정의: 모델이 너무 단순하여 데이터의 패턴을 제대로 학습하지 못하는 현상

원인: 모델이 너무 단순함, 학습 데이터 부족, 학습 시간이 너무 짧음

해결방법: 더 복잡한 모델 사용, 더 많은 feature 추가, 훈련 데이터 증가, 학습 시간 증가

10. Feature Engineering 과 Feature selection의 차이점은?

1. Feature Engineering (특성 공학)

정의: 새로운 특성을 생성하거나 기존 특성을 변형하여 모델의 성능을 향상시키는 과정, 원본 데이터에서 더 유용한 정보를 추출하는 것이 목표

방법: 데이터 변형, 조합, 변환 (다항식 변환, 로그 변환, 원-핫 인코딩)

결과: 데이터의 표현력을 향상

2. Feature selection (특성 선택)

정의: 모델 성능에 영향이 적은 특성을 제거하여 학습 속도를 높이고 과적합을 방지하는 과정, 불필요한 변수를 제거하여 모델을 단순화 하고 일반화 성능을 향상시키는 것이 목표

방법: 통계적 기법, 모델기반 선택 (Lasso, RFE, PCA, 카이제곱 검정)

결과: 계산량 감소, 과적합 방지

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

- 전처리 (Preprocessing)은 머신러닝 모델을 학습시키기 전에 원시 데이터를 처리하고 변환하여 모델이 데이터를 잘 학습할 수 있도록 만드는 과정이다. 데이터 품질을 높이고 모델 성능을 최적화하는 데 중요한 역할

- 노이즈 처리

노이즈는 데이터에 포함된 무의미한 변동이나 불규칙한 값이다.

- 방법: 스무딩: 데이터를 평활화하여 불필요한 변동을 제거 (이동 평균, 로컬 리그레션)

필터링: 주어진 기준에 맞지 않는 데이터를 제거하거나 수정

- 이상치 처리

이상치는 데이터에서 다른 값들과 현저하게 다른 값을 의미한다.

- 방법: IQR 사용, 박스 플롯을 이용해 시각화 → 삭제, 변경 (평균, 중앙값)

- 결측치 처리

결측치는 일부 데이터가 누락된 상태를,

- 방법: 결측치를 포함한 행이나 열 삭제, 평균/중앙값/최빈값으로 대체.

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

정의 : 데이터의 특성을 탐색하고 이해하는 과정 데이터를 분석하기 전에 데이터의 구조, 분포, 상관관계 등을 시각적으로 파악하는 중요한 첫번째 단계이다.

데이터분포확인 : 각 특성의 분포를 확인하기 위해 히스토그램, 박스플랏 등을 활용. 연속형변수는 히스토그램 이 밀도그래프 범주형 변수는 막대 그래프 활용

상관관계 : 두 변수간의 관계를 나타내는 지표, 한 변수의 변화가 다른 변수에 어떻게 영향을 미치는 지를 나타낸다. -1에서 1 사이의 상관계수를 갖고 1에가까울 수록 양의 상관관계 0에 가까울 수록 상관관계가 없음을 나타내고 -1에 가까울 수록 음의 상관관계를 나타낸다.

13 회귀에서 절편과 기울기가 의미하는 바는? 딥러닝과 어떻게 연관되는가?

선형회귀식 : $y = mx + b$

y는 종속변수 (예측값), x는 독립변수 (입력 값), m는 기울기, b는 절편이다.

기울기 (m) : 독립변수가 1 증가할 때 종속변수가 얼마나 변하는 지를 나타낸다.

절편 (b) : x=0 일때 y값. 데이터의 기본점인 위치를 결정하는 역할을 한다.

딥러닝과 연관성 : 딥러닝에서는 선형 회귀의 개념이 뉴런의 가중치와 편향과 직접 연결된다.

뉴런 : 신경망에서 데이터를 처리하는 기본 단위

가중치 : 입력 데이터에 곱해지는 값 → 기울기(m) 역할 (입력 데이터가 모델의 출력에 미치는 영향을 조절)

편향 : 모델이 조정할 수 있는 추가적인 상수 값 → 절편(b) 역할 (뉴런의 활성화 여부를 조정하여 학습의 유연성을 증가시킴)

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

1. 불순도 : 한 노드에 있는 데이터가 서로 다른 클래스로 섞여 있는 정도를 나타내는 지표

- 불순도가 높다 → 클래스가 다양하게 섞여 있음 (분류가 어려움)
- 불순도가 낮다 → 대부분의 데이터가 같은 클래스에 속함 (분류가 쉬움)

2. 지니계수 : 결정트리에서 불순도를 측정하는 대표적인 방법중 하나이다.

(1) 공식

$$Gini = 1 - \sum_{i=1}^{C} P_i^2$$

C는 클래스 개수

P_i는 특정클래스 i에 속하는 데이터 샘플의 비율

$(0 \le Gini \le 0.5)$

- Gini = 0 (불순도 0%) → 한 노드에 한 가지 클래스만 존재. (완전 순수, 최적의 상태)
- Gini = 0.5 (불순도 50%) → 두 개의 클래스가 동일한 비율로 섞여 있음 (최악의 상태)

29. 앙상블이란 무엇인가?

정의: 여러 개의 머신러닝 모델을 조합하여 더 나은 성능을 얻는 기법이다.

- 개별 모델이 가진 한계를 극복하고 일반화 성능을 향상시킬 수 있다.
- 단일 모델보다 과적합(Overfitting)을 방지하고 예측력을 높일 수 있다.
- 여러 모델을 조합하는 방식에 따라 Bagging, Boosting, Stacking 등의 기법이 있다.

30. 부트 스트랩핑(bootstrapping)이란 무엇인가?

정의: 데이터 샘플링 기법으로, 주어진 데이터에서 복원 추출을 통해 서로 다른 데이터셋을 여러개 만드는 방법이다.
즉, 하나의 데이터셋에서 여러개의 샘플을 생성하여 통계적 추정이나 머신러닝 모델 학습에 활용하는 기법이다.

31. 배깅(Bagging)이란 무엇인가?

정의: 부트 스트래핑 기법을 활용하여 여러 개의 모델을 학습한 후, 결과를 평균 또는 투표 방식으로 결합하는 앙상블 기법이다. 즉, 하나의 모델이 아니라 여러 개의 모델을 독립적으로 학습하여 편향은 유지하면서 분산을 줄이는 방식이다.

32. 주성분 분석(PCA)이란 무엇인가?

정의: 고차원의 데이터를 저차원으로 변환하는 차원 축소 기법이다. 즉, 데이터의 중요한 정보는 최대한 유지하면서 차원을 줄이는 방법이다.