인공지능개론 FAQ (2021135253 황인서)

1. 인공지능에서 지능에 해당하는 기능은 무엇인가?

지능의 의미는 음성 또는 문자로 쓰인 언어를 인식하고, 이해하고, 번역하거나, 데이터를 분석하거나, 제안을 하는 등의 기능을 갖는다.

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

- 지도학습은 정답이 있는 데이터를 활용한 훈련을 통해 새로운 데이터의 예측을 한다.
- 비지도학습은 정답이 없는 데이터의 분석으로 패턴을 찾아내거나 그룹화하는 작업을 수행한다.
- 강화학습은 입출력 쌍으로 이루어진 훈련 집합이 제시되지 않으며, 잘못된 행동에 대한 정정이 이루어지지 않고, 행동에 대한 보상을 받으면서 학습하며 특정 환경에서 선택 가능한 행동들 중 보상을 최대화하는 행동 혹은 행동 순서를 학습하는 것이다.

3. 전통적인 프로그래밍 방법과 인공지능 프로그램의 차이점은 무엇인가?

전통적인 프로그래밍 방법은 입력 데이터와 규칙을 사용자가 작성하여 출력 데이터를 얻는 방식인 반면, 인공지능 프로그램은 입출력 데이터만을 사용자가 작성하고 이를 통해 학습한 프로그램이 생성한 규칙을 얻는다. 새로 데이터를 입력할 경우 해당 규칙을 거쳐 새로 출력데이터를 얻을 수 있다.

4. 딥러닝과 머신러닝의 차이점은 무엇인가?

- 머신러닝은 사용자가 직접 특징 (feature)를 선택하여 모델을 훈련시킨다.
- 딥러닝은 머신러닝의 하위분야로 프로그램이 직접 데이터만을 통해 특징(feature)을 설계한다.

5. Classification과 Regression의 주된 차이점은?

Classification은 데이터가 특정 카테고리 (클래스)에 속하도록 예측하는 반면, Regression은 연속적인 숫자 값을 예측하는 차이가 있다.

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

특징 (feature)의 개수가 늘어남에 따라 데이터가 고차원화되어 발생하는 문제이다.

1) 데이터의 희소성이 커져 패턴을 찾지 못하는 경우
2) 점들 사이의 유클리드 거리가 비슷해서 패턴을 찾지 못하는 경우
3) 학습해야 할 파라미터 개수가 증가하면서 계산 복잡도가 증가하여 과적합이 발생한 경우

7. Dimensionality Reduction은 왜 필요한가?

차원의 저주를 극복하고, 과적합 (Overfitting)을 방지할 수 있다. 연산 속도가 개선되고 메모리 사용량이 감소한다. 2차원, 3차원의 경우 데이터의 시각화가 가능하다.

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

Ridge와 Lasso는 모두 선형회귀와 정규화 기법으로 모델이 과적합되는 것을 방지 하기 위해 가중치를 제한한다.

→ 손실함수에 Regularization Term을 추가 = 규제

- Ridge 회귀에 적용하는 규제는 L2규제이다. 모든 변수를 선택하는 대신 가중치를 0에 가깝게 줄인다.
- Lasso회귀에 적용하는 규제는 L1규제이다. 일부 가중치를 완전히 0으로 만들어 불필요한 변수를 제거한 희소모델을 생성한다. 변수 선택이 중요한 경우 도움이 된다.

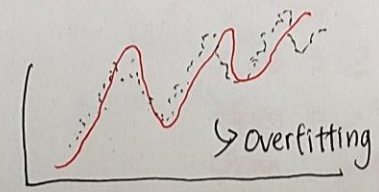
9. Overfitting VS Underfitting

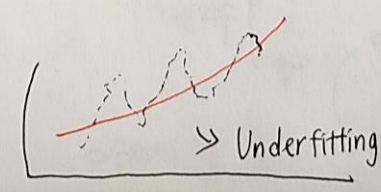
- Overfitting은 모델이 훈련데이터에 과하게 맞춰져서 일반화 성능이 떨어진 경우이다. 훈련데이터는 잘 맞지만 테스트데이터는 오차가 크다. Overfitting은 Noise까지 학습하는데 이는 센서 오류나, 입력 실수, 또는 환경적 요인으로 발생한 무작위적이고 비정상적인 값이다. 데이터가 높은 분산도를 갖는다.

~~해결방법~~ > feature 줄이기 / 학습횟수 (Epoch) 늘이기 ~~좋음~~ 정규화(Ridge와 Loss) 사용 / 데이터 양 늘이기

- Underfitting은 모델이 데이터의 패턴을 충분히 학습시키지 못한 경우이다. 훈련데이터와 테스트 데이터 모두 오차가 크다. 모델이 너무 단순하여 편향도가 높다.

~~해결방법~~ > feature 늘이기 / 학습횟수 (Epoch) 늘이기





10. Feature Engineering과 Feature Selection의 차이점은?

- Feature Engineering은 원본 데이터를 바탕으로 새로운 Feature를 생성하거나 변형하는 과정이다. 전과정으로 Feature가 늘어날 수도 있다.

예로 ① 날짜 특성을 요일과 연도로 분리하거나,
② 수치적으로 로그변환하거나,
③ 범주형 데이터를 수치형으로 변환하는 One-Hot Encoding 이나,
④ 고차원 데이터를 저차원 데이터로 변환하는 (PCA)차원 축소 등이 있다.

- Feature Selection은 기존 Feature에서 좋은 것만 선택하고 그외는 제거하는 과정이다. 전과정으로 Feature가 줄어든다.

예로, ① 상관관계 분석으로 Feature를 제거하거나,
② Lasso Regression에서 가중치가 0인 Feature를 제거하거나,
③ Variance Threshold로 Threshold의 기준보다 낮은 feature를 제거하거나,
④ Recursive Feature Elimination (RFE)으로 feature의 중요도를 계산하며 중요도가 낮은 순으로 제거하면서 최적 모델을 찾는다.

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

전처리의 목적은 노이즈를 제거하여 과적합을 방지하고, 범주형 데이터를 수치형으로 변환하거나 결측치를 제거하고 데이터를 정규화하는 일련의 과정을 말한다. 전처리 방법은

1) 데이터 정리 – 결측치 (NaN) 및 이상치 제거
2) 데이터 변환 – 정규화
3) Feature Selection / Feature Engineering – Feature를 올바른 형태로 변환하고 적합한 feature 수를 찾아 늘리거나 /줄인다.
4) 데이터 균형 조정 – 클래스 데이터가 적으면 증가하거나 (=오버샘플링) / 많으면 감소한다. (=언더샘플링)

12. EDA (Explorary Data Analysis)란?

데이터의 패턴, 분포, 이상치 등을 파악하는 과정이다. 데이터 전처리의 방향 설정을 목적으로 한다.

데이터의 기본 정보 확인 method → .info() .describe() .head() .tail()
결측치 및 이상치 탐색 method → .isnull() .sum()
데이터 분포 분석 → 히스토그램 사용
변수 간 상관관계 분석 method → .corr() 또는 산점도 (Scatter Plot) 사용
범주형 변수 분석 method → .groupby() 또는 막대그래프 (Bar Plot), 피벗 테이블

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

$y = (w)x + (b)$

절편 b : 출력의 기준이 되는 기본값
기울기 w : 데이터의 변화율

⇒ 딥러닝은 회귀의 개념을 다층구조로 활용

$y = w_1x_1 + w_2x_2 + \cdots + w_nx_n + (b)$

절편 b : 입력값이 0이어도 출력값 조정
기울기 w : 입력값의 중요도를 학습

14. 교차검증과 K-fold 교차검증의 의미와 차이

- 교차검증은 동일한 훈련/테스트 데이터를 여러번 학습-검증 하는 과정이다. (일반화 성능을 파악할 수 있다.)
- K-fold 교차검증은 데이터를 K개의 Fold로 분할한 뒤, K번 반복학습하여 학습-검증 하는 과정이다.

K=3 인경우
1회 학습) 학습데이터 = Fold 2, 3 → 테스트 데이터 = Fold 1
2회 학습) 학습데이터 = Fold 1, 3 → 테스트 데이터 = Fold 2
3회 학습) 학습데이터 = Fold 1, 2 → 테스트 데이터 = Fold 3
3회 학습의 평균 성능 확인

15. 하이퍼파라미터 튜닝이란 무엇인가?

하이퍼파라미터는 모델 학습 전에 사용자가 설정하는 값이다.
일반 파라미터는 모델학습을 통해 최적화 된다.
주요 튜닝되는 하이퍼파라미터는 학습률(Learning Rate), 트리의 깊이(max depth), 배치 크기(Batch size), 규제(Regularization)가 있다.
모델의 성능을 극대화하기 위한 하이퍼파라미터 값을 찾는 과정을 하이퍼파라미터 튜닝이라고 한다.

① Grid Search는 가능한 모든 하이퍼파라미터의 값을 지정하여 모든 조합을 실행해 보는 방법이다.

② Random Search는 무작위로 선택한 값을 실행하여 최적의 하이퍼파라미터 값을 찾는 과정이다.

③ Bayesian Optimization은 이전 시행 결과를 바탕으로 다음 실험 값을 선택하는 방법이다.

16. 결정 트리에서 불순도(Impurity) - 지니계수(Gini Index)란 무엇인가?

불순도가 높을수록 다양한 클래스가 섞여 있어 예측이 어려워지고, 낮을수록 예측이 쉬워진다. 결정트리는 불순도를 최소화 하는 방향으로 데이터를 분할하며 학습한다.
지니계수로 불순도를 측정할 수 있다.

$G = 1 - \sum p_i^2$

↳ 클래스 i에 속할 확률 = p_i

17. 앙상블이란 무엇인가?

여러개의 모델을 결합하는 기법이다.

① 배깅(Bagging)은 동일한 여러개의 모델을 독립적으로 학습시킨 후 다수결 혹은 평균으로 최종 예측을 수행한다. 랜덤한 데이터로 모델을 학습시키고, 개별 모델이 과적합되더라도 결합하면 과적합이 줄어든다. → Random Forest

② 부스팅(Boosting)은 동일한 여러 개의 모델을 순차적으로 학습하며, 틀린 샘플에 가중치를 부여하는 과정의 반복으로 더 높은 정확도의 예측을 수행한다.
(단, 과적합이 발생할 수 있다.)

③ 스태킹(Stacking)은 다른 여러개의 모델을 조합하여, 예측값을 다시 입력으로 사용하여 최종 예측을 수행하는 Meta Model을 사용한다.

18. 부트스트래핑(Bootstrapping)이란 무엇인가?

원본 데이터에서 중복을 허용한 복원 추출을 통해 데이터를 샘플링하는 기법이다.
원본 데이터 크기와 동일한 크기의 샘플을 다양하게 추출할 수 있다.
→ 신뢰구간 계산, 평균 및 분산 추정, 머신러닝 모델 평가 등에 활용

19. 배깅(Bagging)이란 무엇인가?

동일한 여러 개의 모델을 독립적으로 학습시킨 후 다수결로 최종 예측을 하는 앙상블 기법의 하나로 중복을 허용하는 복원 추출 방법인 bootstrapping으로 데이터를 샘플링 한다. Random Forest 알고리즘이 Bagging 방식을 사용한다.

RandomForestClassifier(n_estimators=100, bootstrap=True, randomState=42)

- n_estimators=100 : 100개의 결정트리 사용
- bootstrap=True : 복원추출기법 사용 (과적합 방지)
- randomState=42 : 랜덤시드 고정

20. 주성분 분석(PCA)이란 무엇인가?

차원 축소 기법이다. 데이터에서 분산이 가장 큰 정보를 찾고 (데이터의 정보량을 최대한 유지하기 위함이다.) 그 정보를 새로운 축으로 변환하여 데이터의 차원을 줄인다.

PC1 → 가장 큰 분산을 갖는 방향

PC2 → PC1과 직교하면서 그 다음 큰 분산을 갖는 방향

★ 2차원으로 줄임으로써 시각화가 가능하다. feature를 줄임으로써 노이즈가 제거된다.