

(1)

1. 인공지능에서 기능에 해당하는 기능은 무엇인가?
→ 인간의 지능을 도량하는 컴퓨터의 지능이다. 인간의 학습능력, 흡수능력, 추론능력을 컴퓨터 프로그램으로 구현한 것이다.

2. 인공지능의 종류 3가지에 대해서 설명하세요.

1) 지도학습: 분류(Classification) → 주어진 데이터를 기반으로 기대치와 같은 예측치를 예측하는 문제이다.
의전 분류 문제 혹은 다른 분류 문제라고 한다.
회귀(Regression) → 어떤 데이터들의 Feature를 기준으로 연속된 값(예측치)을 예측하는 문제이다.
여전 회귀에는 트리, 회귀를 사용한다. 단지 예측이 정답과 일치하지 않고 어떤 수나 선형은 예측된다.
분류에는 KNN, Naive Bayes, Support vector, Machine learning이 있고,
회귀에는 Linear Regression, Locally weighted linear, Ridge, Lasso 등이 있다.

2) 비지도학습: 지도학습과는 달리 정답 데이터가 없는 데이터를 바탕으로 특징끼리 관계를 찾는다.
예로는 데이터에 대해 정답을 예측하는 방법을 Unsupervised learning이라고 한다.
예시로, K-means, Clustering, Density Estimation, Expectation Maximization, Parzen windows,
DBSCAN 등이 있다.

3) 강화학습: 데이터를 존재하는 것에 대해 고지하거나 그렇지 않은 경우에 행동을 통해 학습하는 것을 의미함.
자신이 한 행동에 대해 보상을 받으면 학습하는 것을 의미함.
개념으로는 에이전트, 환경, 상태, 행동, 보상이 있다.

3. 전통적인 프로그래밍 방식과 인공지능 프로그래밍의 차이점?
전통적인 프로그래밍 방식은 프로그래머가 명시적으로 지침을 제공하는 반면에
인공지능에서는 프로그래머가 문제 자체를 제공하고, 컴퓨터가 방식 자체로 문제를 해결한다.

4. 딥러닝은 머신러닝과 어떤 차이를 보이는가?

딥러닝은 머신러닝에 속하는 한 분야이고, 머신러닝은 전통적인 알고리즘을 사용해서
수동으로 각 데이터의 특징을 알아내고, 다양한 학습방식으로 이를 관리하는 반면에, 딥러닝은
자동 학습방법 기반의 모델을 사용해서 데이터의 특징을 자동으로 알아내고, 학습방법은 딥러닝을 가장 많이
사용한다. 딥러닝의 성능면에서 더 우월하다.

5. Classification과 Regression의 주된 차이점?

Classification은 예측값으로 연속적인 값을 출력하고, Regression은 예측값으로 이산적인 값을 출력한다.

6. 차운의 차운인?

→ 데이터의 차운이 증가할 수록 회귀한 모델의 수가 급격히 증가해요, 예전에는 1000이었어, 그걸 봤을 때 예측 선은 데이터의 유통이 많아.

7. Dimensionality Reduction = 1. 차운이 이득?

→ 비용, 시간, 차운을 아끼고, 고차원의 문제를 없애고, 편리한 회귀를 진행하기 위해서 차운이 줄어든다, 설명하기가 더 쉬워진다.

8. Ridge, Lasso의 공통점

→ Ridge와 Lasso의 공통점으로는 다중회귀라는 점이다. 이는 두개 이상의 특성을 사용한 선형회귀이다.

Pandas의 data frame을 CSV 파일에서 가져와서 다중회귀 모델에 대해서 다룬다.

그리고 이런 선형회귀 모델은 OLS solution을 가진다. 이는 최소 자승법이란으로도 한다.

9. Underfitting, overfitting

→ Underfitting: 모델의 학습 오류를 줄이기 위해선 '성능을' 외쳤다

→ Overfitting: 모델은 학습에 대한 예측에 의존보다는 정부를 외쳤다

10. Feature Engineering vs Feature Selection

→ Feature Engineering은 새로운 특성을 만들거나 선택하는 과정이고, Feature Selection은

기존 Feature 중에서 예측에 도움되는 특성을 선택하는 것이다.

11. 데이터 전처리의 목적 + 예상

데이터 전처리를 하는 이유로는 원시데이터를 분석과 모델링에 적합하도록
정리/구조화하는 행위다. 방범은 데이터 정리, 데이터 변환, 구조화된 작업이라는 세 가지 주로 방범은

방범 가능하다. 그중에서 전처리 정리는 중심 경계값을 대입, 분포가(분포형 추출), 각각(각각의
특성을) 처리하여 특성을 선별하고(여기서 챕터), 그에 따라 기반 값 대입 -
비교 분석 - 유사도와 변환 ... 데이터 변환량이 적은 경우까지 흡수(흡수) - 유사도와 변환량은
변환한다, 혹은 다양한 모델을 여러번 반복해서 대입한다.

**이상치
검출**

이상치 검출은 Variance, Likelihood, Nearest-neighbor, Density, Clustering이 있다.
Variance는 정상분포이면 97.5% 이상, 2.5% 이하에 포함되는 값은 이상으로 판별하고,
Likelihood는 정상 or 이상의 발생 확률을 이용한 방법, Nearest-neighbor는 같은 데이터의 상의
거리로 계산하여 검출, Density는 샘플의 수를 계산하여 같은 가장 큰 데이터를 이상치로 표기,
Clustering 데이터를 여러 Cluster로 분류하고, 같은 크기의 Clusters나, 클러스터간의 거리를 계산하고
반경으로 이상치로 고려

**이상치
제거**

1. 자료값, 상한값 제거 → 자료값보다 작은 미량값으로 대체, 상한값보다 크면 상한값으로 대체
2. 표준화 표준화 (99.7%) ↑ or ↓ 을 이상치로 제거하거나 Outlier
3. 중위수 대체 → 중위수로부터 차이를 계산한 값을 대체
4. 극값제거 → 상위 10%와 하위 10%를 평균으로 대체

작동

작동: Waveform의 일부지만, 예전은 신호가 아님(실제로는 입력되지 않음)

처리

(1) moving average filter, median filter, wavefiltering and spines, Digital filter, pivoting 이 있다.

12. EDA란? 데이터의 분포, 상관관계
EDA는 수치화하고 시각화를 사용하여 데이터를 탐색하고, 변수간 징재적 관계
데이터는 시각화된 데이터로, 통계값을 통해서 데이터의 특성을 분석 가능하고

13. 회귀에서 전편, 기울기의 의미, 임파닝과의 관련성
선형회귀식에서 > 기울기는 x으로 표시하고, y전편은 b로 표시함, 둘의 연관성을 대
주제를 나눠낸다.

28. 지니계수란?

→ 불순도를 측정하는 지표이며, 대푯값의 품질적 분산정도로 정량화해서 흥미한 값이다.
 $G(S) = 1 - \sum_{i=1}^c p_i^2$ (p_i)이거나, '지니계수 < 대푯값의 불순도'이다.

29. 앙상블 기법?

→ 여러개의 개별 모델을 조합하여 최종의 모델로 운영하는 기법이다.

30. 부트스트랩?

→ 확률에서 \hat{S} 같은 표본을 여러번 복원추출하는 것을 의미한다.

31. Bagging 이란?

→ 기존 학습 데이터로부터 랜덤하게 부분집합하여 훈련한 시리즈의 대푯값을 여러개 얻는 것이다
여러 모델을 학습하는 방법

32. PCA란?

→ 여러개의 독립 변수들을 간접영 해석할 수 있는 주된 성분을 분석하는 기법이다.

이는, 전체 변수들의 핵심 특성은 선형이다, 비선형으로 주어질 수 있다.

방법은, 유사성이 낮은 변수 제거, 변수들의 잡다성인 성분을 투영하여 차원 줄이는 법