**< CUAI 4기 BASIC 트랙 야금야금 머신러닝 4회차 >**

공통 교재인 ‘파이썬 머신러닝 완벽 가이드’ 책을 통해 자율적으로 학습하시고,

개념에 대한 질문을 토대로 본인의 답변을 작성해주세요.

야금야금 머신러닝의 모든 질문은 공통 교재로부터 출제됩니다.

답변을 작성하는 과정에서 책을 참고해도 좋고 구글링을 통해 알아오셔도 좋습니다.

다른 Basic 부원분들과 협동해서 풀어도 좋습니다.

다만 답변을 작성하면서 머신러닝 개념들을 본인의 것으로 꼭 만들어 주세요!

|  |  |
| --- | --- |
| 이름 | 홍지중 |
| 학과 | 산업보안학과 |

파일명은 (**야금야금 머신러닝 4회차 Basic\_홍길동)**으로 제출해주세요!

-----------------------------------------------------------------------------------

1) 로지스틱 회귀가 선형 회귀와 다른 점은 무엇일까요?

|  |
| --- |
| 학습을 통해 선형 함수의 회귀 최적선을 찾는 것이 아니라 시그모이드 함수 최적선을 찾고 시그모이드 함수의 반환 값을 확률로 간주해 확률에 따라 분류 |

2) 시그모이드 함수의 정의(수식)를 써주세요.

|  |
| --- |
| y = 1 / 1+e^-x |

3) 회귀 트리와 분류 트리의 차이점은 무엇일까요?

|  |
| --- |
| 리프노드에서 예측 결정 값을 만드는 과정에서 차이가 있음. 분류 트리가 특정 클래스 레이블을 결정하는 것과는 달리 회귀 트리는 리프 노드에 속한 데이터 값의 평균값을 구해 회귀 예측값을 계산 |

4) 회귀 모델을 적용하기 전에 데이터에 대해 처리할 사항 두 가지는 무엇인가요?

|  |
| --- |
| 결과값이 정규분포인지 확인, 원핫인코딩으로 피처 인코딩 |

5) 트리 기반 알고리즘이 분류뿐만 아니라 회귀도 가능하게 해주는

트리 생성 알고리즘은 무엇인가요?

|  |
| --- |
| CART(Classification And Regression Trees) |

6) 컬럼값이 왜곡되어있을 때,

그 값을 정규 분포 형태로 바꾸는 가장 일반적인 방법은 무엇인가요?

|  |
| --- |
| 로그변환 |

7) 데이터 세트의 왜곡된 정도를 추출하기 위해 skew()를 이용할 때,

skew()를 적용하는 숫자형 피처에서 원-핫 인코딩된

카테고리 숫자형 피처를 제외하는 이유는 무엇인가요?

|  |
| --- |
| 카테고리 피처는 코드성 피처이므로 인코딩 시 왜곡될 가능성이 높기 때문이다. |

8) 분류 예측 성능이 뛰어나며 이진 분류의 기본 모델로 사용되고,

텍스트 분류에 자주 사용되는 방법은 무엇인가요?

|  |
| --- |
| 로지스틱 회귀 |

9) 회귀 트리의 동작을 간단하게 설명해주세요.

|  |
| --- |
| 데이터셋의 X피처를 결정 트리 기반으로 분할하면 X값의 균일도를 반영한 지니계수에 따라 분할, 루트 노드를 split 0 기준으로 분할하고 이렇게 분할된 규칙노드에서 다시 split 1과 split 2 규칙 노드로 분할, split2는 다시 재귀적으로 split 3규칙 노드로 분할하며 트리 규칙으로 변환, 리프노드 생성 기준에 부합하는 트리분할이 완료됐다면 리프 노드에 소속된 데이터 값의 평균값을 구해서 최종적으로 리프노드에 결정 값으로 할당l |

10) RMSLE를 수행하는 함수 rmsle()를 만들 때 데이터값의 크기에 따라

오버플로/언더플로 오류가 발생하기 쉬운데,

이를 해결하는 방법과 그 이유는 무엇인가요?

|  |
| --- |
| Log1p()활용, 변환된 로그값에 1을 더해주기 때문이며, 이는 쉽게 원래의 스케일로 변환 가능 |

11) 데이터 세트의 차원이 증가할수록 생기는 문제점은 무엇인가요?

|  |
| --- |
| 차원이 증가할수록 데이터 포인트 간의 거리가 기하급수적으로 멀어져 희소한 구조를 갖게 된다. 그러므로 수백개 이상의 피처로 구성된 경우 적은 차원에서 학습된 것 보다 예측 신뢰도가 떨어진다. 더불어 피처가 많을 경우 피처간 상관관계가 높아져 다중 공선성 문제로 예측 성능이 저하된다. |

12) 차원 축소는 일반적으로 어떤 두 갈래로 나뉘게 되나요?

|  |
| --- |
| 피처 선택과 피처 추출 |

13) PCA(Principal Component Analysis)에 대해 간략히 소개해주세요.

|  |
| --- |
| 여러 변수간 존재하는 상관관계를 이용해 이를 대표하는 주성분을 추출해 차원을 축소 |

14) PCA를 선형대수 관점에서 보면 어떤 의미를 갖나요?

|  |
| --- |
| 입력 데이터의 공분산 행렬을 고유값으로 분해하고, 이게 구한 고유벡터에 입력 데이터를 선형 변환 |

15) PCA에서 공분산 C는 어떤 행렬들로 분해되나요?

|  |
| --- |
| n X n의 직교행렬(P), 정방행렬, P의 전치행렬 |

16) PCA를 import할 때 어떤 라이브러리에서 불러올 수 있나요?

|  |
| --- |
| Sklearn.decomposition |

17) PCA와는 다른 LDA의 특징은 무엇인가요?

|  |
| --- |
| 지도학습의 분류에서 사용하기 쉽도록 개별 클래스를 분별할 수 있는 기준을 최대한 유지하면서 차원 축소 |

18) LDA는 지도학습/비지도학습 중 무엇인가요?

|  |
| --- |
| 지도학습 |

19) SVD가 ‘특이값’ 분해로 불리는 이유는 무엇인가요?

|  |
| --- |
| 행렬 U와 V에 속한 벡터는 특이 벡터이며 모든 특이벡터는 서로 직교하는 성질을 가지기 때문이다. |

20) 넘파이의 SVD 모듈은 무엇인가요?

|  |
| --- |
| Numpy.linalg.svd |

21) Truncated SVD는 SVD와 어떤 점이 다른가요?

|  |
| --- |
| 시그마의 대각원소 중에 상위 몇개만 추출해서 여기에 대응하는 U와 V의 원소도 함께 제거해 더욱 차원을 줄인 형태로 분해 |

22) NMF는 어떤 기법을 지칭하나요?

|  |
| --- |
| 낮은 랭크를 통한 행렬 근사 방식의 변형, 원본 행렬 내의 모든 원소 값이 양수라는게 보장되면 두 개의 기반 양수 행렬로 분해 |

23) SVD와 NMF는 어떤 분야에서 활발하게 사용되나요?

|  |
| --- |
| 이미지 변환 및 압축, 텍스트의 토픽 도출 |