< 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	권송아
학과	소프트웨어학부

파일명은 (**야금야금 머신러닝 4회차 Basic_홍길동)**으로 제출해주세요!

---

1) 로지스틱 회귀가 선형 회귀와 다른 점은 무엇일까요?

시그모이드 함수 최적선을 찾고 시그모이드 함수의 반환 값을 확률로 간주해 확률에 따라 분류를 결정함

2) 시그모이드 함수의 정의(수식)를 써주세요.

1/(1+e^(-x))

3) 회귀 트리와 분류 트리의 차이점은 무엇일까요?

분류 트리가 특정 클래스 레이블을 결정하는 것과 달리 회귀 트리는 리프 노드에 속한 데이터 값의 평균값을 구해 회귀 예측값을 계산함

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

1) 데이터 세트의 X 피처를 결정 트리 기반으로 지니 계수에 따라 분할 2) 리프 노드에 소속된 데이터의 값의 평균값을 구해서 리프 노드에 결정 값으로 할당

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?

로그를 적용해 변환하는 것

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?

인코딩 시 왜곡될 가능성이 높음

8) 분류 예측 성능이 뛰어나며 이진 분류의 기본 모델로 사용되고,

텍스트 분류에 자주 사용되는 방법은 무엇인가요?

로지스틱 회귀

9) 회귀 트리의 동작을 간단하게 설명해주세요.

리프 노드에 속한 데이터 값의 평균값을 구해 회귀 예측값을 계산함

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

오버플로/언더플로 오류가 발생하기 쉬운데,

이를 해결하는 방법과 그 이유는 무엇인가요?

log()대신 log1p()(= 1+log())를 사용 1을 더해줌으로써 언더플로/오버플로 문제를 해결할 수 있음

11) 데이터 세트의 차원이 증가할수록 생기는 문제점은 무엇인가요?

데이터 포인트 간의 거리가 기하급수적으로 멀어지게 되고, 희소한 구조를 가지게 됨

12) 차원 축소는 일반적으로 어떤 두 갈래로 나뉘게 되나요?

피처 선택과 피처 추출

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

여러 변수 간에 존재하는 상관관계를 이용해 이를 대표하는 주성분을 추출해 차원을 축소하는 기법

14) PCA를 선형대수 관점에서 보면 어떤 의미를 갖나요?

입력 데이터의 공분산 행렬을 고유값 분해하고, 이렇게 구한 고유벡터에 입력 데이터를 선형변환 하는 것

15) PCA에서 공분산 C는 어떤 행렬들로 분해되나요?

고유벡터 직교행렬, 고유값 정방행렬, 첫 번째 직교행렬의 전치행렬

16) PCA를 import할 때 어떤 라이브러리에서 불러올 수 있나요?

sklearn.decomposition

17) PCA와는 다른 LDA의 특징은 무엇인가요?

입력 데이터의 결정 값 클래스를 최대한으로 분리할 수 있는 축을 찾음

18) LDA는 지도학습/비지도학습 중 무엇인가요?

지도학습

19) SVD가 '특이값' 분해로 불리는 이유는 무엇인가요?

분해된 행렬 U와 V에 속한 벡터가 특이벡터이기 때문

20) 넘파이의 SVD 모듈은 무엇인가요?

numpy.linalg.svd

21) Truncated SVD는 SVD와 어떤 점이 다른가요?

특이값 중 상위 일부 데이터만 추출해 분해

22) NMF는 어떤 기법을 지칭하나요?

원본 행렬 내의 모든 원소 값이 양수라는 게 보장되면 두 개의 기반 양수 행렬로 분해될 수 있는 기법

23) SVD와 NMF는 어떤 분야에서 활발하게 사용되나요?

토픽 모델링, 추천 시스템