## < CUAI 4기 BASIC 트랙 야금야금 머신러닝 3회차 >

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 3회차 Basic_홍길동)**으로 제출해주세요!

---

1) 회귀 분석이란 무엇인가요?

| |
|---|
| 여러 개의 독립변수와 하나의 종속변수 간의 상관 관계를 모델링하는 기법 |

2) 대표적인 선형 회귀 모델에는 무엇이 있나요?

| |
|---|
| 일반 선형 회귀, 릿지, 라쏘, 엘라스틱넷, 로지스틱회귀 |

3) 실제값과 모델 사이 오류값을 계산할 때 제곱해서 사용하는 이유는 무엇인가요?

미분 등의 계산을 편하게 하기 위해서

4) '데이터를 기반으로 알고리즘이 스스로 학습한다'는 머신러닝 핵심기법은 뭔가요?

경사 하강법

5) 경사하강법에서 학습률을 적용하는 이유는 무엇인가요?

업데이트할 때 쓰이는 비용함수를 편미분 값이 너무 클 수 있기 때문에

6) 경사하강법의 일반적인 프로세스 3단계를 설명해주세요.

(1) w1, w0를 임의의 값으로 설정하고 비용함수 계산 (2) w1, w0의 기울기를 줄여가며 업데이트 (3) 비용함수 값이 감소했으면 2.반복, 감소하지 않으면 반복을 중지하고 그 때의 w1, w0사용

7) 대용량 데이터의 경우 경사하강법을 어떤 방식으로 적용하나요?

확률적 경사 하강법, 미니 배치 확률적 경사 하강법

8) 회귀 평가 지표에는 어떤 것들이 있나요?

MAE, MSE, RMSE, R^2

9) 회귀 평가 지표 중 RMSE에 대해 간략히 소개해주세요.

실제값과 예측값의 차이를 제곱한 것들의 평균을 구해 root를 씌운 것

10) 평가지표 Scoring 함수에 음수를 적용하는 이유는 무엇인가요?

사이킷런의 scoring 함수가 score값이 클수록 좋은 평가 결과로 자동 평가하는데,

오류차가 큰 것은 나쁜 모델을 의미하므로 -를 붙여 의미를 맞춰줌

11) 교재의 코드에서 MSE를 측정하고자 어떤 라이브러리를 임포트해 사용했나요?

sklearn.metrics

12) 다항 회귀가 일반 회귀와 다른 점은 무엇인가요?

회귀가 독립변수의 단항식이 아닌 2,3차 방정식과 같은 다항식으로 표현됨

13) 선형과 비선형 회귀를 나누는 기준은 무엇인가요?

회귀 계수가 선형인가 비선형인가에 따라 나눔

14) 다항 회귀를 사용하는 주된 이유는 무엇인가요?

단순 선형 회귀보다 예측 성능이 높음

15) 다항 회귀에서 과적합이 발생하는 이유는 무엇인가요?

다항회귀의 차수가 높아질수록 학습 데이터에 너무 맞춰진 학습이 이뤄지기 때문

16) 편향-분산 트레이드오프에 대해 간략히 설명해주세요.

한 쪽이 높으면 한 쪽이 낮아지는 경향이 있음

편향이 높으면 분산이 낮아짐(과소적합), 분산이 높으면 편향이 낮아짐(과적합)

17) 편향-분산 그래프에서 과소적합되기 쉬운 부분은 어느 지점인가요?

높은 편향/낮은 분산

18) 최적 모델을 위한 비용 함수의 구성요소에는 무엇이 있나요?

학습데이터 잔차 오류 최소화, 회귀계수 크기제어

19) 규제(Regularization)이란 무엇인가요?

비용함수에 alpha값으로 페널티를 더해 회귀계수의 값을 감소시켜 과적합을 개선하는 방식

20) 릿지(Ridge)와 라쏘(Lasso) 회귀는 어떤 차이가 있나요?

릿지는 L2 Regularization을 이용, 라쏘는 L1 Regularization을 이용

21) 엘라스틱넷(Elastic Net) 회귀는 무엇인가요?

L2 Regularization + L1 Regularization

22) 선형 회귀에서 고려할 중요 사항 3가지는 무엇인가요?

데이터 분포도의 정규화, 인코딩, 최적의 하이퍼 파라미터 찾기

23) 피처 데이터를 정규화하고자 사용할 수 있는 클래스에는 무엇이 있나요?

StandardScaler, MinMaxScaler

24) 타깃값의 경우 로그 변환을 사용하는 이유는 무엇인가요?

많은 사례에서 예측 성능이 향상된 것이 확인됨. 정규분포로 변환하면 변환된 값을 다시 원본 타깃값으로 만들기 어려울 수도 있음.

25) 로그 변환에서 np.log()가 아닌 np.log1p()를 이용하는 이유는 무엇인가요?

np.log()를 사용할 경우 언더 플로우 문제가 발생하기 쉬움