**< 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 | 홍지중 |
| 학과 | 산업보안학과 |

파일명은 (**야금야금 머신러닝 3회차 Basic\_홍길동)**으로 제출해주세요!

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

1) 회귀 분석이란 무엇인가요?

|  |
| --- |
| 여러개의 독립변수와 한개의 종속변수 간의 상관관계를 모델링한 것 |

2) 대표적인 선형 회귀 모델에는 무엇이 있나요?

|  |
| --- |
| 선형회귀, 로지스틱 회귀, 릿지, 라쏘, 엘라스틱 넷 |

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

|  |
| --- |
| RSS – 실제값과 모델 사이의 오류를 계산할때, 음수가 나올 수도 있지만 양수가 나올 수도 있기 때문에 제곱을 사용한다. |

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

|  |
| --- |
| 경사하강법 |

5) 경사하강법에서 학습률을 적용하는 이유는 무엇인가요?

|  |
| --- |
| 경사하강법에서 학습률은 하강하는 보폭을 의미한다. 학습률이 너무 작으면 시간이 오래 걸리고, 골곡이 있는 비용함수의 최소값을 찾을 수 없다. 반면 크게 설정할 경우, 최소값으로 수렴하지 않고 발산한다. |

6) 경사하강법의 일반적인 프로세스 3단계를 설명해주세요.

|  |
| --- |
| 1. w0, w1을 임의의 값으로 설정하고 첫 비용함수의 값을 계산한다.  2. w0, w1을 실제값 – 예측값으로 업데이트 한 뒤 다시 비용함수의 값을 계산한다.  3. 비용함수의 값이 감소했으면 다시 step2를 반복한다. 더 이상 비용 함수의 값이 감소하지 않으면, 그때의 w1 w0를 구하고 반복을 중지한다. |

7) 대용량 데이터의 경우 경사하강법을 어떤 방식으로 적용하나요?

|  |
| --- |
| 대용량 데이터의 경우, 확률적 경사하강법을 이용한다. 전체 입력 데이터로 w가 업데이트되는 값을 계산하는 것이 아니라 일부 데이터에만 이용해 w가 업데이트되는 값을 계산하므로 경사 하강법에 비해서 빠른 속도를 보장한다. |

8) 회귀 평가 지표에는 어떤 것들이 있나요?

|  |
| --- |
| MAE(Means Absolute Error)  MSE(Mean Squered Log Error)  RMSE(Root Mean Square Error)  R 제곱 |

9) 회귀 평가 지표 중 RMSE에 대해 간략히 소개해주세요.

|  |
| --- |
| MSE와 같은 오류의 제곱을 구할 때, 실제 오류 평균보다 더 커니는 특성이 있으므로 MSE에 루트를 씌우는 것 |

10) 평가지표 Scoring 함수에 음수를 적용하는 이유는 무엇인가요?

|  |
| --- |
| Scoring함수에 ‘neg\_mean\_absolute\_error’를 적용해 음수값을 반환한다. 그 이유는 사이킷런의 scoring 함수가 score 값이 클수록 좋은 평가 결과로 장도 평가하기 때문이다. |

11) 교재의 코드에서 MSE를 측정하고자 어떤 라이브러리를 임포트해 사용했나요?

|  |
| --- |
| Sklearn.metrics.mean\_squared\_error |

12) 다항 회귀가 일반 회귀와 다른 점은 무엇인가요?

|  |
| --- |
| 독립변수가 단항식이 아닌 2차, 3차로 표현되는 것을 다항회귀라고 한다. |

13) 선형과 비선형 회귀를 나누는 기준은 무엇인가요?

|  |
| --- |
| 선형회귀와 비선형회귀를 구분하는 것은 회귀 계수의 선형성이다. |

14) 다항 회귀를 사용하는 주된 이유는 무엇인가요?

|  |
| --- |
| 모든 현상을직선으로 표현할 수 없다. 더 복잡하게 나타내면서 최적의 회귀선을 얻을 수 있다. |

15) 다항 회귀에서 과적합이 발생하는 이유는 무엇인가요?

|  |
| --- |
| 실제데이터와 잔차를 줄이는 회귀선을 위해 회귀선의 차수를 변화시킬 수 있다. 하지만, 이때 차수를 높이면 MSE는 줄어들 수 있다. 하지만 과적합의 문제가 발생하기도한다. RSS만 최소화하다 보니 학습데이터에 지나치게 맞추게된다. |

16) 편향-분산 트레이드오프에 대해 간략히 설명해주세요.

|  |
| --- |
| 편향이 높으면 분산은 낮아지고(과소적합), 분산이 높으면 편향이 낮아진다.(과적합) 저편향, 저분산 -> 실제 결과에 매우 잘 접근하면서 예측 변동이 크지 않음 |

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

|  |
| --- |
| 높은 편향, 낮은 분산 |

18) 최적 모델을 위한 비용 함수의 구성요소에는 무엇이 있나요?

|  |
| --- |
| Alpha: 학습 데이터 적합 정도와 회귀 계수 값을 크기 제어를 수행하는 튜닝파라미터  W: 회귀계수 |

19) 규제(Regularization)이란 무엇인가요?

|  |
| --- |
| 비용함수에 alpha 값으로 페널티를 부여해 회귀 계수 값의 크기를 감소시켜 과적합을 개선하는 방식 |

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

|  |
| --- |
| 릿지획귀는 W의 제곱에 대해 패널티를 부여하는 방식, 회귀 계수를 0으로 만들지 않는다.  라쏘는 W의 절대값에 패널티를 부여한다. 회수 계수가 0인 피처는 회귀 식에서 제외되면서 피처 선택의 효과를 얻을 수 있다. |

21) 엘라스틱넷(Elastic Net) 회귀는 무엇인가요?

|  |
| --- |
| L2 규제와 L1 규제를 결합한 회귀이다. RSS(W)+alpha2∗||W||22+alpha1∗||W||1RSS(W)+alpha2∗||W||22+alpha1∗||W||1  식을 최소화 하는 W를 찾는 것이다. 라쏘회귀가 서로 상관관계가 높은 피처들의 경우에 이들 중에서 중요 피처만을 선택하고 다른 피처들은 모두 회귀 계수를 0으로 만드는 성향이 강하지만, 엘라스틱넷은 이에 L2 규제를 넣으면서 완화했다. |

22) 선형 회귀에서 고려할 중요 사항 3가지는 무엇인가요?

|  |
| --- |
| 데이터에 대한 스케일링 정규화 작업 수행  중요한 피저들이나 타겟값의 분포도가 심하게 왜곡됐을 경우 변환 작업을 수행한다. |

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

|  |
| --- |
| StandardScaler, MinMaxScaler |

24) 타깃값의 경우 로그 변환을 사용하는 이유는 무엇인가요?

|  |
| --- |
| 정규분포, 다른 정규값으로 변환하면 원본으로 돌리기 어렵기 때문이다.  왜곡된 분포도 형태의 타겟값을 로그 변환하여 예측 성능 향상이 된 경우가 많은 사례에서 검증 |

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

|  |
| --- |
| X가 0일 경우 y는 -무한대의 값을 가진다. 이러한 경우에 오류 메세지를 방지하기 위해 x+1을 해줘서 0->1로 바꾸어서 문제가 없어지게한다. |