서울대학교 데이터사이언스 대학원 지필고사

제한시간 : 25분

데이터 사이언티스트 A씨는 서울병원의 100명의 인원을 대상으로 암환자를 판별할 수 있는 모델을 고안하기 위해서 다음과 같은 방법으로 분석을 진행하였다. 서울병원에서는 100명의 인원에 대한 2000 개의 feature 를 가지고 있으며, 해당 feature 를 통해 환자의 암 발병 여부를 예측 및 판단 하려고 한다.

$$< X_1, X_2, \cdots X_{2000} > \quad \left(\textit{암환자를 예측하기 위한 } 2000\textit{차원의 } feature\ Matrix\right)$$

$$Y_i \quad (1 \leq i \leq 100, \qquad \textit{암환자면 } y_i = 1, \qquad \textit{그렇지 않으면} y_i = 0\ )$$

데이터 사이언티스트 A씨는 고차원 데이터의 분석에는 어려움이 있다고 판단하여, 100명의 sample 을 이용하여 2000개의 feature 중 Y 와 연관성이 높은 feature 50개를 선택하였다. 선택된 50 개의 feature 만 활용하여 암환자 예측 모델의 학습을 진행하였다.

Q1) 고차원 데이터의 분석에 어떤 어려움이 있는가?

데이터 사이언티스트 A씨는 선택된 50개의 feature 를 이용하여 모델의 학습 및 테스트를 진행하기 위해서 100명의 인원 중 무작위로 10명의 인원을 선택하였다. 선택된 10명의 인원을 test 집단으로 분류하고, 나머지 90명을 train 집단으로 분류하여 학습을 진행하였다. 학습된 모델을 test 집단을 이용하여 성능을 평가한 결과, 이 모델은 매우 높은 정확도를 보였다.

Q2) 해당 분석 방법에 잘못된 점이 무엇이라고 생각하는가?

Q3) 그렇다면 더 나은 분석 방법을 제시하고, 그 이유를 설명하여라.