19101185 강동현 데이터분석 Assignment 06

5. LDA와 QDA의 차이를 살펴보는 문제이다.

(a). Bayes decision boundary가 linear 하다면 training set과 test set에서 어느 방법이 더 좋은 성과를 내는지 답하는 문제이다. 일단 training set 에서는 주어진 데이터에 fitting을 얼마나 잘하는지에 대해 보는 것 이므로 linear 하지 않은 , 즉 이차 함수처럼 구부러진 부분이 등장 할 수 있는 유연한 QDA가 LDA보다 training set 에서는 좋은 성과를 보일 것이다. 하지만 QDA는 fitting을 하는 과정에서 overfit이 발생할 수 있기 때문에 test set에서는 LDA가 QDA 보다는 좋은 성과를 낼 것이다. 이 판단의 근거에는 모두 Bayes decision boundary가 linear 하다는 가정이 들어 있다.

(b). Bayes decision boundary가 non-linear 하다면 training set과 test set에서 어느 방법이 더 좋은 성과를 내는지 답하는 문제이다. 위 (a)번 문제에서 test set에서 LDA가 QDA 보다 좋은 성능을 보여준다고 답한 것은 본래의 Bayes decision boundary가 linear 하다는 가정 때문이었다. 하지만 이 경우 Bayes decision boundary가 non-linear 하므로 test set에서 LDA가 QDA 보다 굳이 좋을 이유가 없다. 따라서 training set과 test set에서 QDA가 LDA 보다 더 좋은 성능을 보여줄 것이다.

(c). 보통 sample size인 n이 증가하면 LDA에 비해 QDA의 prediction accuracy이 증가하는지 , 감소하는지 , 혹은 변하지 않는지에 대해서 묻고 있다. sample size 즉 , n이 증가한다는 것은 training set의 data가 증가한다고도 볼 수 있다. 따라서 n이 증가 함에 따라서 QDA의 유연성 때문에 training data에 더 잘 fit 할 수 있는 것이다. 따라서 LDA에 비해 QDA의 prediction accuracy는 증가할 것이라고 예상된다.

(d). Bayes decision boundary가 linear로 주어졌다. error rate에서 QDA를 사용하는 경우가 LDA를 사용하는 경우보다 좀 더 좋을 것이라고 문제는 이야기 한다. 이유에 대해서는 QDA의 우연성을 근거로 들고 있다. 근거를 들어 참 , 거짓을 밝히는 문제이다. 일단 우리는 Bayes decision boundary가 linear로 주어졌다는 점에 주목해야 한다. 즉 QDA의 유연성이 좋은 방향으로 적용되는지 아닌지를 살펴보아야 하는 것이다. 이 경우 Bayes decision boundary가 linear 이므로 QDA의 유연성은 좋은 방향으로 적용 될 수 없다. 즉 QDA는 overfitting 될 가능성이 있다는 것이다. 따라서 정답은 거짓 이다.

6. 통계 수업의 학생들에 대해서 데이터를 수집 하였다. X1 = 공부한 시간 , X2 = undergrad GPA , Y = A를 받는 것 이다. 문제는 logistic regression을 통해 추정한다. 베타햇0 = -6 , 베타햇1 = 0.05 , 베타햇2 = 1로 문제에서 주어졌다.

(a). 40시간 공부하는 학생이 undergrad GPA에서 3.5로 A를 받을 확률을 구하는 문제이다. 확률을 구하는 함수 P(X) = e^ (베타0 + 베타1X1 + 베타2X2 ) / 1 + e^ (베타0 + 베타1X1 + 베타2X2 ) 이다. 따라서 X1에 40 , X2에 3.5를 대입하고 문제에서 주어진 베타값 들을 활용하면 확률을 구할 수 있다. 정리하면 e^ (-6 + 2 + 3.5 ) / 1 + e^ (-6 + 2 + 3.5 ) 이고 이는 e^( -0.5 ) / 1 + e^( -0.5 ) 이다. 계산하면 0.375 이고 즉 37.5%를 의미한다. 따라서 답은 37.5% 이다.

(b). 문제 (a)에서 얼마나 많이 공부해야 반에서 A를 받을 확률이 50%가 되는지 구하는 문제이다. 이전의 식에서 X1만을 변수로 두고 식을 정리해 계산하면 된다. 주어진 조건에 따라서 P(X1)을 구하면 P(X1) = e^ ( -6 + 0.05X1 + 3.5 ) / 1 + e^ ( -6 + 0.05X1 + 3.5 ) 이다. 따라서 P(X1) = 50% 즉 0.5 라고 두고 방정식을 풀면 X1에 대한 답을 구할 수 있다. e^ ( -6 + 0.05X1 + 3.5 )를 K로 두고 식을 정리하면 0.5= K / 1 + K 이고 이를 정리하면 0.5 + 0.5K = K 즉 0.5K = 0.5 따라서 K = 1 이다. 즉 e^ ( -6 + 0.05X1 + 3.5 ) 가 1 이므로 log를 취해 살펴보면 log(1) = 0.05X1 – 2.5 이고 정리하면 X1 = 50임을 알 수 있다. 따라서 정답은 50시간 이다.