19101185 강동현 데이터분석 Assignment 01

1. (a)부터 (d)를 보고 각 상황의 flexible statistical learning 방법이 inflexible method 보다 더 좋을지 좋지 않을지 , 판단하는 문제이다.

1-(a). Sample size의 크기는 매우 크고 , predictors의 개수는 작은 경우이다. 데이터의 수가 많은 경우는 유연한 statistical learning이 inflexible method 보다 좋은 효과를 보일 것이다. 따라서 flexible method가 더 좋다.

1-(b). Sample size의 크기는 작고 , predictors의 개수는 많은 경우이다. 데이터의 수가 많은 경우에는 flexible method가 더 좋은 효과를 보였지만 , 반대로 데이터의 수가 적은 경우에는 inflexible method가 더 좋은 효과를 보인다. 따라서 inflexible method가 더 좋다.

1-(c). predictors 와 response 간의 관계가 non-linear한 경우이다. non-linear 하다는 것은 비선형적 이라는 뜻이고 , predictors 와 response 간의 관계를 그래프로 나타내었을 때 , 그래프가 비선형적이라는 의미이다. 이런 경우 flexibility를 늘리는 것이 도움이 된다. 따라서 flexible method가 더 좋다.

1-(d). 줄일수 없는 에러 , 즉 Var(입실론)이 너무 큰 경우이다. 이 경우 flexibility가 너무 큰 경우 , 에러 조차도 fitting 해버려 over-fitting이 일어날 수 있다. 따라서 inflexible method가 더 좋다.

2. 각 시나리오에 대해 classification 혹은 regression 문제인지 정의하고 , inference 와 prediction 중 더 어떤 것에 관심있는지 제시한다. 또한 n , p를 제시하는 문제이다.

2-(a). 미국의 500대 기업에 대해 데이터를 수집하고 , 각 회사에 대하 profit, number of employees, industry , CEO salary 를 조사한다. 그리고 CEO salary에 대해 어떤 요소가 영향을 미치는 지 알고 싶어한다. 이 경우 regression에 관한 시나리오이다. 또한 inference를 통해 CEO salary를 조사하려는 시나리오이다. n 즉 데이터 케이스의 경우 미국의 500대 기업이 될 것이고 , p 의 경우는 profit, number of employees, industry가 될 것이다.

2-(b). 새로운 제품을 출시하려 하고 , 20개의 비슷한 제품들에 대해 데이터를 수집하였다. 조사한 내용은 success or failure, price charged for the product, marketing budget, competition price, ten other variables 이다. 이 경우 , classification과 prediction이 필요한 시나리오 이다. n 의 경우 이전에 출시된 비슷한 20개의 제품이 될 것이고 , p 의 경우는 price charged, marketing budget 등등 다른 요소들이 될 것이다.

2-(c). 세계 증시에 따라 USD/Euro 환율에 어떻게 변하는지 예측하는 시나리오다. 따라서 weekly data for all of 2012 를 조사 및 수집하고 , 매주 USD/Euro의 변화를 기록하고 , 미국 , 영국 , 독일 시장의 변화를 기록한다. 이 경우 regression과 prediction이 필요한 시나리오 이다. n 의 경우 2012년의 모든 주의 데이터 일 것이고 , p 의 경우는 미국 , 영국 , 독일 시장의 변화율이 될 것 이다.

6. parametric 과 a non-parametric statistical learning의 차이점에 대해서 설명하는 문제이다. 또한 regression 와 classification에 대하여 parametric approach가 가지는 장단점에 대하여 설명하는 문제이다. Parametric의 경우 함수 f를 가정하기 때문에 , 함수 f를 구하는 문제로 문제를 축소 시킬 수 있다. 반면 non-parametric의 경우 , 함수 f를 가정하지 않으므로 , 문제를 축소시키지 않고 , 이에 따라 많은 수의 데이터가 필요하게 된다. 또한 Parametric의 장점으로는 non-Parametric에 비해 많은 수의 데이터 케이스를 요구하지 않는 다는 것이다. 단점의 경우 f의 flexibility에 따라 over-fitting될 가능성이 존재한다는 것이다.

7. 주어진 데이터 set table을 보고 물음에 답하는 문제이다. X1 = X2 = X3 = 0 일때, K-nearest neighbors를 사용하여 Y를 예측한다고 가정한다.

7-(a). 각 데이터 간 사이의 유클리드 거리를 계산하는 문제이다.

1번 -> 0^2 + 3^2 + 0^2 에 루트를 씌운 값이므로 거리는 3 이다.

2번 -> 2^2 + 0^2 + 0^2 에 루트를 씌운 값이므로 거리는 2 이다.

3번 -> 0^2 + 1^2 + 3^2 에 루트를 씌운 값이므로 거리는 루트 10 이다.

4번 -> 0^2 + 1^2 + 2^2 에 루트를 씌운 값이므로 거리는 루트 5 이다.

5번 -> (-1)^2 + 0^2 + 1^2 에 루트를 씌운 값이므로 거리는 루트 2 이다.

6번 -> 1^2 + 1^2 + 1^2 에 루트를 씌운 값이므로 거리는 루트 3 이다.

7-(b). k=1인 경우에 대해 예측을 묻는 문제이다. 이 경우 X1 = X2 = X3 = 0 에서 가장 가까운점 하나만을 고려하는 케이스 이므로 , 가장 가까운점인 5번 점을 고려하게 된다. 5번 점의 경우 Y = Green 이기 때문에 이경우 , Green으로 예측할 수 있다.

7-(c). k=3인 경우에 대해 예측을 묻는 문제이다. 이경우 X1 = X2 = X3 = 0 에서 순서대로 가까운점 3개를 고려하는 경우이다. 따라서 2번 , 5번 , 6번을 고려하게 되고 2번 , 6번은 Red , 5번은 Green 이므로 , 이 경우 Red가 Green 보다 많기 때문에 Red로 예측된다.

7-(d). 이 문제의 Bayes decision boundary가 매우 비선형적이라면 , k의 best value가 클 것이라고 예측하는지 , 작을 것이라고 예측하는지 , 그렇다면 이유는 무엇인지 묻는 문제이다. 비선형적 Bayes decision boundary를 가질수록 k가 작은 것이 도움이 된다. 왜냐하면 k가 크다면 boundary가 매우 선형적으로 나타나기 때문이다. K가 작은 경우 불규칙한 다각형이 boundary에 나타날 수도 있지만 , 이를 통해 비선형적인 boundary를 잘 나타낼 수 있기 때문에 k의 best value는 작은 값이 될 것이다.