19101185 강동현 데이터분석 Assignment 07

2. Bootstrap sample을 통해 물음에 답하는 문제이다.

(a). 첫번째 관찰된 Bootstrap이 원래 샘플의 j번째 관찰된 Bootstrap이 아닐 확률을 구하는 문제이다. 즉 관찰된 Bootstrap은 첫번째로 순서가 고정되어 있다. 따라서 j번째 Bootstrap과 같을 확률은 j=1이고 이는 전체 Bootstrap수가 N이라고 할때 1/N이다. 따라서 다를 확률은 전체확률에서 같을 확률을 빼면 되므로 다를 확률을 (N-1)/N 이다.

(b). 두번째 관찰된 Bootstrap이 원래 샘플의 j번쨰 관찰된 Bootstrap이 아닐 확률을 구하는 문제이다. 이 역시 a번과 같은 문제이다. 순서가 고정되어 있으므로 j=2인 경우를 구해 전체에서 뺴면 된다. 따라서 다를 확률은 (N-1)/N 이다.

(c). 우리는 위쪽에서 j번째의 관찰에 대해서 같지 않을 확률이 (N-1)/N의 확률을 가진다는 것을 알았다. 이 확률이 독립이기 떄문에 전체 N개에 대해서 개수만큼 곱해주면 된다. 따라서 전체 N개이기 때문에 Bootstrap 샘플에 j번째 관찰된 요소가 없을 확률을 ((N-1)/N)^N 이다.

(d). n=5일 때 , j번째 관측치가 bootstrap 샘플안에 있을 확률을 구하는 문제이다. 우리는 위쪽에서 없을 확률을 구했으므로 전체 즉 1 에서 없을 확률을 빼면 된다. 없을 확률은 (4/5)^5 이고 1 – (4/5)^5 = 0.67232 이다. 따라서 같을 확률은 0.67232이다.

(e). n=100일 때 , j번째 관측치가 bootstrap 샘플안에 있을 확률을 구하는 문제이다. 위쪽에서 푼것과 같이 n=100을 대입하여 확률을 구한후 전체 확률에서 빼면 구할 수 있다. 없을 확률은 (99/100)^100 이고 1 – (99/100)^100 = 약 0.63이다. 따라서 확률은 0.63이다.

(f). 이번에는 n=10000인 경우이다. 이 역시 같은 과정으로 풀면 1 – (9999/10000)^10000 이고 , 이는 약 0.6321 이다. 따라서 답은 0.6321 이다.

3-(b). k-fold의 장점과 단점을 각 경우에 따라 설명하는 문제이다.

i) The validation set approach에 관하여 장점과 단점을 설명하는 문제이다. The validation set approach의 경우 다른 방법보다 구현하기 쉽다는 장점이 있다. 하지만 추정된 test 에러는 나누어진 set들에 따라서 매우 다를 수 있다는 단점이 있다.

ii) LOOCV의 장점으로는 여러 번 validation을 진행하기 떄문에 정확도가 높고bias가 낮다는 점이 있다. LOOCV의 단점으로는 매우 많은 수의 validation을 진행하기 때문에 다른 방법들보다 계산 복잡도가 매우 높은 점이 단점이다.