저희는 데이터 학습에 도움이 될 수 있는 추가적 데이터 column을 생성하기 위해 유입량과 E지역의 수위데이터 간 관계성에 집중하였습니다. E지역에 대한 수위 데이터를 추세에 따라 클러스터링한 결과를 데이터에 포함시키면 유입량의 전체적 추세를 더욱 원활히 예측할 수 있을지 알아보고자 하였습니다. E지역 수위 데이터를 각 홍수사상번호 별 25개의 묶음으로 만들고, 이 25개의 묶음 중 어느 묶음 끼리 유사한 추세를 보이는지 클러스터로 묶는 것을 목적으로 하였습니다.

따라서, 클러스터의 개수 K 값이 3부터 9까지의 수 중 어느 값이 제일 알맞을지 확인하고, 정해진 클러스터 값에 따라 one-hot 인코딩으로 라벨링하여 데이터 column을 생성하는 과정을 거쳤습니다.

-  
클러스터링을 진행할 때는 tslearn.clustering 모듈을 이용하였으며, k-means clustering을 이용하였습니다. 이때 이용한 metric은 ‘soft DTW’인데, 기존 유클리디언 distance와는 달리 각 홍수사상번호에 따라 제공되는 시계열 데이터의 길이가 다르더라도 패턴에 따라 유사성을 확인할 수 있다는 강점을 가지고 있어 이를 사용하게 되었습니다.

\_

클러스터링에 적당한 k값을 선정하는 지표로 실루엣 점수를 이용하였습니다.

실루엣 점수는 1에서 -1까지의 수치를 가지는데, +1에 가까우면 자신이 클러스터 안에 잘 속해있고 다른 클러스터와는 멀리 떨어져 있다는 뜻입니다. 0과 가까울 수록 클러스터의 경계에 데이터가 위치해 있다는 뜻이고, -1에 가까우면 샘플이 잘못된 클러스터에 할당되었다는 뜻입니다. 위 그래프에 따라 k=3 또는 k=4일때 비교적 클러스터링이 잘 되었고, k=5, k=6 또한 클러스터 개수로 괜찮은 값임을 알 수 있습니다.

\_

K=3. 4. 5. 6일 때 각 클러스터링 결과를 첨부해보았습니다.

\_

이 결과에 따라 라벨링된 숫자 column을 더미화시켜 데이터 column에 추가해 학습하는 과정을 거쳐보았습니다. 자세한 성능은 실험결과에서 함께 보여드리도록 하겠습니다.