기존에 feature 들을 고를 때 주로 사용하는 방법에는 feature 들 간의 correlation 을 살펴보고 correlation 이 높게 나오는 feature 들은 독립성이 약하므로 제거를 해주는 방법이 있다. 하지만 이 방법은 다음과 같은 단점이 있다.

1. correlation 이 크다고 할 명확한 기준없이 주관적인 판단에 의존함
2. 두개의 feature 끼리 밖에 비교를 할 수 없음

이 두 문제점을 해결하기 위해 correlation 대신 feature 들 간의 distance 를 고려했다. Distance 를 다룰 때 생기는 이점으로는 Hausdorff distance (또는 linkage distance) 를 이용해 점집합들 간의 거리를 생각해 2번 문제를 해결할 수 있다.

Distance 를 이용해 1번 문제를 해결하기 위해 max-min selection 을 생각했다. Max-min selection 은 TDA 의 한 종류인 persistent homology 의 계산량이 너무 커서 계산량을 줄이기 위한 data approximation 방법 중 하나인 witness complex 를 만들 때 사용하는 방법이다. Witness complex 는 주어진 데이터 점들 중 적정 갯수의 기준점(landmark)을 정한 뒤 Delaunay triangulation 의 방법을 응용해 graded simplicial complex 를 만든 것이다. 이 때 기준점을 잡는 방법에는 크게 random selection 과 max-min selection 이 있다.

Random selection 은 이름에서도 알 수 있듯이 주어진 점들에서 임의로 뽑는 방법이다. Max-min selection 은 최대한 기준점들이 퍼져있도록 뽑는 방법인데, 처음 한 점을 임의로 고정한 뒤 귀납적으로

$$p = \arg\max_{p \in S-L} \min\{d(p,q) : q \in L\}$$

를 계속해서 추가하는 방식으로 기준점들을 잡는다. 여기서 S 는 전체 데이터 집합이고 L 은 이 전 과정에서 뽑아 놓은 기준점 집합을 뜻한다. 다시말해 max-min selection 은 single linkage distance 를 이용해 계속해서 가장 먼 점들을 뽑아나가는 방법이다.

Max-min selection 은 최대한 점들간의 거리가 멀게 기준점들을 뽑아주므로 각 점들의 local density 에 영향을 받지 않는다는 장점이 있지만 기준점으로 outlier 가 뽑힐 가능성이 높다는 치명적인 단점이 있다. 따라서 outlier 를 미리 제거하는 과정이 필수적으로 선행돼야하지만 feature 들을 고를 때는 outlier 를 고려할 필요가 없다.

Feature 들의 거리를 정의하기 앞서 우선 feature 들 간의 scale 을 맞춰 줄 필요가 있다. 따라서 feature 들을 normalize 를 해주었는데 여기서 살펴볼 중요한 사실이 있다. Feature vector $f \in \mathbb{R}^n$ 를 다음과 같이 normalize 했다고 하자.

$$\hat{f} = \frac{f - \mu_f}{\sigma_f}.$$

여기서 μ_f 를 빼주는 과정에서 feature vector 는 $\{(x_1, \ldots, x_n) : x_1 + \cdots + x_n = 0\}$ 인 초평면으로 사영된다. 다시 σ_f 로 그 값을 나눠주는 과정에서 초평면 내의 S^{n-2} 로 사영이 된다. 이 사실을 이용해 feature 들 간의 distance 를 S^{n-2} 위에서의 geodesic distance 인 arc length 로 설정하였다.

Feature 간의 distance 를 arc length 로 설정한 데에는 계산량을 줄이기 위한 목적도 있다. 실제로 arc length 는 두 feature vector 들의 각도를 구해서 얻을 수 있는데 이는 feature vector 의 내적을 한 값에 arccos 을 취해 얻을 수 있다. 여기서 feature vector 의 내적은 행렬곱 알고리즘으로 빠르게 구할 수 있다.

실제로 max-min selection 을 이용해 feature 들을 뽑았을 때 아쉬운 점이 있었다. 성능은 좋았지만 실제 feature vector 들을 2차원으로 축소해 살펴보았을 때 눈에 띄는 cluster 들이 있었는 데 feature 가 하나도 안뽑힌 cluster 들도 존재하였다. 이 때 각 cluster 들은 비슷한 feature 들이 모여있는 것으로 해석해줄 수 있는 데 cluster 가 확실하게 형성되었다는 것을 그 feature 들이 갖고 있는 강한 경향성이 있다는 것으로 보았다. 따라서 이런 경향성들을 위주로 feature 들을 뽑는 clustering selection 방법도 고안해보았다. 방법은 단순하게 single linkage clustering 을 시행한 뒤 각 cluster 마다 대표 feature 를 하나씩 뽑았다.