자료구조 (Homework 2)

Singly Linked List를 이용한 K-Medoids Clustering 알고리즘 구현

제출기한 : 2021. 05. 12., 5p.m. (늦은 제출은 받지 않음)

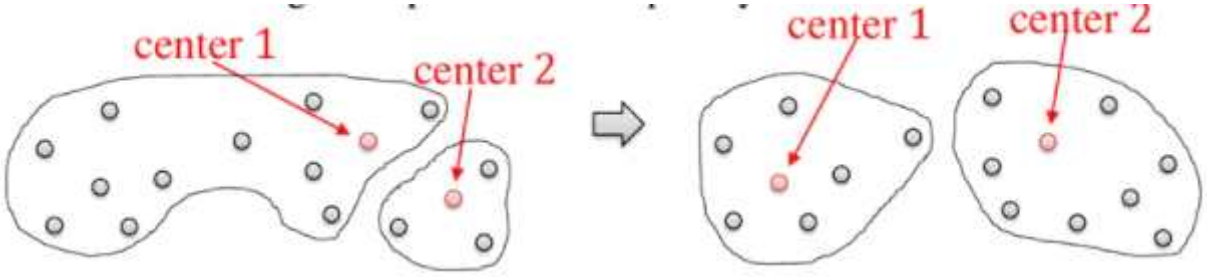
1. K-Medoids Clustering 알고리즘

K-Medoids Clustering은 주어진 데이터셋을 K개의 군집으로 묶는 clustering 알고리즘이다. clustering은 기계학습 알고리즘 중 비지도학습(Unsupervised learning)의 일종으로, 비지도학습은 지도학습과 달리 클래스 레이블(Weighted KNN 숙제에서의 정보(고등학생/유치원))이 존재하지 않는 데이터들을 가지고 유의미한 정보를 도출하는 학습 방법이다.

K-Medoids Clustering 알고리즘은 다음과 같이 동작한다.

1. 임시로 K개의 medoid가 될 데이터들을 주어진 데이터에서 임의로(randomly) 선택한다.
2. 모든 데이터와 K개 각각의 medoids ($m_1, m_2, ..., m_K$)와의 거리를 계산하여 가장 가까운 medoid에 해당 데이터를 할당한다. 즉, m_i과 가장 가까운 데이터들을 m_i의 군집(cluster)으로 할당한다. (K개의 cluster 생성)
3. 각 cluster에 대해 cluster 내 거리의 합이 최소화 되는 데이터를 찾아 그 데이터를 새로운 medoid로 업데이트 한다.
4. 2단계, 3단계를 더 이상 군집 분류에 변화가 없거나 사용자가 정해놓은 maximum iteration 수에 도달할 때 까지 반복한다.

K-Medoids Clustering 알고리즘의 과정을 시각화 한 예시이다. (K=2)



위의 그림에서 각 cluster의 medoid는 center 1, 2로 표시된 데이터이다.

이번 과제에서는 각각의 cluster를 singly linked list 자료구조를 사용하여 구현한다. K값과 데이터셋이 주어졌을 때, 데이터들을 K개의 군집으로 묶어 결과를 output.txt 파일로 출력한다. K개의 singly linked list의 head pointer를 저장하고 있는 pointer array가 있고, 각 linked list는 첫 번째 node에 medoid 정보가, 그 이후의 노드들에 해당 군집에 속하는 데이터들의 정보가 저장되도록 구현해야 한다. 예를 들어, 어떤 군집의 medoid가 (1.5, 3.1)이고, 해당 군집에 속해있는 데이터들이 (1.4, 3.41), (1.12, 3.21), (1.6, 3.56) 이라고 할 때, 해당 군집의 linked list는 [(1.5, 3.1)] -> [(1.4, 3.41)] -> [(1.12, 3.21)] -> [(1.6, 3.56)] 이어야 한다. 만약 알고리즘의 iteration 과정에서 특정 데이터의 군집이 바뀌었다면, 원래 군집 linked list에서 해당 데이터를 삭제한 뒤 새로운 군집의 linked list에 데이터를 추가하는 작업이 필요하다.

2. 문제 설명

- input.txt 파일의 포맷은 다음과 같다.

<데이터 개수> <데이터 feature 개수> <K값>
<첫 번째 데이터의 첫 번째 feature값> ... <첫 번째 데이터의 마지막 feature값>
<두 번째 데이터의 첫 번째 feature값> ... <두 번째 데이터의 마지막 feature값>
...
<마지막 데이터의 첫 번째 feature값> ... <마지막 데이터의 마지막 feature값>

- output.txt 파일의 포맷은 다음과 같아야 한다. (output.txt에 있는 대괄호 표시는 안나와도 무방함.)

<군집 번호(0부터 시작)>
<군집 0의 첫 번째 데이터의 첫 번째 feature값> ... <군집 0의 첫 번째 데이터의 마지막 feature값>
...
<군집 0의 마지막 데이터의 첫 번째 feature값> ... <군집 0의 마지막 데이터의 마지막 feature값>
<군집 번호(1)>
....
<군집 번호(K-1)>
<군집 K-1의 첫번째 데이터의 첫번째 feature값> ... <군집 K-1의 첫번째 데이터의 마지막 feature값>
...
<군집 K-1의 마지막 데이터의 첫번째 feature값> ... <군집 K-1의 마지막 데이터의 마지막 feature값>

- 데이터 개수, 데이터 feature 개수, K값과 군집 번호의 자료형은 int형, 데이터의 feature값은 float형이다.

- linked list에 연결돼있는 데이터들의 순서는 채점시 고려하지 않는다.

- output.txt에서 각 군집별로 출력되는 데이터들의 순서 또한 채점시 고려하지 않는다.

- medoids의 초기값을 랜덤하게 샘플링하기 때문에 clustering의 결과가 달라질 수 있다. 따라서 채점을 위해 코드의 random seed 값을 1000으로 고정한 후 제출한다.

- 코드 뿐만 아니라 K-medoids Clustering 알고리즘, linked list 자료구조와 관련된 코드의 중요한 부분에 대한 설명이 담긴 보고서를 제출한다.

3. 제출

kmedoids_학번.c 파일과 kmedoids_보고서_학번.pdf을 사이버 캠퍼스 상에 제출한다.