Notes 1 – Clustering

Instructor: *Guoqiang Li* $\qquad$ Scribes: *Haotian Wang, Yiyan Li*

1 Introduction

Clustering can be considered the most important unsupervised learning problem; As every other problem of this kind, it deals with finding a structure in a collection of unlabeled data.

Definition 1. *A cluster is a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. Clustering is the algorithm that recognizes clusters from a given data set.*

Note that this is a very rough definition. Part of common application domains in which the clustering problem arises are as follows:

- **Multimedia Data Analysis:** Learning image or video representations without manual annotations. e.g. When using Streaming media platform, face clustering can recognize all the actors in any frame. [1, 2]

- **Responding to public health crises:** With the increasing number of samples, the manual clustering of COVID-19 data samples becomes time-consuming. Clustering helps classify medical datasets deterministically.[3]

- **Social Network Analysis:** Clustering provides an important understanding of the community structure in the network. Results can be used for customer segmentation and sending ads. Put it in a formal way, **clustering groups the nodes of the graph into clusters**, taking into account the edge structure of the graph in such a way that there are several edges within each cluster and very few between clusters. [4]

- **Intermediate Step for other fundamental data mining problems:** Clustering can be considered as a form of data summarization. Many clustering methods are closely related to dimensionality reduction methods. Such methods can be considered a form of data summarization.

- **Intelligent Transportation:** Under **the online scenario**, data is in the form of streams, i.e., the whole dataset could not be accessed at the same time. In future intelligent transportation, low-latency online vehicle tracking is essential and can be solved by online clustering.[5]

Today we'll start from the naive K-means clustering and improve the algorithm step by step. The lecture has X main topics that we'll go through, i.e. TODO AL LAST!!

2   Problem Description

The k-means clustering problem is one of the oldest and most important questions in all of computational geometry. Given an integer k and a set of n data points in $\mathbb{R}^d$, the goal of this problem is to choose k centers so as to minimize the total squared distance between each point and its closest center.[6]

There are several kinds of k-means algorithms among which the most common algorithm , also called naive k-means algorithm, was first proposed by Stuart Lloyd[7] of Bell Labs in 1957.

For a k-means problem, we are given an integer k and a set of data vector $(x_1, x_2, x_3 \ldots x_n)$ in d-dimension. And we need to choose k centroids to partition the n vectors into k types T $(T_1, T_2 \ldots T_k)$ with the minimum within-cluster sum of squares ($WCSS$)

$$WCSS = \arg\min \sum_{i=1}^{k} \sum_{x \in S_i} \|x - c_i\|^2$$

where μ_i is the mean of vector in set S_i.

Lemma 2. *Let S be a set of points with its mean to be μ, and let c to be and arbitrary point. Then* $\sum_{x \in S} \|x - c\|^2 = \sum_{x \in S} \|x - \mu\|^2 + |S| \cdot \|z - \mu\|^2$

So we minimize the function only when $c_i = \mu_i$

$$
\begin{aligned}
WCSS &= \arg\min \sum_{i=1}^{k} \sum_{x \in S_i} \|x - c_i\|^2 \\
&= \arg\min \sum_{i=1}^{k} (\sum_{x \in S_i} \|x - \mu_i\|^2 + |S_i| \cdot \|c_i - \mu_i\|^2) \\
&= \arg\min \sum_{i=1}^{k} |S_i| \cdot Var S_i \\
&= \arg\min \sum_{i=1}^{k} \frac{1}{2 \cdot |S_i|} \sum_{x,y \in S_i} \|x_i - y_i\|^2
\end{aligned}
$$

3   Algorithms

3.1   The K-means algorithm

The k-means algorithm is a simple and fast algorithm for this problem, although it offers no approximation guarantees at all. It iteratively calculates the sum of distance within a cluster and updates the partition. The details are as follows.[**?**]

3.2   Reduce cost of single iteration

In both K-means and K-means++ algorithm, if there are n data points in $\mathbb{R}^d$ space and k clusters for partition, each iteration involves $n * k$ distance computations, which would significantly slow down algorithm. One contribution from Siddhesh Khandelwal[**?**] helps reduce this cost to $n * k'$ $(k' << k)$ by

Algorithm 1 K-means

Input: k: number of output cluster; Data: input data
Output: $\mathcal{S}$: set of all clusters S_i
  Arbitrarily initialize k centroids $C = \{c_1, c_2 \ldots c_k\}$
  repeat
    for each point x in Data S **do**
      for $i = 0 \to k$ **do**
        for $j = 0 \to k$ **do**
          set x to be a member of cluster S_i where $\|x - c_i\|^2 < \|x - c_j\|^2$
        end for
      end for
    end for
    for $i = 0 \to k$ **do**
      $c_i \leftarrow \frac{1}{|S_i|} \sum_{x \in S_i} x$
      set c_i to be the centroid of all points in cluster S_i
    end for
  until S stays unchanaged
Output: $\mathcal{S}$: set of all clusters S_i

---

generating candidate cluster list (CCL) of size k' for each data point. We'll show how this heuristic works in detail.

3.2.1 Inspiration

The optimization is based on the observation that across all iterations of K-means or K-means++, a data point changes its membership only among a small subset of clusters. The heuristic considers only a subset of **nearby cluster as candidates** for deciding membership for a data point. This heuristic has advantage of speeding up K-means and K-means++ clustering with marginal increase in loss functuion(MSE in our case). Note that the optimization can be applied in any algorithm to solve K-means problem with steps to caculate distance between points and centers, acting as augmentation.

3.2.2 Augmentation target

Let A be any variant algorithm of K-means problem and B be the same variant augmented with this geuristic. Let T be the time required for A to converge to MSE value of E. Let T' be the time required for B to converge to MSE value of E'. Our target is:

- **For convergence time:** $T' < T$

- **For loss:** $E' \leq E$ or $E' \overset{marginally}{>} E$

3.2.3 Augmentation detail

We assume that k' is significantly smaller than k. We will show how to choose k' later.We build candidate cluster list (CCL) based on top k' nearest clusters to the data point after first iteration of K-means.

Definition 3. *Cluster centroid is the middle of a cluster. A centroid is a vector that contains one number for each variable, where each number is the mean of a variable for the observations in that cluster. The centroid can be thought of as the multi-dimensional average of the cluster.*

Consider a data point p1 and cluster centroids represented as c_1, c_2..., c_k. We assume that $k' = 4$, and $k' << k$. After first iteration of K-means c_5, c_6, c_8, and c_{11} are the top four closest centroids to p_1 in the increasing order of distance. This is the candidate cluster list for p_1. If we run K-means for second iteration, p_1 will compute distance to all k centroids. After second iteration, there are two possible cases:

1. The list do not change but only members'ranking changes.

2. Several members of the centroids in the previous list are replaced with other centroids which were not in the list.

It seems that the augmentation makes no sense. But what makes this method succeed is **real world data rarely makes case 2 happen**. That is, the set of top few closest centroids for a data point **remains almost unchanged even though order among them might change**.

3.2.4 Augmentation analysis

Overhead analysis:

- **Computation overhead:** $O(nklog(k))$ for creating CCL at first. We have to compute the distance to each cluster's centroid for each point and sort them to create CCL.

- **Memory overhead:**$O(nk')$ to maintain CCL.

4 Key properties

K means problem is an NP Hard problem, and two teams have proved them using 3-SAT and Exact Cover by 3-Sets respectively.[8, 9] Next, I'll try to describe the reduction from 3-SAT to k-means since NP-Complete problem is an inescapable topic in algorithm course and try my best to get rid of copying original material.

4.1 Reduction from 3-SAT to K-means

Let F be the given planar 3-SAT instance with n variables and m clauses. We construct an instance I of planar k-means corresponding to F. Properties of layout I are listed below:

1. Each variable x_i corresponds to a simple circuit s_i in the plane and each circuit has an even number Q of vertices. Each vertex on such a circuit have M copies of a point. Note that M and Q will be stricted below. Now we can partiution

2. b

3. cc

5 Median Trick

So far, we have an algorithm A which estimates in correct range of ϵ with probability ≥ 0.9. Our new algorithm A^* will output in range of ϵ with probability $1 - \delta$. Algorithm:

- Repeat A for $m = O(log(1/\delta))$ times

- Take median of all the m answers.

To prove the correctness, we'll use Chernoff/Hoeffding bounds.

Definition 4 (Chernoff/Hoeffding Bound)**.** *Let X_1, X_2, ..., X_m be independent random variables $\in \{0,1\}$, $\mu = E[\Sigma_i X_i], \epsilon \in [0,1]$. Then $Pr[|\Sigma_i X_i - \mu| > \epsilon\mu] \leq 2e^{-\epsilon^2\mu/3}$*

Define $X_i = 1$ iff the ith answer of A is correct (i.e. estimated value of A lies in correct range).

Claim 5. $E[X_i] = 0.9$, and $E[\mu] = 0.9m$

证明. Since A is correct with probability 0.9, $E[X_i] = 0.9$. And $E[\mu] = 0.9m$ due to linearity of expectation. $\square$

Claim 6. *New algorithm A^* is correct when $\Sigma_i X_i > 0.5m$*

证明. Since we are considering median value to be our answer, if more than half the trials of A are correct, algorithm A^* is also correct. $\square$

Claim 7. *To prove, $Pr[\Sigma_i X_i \geq 0.5m] \geq 1 - \delta$ or $Pr[\Sigma_i X_i < 0.5m] < \delta$*

证明.

$$\begin{aligned}
Pr[\Sigma_i X_i < 0.5m] &= Pr[\Sigma_i X_i - 0.9m < -0.4m] \\
&\leq Pr[|\Sigma_i X_i - \mu| > 0.4m] \\
&= Pr[|\Sigma X_i - \mu| > 0.4/0.9\mu]
\end{aligned} \tag{1}$$

Using Chernoff bound,

$$\begin{aligned}
&\leq e^{-c*0.9m} \\
&< \delta
\end{aligned} \tag{2}$$

Above equation holds for $m = O(log(1/\delta))$ $\square$

6 Distinct Elements

Given, a stream of size m containing numbers from $[n]$, we have to approximate the number of elements with non-zero frequency. To calculate the exact value the space required:

- $O(n)$ bits. (maintain a vector of length n).

- $O(m \log(n))$ bits. (save m numbers, each taking $log(n)$ bits).

Since, this complexity is not feasible as m,n can be very large, we'll look at algorithm for approximating the distinct count value.

6.0.1 Hash Function

- $h : [n] \to [0, 1]$

- $h(i)$ is uniformly distributed in $[0, 1]$.

6.1 Algorithm [Flajolet-Martin 1985]

We maintain a variable z.

1. Initialize $z = 1$.

2. Whenever i is encountered: $z = \min(z, h(i))$

3. When done, output $1/z - 1$.

Now, we'll prove the algorithm works in a similar fashion followed in previous lecture. Let d be number of distinct elements.

Claim 8. $E[z] = d + 1$

证明. z is the minimum of d random numbers in $[0, 1]$. Pick another random number $a \in [0, 1]$. The probability $a < z$:

1. exactly z

2. probability it's smallest among $d + 1$ reals : $1/(d + 1)$

Equating these two, one can prove the claim. $\qquad \square$

Claim 9. $var[z] \leq 2/d^2$

证明. It can be done in a similar fashion described in previous lecture. $\qquad \square$

6.1.1 $(1 + \epsilon)$ approximation Algorithm

We can take $Z = (z_1 + z_2 + ...z_k)/k$ for independent $z_1, ...z_k$

6.2 Alternate Algorithm: Bottom-k

Instead of just use the minimum value of hash function for i inputs, we'll maintain the k smallest hashes seen.

1. Initialize $(z_1, z_2, ...z_k) = 1$.

2. Keep k smallest hashes seen, s.t. $z_1 \leq z_2 \leq ...z_k$

3. When done, output $\hat{d} = k/z_k$

Claim 10. *The following claims are stated:*

- $Pr[\hat{d} > (1 + \epsilon)d] \leq 0.05$

- $Pr[\hat{d} < (1 - \epsilon)d] \leq 0.05$

- *Overall probability that $\hat{d}$ outside range is at most 0.1*

证明. To compute $Pr[\hat{d} > (1 + \epsilon)d]$:

- Define $X_i = 1$ iff $h(i) < \dfrac{k}{(1 + \epsilon)d}$

- Then $\hat{d} > (1 + \epsilon)d$ iff $\Sigma_i X_i > k$

- if $\Sigma_i X_i > k$
  $\iff \exists$ at least k numbers for which $h(i) < \dfrac{k}{(1 + \epsilon)d}$

$$\iff z_k < \frac{k}{(1 + \epsilon)d} \iff \frac{k}{z_k} > (1 + \epsilon)d \iff \hat{d} > (1 + \epsilon)d \tag{3}$$

- $E[X_i] = \dfrac{k}{(1 + \epsilon)d}$

  $E[\Sigma_i X_i] = dE[X_i] = \dfrac{k}{1 + \epsilon}$

  $\text{var}[\Sigma_i X_i] = d\text{var}[X_i] \leq dE[X_1^2] \leq \dfrac{k}{1 + \epsilon} \leq k$

  (Since $X_1 \in \{0, 1\}$, $E[X_1^2] = E[X_i]$)

- By Chebyshev: $Pr[|\Sigma X_i - \dfrac{k}{1 + \epsilon}| > \sqrt{20k}] \leq 0.05 \implies Pr[\Sigma X_i > \dfrac{k}{1 + \epsilon} + \sqrt{20k}] \leq 0.05$

  - (For $\epsilon < 1/2$ and $k = c/\epsilon^2$)
    $\dfrac{k}{1 + \epsilon} + \sqrt{20k} \leq k(1 - \epsilon + \epsilon^2) + \sqrt{20k}$ (Taylor Series Expansion)
    $\leq k - k\epsilon/2 + 5\sqrt{c}/\epsilon = k - c/2\epsilon + 5\sqrt{c}/\epsilon$
    $< k$ where $c > 100$
  - Since $k > \dfrac{k}{1 + \epsilon} + \sqrt{20k}$ in our case and ΣX_i is monotonically increasing, $Pr[\Sigma X_i > k] \leq$
    $Pr[\Sigma X_i > \dfrac{k}{1 + \epsilon} + \sqrt{20k}] \leq 0.05$

$\square$

6.3 Hash functions in stream

The hash function we used has two practical issues: (1) the return value should be a real number. (2) how do we store it?

Discretization can solve the first issue. Instead of all the real numbers in $[0, 1]$, we use hash function with range $\{0, \frac{1}{M}, \frac{2}{M}, \frac{3}{M}, \ldots, 1\}$. For large $M \gg n^3$, the probability that $d \leq n$ random numbers collide is at most $\frac{1}{n}$.

For the second issue, we use pairwise independent function instead of independent function.

Definition 11. $h : [n] \to \{1, 2, \ldots M\}$ *is pairwise independent if for all $i \neq j$ and $a, b \in [M]$, $Pr[h(i) = a \wedge h(j) = b] = \frac{1}{M^2}$*

It works because in previous calculation, we only care about pairs. We defined $X_i = 1$ iff $h(i)$ is small than a threshold, then we computed $\text{var}[\Sigma X_i] = E[(\Sigma X_i)^2] - E[(\Sigma X_i)^2] = E[X_1 X_1 + X_1 X_2 + \ldots] - E[(\Sigma X_i)^2]$. Notice that $E[X_i X_j]$ is the same for fully random h and pairwise independent h.

Example 12 (Construct a pairwise independent hash). *Assume M is a prime number (if not, we can always pick a larger M that is a prime number). We pick $p, q \in \{0, 1, 2, \ldots M - 1\}$ and the hash function $h(i) = pi + q \mod M$. In this construction we only need $O(\log M) = O(\log n)$ space (to store p, q, M).*

证明. $h(i) = a, h(j) = b$ is equivalent to $pi + q \equiv a, pj + q \equiv b$. So $p(i - j) \equiv a - b$ and $p \equiv (a - b)(i - j)^{-1}, q \equiv a - pi$. Since M is a prime number, the unique inverse implies that there is only one pair (p, q) satisfies it. And the probability that pair is chosen is exactly $\frac{1}{M^2}$. $\qquad\square$

7 Impossibility Results

We have used both approximation and randomization to solve the distinct counting problem with space much less than $\min(m, n)$. Now we are wondering: can we omit either approximation or randomization to achieve the same space efficiency? The answer is no.

7.1 Deterministic Exact Won't Work

First, we will show that there is no deterministic (no randomization) and exact (no approximation) way to solve it.

Suppose there do exists a deterministic and exact algorithm A and an estimator function R that use space $s \ll n, m$. That is, for a given integer stream, we first run the algorithm A on the stream. As the stream goes A will return middle memory steps, and we obtain the final memory state σ after the stream ends. Then we apply R on σ to obtain our estimator $\hat{d}$. Since both A and R are deterministic and exact, $\hat{d}$ must equals to the distinct count for the stream.

We now build a binary representation x of the stream with the following rules: (1) $x \in \{0, 1\}^n$, (2) i in stream iff $x_i = 1$. For example, if 1, 3, 5, 6, 7 are in the stream and 2, 4 are not, x will start with 1, 0, 1, 0, 1, 1, 1. Notice that each stream has a corresponding representation and streams containing different numbers have different representations.

Claim 13. *We can recover the x of the stream given the memory state σ*

证明. Denote $d = R(\sigma)$ be the original estimator. Now we treat σ as a middle snapshot of the memory and add integer i as the next element of the stream. Now A will return another memory state σ', and $d' = R(\sigma)'$ will be our new estimator. If $d' = d$, i must have appeared in the stream before since A and R are deterministic and exact. Similarly, if $d' > d$, i must have not appeared in the stream before. Using this method with $i = 1, 2, 3 \ldots$ and we can recover the x. $\qquad\square$

Since we can recover x from σ, we can treat σ as an encoding of a string x of length n. But σ has only $s \ll n$ bits! Furthermore, we can treat A, the function that produces σ, as a function with domain $\{0, 1\}^n$ and $\{0, 1\}^s$. We can see that A must be injective because if $A(x) = A(x') = \sigma$, the recoverability implies $x = x'$.

Hence $s \geq n$. Which implies that there is no deterministic and exact algorithm A and an estimator function R that use space $s \ll n, m$.

7.2 Deterministic Approx. Won't Either

We can use the similar strategy to prove that deterministic approx. won't work. We pick $T \subset \{0,1\}^n$ that satisfies the following conditions: (1) for all distinct $x, y \in T$, the number of digits i that $y_i = 1$ and $x_i = 0$ should $\geq \frac{n}{6}$. (2) $|T| \geq 2^{\Omega(n)}$. Now we use algorithm A to encode an input x into $\sigma = A(x)$ and our estimator would be $\hat{d} = R(\sigma)$.

Now we want to recover x based on σ, as what we have done in the last section. For a given σ and any $y \in T$, we append y to the stream and apply A on it, and A will return a memory state σ'. Using σ' we have new estimator $\hat{d}' = R(\sigma')$.

Claim 14. *If $\hat{d}' > 1.01\hat{d}$, then $x \neq y$, else $x = y$.*

证明. The idea is that when $x = y$, $\hat{d}$ would be really close to $\hat{d}'$ (up to $(1 + \epsilon)^2$ because both of them are ϵ-approximated) and when $x \neq y$, the construction of T guarantee that $\hat{d} \geq \hat{d} + \frac{n}{6}$. So we can pick an ϵ that works for our claim. $\square$

We can use this method to check every element $y \in T$ to see if $y = x$, and eventually we can recover x from it. Similar to last section, we can show that A is an injective function and it implies that $2^s \geq |T|$ or $s = \Omega(n)$.

8 Concluding Remarks

- We can use median trick and Chernoff bound to improve the probability of an existing algorithm.

- For distinct elements problem, we can also store the hashes $h(i)$ approximately. One example is to store the number of leading zeros, and it only cost $O(\log \log n)$ bits per hash value, and that is the idea behind another algorithm called HyperLogLog.

- For the impossibility results, we can also prove that randomized exact algorithm won't work.

参考文献

[1] Z. Wang, L. Zheng, Y. Li, and S. Wang, "Linkage based face clustering via graph convolution network," 2019.

[2] Y. Xing, T. He, T. Xiao, Y. Wang, Y. Xiong, W. Xia, D. Wipf, Z. Zhang, and S. Soatto, "Learning hierarchical graph neural networks for image clustering," 2021.

[3] B. A. Hassan, T. A. Rashid, and H. K. Hamarashid, "A novel cluster detection of covid-19 patients and medical disease conditions using improved evolutionary clustering algorithm star," *Computers in Biology and Medicine*, vol. 138, p. 104866, Nov 2021. [Online]. Available: http://dx.doi.org/10.1016/j.compbiomed.2021.104866

[4] B.-J. Sun, H. Shen, J. Gao, W. Ouyang, and X. Cheng, "A non-negative symmetric encoder-decoder approach for community detection," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, ser. CIKM '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 597–606. [Online]. Available: https://doi.org/10.1145/3132847.3132902

[5] E. Luna, J. C. SanMiguel, J. M. Martínez, and M. Escudero-Viñolo, "Online clustering-based multi-camera vehicle tracking in scenarios with overlapping fovs," 2021.

[6] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding," in *SODA '07*, 2007.

[7] S. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[8] M. Mahajan, P. Nimbhorkar, and K. Varadarajan, "The planar k-means problem is np-hard," in *WALCOM: Algorithms and Computation*, S. Das and R. Uehara, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 274–285.

[9] A. Vattani, "The hardness of k-means clustering in the plane."

Appendix

A K-means Algorithm Code in Python

```python
import math
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np


def loadData():
    df = pd.read_csv("./data/data.csv")
    return df.values


def euclideanDistance(vector1, vector2):
    return math.sqrt(sum(np.power(vector1 − vector2, 2)))


def initRandomCentroids(data, k):
    count, dim = data.shape
    centroids = np.zeros((k, dim))
    colMax = np.max(data, axis=0)
    colMin = np.min(data, axis=0)
    colRange = colMax − colMin
    for i in range(k):
        centroid = colMin + np.random.rand(dim) * colRange
        centroids[i, :] = centroid
    print(centroids)
    return centroids

```

```python
28
29  def kmeans(k):
30      data = loadData()
31      count = data.shape[0]
32      centroids = initRandomCentroids(data, k)
33      clusterBound = np.zeros((count, 2))
34      index = np.zeros((count, 1))
35      processing = True
36      while processing:
37          processing = False
38          for i in range(count):
39              minIndex = 0
40              minDist = float("inf")
41              for j in range(k):
42                  distance = euclideanDistance(centroids[j, :], data[i, :])
43                  if distance < minDist:
44                      minDist = distance
45                      minIndex = j
46
47              if clusterBound[i, 0] != minIndex:
48                  processing = True
49                  clusterBound[i, :] = minIndex, minDist ** 2
50          index[:, 0] = clusterBound[:, 0]
51          for j in range(k):
52              newCentroid = data[np.all(index == j, axis=1), :]
53              centroids[j, :] = np.mean(newCentroid, axis=0)
54      print("k means finished!")
55      visualization(centroids, clusterBound, data)
56
57
58  def visualization(centroids, clusterBound, data):
59      plotMarkList = ['oy', 'og', 'or', 'oc', '^m', '+y', 'sk', 'dw', '<b', 'pg']
60      centroidMarkList = ['Dr', 'Dc', 'Dm', 'Dy', '^k', '+w', 'sb', 'dg', '<r', 'pc']
61      k = centroids.shape[0]
62      count = data.shape[0]
63      if data.shape[1] != 2:
64          print("too many dimensions to draw :(")
65          return
66      if k > len(plotMarkList):
67          print("too many centroids to draw :(")
68          return
69      for i in range(count):
70          mark = plotMarkList[int(clusterBound[i, 0])]
71          plt.plot(data[i, 0], data[i, 1], mark)
```

```python
72        for i in range(k):
73            mark = centroidMarkList[i]
74            plt.plot(centroids[i, 0], centroids[i, 1], mark)
75        plt.show()
76
77
78 if __name__ == "__main__":
79     kmeans(3)
```