COMS 4995: Unsupervised Learning (Summer'18) May 24, 2018

Lecture 2 – Clustering Part II

Instructor: *Nakul Verma* Scribes: *Jie Li, Yadin Rozov*

Today, we will be talking about the hardness results for k-means. More specifically, we will develop tools and complete a proof that the 2-means problem is NP-hard along the lines of [3].

1 k-means – overview

1.1 k-means problem - definition I

The definition of the k-means from the previous class:

- Input: A set of n points $x_1, ...x_n \in \mathbb{R}^d$ and a positive integer $k < n$.
- Output: $T \subset \mathbb{R}$ s.t. $|T| = k$.
- Goal: minimize "cost" of T where: $cost(T) := \sum_{i=1}^{n} \min_{\mu_j \in T} \|x_i - \mu_j\|^2$. $\mu_j = \frac{\sum_{xi \in C_j} x_i}{|C_j|}$ and $C_1, ..., C_k$ are the clusters (specific partition of the n points).

1.2 k-means problem - definition II

Alternative definition of the problem that is more useful for today's proof:

- Input: A set of n points $\mathcal{X} = x_1, ...x_n \in \mathbb{R}$ and a positive integer $k < n$.
- Output:
  (a) $P_1, P_2, ...P_k \subset \mathcal{X}$, "partitions" s.t. $\cup_i P_i = \mathcal{X}$, $P_i \cap P_j = \emptyset$
  (b) $\mu_1, \mu_2, ..., \mu_k$ centroids
- Goal: minimize "cost" of P where cost is defined as:
  (a) $\sum_{j=1}^{k} \sum_{i \in P_j} ||x_i - \mu_j||^2$ where $P_1, ..., P_k$ are the clusters (specific partition of the n points) [k-means cost]

1.3 Observations

- The obvious way to find the optimal solution to k-means is through exhaustive search which is untenable, as that takes a long time and has exponential complexity. While there are only $O(n^k)$ combinations of possible choices for centroids (assuming only points of $\mathcal{X}$ are admissible) there are k^n possible partitions, which for $k = 10$ and $n = 100$ equals the number of atoms in the universe!

- The identity $\mathbb{E}\,||X-Y||^2 = 2\,\mathbb{E}\,||X-\mathbb{E}\,X||^2$ implies that the cost function in the second definition above can be re-written as:

$$\sum_{j=1}^{k}\sum_{i\in P_j}||x_i-\mu_j||^2 = \sum_{j=1}^{k}\frac{1}{2|P_j|}\sum_{i,k\in P_j}||x_i-x_k||^2$$

- The first of these is true because (assuming X and Y to be I.I.D.):

$$\begin{aligned}\mathbb{E}\,||X-Y||^2 &= \mathop{\mathbb{E}}_{x}\mathop{\mathbb{E}}_{y}[||X||^2+||Y||^2-2XY] = \mathop{\mathbb{E}}_{x}||X||^2+\mathop{\mathbb{E}}_{y}||Y||^2-2\mathop{\mathbb{E}}_{x}\mathop{\mathbb{E}}_{y}[XY]\\ &= 2[\mathbb{E}\,X^2-(\mathbb{E}\,X)^2] = 2\,\mathbb{E}[X-\mathbb{E}\,X]^2 = 2\,\mathbb{E}\,||X-\mathbb{E}\,X||^2\end{aligned}$$

- And the second of these is true because by using the first identity and since $\mu_j = \mathbb{E}\,X = \frac{1}{|P_j|}\sum_{x_i\in P_j}x_i$:

$$\sum_{j=1}^{k}\sum_{i\in P_j}||x_i-\mu_j||^2 = \sum_{j=1}^{k}\sum_{i\in P_j}||x_i-\frac{1}{|P_j|}\sum_{x_k\in P_j}x_k||^2 = \sum_{j=1}^{k}\frac{1}{2|P_j|}\sum_{i,k\in P_j}||x_i-x_k||^2$$

1.4 Review of NP-hard problems

For a more complete review of complexity and hardness please go to reference [4] chapter 34.

- problems that are **NP-hard** admit polynomial time reductions from all other problems $\in NP$
- to carry out such a necessary reduction that proves a problem (B below) is **NP-hard** the following steps can used (based off page 1052 from reference [4]):
  (a) Given an instance α of a problem A that has previously been proven to be $\in NP$, use a polynomial time reduction algorithm to transform it to an instance β of problem B
  (b) Run a decision algorithm for B on instance β
  (c) Use the answer for β to get α

1.5 2-means hardness - statement of main theorem and discussion of approach

Theorem 1. *2-means clustering is an NP-hard optimization problem*

Approach to the problem is based on Dasgupta from 2008 [3]. To prove this we will start with the known NP-hard problem of 3SAT and show a reduction from it to the NAE-3SAT* problem. From that problem we will show a reduction to the Generalized 2-means problem and finally show a reduction from that to the 2-means problem. In each reduction as above we need to show how an instance of the known NP-hard problem is polynomially modified cleverly into an instance of the problem we want to show is NP-hard and back (to show that the reduction maps a 'yes' instance of the known problem to a 'yes' instance of the new problem and 'no' instance of the known problem to a 'no' instance of the new problem). Note since we are dealing with 'decision problems', the input of an instance of a problem must include the decision threshold for the problem. We begin by defining the various problems before proving hardness and properties of the reductions.

We'll briefly review NP-completeness, only to the extent necessary to set the stage for this proof. A more thorough treatment can found in a computational complexity course. As a consequence of the Cook-Levin Theorem, which pointed to the first NP-hard problem, we know that SAT and variations, such as 3SAT and NAE 3-SAT, are NP-hard.

1.6 Definitions of various problems required for proving the main theorem

Definition 2 (3SAT)**.**

Input: A Boolean formula in 3CNF-form: a formula of m clauses, each containing 3 literals, connected by 'and' operator.

Output: true if formula is satisfiable, false if not

Definition 3 (NAE 3-SAT)**.** *Not-all-equal-3SAT. A 3SAT formula, with the additional requirement that, in each clause at least one literal is true and at least one literal is false. This removes the case where all three literals in a clause are true.*

Definition 4 (NAE 3-SAT*)**.** *A boolean formula ϕ containing n literals $x_1, ...x_n$. Exactly 3 literals for each of m clauses. Each pair of variables x_i, x_j appears in at most 2 clauses. Once as (x_i, x_j) or $(\neg x_i, \neg x_j)$ and once as $(x_i, \neg x_j)$ or $(\neg x_i, x_j)$*

Definition 5 (Generalized k-means)**.**

Input: nxn matrix, "distance matrix" with elements D_{ij} = distance between object i and object j.

Output: Partition of objects into P_1 and P_2

Goal: minimize cost$(P_1, P_2) = \sum_{j=1}^{2} \frac{1}{2|p_j|} \sum_{i,j \in p_j} D_{ij}$

1.7 Hardness of NAE-3SAT*

Lemma 6. *see [3]*

1.8 Hardness of Generalized 2-means

For any instance ϕ of $x_1, ...x_n$ of NAE-3SAT* we construct a $2n$ x $2n$ distance matrix $D_{\alpha,\beta}$ as below where $\alpha, \beta \in x_1, ...x_n, \neg x_1, ..., \neg x_n$. Note that because the definition of NAE-3SAT* requires that each pair of variables x_i, x_j appears in at most 2 clauses, once as (x_i, x_j) or $(\neg x_i, \neg x_j)$ and once as $(x_i, \neg x_j)$ or $(\neg x_i, x_j)$, the matrix is uniquely defined for a given ϕ.

Definition 7 (Distance matrix for Generalized 2-means - $D(\phi)$)**.**

$$D_{\alpha,\beta} = \begin{cases} 0 & \text{if } \alpha = \beta \\ 1 + \Delta & \text{if } \alpha = \overline{\beta} \\ 1 + \delta & \text{if } \alpha \sim \beta \\ 1 & \text{otherwise} \end{cases} \tag{1}$$

Where $\alpha \sim \beta$ means that either α and β occur together in a clause or α and $\overline{\beta}$ occur together in a clause Where:

$$\Delta = \frac{5m}{5m + 2n} \tag{2}$$

And:

$$\delta = \frac{1}{5m + 2n} \tag{3}$$

Note that above implies that $0 < \delta < \Delta < 1$ and by using algebra we get that:

$$4\delta m < \Delta \leq 1 - 2\delta n \tag{4}$$

Lemma 8. *If ϕ is NAE-3SAT* satisfiable, then $D(\phi)$ admits to a generalized 2-means cost of $cost(\phi) = n - 1 + \frac{2\delta m}{n}$*

Proof. Partition the corresponding matrix object ($2n$ object) for the NAE-3SAT* satisfied ϕ into two partitions; one for all the literals that are assigned *true* and a second for all literals that are assigned *false*. Since each literal is represented twice we have $|P_1| = |P_2| = n$. By definition of the NAE-3SAT*, each clause contributes one pair to P_1 and pair to P_2. Also this leads to the fact that the distances between pairs can only be 1, $1 + \delta$, with m instances of the later and the fact that the two clusters have identical costs. So we get that

$$\begin{aligned} cost(P_1, P_2) &= \sum_{j=1}^{2} \frac{1}{2|P_j|} \sum_{i,j \in P_j} D_{ij} \\ &= \frac{1}{2n}(2\binom{n}{2} + 2m\delta) + \frac{1}{2n}(2\binom{n}{2} + 2m\delta) \\ &= \frac{n(n-1)}{n} + \frac{2m\delta}{n} \\ &= n - 1 + \frac{2m\delta}{n} \end{aligned}$$

□

Lemma 9. *For any partition P_1 and P_2, WLOG P_1 contains a variable and its negation, with $cost(P_1, P_2) \geq n - 1 + \frac{\Delta}{2n} > n - 1 + \frac{2m\delta}{n} = cost(\phi)$.*

Proof. Let $n' = |P_1|$. Note

$$\begin{aligned} cost(P_1, P_2) &\geq \frac{1}{n'}(\binom{n'}{2} + \Delta) + \frac{1}{2n - n'}\binom{2n - n'}{2} \\ &= n - 1 + \frac{\Delta}{n'} \geq n - 1 + \frac{\Delta}{2n} \end{aligned}$$

□

Lemma 10. *If $D(\phi)$ admits a generalized 2-means cost of $cost(\phi) \leq n - 1 + \frac{2\delta m}{n}$, then ϕ is a satisfiable instance of NAE-3SAT*.*

Proof. Let P_1 and P_2 be the partition with cost $\leq n - 1 + \frac{2\delta m}{n}$. First note that P_1 and P_2 do not contain a variable and its negation and $|P_1| = |P_2| = n$. The cost of clustering P_1 and P_2

$$= \frac{2}{n}(\binom{n}{2} + \delta \sum_{clauses} \begin{cases} 1 & \text{if clause is split across } P_1 \text{ and } P_2 \\ 3 & \text{otherwise} \end{cases})$$

Since cost $\leq n - 1 + \frac{2\delta m}{n}$, it follows that all clauses are split between P_1 and P_2. That is, every clause has at least one literal in P_1 and one literal in P_2. Therefore, the assignment that sets all of the P_1 to true and all of P_2 to false is a valid NAE-3SAT* assignment. □

1.9 From Generalized 2-means to 2-means - Embedding of $D(\phi)$

Fact 11. *Note that any $n \times n$ symmetric matrix D can be embedded in l_2^2 iff $u^T D u \leq 0$ for all $u \in \mathbb{R}^n$ s.t. $\sum u_i = 0$.*

Proof. Homework 1 □

Fact 12. *For $D(\phi)$, note*

$$\begin{aligned}
u^T D u &= \sum_{\alpha,\beta} u_\alpha u_\beta D_{\alpha\beta} \\
&= \sum_{\alpha,\beta} u_\alpha u_\beta (1 - \mathbf{1}_{(\alpha=\beta)} + \Delta \mathbf{1}_{(\alpha=\bar{\beta})} + \delta \mathbf{1}_{(\alpha\sim\beta)}) \\
&= \sum_{\alpha,\beta} u_\alpha u_\beta - \sum_{\alpha} u_\alpha^2 + 2\Delta(u^+ \cdot u^-) + \delta \sum_{\alpha,\beta} u_\alpha u_\beta \mathbf{1}_{(\alpha\sim\beta)} \\
&\leq (\sum u_\alpha)^2 - \|u\|^2 + 2\Delta(u^+ \cdot u^-) + \delta \sum_{\alpha,\beta} |u_\alpha||u_\beta|, \text{ and use: } 2ab \leq a^2 + b^2 \\
&\leq -\|u\|^2 + \Delta(\left\|u^+\right\|^2 + \left\|u^-\right\|^2) + \delta(\sum |u_\alpha|)^2 \\
&\leq -(1-\Delta)\|u\|^2 + \delta 2n \|u\|^2, \text{ and since } (1-\Delta) \geq \delta 2n \\
&\leq 0
\end{aligned}$$

1.10 Proof of Theorem 1

Proof. NAE-3SAT* is NP hard from Lemma 6. From Definition 7 and Lemmas 8,9,10 we have that any instance of the NAE-3SAT*, ϕ of $x_1, ...x_n$ can be reduced to an instance of the (decision version of the) Generalized 2-means problem with $D(\phi)$ and threshold $cost(\phi)$. We also have from the Lemma that with these specific instances that NAE-3SAT* is solved, **if and only if** the Generalized 2-means problem is solved. This combined with the fact that the reduction steps take polynomial time in n and Fact 12 that $D(\phi)$ can be embedded into l_2, completes the proof for 2-means. □

References

[1] Gonzalez, F. "Clustering to minimize the maximum intercluster distance." *Theoretical Computer Science 38* (1985): 293-306.

[2] Hartigan, John A. "Clustering Algorithms" *John wiley & sons* (1977).

[3] Sanjoy Dasgupta. "The hardness of k-means clustering" *Department of Computer Science and Engineering University of California, San Diego* (2008): Technical Report CS2008-0916.

[4] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Clifford Stein "Introduction to Algorithms, Third Edition" The MIT Press (2009)