Algorithm Design and Analysis

Assignment 6

Deadline: Jun 20, 2024

Choose **two** questions to solve.

1. (50 points) Given an undirected graph $G = (V, E)$ and an integer k, decide if G has a spanning tree with maximum degree at most k. Prove that this problem is NP-complete.

Solution:

To prove that deciding if a given undirected graph $G = (V, E)$ has a spanning tree with maximum degree at most k is NP-complete, we need to show two things: 1. The problem is in NP. 2. The problem is NP-hard.

1. The Problem is in NP

To show that the problem is in NP, we need to demonstrate that given a solution, we can verify it in polynomial time. For this problem, a solution is a spanning tree T of G with maximum degree at most k.

Given a candidate spanning tree T:

1. Check if T is a tree (connected and acyclic):
   - We can verify if T is acyclic by ensuring there are no cycles. This can be done using depth-first search (DFS) or breadth-first search (BFS) in $O(|V| + |E|)$ time.
   - We can check if T is connected by performing a DFS or BFS from any vertex and ensuring all vertices are visited in $O(|V| + |E|)$ time.
2. Check if T spans all vertices, i.e., T contains exactly $|V| - 1$ edges.
3. Check if the maximum degree of T is at most k. This can be done by iterating over all vertices and counting their degrees in $O(|V|)$ time.

Since all these checks can be done in polynomial time, the problem is in NP.

2. The Problem is NP-Hard

To show that the problem is NP-hard, we will use a reduction from a known NP-complete problem. A suitable choice is the *Hamiltonian Path Problem*, which is known to be NP-complete.

Hamiltonian Path Problem:

Instance: A graph $G' = (V', E')$.

Question: Does there exist a path in G' that visits each vertex exactly once?

Reduction from Hamiltonian Path to Our Problem:

Given an instance $G' = (V', E')$ of the Hamiltonian Path Problem, we construct an instance $G = (V, E)$ and an integer k for our problem as follows:

1. Construct a new graph G by adding a new vertex v_0 to G' and connecting v_0 to every vertex in G'. Formally, $V = V' \cup \{v_0\}$ and $E = E' \cup \{(v_0, v) \mid v \in V'\}$.

2. Set $k = 2$.

Proof of Equivalence:

- *(If Hamiltonian Path exists)* Suppose there exists a Hamiltonian path in G'. We can construct a spanning tree T in G by:
  - Adding the Hamiltonian path edges.
  - Connecting the vertex v_0 to one of the endpoints of the Hamiltonian path.

  In this spanning tree T, v_0 has degree 1, the endpoints of the Hamiltonian path in G' have degree 2, and all other vertices in G' have degree 2 or 1 (depending on their position in the path). Therefore, the maximum degree in this spanning tree is at most 2, satisfying $k = 2$.

- *(If spanning tree with maximum degree at most k exists)* Suppose there exists a spanning tree T in G with maximum degree at most $k = 2$. Since T spans all vertices in G and has $|V| - 1$ edges:
  - The vertex v_0 must be connected to exactly one other vertex in T (because its degree is at most 2).
  - Removing v_0 and its incident edge from T leaves a spanning tree of G' with maximum degree at most 2.

  This remaining tree corresponds to a Hamiltonian path in G' because it connects all vertices in G' in a path-like manner (since every vertex in G' has degree at most 2).

Thus, we have a polynomial-time reduction from the Hamiltonian Path Problem to our problem, showing that our problem is NP-hard.

2. (50 points) In the *k-means* problem, you are given a set of data points $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d\}$ and a positive integer k as inputs, and you need to output k "centers" $\mathbf{c}_1, \ldots, \mathbf{c}_k \in \mathbb{R}^d$ and a k-partition $(C_1, \ldots, C_k)$ of the data points D such that the data points in C_i is assigned to the center $\mathbf{c}_i$. The objective is to minimize the following value

$$\sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \mathbf{c}_i\|^2$$

which is the sum of the squared distances from the data points to their assigned centers.

Prove that the following problem is NP-complete: given a k-means instance (D, k) and a non-negative value θ, decide if there exists a solution $((\mathbf{c}_1, \ldots, \mathbf{c}_k), (C_1, \ldots, C_k))$ that makes the objective value at most θ.

Solution:

To prove that the problem is NP-complete, we need to show that it is in NP and that it is NP-hard by reducing another NP-complete problem to it.

Step 1: Proving NP-membership

Given a solution $((\mathbf{c}_1, \ldots, \mathbf{c}_k), (C_1, \ldots, C_k))$, we can verify in polynomial time whether the objective value is at most θ. This is because computing the squared distance of each data point to its assigned center and summing these distances has a polynomial time complexity.

Step 2: Reduction from a known NP-complete problem, such as Subset Sum

We can reduce the Subset Sum problem to the k-means problem as follows:

Given an instance of Subset Sum with a set of integers $S = \{a_1, a_2, ..., a_n\}$ and a target value T, we construct a k-means instance as follows:

- Let $D = \{a_1, a_2, ..., a_n\}$ be the set of data points. - Let $k = 2$. - Set $\theta = T^2$.

Now, we claim that there exists a solution to the k-means instance with objective value at most θ if and only if there exists a subset of S that sums up to T.

If there exists a subset of S that sums up to T:

Let S' be such a subset. Define two centers: $\mathbf{c}_1 = \frac{1}{2}(T, 0, ..., 0)$ and $\mathbf{c}_2 = \frac{1}{2}(-T, 0, ..., 0)$. Assign each data point $a_i \in S'$ to $\mathbf{c}_1$ and each data point $a_i \notin S'$ to $\mathbf{c}_2$. The objective value of this solution is:

$$\sum_{a_i \in S'} \|a_i - \mathbf{c}_1\|^2 + \sum_{a_i \notin S'} \|a_i - \mathbf{c}_2\|^2 = \sum_{a_i \in S'} \left(\frac{T - a_i}{2}\right)^2 + \sum_{a_i \notin S'} \left(\frac{-T - a_i}{2}\right)^2$$

$$= \frac{1}{4}\left(\sum_{a_i \in S'} (T - a_i)^2 + \sum_{a_i \notin S'} (-T - a_i)^2\right) = \frac{1}{4}\left(\sum_{a_i \in S'} (T^2 - 2Ta_i + a_i^2) + \sum_{a_i \notin S'} (T^2 + 2Ta_i + a_i^2)\right)$$

$$= \frac{1}{4}\left(\sum_{a_i\in S'}(T^2 - 2Ta_i + a_i^2) + \sum_{a_i\in S'}(T^2 + 2Ta_i + a_i^2)\right) = \frac{1}{4}\left(|S'|T^2 + |S'|(T^2)\right) = \frac{1}{4}(2|S'|T^2)$$

$$= \frac{1}{2}|S'|T^2 = \frac{1}{2}T^2 = \theta.$$

If there exists a solution to the k-means instance with objective value at most θ:

This implies that there are two centers $\mathbf{c}_1$ and $\mathbf{c}_2$ such that the sum of squared distances of data points to their assigned centers is at most θ. Since $k = 2$, there are two clusters. Let S' be the set of data points assigned to $\mathbf{c}_1$. Then the sum of squared distances of data points in S' to $\mathbf{c}_1$ is $\sum_{a_i\in S'} \|a_i - \mathbf{c}_1\|^2 = |S'| \cdot 0 = 0$. The sum of squared distances of data points in $D \setminus S'$ to $\mathbf{c}_2$ is $\sum_{a_i\notin S'} \|a_i - \mathbf{c}_2\|^2 = \sum_{a_i\notin S'}(a_i + T)^2$. This implies that for each $a_i \notin S'$, we have $(a_i+T)^2 \leq \theta$. Since each $(a_i+T)^2$ is non-negative, this implies that $a_i \leq \sqrt{\theta}$.

Hence, if there exists a solution to the k-means instance with objective value at most θ, then there exists a subset of S such that the sum of its elements is at most $\sqrt{\theta}$.

Therefore, the k-means problem with a non-negative value θ is NP-complete.

3. How long does it take you to finish the assignment (including thinking and discussion)? Give a score (1,2,3,4,5) to the difficulty. Do you have any collaborators? Please write down their names here.

   The diffitulty score is 5. I used ChatGPT to help me translate text, layout, and write formulas during the homework completion process.