1 Clustering

We have \( n \) objects, each with a number of attributes. We wish to group similar objects into clusters. There is no limit on the number of clusters, or on the size of each cluster. We have a measure \( c_{ij} \) of the difference between two objects \( i \) and \( j \); the larger this measure, the less similar the objects. This measure can take positive or negative values.

We model this by introducing variables
\[
x_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ in same cluster} \\ 0 & \text{if } i \text{ and } j \text{ in different clusters} \end{cases} \quad \text{for } 1 \leq i < j \leq n
\]

Let \( S \subseteq \mathbb{B}^{\binom{n}{2}} \) be the set of feasible solutions. We have the following results regarding \( \text{conv}(S) \):

**Proposition 1** The set \( S \) is full-dimensional.

One way to prove this is to note that the origin and all the unit vectors are in \( S \).

**Proposition 2** The lower bound constraints \( x_{ij} > 0 \) define facets of \( \text{conv}(S) \).

**Proposition 3** Let \( 1 \leq i < j < k \leq n \). The triangle inequalities
\[
-x_{ij} + x_{ik} + x_{jk} \leq 1
\]
define facets of \( \text{conv}(S) \).

These inequalities enforce consistency. For example, the first one says that if \( i \) and \( j \) are in the same cluster and also \( i \) and \( k \) are in the same cluster then \( j \) and \( k \) must be in the same cluster. The only binary solution violating this constraint is \( x_{ij} = x_{ik} = 1, x_{jk} = 0 \).

**Proposition 4** Any binary vector satisfying all the triangle inequalities is the incidence vector of a clustering.

The upper bound constraints \( x_{ij} \leq 1 \) do not define facets of \( \text{conv}(S) \). In particular, if \( x_{ij} = 1 \) then we must also have \( x_{ij} + x_{ik} - x_{jk} = 1 \) and \( x_{ij} - x_{ik} + x_{jk} = 1 \) for each other \( k \).

The following proposition generalizes the lower bound and triangle inequalities.

**Proposition 5** (2-partition inequalities) Let \( U \) and \( W \) be disjoint collections of objects with \( |U| > |W| \). The following inequality defines a facet of \( \text{conv}(S) \):
\[
\sum_{i \in U, j \in W} x_{ij} - \sum_{i \in U, j \in U} x_{ij} - \sum_{i \in W, j \in W} x_{ij} \leq |W|.
\]

This gives the lower bound constraints when \( |U| = 2 \), \( |W| = 0 \). It gives the triangle constraints when \( |U| = 2 \), \( |W| = 1 \).

Note that if all the \( c_{ij} \) are nonnegative then the optimal solution is to place each object in its own cluster, so all \( x_{ij} = 0 \). Thus, our measure \( c_{ij} \) cannot simply be the distance between two objects, but must allow negative values if we are to have an interesting problem.

For more details, see [3, 4].
2 Equipartition

Given a graph \( G = (V, E) \) with \( n = |V| = 2q \) for some integer \( q \), we partition \( V \) into two sets of size \( q \). We define the variables \( x_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ in same partition} \\ 0 & \text{if } i \text{ and } j \text{ in different partitions} \end{cases} \) for \( 1 \leq i < j \leq n \).

Let \( S \) be the set of feasible incidence vectors of equipartitions.

We have the following results:

**Proposition 6** The dimension of \( \text{conv}(S) \) is \( \frac{1}{2}n(n-3) \). If \( C \) is a cycle with \( q + 1 \) vertices then the inequality \( x(E(C)) \leq q - 1 \) is facet defining. If \( U \subseteq V \) with \( |U| \geq 3 \) and odd, the clique inequality \( x(E(U)) \geq \lfloor \frac{1}{2}|U| \rfloor^2 \) is facet-defining.

Other inequalities are known [1, 2].

3 Clustering with lower bound

Now consider a clustering problem where we require each cluster to contain at least \( q \) elements, for some positive integer \( q \). For example, this problem arises in the following settings:

- allocating teams to divisions in a sports league. In this case, often require each division to have the same cardinality.

- microaggregation in the release of data: in order to preserve privacy, clusters with tiny sizes must be avoided.

Let \( S \subseteq \mathbb{B}^\frac{1}{2}n(n-1) \) be the set of incidence vectors of clusterings where each cluster contains at least \( q \) elements. We have the following results regarding \( \text{conv}(S) \):

**Proposition 7** If \( q < n/2 \) then \( \dim(\text{conv}(S)) = \frac{1}{2}n(n-1) \), so \( S \) is full-dimensional.

**Proposition 8** The nonnegativity constraints and the triangle constraints of Proposition 3 define facets of \( \text{conv}(S) \), provided \( q < n/3 \). The 2-partition inequalities of Proposition 5 define facets of \( \text{conv}(S) \) provided \((|W|+2)q < n\).

Other families of valid inequalities are also known [5, 6].

References


