Integer Programming Model for the k-means Problem

Davide Beltrame – 3306906

Problem 2, First Assignment
Algorithms for Optimisation and Inference 2024

We formulate the k-means clustering problem as an IP by precomputing possible clusters and their costs, then selecting the optimal subset of these clusters. This approach reduces computational complexity by considering only ”promising” clusters.

First, we generate a set of candidate clusters. For each combination of up to three colours:

- we compute the centroid as the weighted average of the colours in the cluster;
- we calculate the total cost of the cluster as the sum of weighted squared distances from each colour to the centroid;
- we keep the cluster only if its cost is less than or equal to the best known solution from k-means.

0.1 Variables

- n:= # of unique colours in the image
- P:= # of precomputed candidate clusters
- k:= # of clusters (= 8 in our case)
- c_j:= total cost of using cluster j, for $j = 1, \dots, P$
- C_{ji}:= binary parameter indicating if color i belongs to cluster j (1 if yes, 0 if no)
- $y_j \in \{0, 1\}$: indicates whether cluster j is selected ($y_j = 1$) or not ($y_j = 0$)

0.2 IP Model

The objective is to minimise the total cost of selected clusters:

$$\min \sum_{j=1}^{P} c_j y_j$$

subject to:

$$\sum_{j=1}^{P} C_{ji} y_j = 1, \qquad \forall i = 1, \dots, n \quad \text{(each colour must be covered)} \tag{1}$$

$$\sum_{j=1}^{P} y_j = k \qquad \text{(select exactly } k \text{ clusters)} \tag{2}$$

$$y_j \in \{0, 1\}, \qquad \forall j = 1, \dots, P \tag{3}$$

0.3 Model Explanation

- The objective function minimises the total cost of the selected clusters, where c_j represents the precomputed cost of cluster j.
- Constraint (1): ensures each colour is included in at least one selected cluster.
- Constraint (2) enforces that exactly k clusters are selected.
- Constraint (3) defines the variables as binary.

0.4 Computational Efficiency

This formulation is more computationally tractable than the standard k-means IP model because:

- clusters are precomputed, eliminating the need for assignment variables;
- only promising clusters (those with cost $\leq$ best known solution) are considered;
- the cluster size is limited to at most 3 colours if we want to assume at most 3 colours four our 8 centres (since we have 20 colours, however if we relax this assumption the model works nevertheless);
- the total number of variables is reduced from $O(n^2)$ to $O(P)$, where P is typically much smaller than n^2.