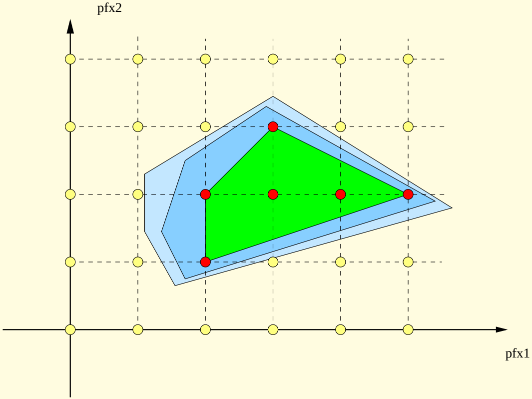
(2/5/2024)

Recap: Formulating IPs



There are different ways to formulate the same integer program:

$$P^1 \supsetneq P^2 \supsetneq P^3 \text{ where as } K^1 = K^2 = K^3$$

where $K^i = P^i \cap \mathbb{Z}^n$ for $i = 1, 2, 3$

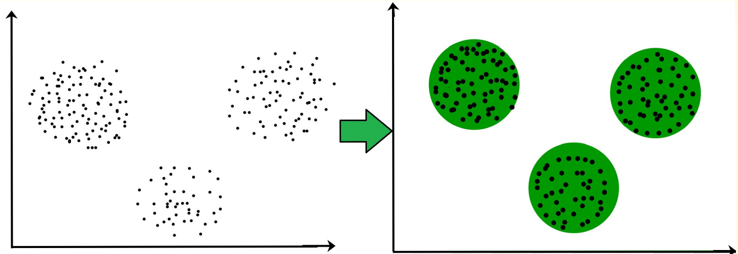
Recap: Clustering problem

Given: An integer $k > 1$ and a collection of points $X = \{x^1, x^2, \ldots\}$ together with distances between pairs of these points.

Goal: Partition X into k clusters $C_1, \ldots, C_k$, such that minimum distance between pairs of points in different clusters:

$$\min_{i \in C_p,\, j \in C_q,\, p \neq q} d(i, j)$$

is maximized. ($d(i, j)$ measures the distance between points i and j)



Lloyd's algorithm: K-Means clustering

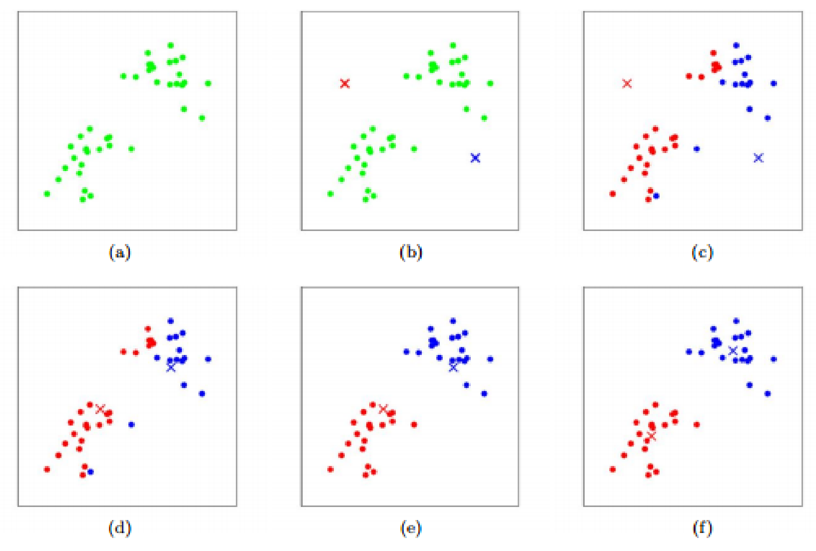
Clustering problem: Partition X into clusters $C_1, \ldots, C_k$ so as to maximize minimum distance between clusters :

$$d^* = \max_{C_1 \ldots, C_k \text{ is a partition}} \left(\min_{i \in C_p,\, j \in C_q,\, p \neq q} d(i,j) \right)$$

- K-Means is the most popular clustering algorithm (it is a heuristic).
  1. Randomly pick k seed points (one for each cluster).
  2. Assign points to the closest seed to form the clusters.
  3. Change the seed points to a "central" point in each cluster
  4. Repeat until clusters do not change much.
  5. Return the best solution found during the search
- Easy to understand and implement.
- It is a good heuristic for the clustering problem (practical performance).

K-Means example

Data points | Initial seeds | A clustering



Clustering when $d(i, j) \in \{0, 1\}$

(points are either similar or dissimilar)

Clustering Problem

Consider the following clustering problem:

- There are n objects $N = \{1, \ldots, n\}$.
- Any pair of objects $i, j \in N$ is either similar or dissimilar

  [$d(i, j)$ is either 0 (if similar) or 1 (if dissimilar)]
- We are given a set D that contains pairs of dissimilar objects (the other pairs are similar)
- We want to cluster the objects in exactly k clusters so that each cluster $C_1, \ldots, C_k$ consists of items that are mostly similar to each other. $K = \{1, \ldots, k\}$
- In addition, each cluster must contain at least ℓ objects.

- Model this as an IP where we the objective is to minimizing the total number of pairs of dissimilar objects put in the same cluster.

Clustering problem

Input:

- n objects numbered $1, 2, \ldots, n$
- Desired number of clusters k, and a lower bound ℓ on the number of objects in a cluster
- A set D of pairs of dissimilar objects

  (i,e. $\{i, j\} \in D$ means that objects i and j are dissimilar)

Output:

- A partitioning of the objects into cluster $C_1, C_2, \ldots, C_k$

  Partitioning means:

  (i) $C_1 \cup C_2 \cup \cdots \cup C_k = \{1, 2, \ldots, n\}$, and

  (ii) $C_s \cap C_t = \varnothing$ for all $s \neq t$.

Goal:

- Minimize the total number of pairs $\{i, j\}$ where i and j are clustered in the same cluster, but are dissimilar (meaning, $\{i, j\} \in D$)

Clustering Problem

Decision variables

$$y_{is} \overset{\text{interpretation}}{=} \begin{cases} 1 & \text{if object } i \text{ is put in cluster } C_s \\ 0 & \text{otherwise.} \end{cases}$$

$$x_{ijs} \overset{\text{interpretation}}{=} \begin{cases} 1 & \text{if both objects } i \text{ and } j \text{ are put in cluster } C_s \\ 0 & \text{otherwise.} \end{cases} \qquad \text{(only for } i < j\text{)}$$

IP Formulation:

$$\begin{array}{llll}
\min & \sum_{\{i,j\}\in D} \sum_{s\in K} x_{ijs} & \longleftarrow & \text{dissimilar pairs in the same cluster} \\
\text{s.t.} & \sum_{s\in K} y_{is} = 1 & \forall i \in N & \longleftarrow \quad \text{objects} \\
& \sum_{i\in N} y_{is} \geq \ell & \forall s \in K & \longleftarrow \quad \text{clusters} \\
& x_{ijs} \in \{0,1\} & \forall i < j \in N, \forall s \in K & \\
& y_{is} \in \{0,1\} & \forall i \in N, \forall s \in K &
\end{array}$$

Clustering Problem

Decision variables

$$y_{is} \overset{\text{interpretation}}{=} \begin{cases} 1 & \text{if object } i \text{ is put in cluster } C_s \\ 0 & \text{otherwise.} \end{cases}$$

$$x_{ijs} \overset{\text{interpretation}}{=} \begin{cases} 1 & \text{if both objects } i \text{ and } j \text{ are put in cluster } C_s \\ 0 & \text{otherwise.} \end{cases}$$

(only for $i < j$)

IP Formulation:

$$\begin{array}{llll} \min & \sum_{\{i,j\} \in D} \sum_{s \in K} x_{ijs} & \longleftarrow & \text{dissimilar pairs in the same cluster} \\ \text{s.t.} & \sum_{s \in K} y_{is} = 1 & \forall i \in N & \longleftarrow \quad \text{objects} \\ & \sum_{i \in N} y_{is} \geq \ell & \forall s \in K & \longleftarrow \quad \text{clusters} \\ & & & \\ & x_{ijs} \in \{0,1\} & \forall i < j \in N, \forall s \in K & \\ & y_{is} \in \{0,1\} & \forall i \in N, \forall s \in K & \end{array}$$

Clustering Problem

Decision variables

$$y_{is} \stackrel{\text{interpretation}}{=} \begin{cases} 1 & \text{if object } i \text{ is put in cluster } C_s \\ 0 & \text{otherwise.} \end{cases}$$

$$x_{ijs} \stackrel{\text{interpretation}}{=} \begin{cases} 1 & \text{if both objects } i \text{ and } j \text{ are put in cluster } C_s \\ 0 & \text{otherwise.} \end{cases} \qquad \text{(only for } i < j\text{)}$$

IP Formulation:

$$\min \quad \sum_{\{i,j\} \in D} \sum_{s \in K} x_{ijs} \quad \longleftarrow \quad \text{dissimilar pairs in the same cluster}$$

$$\text{s.t.} \quad \sum_{s \in K} y_{is} = 1 \qquad \forall i \in N \qquad \longleftarrow \quad \text{objects}$$

$$\sum_{i \in N} y_{is} \geq \ell \qquad \forall s \in K \qquad \longleftarrow \quad \text{clusters}$$

How do we say: $x_{ijs} = \begin{cases} 1 & \text{if } y_{is} = 1 \text{ and } y_{js} = 1 \\ 0 & \text{otherwise.} \end{cases}$

$$x_{ijs} \in \{0, 1\} \qquad \forall i < j \in N, \forall s \in K$$

$$y_{is} \in \{0, 1\} \qquad \forall i \in N, \forall s \in K$$

- How do we say:

$$x_{ijs} = \begin{cases} 1 & \text{if } y_{is} = 1 \text{ and } y_{js} = 1 \\ 0 & \text{otherwise.} \end{cases}$$

using the fact that x and y take $\{0, 1\}$ values?

- First idea: We can write

$$x_{ijs} \geq y_{is} + y_{js} - 1$$

and

$$x_{ijs} \leq \tfrac{1}{2}(y_{is} + y_{js})$$

  - When both $y_{is} = 1$ and $y_{js} = 1$ the we have

$$x_{ijs} \geq 1 + 1 - 1 = 1 \quad \text{and} \quad x_{ijs} \leq \tfrac{1}{2}(1 + 1) = 1 \implies x_{ijs} = 1$$

  - If not, then we must have $y_{is} + y_{js} \leq 1$ and

$$\underbrace{x_{ijs} \geq y_{is} + y_{js} - 1}_{x_{ijs} \geq 0} \quad \text{and} \quad \underbrace{x_{ijs} \leq \tfrac{1}{2}(y_{is} + y_{js})}_{x_{ijs} \leq 1/2} \implies x_{ijs} = 0$$

  - It works! (because $x_{ijs} \in \{0, 1\}$)

- How do we say:

$$x_{ijs} = \begin{cases} 1 & \text{if } y_{is} = 1 \text{ and } y_{js} = 1 \\ 0 & \text{otherwise.} \end{cases}$$

using the fact that x and y take $\{0, 1\}$ values?

- First idea: We can write

$$x_{ijs} \geq y_{is} + y_{js} - 1$$

and

$$x_{ijs} \leq \tfrac{1}{2}(y_{is} + y_{js})$$

  - When both $y_{is} = 1$ and $y_{js} = 1$ the we have

$$x_{ijs} \geq 1 + 1 - 1 = 1 \quad \text{and} \quad x_{ijs} \leq \tfrac{1}{2}(1 + 1) = 1 \implies x_{ijs} = 1$$

  - If not, then we must have $y_{is} + y_{js} \leq 1$ and

$$\underbrace{x_{ijs} \geq y_{is} + y_{js} - 1}_{x_{ijs} \geq 0} \quad \text{and} \quad \underbrace{x_{ijs} \leq \tfrac{1}{2}(y_{is} + y_{js})}_{x_{ijs} \leq 1/2} \implies x_{ijs} = 0$$

  - It works! (because $x_{ijs} \in \{0, 1\}$)

- How do we say:

$$x_{ijs} = \begin{cases} 1 & \text{if } y_{is} = 1 \text{ and } y_{js} = 1 \\ 0 & \text{otherwise.} \end{cases}$$

using the fact that x and y take $\{0,1\}$ values?

- First idea: We can write

$$x_{ijs} \geq y_{is} + y_{js} - 1$$

and

$$x_{ijs} \leq \tfrac{1}{2}(y_{is} + y_{js})$$

  - When both $y_{is} = 1$ and $y_{js} = 1$ the we have

$$x_{ijs} \geq 1 + 1 - 1 = 1 \quad \text{and} \quad x_{ijs} \leq \tfrac{1}{2}(1+1) = 1 \implies x_{ijs} = 1$$

  - If not, then we must have $y_{is} + y_{js} \leq 1$ and

$$\underbrace{x_{ijs} \geq y_{is} + y_{js} - 1}_{x_{ijs} \geq 0} \quad \text{and} \quad \underbrace{x_{ijs} \leq \tfrac{1}{2}(y_{is} + y_{js})}_{x_{ijs} \leq 1/2} \implies x_{ijs} = 0$$

  - It works! (because $x_{ijs} \in \{0,1\}$)

Clustering Problem: Formulation 0

Decision variables

$$y_{is} \overset{\text{interpretation}}{=} \begin{cases} 1 & \text{if object } i \text{ is put in cluster } C_s \\ 0 & \text{otherwise.} \end{cases}$$

$$x_{ijs} \overset{\text{interpretation}}{=} \begin{cases} 1 & \text{if both objects } i \text{ and } j \text{ are put in cluster } C_s \\ 0 & \text{otherwise.} \end{cases} \qquad \text{(only for } i < j\text{)}$$

IP Formulation:

$$\begin{array}{llll}
\min & \sum_{\{i,j\} \in D} \sum_{s \in K} x_{ijs} & & \\
\text{s.t.} & \sum_{s \in K} y_{is} = 1 & \forall i \in N & \longleftarrow \text{ objects} \\
& \sum_{i \in N} y_{is} \geq \ell & \forall s \in K & \longleftarrow \text{ clusters} \\
& x_{ijs} \geq y_{is} + y_{js} - 1 & \forall i < j \in N, s \in K & \\
& x_{ijs} \leq \frac{1}{2}(y_{is} + y_{js}) & \forall i < j \in N, s \in K & \\
& x_{ijs} \in \{0, 1\} & \forall i < j \in N, \forall s \in K & \\
& y_{is} \in \{0, 1\} & \forall i \in N, \forall s \in K &
\end{array}$$

- We want to say:

$$x_{ijs} = \begin{cases} 1 & \text{if } y_{is} = 1 \text{ and } y_{js} = 1 \\ 0 & \text{otherwise.} \end{cases}$$

  using the fact that x and y take $\{0,1\}$ values?

- This is same as

$$x_{ijs} \ = \ y_{is} y_{js}$$

- But we cannot write this in a linear integer program.
- So instead we wrote 2 linear inequalities

$$x_{ijs} \geq y_{is} + y_{js} - 1$$

  and

$$x_{ijs} \leq \tfrac{1}{2}(y_{is} + y_{js})$$

- Question: Can we do better?

Taking a step back: Multiplying binary variables

- Let $y_1 \in \{0, 1\}$ and $y_2 \in \{0, 1\}$ be two binary variables.
- Assume we are interested in their product $y_1 \cdot y_2$.
- How can we express their product $x = y_1 y_2$ using linear inequalities?

$$x = \begin{cases} 1 & \text{if } y_1 = 1 \text{ and } y_2 = 1 \\ 0 & \text{otherwise.} \end{cases}$$

- Consider the following inequalities:

$$\begin{aligned} x &\leq y_1 \\ x &\leq y_2 \\ x &\geq 0 \\ x &\geq y_1 + y_2 - 1 \end{aligned}$$

Claim

If $y_1, y_2 \in \{0, 1\}$ and x satisfies the constraints above, then

$$x = y_1 y_2$$

Claim

If x, y_1, y_2 satisfy the McCormick constraints

$$\begin{aligned} x &\leq y_1, \\ x &\leq y_2, \\ x &\geq y_1 + y_2 - 1, \\ x &\geq 0, \end{aligned}$$

and

$$y_1, y_2 \in \{0, 1\}$$

then $x = y_1 y_2$.

Proof :

$\boldsymbol{y}_1$	$\boldsymbol{y}_2$	constraints		$\boldsymbol{x}$
0	0	$x \leq y_1,$	$x \geq 0$	0
0	1	$x \leq y_1,$	$x \geq 0$	0
1	0	$x \leq y_2,$	$x \geq 0$	0
1	1	$x \geq y_1 + y_2 - 1,$	$x \leq y_1$	1

Back to Formulation 0 for the Clustering Problem

Decision variables

$$y_{is} = \begin{cases} 1 & \text{if } i \text{ in } C_s \\ 0 & \text{otherwise.} \end{cases} \qquad x_{ijs} = \begin{cases} 1 & \text{if both } i \text{ and } j \text{ are in } C_s \\ 0 & \text{otherwise.} \end{cases}$$

(Notice that we want x_{ijs} variable to be equal to $y_{is} \cdot y_{js}$)

IP Formulation:

$$\begin{array}{lrll}
\min & \sum_{\{i,j\} \in D} \sum_{s \in K} x_{ijs} & & \\
\text{s.t.} & \sum_{s \in K} y_{is} = 1 & \forall i \in N & \longleftarrow \quad \text{objects} \\
& \sum_{i \in N} y_{is} \geq \ell & \forall s \in K & \longleftarrow \quad \text{clusters} \\
& x_{ijs} \geq y_{is} + y_{js} - 1 & \forall i < j \in N \text{ , } s \in K & \\
& x_{ijs} \leq \frac{1}{2}(y_{is} + y_{js}) & \forall i < j \in N \text{ , } s \in K & \\
& x_{ijs} \in \{0, 1\} & \forall i < j \in N \text{ , } \forall s \in K & \\
& y_{is} \in \{0, 1\} & \forall i \in N \text{ , } \forall s \in K &
\end{array}$$

Clustering Problem: Formulation 1

Decision variables

$$y_{is} = \begin{cases} 1 & \text{if } i \text{ in } C_s \\ 0 & \text{otherwise.} \end{cases} \qquad x_{ijs} = \begin{cases} 1 & \text{if both } i \text{ and } j \text{ are in } C_s \\ 0 & \text{otherwise.} \end{cases}$$

IP Formulation:

$$\begin{array}{llll}
\min & \sum_{\{i,j\}\in D} \sum_{s\in K} x_{ijs} & & \\
\text{s.t.} & \sum_{s\in K} y_{is} = 1 & \forall i \in N & \longleftarrow \quad \text{objects} \\
& \sum_{i\in N} y_{is} \geq \ell & \forall s \in K & \longleftarrow \quad \text{clusters} \\
& x_{ijs} \geq y_{is} + y_{js} - 1 & \forall i < j \in N, s \in K & \\
& x_{ijs} \leq y_{is} & \forall i < j \in N, s \in K & \\
& x_{ijs} \leq y_{js} & \forall i < j \in N, s \in K & \\
& x_{ijs} \in \{0,1\} & \forall i < j \in N, s \in K & \\
& y_{is} \in \{0,1\} & \forall i \in N, s \in K &
\end{array}$$

Comparing formulations 0 and 1

- Size of the formulation for $n = 40$ objects and $k = 3$ clusters:

	variables	constraints	nonzeros
Formulation 0	2,460	4,723	14,280
Formulation 1	2,460	7,063	16,620

  – y_{is} variables: $40 \times 3 = 120$

  – x_{ijs} variables: $\binom{40}{2} \times 3 = 780 \times 3 = 2340$

- Solution time:

	B& B nodes	Simplex iterations	Solution time
Formulation 0	40,560	11,210,558	327.78 seconds
Formulation 1	23,050	4,033,965	159.24 seconds

  – Form. 1 needs much fewer nodes to solve the IP.

  – Surprisingly, Form. 1 is ≈ 20% faster per B&B node (# of LPs)

  – In both formulations x and y variables are declared binary

  – Form.0 has 2 constraints for each x_{ijs} variable, Form.1 has 3.

Comparing formulations 0 and 1

- Size of the formulation for $n = 40$ objects and $k = 3$ clusters:

	variables	constraints	nonzeros
Formulation 0	2,460	4,723	14,280
Formulation 1	2,460	7,063	16,620

  – y_{is} variables: $40 \times 3 = 120$

  – x_{ijs} variables: $\binom{40}{2} \times 3 = 780 \times 3 = 2340$

- Solution time:

	B& B nodes	Simplex iterations	Solution time
Formulation 0	40,560	11,210,558	327.78 seconds
Formulation 1	23,050	4,033,965	159.24 seconds

  – Form. 1 needs much fewer nodes to solve the IP.

  – Surprisingly, Form. 1 is ≈ 20% faster per B&B node (# of LPs)

  – In both formulations x and y variables are declared binary

  – Form.0 has 2 constraints for each x_{ijs} variable, Form.1 has 3.

Why fewer nodes?

LP relaxation of these two formulations look like:

$$
\begin{array}{llll}
\min & \sum_{\{i,j\}\in D} \sum_{s\in K} x_{ijs} & & \\
\text{s.t.} & \sum_{s\in K} y_{is} = 1 & \forall i \in N & \longleftarrow \quad \text{objects} \\
 & \sum_{i\in N} y_{is} \geq \ell & \forall s \in K & \longleftarrow \quad \text{clusters} \\
 & x_{ijs} \text{ constraints} & \forall i < j \in N\ , s \in K & \\
 & 1 \geq x_{ijs} \geq 0 & 1 \geq y_{is} \geq 0 &
\end{array}
$$

Formulation 0:

$$x_{ijs} \geq y_{is} + y_{js} - 1$$

$$x_{ijs} \leq \frac{1}{2}(y_{is} + y_{js})$$

Formulation 1:

$$x_{ijs} \geq y_{is} + y_{js} - 1$$

$$x_{ijs} \leq y_{is}$$

$$x_{ijs} \leq y_{js}$$

- Now consider a feasible solution to the LP relaxation of F1.

$$\underbrace{(x_{ijs} \leq y_{is}) \quad \text{AND} \quad (x_{ijs} \leq y_{is})}_{\text{solution feasible for F1}} \quad \Rightarrow \quad \underbrace{\left(x_{ijs} \leq \frac{1}{2}(y_{is} + y_{js}\right)}_{\text{also feasible for F0}}$$

- Therefore, Formulation 1 is better as its LP feasible region is smaller.

Why fewer nodes?

LP relaxation of these two formulations look like:

$$\begin{array}{lrlll}
\min & \sum_{\{i,j\}\in D} \sum_{s\in K} x_{ijs} & & & \\
\text{s.t.} & \sum_{s\in K} y_{is} = 1 & \forall i \in N & \longleftarrow & \text{objects} \\
 & \sum_{i\in N} y_{is} \geq \ell & \forall s \in K & \longleftarrow & \text{clusters} \\
 & x_{ijs} \text{ constraints} & \forall i < j \in N\ , s \in K & & \\
 & 1 \geq x_{ijs} \geq 0 & 1 \geq y_{is} \geq 0 & &
\end{array}$$

Formulation 0:

$$x_{ijs} \geq y_{is} + y_{js} - 1$$

$$x_{ijs} \leq \frac{1}{2}(y_{is} + y_{js})$$

Formulation 1:

$$x_{ijs} \geq y_{is} + y_{js} - 1$$

$$x_{ijs} \leq y_{is}$$

$$x_{ijs} \leq y_{js}$$

- Now consider a feasible solution to the LP relaxation of F1.

$$\underbrace{(x_{ijs} \leq y_{is}) \quad \text{AND} \quad (x_{ijs} \leq y_{is})}_{\text{solution feasible for F1}} \quad \Rightarrow \quad \underbrace{(x_{ijs} \leq \frac{1}{2}(y_{is} + y_{js})}_{\text{also feasible for F0}}$$

- Therefore, Formulation 1 is better as its LP feasible region is smaller.

Why fewer nodes?

LP relaxation of these two formulations look like:

$$
\begin{array}{lrll}
\min & \sum_{\{i,j\}\in D}\sum_{s\in K} x_{ijs} & & \\
\text{s.t.} & \sum_{s\in K} y_{is} = 1 & \forall i \in N & \longleftarrow \quad \text{objects} \\
& \sum_{i\in N} y_{is} \geq \ell & \forall s \in K & \longleftarrow \quad \text{clusters} \\
& x_{ijs} \text{ constraints} & \forall i < j \in N\ , s \in K & \\
& 1 \geq x_{ijs} \geq 0 & 1 \geq y_{is} \geq 0 &
\end{array}
$$

Formulation 0:

$$x_{ijs} \geq y_{is} + y_{js} - 1$$

$$x_{ijs} \leq \frac{1}{2}(y_{is} + y_{js})$$

Formulation 1:

$$x_{ijs} \geq y_{is} + y_{js} - 1$$

$$x_{ijs} \leq y_{is}$$

$$x_{ijs} \leq y_{js}$$

- Now consider a feasible solution to the LP relaxation of F1.

$$\underbrace{(x_{ijs} \leq y_{is}) \quad \text{AND} \quad (x_{ijs} \leq y_{is})}_{\text{solution feasible for F1}} \quad \Rightarrow \quad \underbrace{(x_{ijs} \leq \tfrac{1}{2}(y_{is} + y_{js})}_{\text{also feasible for F0}}$$

- Therefore, Formulation 1 is better as its LP feasible region is smaller.

Observation 0: If x, y_1, y_2 satisfy the constraints

$$x \geq y_1 + y_2 - 1$$
$$x \leq \frac{1}{2}(y_1 + y_2)$$
$$y_1, y_2 \in \{0,1\} \text{ and } x \in \{0,1\}$$

then $x = y_1 y_2$. (i.e., $x = 1$ only when both $y_1 = 1$ and $y_2 = 1$)

Without $x \in \{0,1\}$, the point $\underbrace{(1, 0, \frac{1}{2})}_{(y_1, y_2, x)}$ is feasible to the system above.

Observation 1: If x, y_1, y_2 satisfy the constraints

$$x \leq y_1, \quad x \leq y_2,$$
$$x \geq 0, \quad x \geq y_1 + y_2 - 1,$$
$$y_1, y_2 \in \{0,1\}$$

then $x = y_1 y_2$. (Notice that x variable is not declared to be binary)

Observation 0: If x, y_1, y_2 satisfy the constraints

$$x \geq y_1 + y_2 - 1$$
$$x \leq \frac{1}{2}(y_1 + y_2)$$
$$y_1, y_2 \in \{0,1\} \text{ and } x \in \{0,1\}$$

then $x = y_1 y_2$. (i.e., $x = 1$ only when both $y_1 = 1$ and $y_2 = 1$)

Without $x \in \{0,1\}$, the point $\underbrace{(1, 0, \frac{1}{2})}_{(y_1, y_2, x)}$ is feasible to the system above.

Observation 1: If x, y_1, y_2 satisfy the constraints

$$x \leq y_1, \quad x \leq y_2,$$
$$x \geq 0, \quad x \geq y_1 + y_2 - 1,$$
$$y_1, y_2 \in \{0,1\}$$

then $x = y_1 y_2$. (Notice that x variable is not declared to be binary)

Clustering Problem: Formulation 1^+

Decision variables

$$y_{is} = \begin{cases} 1 & \text{if } i \text{ in } C_s \\ 0 & \text{otherwise.} \end{cases} \qquad x_{ijs} = \begin{cases} 1 & \text{if both } i \text{ and } j \text{ are in } C_s \\ 0 & \text{otherwise.} \end{cases}$$

IP Formulation:

$$\begin{array}{llll}
\min & \sum_{\{i,j\}\in D} \sum_{s\in K} x_{ijs} & & \\
\text{s.t.} & \sum_{s\in K} y_{is} = 1 & \forall i \in N & \longleftarrow \quad \text{objects} \\
& \sum_{i\in N} y_{is} \geq \ell & \forall s \in K & \longleftarrow \quad \text{clusters} \\
& x_{ijs} \geq y_{is} + y_{js} - 1 & \forall i < j \in N\ , s \in K & \\
& x_{ijs} \leq y_{is},\ x_{ijs} \leq y_{js} & \forall i < j \in N\ , s \in K & \\
& \cancel{x_{ijs} \in \{0,1\}}\ \ 1 \geq x_{ijs} \geq 0 & \forall i < j \in N\ , s \in K & \\
& y_{is} \in \{0,1\} & \forall i \in N\ , s \in K &
\end{array}$$

Comparing formulations 0, 1 and 1^+

- Size of the formulation for $n = 40$ objects and $k = 3$ clusters:

	variables	constraints	nonzeros
Formulation 0	2,460	4,723	14,280
Formulation 1	2,460	7,063	16,620
Formulation 1^+	2,460	7,063	16,620

  - y_{is} variables: $40 \times 3 = 120$ $\longleftarrow$ binary
  - x_{ijs} variables: $\binom{40}{2} \times 3 = 780 \times 3 = 2340$ $\longleftarrow$ continuous in F 1^+

- Solution time:

	B& B nodes	Simplex iterations	Solution time
Formulation 0	40,560	11,210,558	327.78 seconds
Formulation 1	23,050	4,033,965	159.24 seconds
Formulation 1^+	4,715	842,274	32.87 seconds

  - Formulation 1^+ is 5x faster than Formulation 1, that's pretty good!

Comparing formulations 0, 1 and 1^+

- Size of the formulation for $n = 40$ objects and $k = 3$ clusters:

	variables	constraints	nonzeros
Formulation 0	2,460	4,723	14,280
Formulation 1	2,460	7,063	16,620
Formulation 1^+	2,460	7,063	16,620

  - y_{is} variables: $40 \times 3 = 120$ $\longleftarrow$ binary
  - x_{ijs} variables: $\binom{40}{2} \times 3 = 780 \times 3 = 2340$ $\longleftarrow$ continuous in F 1^+

- Solution time:

	B& B nodes	Simplex iterations	Solution time
Formulation 0	40,560	11,210,558	327.78 seconds
Formulation 1	23,050	4,033,965	159.24 seconds
Formulation 1^+	4,715	842,274	32.87 seconds

  - Formulation 1^+ is 5x faster than Formulation 1, that's pretty good!

Lets look at Formulation 1^+ again

(y_{is} : item i is in cluster C_s, x_{ijs} : both i and j are in cluster C_s)

Formulation 1^+ :

min	$\sum_{\{i,j\}\in D} \sum_{s\in K} x_{ijs}$			
s.t.	$\sum_{s\in K} y_{is} = 1$	$\forall i \in N$	$\longleftarrow$	objects
	$\sum_{i\in N} y_{is} \geq \ell$	$\forall s \in K$	$\longleftarrow$	clusters
	$x_{ijs} \geq y_{is} + y_{js} - 1$	$\forall i < j \in N\ , s \in K$		
	$x_{ijs} \leq y_{is},\ x_{ijs} \leq y_{js}$	$\forall i < j \in N\ , s \in K$		
	$\cancel{x_{ijs} \in \{0,1\}}\ \ 1 \geq x_{ijs} \geq 0$	$\forall i < j \in N\ , s \in K$		
	$y_{is} \in \{0,1\}$	$\forall i \in N\ , s \in K$		

Question: Even if we did not have the constraints $x_{ijs} \leq y_{is},\ x_{ijs} \leq y_{js}$, would $x_{ijs} = 1$ in an optimal sol. if either $y_{is} = 0$ or $y_{js} = 0$?

Lets look at Formulation 1^+ again

(y_{is} : item i is in cluster C_s, x_{ijs} : both i and j are in cluster C_s)

Formulation 1^+ :

$$
\begin{array}{llll}
\min & \sum_{\{i,j\}\in D} \sum_{s\in K} x_{ijs} & & \\
\text{s.t.} & \sum_{s\in K} y_{is} = 1 & \forall i \in N & \longleftarrow \text{ objects} \\
 & \sum_{i\in N} y_{is} \geq \ell & \forall s \in K & \longleftarrow \text{ clusters} \\
 & x_{ijs} \geq y_{is} + y_{js} - 1 & \forall i < j \in N\ , s \in K & \\
 & x_{ijs} \leq y_{is},\ x_{ijs} \leq y_{js} & \forall i < j \in N\ , s \in K & \\
 & \cancel{x_{ijs} \in \{0,1\}}\ \ 1 \geq x_{ijs} \geq 0 & \forall i < j \in N\ , s \in K & \\
 & y_{is} \in \{0,1\} & \forall i \in N\ , s \in K &
\end{array}
$$

Question: Even if we did not have the constraints $x_{ijs} \leq y_{is},\ x_{ijs} \leq y_{js}$, would $x_{ijs} = 1$ in an optimal sol. if either $y_{is} = 0$ or $y_{js} = 0$?

Clustering Problem: Formulation 1^{++}

$$y_{is} = \begin{cases} 1 & \text{if } i \text{ in } C_s \\ 0 & \text{otherwise.} \end{cases} \qquad x_{ijs} = \begin{cases} 1 & \text{if both } i \text{ and } j \text{ are in } C_s \\ 0 & \text{otherwise.} \end{cases}$$

IP Formulation:

$$\begin{array}{llll}
\min & \sum_{\{i,j\} \in D} \sum_{s \in K} x_{ijs} & & \\
\text{s.t.} & \sum_{s \in K} y_{is} = 1 & \forall i \in N & \longleftarrow \quad \text{objects} \\
& \sum_{i \in N} y_{is} \geq \ell & \forall s \in K & \longleftarrow \quad \text{clusters} \\
& x_{ijs} \geq y_{is} + y_{js} - 1 & \forall i < j \in N, s \in K & \\
& \cancel{x_{ijs} \leq y_{is}},\ \cancel{x_{ijs} \leq y_{js}} & \cancel{\forall i < j \in N}, \cancel{s \in K} & \\
& x_{ijs} \geq 0 & \forall i < j \in N, s \in K & \\
& y_{is} \in \{0,1\} & \forall i \in N, s \in K &
\end{array}$$

Note: This formulation allows $x_{ijs} = 1$ even when items i and j are in different clusters but this would never happen in an opt. solution.

Comparing the Formulations 0, 1, 1^+ and 1^{++}

- Size of the formulations:

	variables	constraints	nonzeros
Formulation 0	2,460	4,723	14,280
Formulation 1	2,460	7,063	16,620
Formulation 1^+	2,460	7,063	16,620
Formulation 1^{++}	2,460	2,383	7,260

- Solution time:

	B& B nodes	Simplex iterations	Solution time
Formulation 0	40,560	11,210,558	327.78 seconds
Formulation 1	23,050	4,033,965	159.24 seconds
Formulation 1^+	4,715	842,274	32.87 seconds
Formulation 1^{++}	7,471	542,972	5.41 seconds

  – Form. 1^{++} enumerates more nodes but the speed up is 6x.

  – LPs are now much easier to solve (fewer constraints and non-zeroes).

Comparing the Formulations 0, 1, 1^+ and 1^{++}

- Size of the formulations:

	variables	constraints	nonzeros
Formulation 0	2,460	4,723	14,280
Formulation 1	2,460	7,063	16,620
Formulation 1^+	2,460	7,063	16,620
Formulation 1^{++}	2,460	2,383	7,260

- Solution time:

	B& B nodes	Simplex iterations	Solution time
Formulation 0	40,560	11,210,558	327.78 seconds
Formulation 1	23,050	4,033,965	159.24 seconds
Formulation 1^+	4,715	842,274	32.87 seconds
Formulation 1^{++}	7,471	542,972	5.41 seconds

  - Form. 1^{++} enumerates more nodes but the speed up is 6x.
  - LPs are now much easier to solve (fewer constraints and non-zeroes).

Clustering Problem: Formulation 2

Decision variables

$$y_{is} \overset{\text{interpretation}}{=} \begin{cases} 1 & \text{if object } i \text{ is put in cluster } C_s \\ 0 & \text{otherwise.} \end{cases}$$

$$z_{ij} \overset{\text{interpretation}}{=} \begin{cases} 1 & \text{if objects } i \text{ and } j \text{ are put in the same cluster} \\ 0 & \text{otherwise.} \qquad \text{(only for } i < j\text{)} \end{cases}$$

IP Formulation:

$$\begin{array}{lrll}
\min & \sum_{\{i,j\} \in D} z_{ij} & & \\
\text{s.t.} & \sum_{s \in K} y_{is} = 1 & \forall i \in N & \longleftarrow \quad \text{objects} \\
& \sum_{i \in N} y_{is} \geq \ell & \forall s \in K & \longleftarrow \quad \text{clusters}
\end{array}$$

How to say $z_{ij} = 1$ when both $y_{is}, y_{js} = 1$ for some $s \in K$

$$\begin{array}{rl}
z_{ij} \geq 0 & \forall i < j \in N \\
y_{is} \in \{0, 1\} & \forall i \in N\ , s \in K
\end{array}$$

Clustering Problem: Formulation 2

Decision variables

$$y_{is} \overset{\text{interpretation}}{=} \begin{cases} 1 & \text{if object } i \text{ is put in cluster } C_s \\ 0 & \text{otherwise.} \end{cases}$$

$$z_{ij} \overset{\text{interpretation}}{=} \begin{cases} 1 & \text{if objects } i \text{ and } j \text{ are put in the same cluster} \\ 0 & \text{otherwise.} \qquad \text{(only for } i < j\text{)} \end{cases}$$

IP Formulation:

$$\begin{array}{lrll} \min & \sum_{\{i,j\} \in D} z_{ij} & & \\ \text{s.t.} & \sum_{s \in K} y_{is} = 1 & \forall i \in N & \longleftarrow \text{ objects} \\ & \sum_{i \in N} y_{is} \geq \ell & \forall s \in K & \longleftarrow \text{ clusters} \\ & z_{ij} \geq y_{is} + y_{js} - 1 & \forall i < j \in N\ , s \in K & \\ & z_{ij} \geq 0 & \forall i < j \in N & \\ & y_{is} \in \{0,1\} & \forall i \in N\ , s \in K & \end{array}$$

Comparing the formulations

- Size of the formulation:

	variables	constraints	nonzeros
Formulation 0	2,460	4,723	14,280
Formulation 1	2,460	7,063	16,620
Formulation 1^+	2,460	7,063	16,620
Formulation 1^{++}	2,460	2,383	7,260
Formulation 2	900	2,383	7,260

- Solution time:

	B& B nodes	Simplex iterations	Solution time
Formulation 0	40,560	11,210,558	327.78 seconds
Formulation 1	23,050	4,033,965	159.24 seconds
Formulation 1^+	4,715	842,274	32.87 seconds
Formulation 1^{++}	7,471	542,972	5.41 seconds
Formulation 2	4,392	369,175	3.98 seconds

(Formulation 2 is 35% faster than Formulation 1^{++})

Comparing the formulations

- Size of the formulation:

	variables	constraints	nonzeros
Formulation 0	2,460	4,723	14,280
Formulation 1	2,460	7,063	16,620
Formulation 1^+	2,460	7,063	16,620
Formulation 1^{++}	2,460	2,383	7,260
Formulation 2	900	2,383	7,260

- Solution time:

	B& B nodes	Simplex iterations	Solution time
Formulation 0	40,560	11,210,558	327.78 seconds
Formulation 1	23,050	4,033,965	159.24 seconds
Formulation 1^+	4,715	842,274	32.87 seconds
Formulation 1^{++}	7,471	542,972	5.41 seconds
Formulation 2	4,392	369,175	3.98 seconds

(Formulation 2 is 35% faster than Formulation 1^{++})

Gurobi Output for Formulation 2

```
Gurobi log file for last model:
===============================

900 variables, all binary
2383 constraints, all linear; 7260 nonzeros
40 equality constraints
2343 inequality constraints
1 linear objective; 226 nonzeros.

Gurobi 9.1.1: outlev=1
threads=4
Gurobi Optimizer version 9.1.1 build v9.1.1rc0 (linux64)
Thread count: 32 physical cores, 64 logical processors, using up
Optimize a model with 2383 rows, 900 columns and 7260 nonzeros
Model fingerprint: 0xae721739
Variable types: 0 continuous, 900 integer (900 binary)
Coefficient statistics:
Matrix range     [1e+00, 1e+00]
Objective range  [1e+00, 1e+00]
Bounds range     [1e+00, 1e+00]
RHS range        [1e+00, 1e+01]
```

(continued....)

```
Found heuristic solution: objective 81.0000000
Presolve removed 1662 rows and 554 columns
Presolve time: 0.00s
Presolved: 721 rows, 346 columns, 2274 nonzeros
Variable types: 0 continuous, 346 integer (346 binary)

Root relaxation: objective 0.000000e+00, 161 iterations, 0.00 se

Nodes|Current Node|Obj. Bounds|Work |Expl Unexpl |  Obj | Depth
IntInf | Incumbent BestBd|Gap | It/Node Time

0    0   0.0  0   58   81.00  0.00  100%     -    0s
H    0   0             34.00  0.00  100%     -    0s
0    0   0.0  0   85   34.00  0.00  100%     -    0s
0    0   1.5  0  127   34.00  1.50   95.6%   -    0s
0    0   1.5  0  124   34.00  1.50   95.6%   -    0s
0    2   1.5  0  121   34.00  1.50   95.6%   -    0s
*  271  239        17   32.00  9.04   71.7%  103   0s
H  494  297             29.00 10.35   64.3%   93.5 0s
```

(continued....)

```
H  630  351              28.00 11.21   59.9%   94.5 0s
*  633  335       18     27.00 11.21   58.5%   94.3 0s
H  691  316              25.00 12.30   50.8%   95.0 0s
H  974  354              24.00 13.91   42.0%   95.5 1s

Explored 4392 nodes (369175 simplex iterations) in 3.98 seconds

Optimal solution found (tolerance 1.00e-04)
Best objective 2.400e+01, best bound 2.400e+01, gap 0.0000%

369175 simplex iterations
4392 branch-and-cut nodes

Cutting planes:
Gomory: 3
MIR: 7
Zero half: 26
RLT: 128
BQP: 60
```

Solving IPs: computation time

- Consider the following LP formulation

$$\begin{aligned} \text{minimize} \quad & c^T x \\ \text{subject to} \quad & A^1 x \geq b^1, \\ & A^2 x = b^2, \\ & x \geq 0 \end{aligned}$$

- The non-zeroes of this formulation is the number of nonzero entries in the matrices A^1 and A^2.
- LPs are solved using either simplex or interior point algorithms,
- In both cases one has to solve (many, many) linear equations
- The computational burden per iteration typically grows with the number of non-zero entries of the constraint matrices A^1 and A^2
- It also grows with the number of rows of A^1 and A^2.
- IP solution time depends on the number of B&B nodes and the LP solution time at each node.