L'indice di Davies-Bouldin.

Sono dati N punti nel piano Euclideo e si vuole definire una loro partizione in un dato numero K di clusters, in modo da ottimizzare l'indice di Davies Bouldin, che è definito come segue.

Per ogni cluster C_k, sia T_k il numero di punti assegnati al cluster. Usando la norma Euclidea, si definisce un indicatore di compattezza del cluster C_k come

$$S_k = \sqrt{\frac{1}{T_k} \sum_{j \in C_k} ((x_j - x_k)^2 + (y_j - y_k)^2)}$$

dove si sono indicate con indice j le coordinate dei punti del cluster e con indice k le coordinate del centroide del cluster.

Si definisce poi per ogni coppia di clusters C_h e C_k un indicatore di separazione, M_{hk}, che è dato dalla distanza Euclidea tra i due centroidi.

La misura di quanto bene due clusters C_h e C_k sono distinti è data da $R_{hk} = \dfrac{S_h + S_k}{M_{hk}}$, che si vorrebbe basso; in tal modo, infatti, la partizione risulta buona quando i due clusters sono compatti (valori di S bassi) e distanti tra loro (valore di M alto).

L'indice di Davies Bouldin è definito come $DB = \dfrac{1}{N} \sum_{k=1}^{K} \max_{h \neq k} \{R_{kh}\}$.

Formulare il problema, classificarlo e risolvere l'esempio riportato con un numero a scelta di clusters (ad es. $K = 5$). Discutere ottimalità e unicità della soluzione ottenuta.

Punto	x	y
1	24	9
2	16	33
3	8	32
4	42	31
5	40	45
6	41	89
7	13	71
8	37	64
9	34	66
10	50	58
11	91	43
12	68	27
13	63	29
14	61	45
15	54	50
16	62	79
17	65	75
18	80	81
19	85	67
20	51	56

Tabella 1: Posizioni dei punti.

Soluzione. I dati del problema sono il numero N di punti in posizione $(x_i, y_i)\ \forall i = 1, \ldots, N$ ed il numero K di clusters.

Una soluzione è definita da una partizione dei punti in clusters e dalla posizzione dei centroidi dei clusters. La partizione può essere rappresentata da variabili binarie di assegnamento w_{ik} per ogni punto i ed ogni cluster k. La posizione dei centroidi viene rappresentata da variabili continue e libere $(\overline{x}_k, \overline{y}_k)$ per ogni cluster k.

Per comodità si possono definire le variabili ausiliarie implicate nella definizione dell'indice.

$$S_k = \sqrt{\frac{1}{\sum_{j=1}^{N} w_{jk}} \sum_{i=1}^{N} w_{ik}((\overline{x}_k - x_i)^2 + (\overline{y}_k - y_i)^2)} \ \ \forall k = 1, \ldots, K.$$

$$M_{k'k''} = \sqrt{(\overline{x}_{k'} - \overline{x}_{k''})^2 + (\overline{y}_{k'} - \overline{y}_{k''})^2} \ \ \forall k' \neq k'' = 1, \ldots, K.$$

$$R_{k'k''} = \frac{S_{k'} + S_{k''}}{M_{k'k''}} \ \ \forall k' \neq k'' = 1, \ldots, K.$$

Infine si può introdurre una variabile continua D_k per ogni cluster ed imporre il vincolo seguente:

$$D_{k'} \geq R_{k'k''} \ \ \forall k' \neq k'' = 1, \ldots, K$$

per linearizzare la definizione che implica la funzione $\max$.

Gli altri vincoli del problema sono quelli di assegnamento

$$\sum_{k=1}^{K} w_{ik} = 1 \ \ \forall i = 1, \ldots, N.$$

Per evitare che il denominatore nella definizione di S_k possa valere 0 provocando problemi numerici al solutore, può essere utile aggiungere vincoli sullla cardinalità minima dei clusters:

$$\sum_{i=1}^{N} w_{ik} \geq 1 \ \ \forall k = 1, \ldots, K.$$

L'obiettivo è

$$\text{minimize } z = \frac{1}{N} \sum_{k=1}^{K} D_k.$$

Il modello risultante è di programmazione non-lineare intera.

Fissate le variabili binarie, la minimizzazione di S_k in ogni cluster dà luogo ad un sotto-problema convesso, ma la presenza dei termini M inficia questa proprietà. Per studiare la convessità sarebbe necessario studiare le derivate dell'indice di Davies-Bouldin rispetto alla coordinata (non importa se $\overline{x}$ o $\overline{y}$, data la simmetria) del generico centroide. L'obiettivo z è una funzione convessa delle variabili D. Ogni variabile D è una funzione convessa delle variabili R. Si tratta di determinare se ogni $R_{k'k''}$ è una funzione convessa di $\overline{x}_i$ (o $\overline{y}_i$) quando $w_{ik'} = 1$ (o $w_{ik''} = 1$). Ai fini della soluzione dell'esercizio non è richiesto di addentrarsi in calcoli ulteriori.