

Separazione con iperpiani.

Un dataset per il training di un sistema di apprendimento automatico è dato da un insieme di punti con coordinate note in due dimensioni. Le due coordinate di ogni punto rappresentano i valori di due features, in base alle quali i punti sono stati classificati in Sani e Malati.

I valori delle features dei punti classificati come Sani sono vicini ad un "punto ideale", mentre quelli classificati come "Malati" sono più lontani, in varie direzioni.

Per costruire un classificatore automatico lineare, si vuole trovare un insieme di rette tali da descrivere una regione convessa che comprenda tutti i punti Sani e nessuno dei punti Malati.

L'obiettivo è di usare il minimo numero di rette.

Formulare il problema e classificarlo.

Risolvere l'esempio con i dati riportati nel file SEPARAZIONE.TXT. Discutere l'ottimalità e l'unicità della soluzione ottenuta.

Esempio.

Punto	x	y
1	4.18	4.38
2	4.72	5.65
3	4.92	6.18
4	5.61	2.25
5	4.90	1.97
6	6.83	4.15
7	2.12	4.18
8	8.13	6.40
9	3.77	5.67
10	2.46	4.26
11	2.94	6.37

Tabella 1: Valori delle features dei punti Sani.

Punto	x	y
1	5.87	9.16
2	0.07	6.32
3	8.11	2.19
4	9.42	6.71
5	2.01	7.80
6	9.41	3.59
7	7.24	1.71
8	0.55	5.33
9	10.0	6.36
10	2.00	1.50

Tabella 2: Valori delle features dei punti Malati.

Soluzione commentata.

Sia S il set dei punti Sani, con coordinate date (x^S, y^S) e M il set dei punti Malati, con coordinate date (x^M, y^M) .

Sia n il massimo numero di rette consentito (da dimensionare) e sia R il set delle rette, ciascuna caratterizzata dai coefficienti (a_r, b_r, c_r) per ogni $r \in R$ (variabili decisionali).

Utilizziamo una variabile binaria k_r per ogni retta $r \in R$ per indicare se la retta è utilizzata (1) o no (0).

Così è immediato esprimere l'obiettivo:

$$\text{minimize } z = \sum_{r \in R} k_r.$$

Per le rette usate deve valere la condizione di normalizzazione $a_r^2 + b_r^2 = 1$, mentre per le rette non usate si vuole che sia $a_r = b_r = c_r = 0$ in modo che ogni punto soddisfi il vincolo corrispondente alla loro disequazione, ovunque esso si trovi e comunque sia orientata la disequazione. Questo si ottiene facilmente proprio utilizzando la variabili k , imponendo

$$a_r^2 + b_r^2 = k_r \quad \forall r \in R.$$

Ogni retta corrisponde ad un vincolo di disuguaglianza lineare $ax + by + c \leq 0$. Si vuole che ogni punto di S soddisfi tutte le disuguaglianze e che ogni punto di M ne violi almeno una.

Quindi sui punti di S si impongono i vincoli

$$a_r x_i^S + b_r y_i^S + c_r \leq 0 \quad \forall i \in S \quad \forall r \in R$$

mentre sui punti di M i vincoli hanno verso opposto e possono essere violati e quindi devono essere disattivabili. A questo scopo introduciamo una variabile binaria per ogni punto di M e per ogni retta: $w_{ir} = 1$ indica che il vincolo corrispondente è disattivato, cioè può essere violato. I vincoli sui punti di M hanno quindi l'espressione seguente

$$a_r x_i^M + b_r y_i^M + c_r \geq -M w_{ir}.$$

Quando un punto $i \in M$ non è all'esterno della retta r , la variabile w_{ir} è forzata a 1. Per imporre che ogni punto di M sia esterno ad almeno una delle rette, si impone quindi che per ogni punto di M il numero di vincoli disattivati sia strettamente inferiore al numero di rette usate.

$$\sum_{r \in R} w_{ir} \leq \sum_{r \in R} k_r - 1.$$

Il modello è non-lineare e con variabili binarie. A parità di numero di rette usate, può ammettere diversi ottimi locali (non è convesso).

Nell'esempio fornito, il minimo numero di rette necessarie è pari a 3.

Per aiutare il solutore a convergere, può essere utile inizializzare i coefficienti delle rette, in modo che descrivano approssimativamente un poligono centrato intorno ai punti di S (che sono stati generati casualmente a distanza inferiore a 3.5 dal punto di coordinate (5,5)). Una possibile inizializzazione con 6 rette è illustrata nel file `Separazione.mod`. Una soluzione calcolata da Knitro è illustrata nella tabella.

Retta	a	b	c
1	-0.581941	0.813231	-3.97622
2	0.871924	-0.489641	-4.52044
3	-0.882499	-0.470314	3.44649

Tabella 3: Coefficienti delle tre rette nella soluzione.