CLUSTERING

Dato un set S, un numero K, l'idea è trovare le K partizioni di S: $S_1, \dots, S_K$ che sono omogenee e ben separate. È un problema non supervisionato.

Dati i punti $p_1 \dots p_\ell \in \mathbb{R}^n$
Data la funzione distanza $d: \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}$, dove va definita una distanza fra due vettori: $d(x, y) = \|x - y\|_{1/2/\infty}$
Definiamo per ogni cluster un centroide $x_j \in \mathbb{R}^n$.
Il problema di trovare i migliori cluster:

$$\begin{cases} \min \sum_{i=1}^{\ell} \min_{j=1 \dots K} d(p_i, x_j) \\ x_j \in \mathbb{R}^n \quad \forall j = 1 \dots K \end{cases}$$

Norma 2 : $\|\cdot\|_2$

$$\begin{cases} \min \sum_{i=1}^{\ell} \min_{j=1 \dots K} \| p_i - x_j \|_2^2 \\ x_j \in \mathbb{R}^n \quad \forall j = 1 \dots K \end{cases}$$

Per trovare l'ottimo per $K=1$, calcoliamo $\nabla f(x) = 0$.
Notiamo che per $K=1$ è inutile l'operazione di $\min_{j=1 \dots K}(\cdot)$:

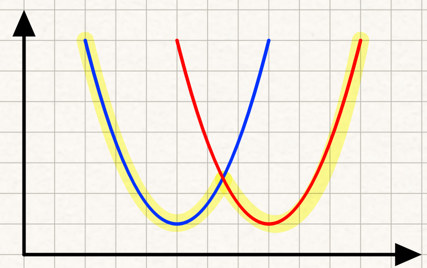
$$\min \sum_{i=1}^{\ell} \|p_i - x\|_2^2 = \sum_{i=1}^{\ell} (x - p_i)^T (x - p_i)$$

$$\nabla \left[\sum_{i=1}^{\ell} (x - p_i)^T (x - p_i) \right] = 2\ell x - 2 \sum_{i=1}^{\ell} p_i = 0 \implies x = \underbrace{\frac{\sum_{i=1}^{\ell} p_i}{\ell}}_{\text{media}}$$

Nota: questo problema è convesso solo per $K=1$. Per $K>1$ il problema diventa **non-convesso** e **non differenziabile**.

Perché è non convesso se $K>1$?

Trovare il miglior centroide per un singolo cluster equivale a minimizzare un quadrato, convesso per definizione ($\min \|p_i - x_i\|^2$). Per più cluster, diciamo 2, dobbiamo minimizzare contemporaneamente più parabole:



Fissando p_i, e per i centroidi $x_j : j=1 \dots K$, possiamo riscrivere che:

$$\min_{j=1\dots K} \|p_i - x_j\|_2^2 = \begin{cases} \min \sum_{j=1}^{K} \alpha_{ij} \|p_i - x_j\|_2^2 \\ \sum_{j=1}^{K} \alpha_{ij} = 1 \\ \alpha_{ij} \geq 0 \quad \forall j = 1 \dots K \end{cases}$$

Dove $\alpha_{ij} = (0, \dots, 0, 1, 0, \dots, 0)$ ovvero con 1 al j-esimo elemento. La soluzione ottima dell'ultimo sistema è:

$$\alpha^*_{ij} = \begin{cases} 1 & \|p_i - x_j\|_2 = \min_{h=1\dots K} \|p_i - x_h\|_2 \\ 0 & \text{altrimenti} \end{cases}$$

TEOREMA

Il problema con $\|\cdot\|_2$, equivale a risolvere il seguente sistema non-convesso ma differenziabile:

$$\begin{cases} \min\limits_{x,\alpha} f(x,\alpha) := \sum\limits_{i=1}^{e} \sum\limits_{j=1}^{k} \alpha_{ij} \|p_i - x_j\|_2^2 \\ \sum\limits_{j=1}^{k} \alpha_{ij} = 1 \qquad \forall i = 1 \dots e \\ \alpha_{ij} \geq 0 \qquad \forall i = 1 \dots e \quad e \quad \forall j = 1 \dots k \\ x_j \in \mathbb{R}^n \qquad \forall j = 1 \dots k \end{cases}$$

Dove bisogna cercare un solo minimo, ma essendo differenziabile possiamo usare il KKT, o altri metodi di risoluzione.

Nota:

Nonostante abbiamo definito $\alpha_{ij} = 1$ o $\alpha_{ij} = 0$, nel sistema non c'è nessuna condizione che esplicitamente dice che sia così. Abbiamo solo la condizione che $\sum_{j=1}^{k} \alpha_{ij} = 1$ e $\alpha \geq 0$. Quindi una soluzione ammissibile può essere $\alpha_i = \left(\frac{1}{2}, \frac{1}{2}\right)$ per $k=2$, fissato un punto i. Questo però non succede perché fissato un punto, il problema diventa convesso su una regione ammissibile che è un poliedro, per cui l'ottimo è necessariamente su un vertice del poliedro, per cui $\alpha_{ij} = (1, 0)$ o $(0, 1)$.

K-MEANS

L'algoritmo di K-means si basa sulla seguente proprietà del sistema definito sopra nel teorema, ovvero che se fissiamo il centroide x_i, possiamo dividere il problema in ℓ sotto problemi di tipo LP, nella forma:

$$\begin{cases} \min \sum_{j=1}^{k} \alpha_{ij} \|p_i - x_j\|_2^2 \\ \sum_{j=1}^{k} \alpha_{ij} = 1 \\ \alpha_{ij} \geq 0 \quad \forall j = 1 \dots k \end{cases}$$

Dove la soluzione ottima è

$$\alpha_{ij}^* = \begin{cases} 1 & \text{se } j : \|p_i - x_j\|_2 = \min_{h=1\dots K} \|p_i - x_h\|_2 \\ 0 & \text{altrimenti} \end{cases}$$

Se invece fissiamo α_{ij}, il problema si riduce in K sottoproblemi di tipo QP convessi

$$\begin{cases} \min \sum_{i=1}^{\ell} \alpha_{ij} \|p_i - x_j\|_2^2 \\ x_j \in \mathbb{R}^n \end{cases}$$

Dove l'ottimo è: $x_j = \dfrac{\sum_{i=1}^{\ell} \alpha_{ij} p_i}{\sum_{i=1}^{\ell} \alpha_{ij}}$ media

Objective Function

$$f(x, \alpha) = \sum_{i=1}^{e} \sum_{j=1}^{k} \alpha_{ij} \| p_i - x_j \|_2^2$$

TEOREMA

L'algoritmo di K-means si ferma dopo un numero finito di iterazioni, a una soluzione (x^*, α^*) del KKT del problema

$$\begin{cases} \min\limits_{x, \alpha} f(x, \alpha) := \sum_{i=1}^{e} \sum_{j=1}^{k} \alpha_{ij} \| p_i - x_j \|_2^2 \\ \sum_{j=1}^{k} \alpha_{ij} = 1 \qquad \forall i = 1 \dots e \\ \alpha_{ij} \geq 0 \qquad \forall i = 1 \dots e \text{ e } \forall j = 1 \dots k \\ x_j \in \mathbb{R}^n \qquad \forall j = 1 \dots k \end{cases}$$

Tale che:

$$f(x^*, \alpha^*) \leq f(x^*, \alpha) \qquad \forall \alpha \geq 0 : \sum_{j=1}^{k} \alpha_{ij} = 1 \quad \forall i = 1 \dots e$$

$$f(x^*, \alpha^*) \leq f(x, \alpha^*) \qquad \forall x \in \mathbb{R}^{kn}$$

Nota: K-means non garantisce di trovare l'ottimo globale, perché il problema è convesso in una sola direzione e non in tutte.

CLUSTER CON $\|\cdot\|_1$

È lo stesso sistema, ma con la norma 1:

$$\begin{cases} \min \sum_{i=1}^{l} \min_{j=1\dots k} \|p_i - x_j\|_1 \\ x_j \in \mathbb{R}^n \quad \forall j = 1 \dots k \end{cases}$$

Prendiamo ora l numeri reali $a_1 < \dots < a_l$, e vogliamo cercare la soluzione ottima del seguente sistema, con $k = 1$.

$$\begin{cases} \min \sum_{i=1}^{l} |x - a_i| = f(x) \\ x \in \mathbb{R} \end{cases}$$

È il mediano fra $(a_1 \dots a_l)$: $\begin{cases} a_{l+1/2} & l \text{ dispari} \\ \dfrac{a_{l/2} + a_{1+l/2}}{2} & l \text{ pari} \end{cases}$

Objective function:

$$f(x) = \begin{cases} -lx + \sum_{i=1}^{l} a_i & \text{se } x < a_i \\ (2-l)x + \sum_{i=2}^{l} a_i - a_1 & \text{se } x \in [a_1, a_2] \\ \dots \\ (2r-l)x + \sum_{i=r+1}^{l} a_i - \sum_{i=1}^{r} a_i & \text{se } x \in [a_r, a_{r+1}] \\ \dots \\ (2-l)x + a_l + \sum_{i=1}^{l-1} a_i & \text{se } x \in [a_{l-1}, a_l] \\ lx - \sum_{i=1}^{l} a_i & \text{se } x > a_l \end{cases}$$

Questa funzione è convessa e lineare a tratti

Esempio

$f(x) = |x-2| + |x-4| + |x-10| \quad \text{con} \quad \ell = 3$

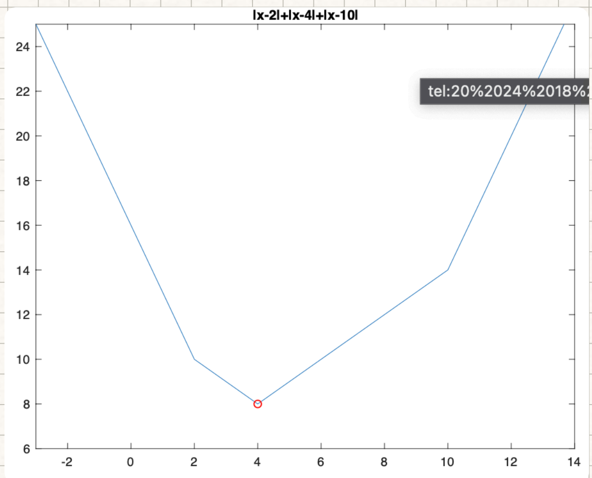
1) se $x < 2$: $2 - x + 4 - x + 10 - x = -3x + 16 = f(x)$

se $x \in [2,4]$: $x - 2 + 4 - x + 10 - x = -x + 12 = f(x)$

se $x \in [4,10]$: $x - 2 + x - 4 + 10 - x = x + 4 = f(x)$

se $x > 10$: $x - 2 + x - 4 + x - 10 = 3x - 16 = f(x)$

Usando la regola, non facciamo conti.



If $k > 1$ (at least two clusters), then the problem is nonconvex and nonsmooth:

$$\begin{cases} \min\limits_{x} \sum\limits_{i=1}^{\ell} \min\limits_{j=1,\dots,k} \|p_i - x_j\|_1 \\ x_j \in \mathbb{R}^n \qquad \forall\, j = 1,\dots,k \end{cases} \tag{8}$$

Theorem

Problem (8) is equivalent to the following problem:

$$\begin{cases} \min\limits_{x,\alpha} \sum\limits_{i=1}^{\ell} \sum\limits_{j=1}^{k} \alpha_{ij} \|p_i - x_j\|_1 \\ \sum\limits_{j=1}^{k} \alpha_{ij} = 1 \quad \forall\, i = 1,\dots,\ell \\ \alpha_{ij} \ge 0 \qquad \forall\, i = 1,\dots,\ell,\ j = 1,\dots,k \\ x_j \in \mathbb{R}^n \qquad \forall\, j = 1,\dots,k. \end{cases} \tag{9}$$

Il problema (9) del teorema è equivalente al seguente problema bilineare (ovvero differenziabile) e non-convesso

$$
\begin{cases}
\min\limits_{x,\alpha,u} \sum\limits_{i=1}^{\ell} \sum\limits_{j=1}^{k} \sum\limits_{h=1}^{n} \alpha_{ij} u_{ijh} \\
u_{ijh} \geq (p_i)_h - (x_j)_h \qquad \forall\ i = 1,\dots,\ell,\ j = 1,\dots,k,\ h = 1,\dots,n \\
u_{ijh} \geq (x_j)_h - (p_i)_h \qquad \forall\ i = 1,\dots,\ell,\ j = 1,\dots,k,\ h = 1,\dots,n \\
\sum\limits_{j=1}^{k} \alpha_{ij} = 1 \qquad \forall\ i = 1,\dots,\ell \\
\alpha_{ij} \geq 0 \qquad \forall\ i = 1,\dots,\ell,\ j = 1,\dots,k \\
x_j \in \mathbb{R}^n \qquad \forall\ j = 1,\dots,k.
\end{cases}
$$