Clustering di misure di probabilità

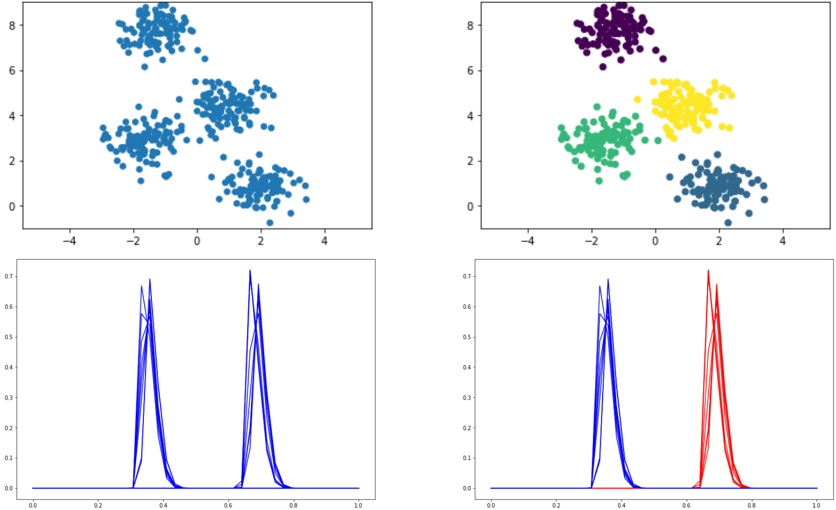
Candidati: Pietro Masini, Maria Chiara Menicucci
Relatore: Prof. Mario Beraha

Politecnico di Milano

18 luglio 2024

Indice

Introduzione



Modello

Sia $X = \{x_1, ..., x_n\}$ un insieme di osservazioni, con x_i elemento di uno spazio metrico (A, d) per $i = 1, ..., n$, e $k \leq n$ fissato.

Consideriamo una partizione $S = \{S_1, ..., S_k\}$ di X (ciascun S_i ha il significato di cluster), che può essere rappresentata equivalentemente con una matrice $C \in M_{\mathbb{R}}(n, k)$ tale che $c_{ij} = \begin{cases} 1 & \text{se } x_i \in S_j \\ 0 & \text{altrimenti} \end{cases}$, detta di assegnamento. Consideriamo un insieme di vettori $M = (\mu_1, ...\mu_k)$, detti centroidi.

Definiamo la funzione perdita

$$L(\mu, C) = \sum_{j=1}^{k} \sum_{\mathrm{x}_i \in S_j} d^2(\mathrm{x}_i, \mu_j) = \sum_{j=1}^{k} \sum_{i=1}^{n} c_{ij} d^2(\mathrm{x}_i, \mu_j)$$

Vogliamo trovare (μ, C) che minimizzino L.

Algoritmo K-Means

Data una stima iniziale dei centroidi M_0, si alternano due passi:

- data una stima dei centroidi M_t, assegna l'osservazione x_i al cluster S_j se e solo se $\mu_j^{(t)}$ è il centroide più vicino a x_i e ottieni la matrice di assegnamento C_t
- data una matrice di assegnamento C_t, per ogni cluster ricalcola il centroide secondo la regola $\mu_j^{(t+1)} = \frac{1}{\left|S_j^{(t)}\right|} \sum_{x \in S_j^{(t)}} x$ e ottieni M_{t+1}.

Si dimostra che con tale algoritmo la funzione di perdita decresce a ogni passo.

K-Means

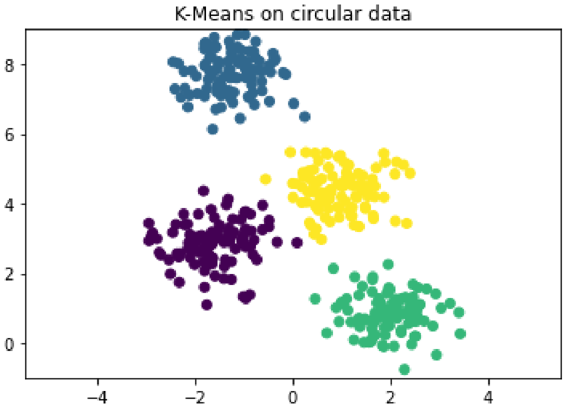


Figura: K-Means on circular data, 4 clusters

L'algoritmo K-Means ha due difetti:

- nessuna indicazione del grado di sicurezza nell'assegnamento finale dei cluster;
- errori nel clustering di dati che sono raggruppati in forme non circolari.

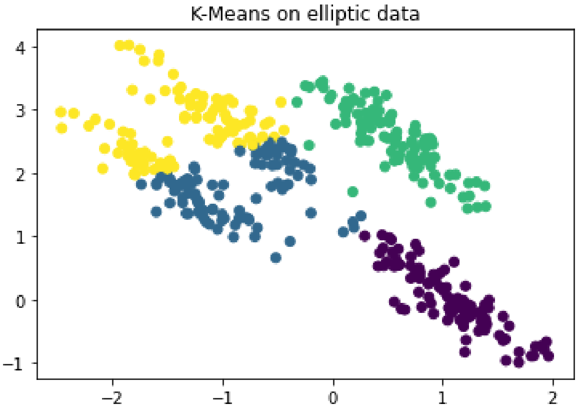


Figura: K-Means on elliptic data, 4 clusters

I GMM (gaussian mixture models) risolvono entrambi i problemi osservati.

Modello mistura di gaussiane

Supponiamo che le osservazioni $x_1, ..., x_n$ siano realizzazioni di vettori aleatori $X_1, ..., X_n : \Omega \to \mathbb{R}^d$ tali che

$$X_1, ..., X_n \overset{\text{i.i.d.}}{\sim} \sum_{j=1}^{k} \tau_j \mathcal{N}\left(\mu_j, \Sigma_j\right)$$

Se $Z : \Omega \to \{1, ..., k\}^n$ è un vettore aleatorio discreto tale che $\mathbb{P}(Z_i = j) = \tau_j$ per $i = 1, ..., n, j = 1, ..., k$, allora

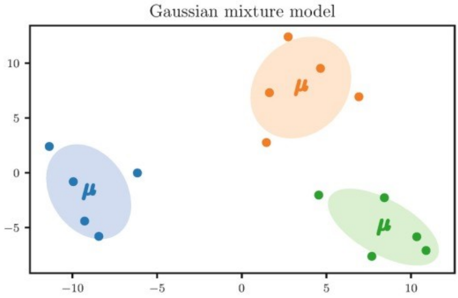
$$X_i | Z_i = j \sim \mathcal{N}\left(\mu_j, \Sigma_j\right).$$

Avendo i dati, l'obiettivo è stimare i parametri delle k leggi gaussiane e il vettore $\boldsymbol{\tau}$ legge di Z:

$$\theta = (\boldsymbol{\tau}, \mathrm{M}, \boldsymbol{\Sigma})$$

dove $\boldsymbol{\tau} = (\tau_1, ..., \tau_k) \in \mathbb{R}^k$, $\mathrm{M} = (\mu_1, ..., \mu_k) \in M_{\mathbb{R}}(d, k)$ e $\boldsymbol{\Sigma} \in M_{\mathbb{R}}(k, d, d)$ tensore tale che $\boldsymbol{\Sigma}(j, \cdot, \cdot) = \Sigma_j$.

Vedere i dati come realizzazioni di misture di gaussiane permette di tenere conto anche di gruppi di dati con forme non circolari:



La stima dei parametri $\theta = (\boldsymbol{\tau}, \mathrm{M}, \boldsymbol{\Sigma})$ permette di stimare anche $\mathbb{P}(Z_i = j | \mathrm{X_i} = \mathrm{x_i}; \theta)$: in base a queste probabilità è determinato l'assegnamento ai cluster.

Vogliamo trovare un MLE per θ.
La funzione di verosimiglianza è

$$L(\theta; \mathrm{x}, \mathrm{z}) = p(\mathrm{x}, \mathrm{z}|\theta) = \prod_{i=1}^{n} \prod_{j=1}^{k} \left[f\left(\mathrm{x}_i; \mu_j, \Sigma_j\right) \tau_j \right]^{I_{(z_i=j)}}$$

con f densità gaussiana.

Stimiamo i parametri usando l'algoritmo Expectation-Maximization, che, data una stima iniziale $\theta^{(0)}$ dei parametri, consiste nell'alternare l'E-step e l'M-step.

E-step

Data una stima attuale dei parametri $\theta^{(t)} = (\tau^{(t)}, M^{(t)}, \mathbf{\Sigma}^{(t)})$, definiamo

$$T_{ji}^{(t)} = \mathbb{P}\left(Z_i = j | \mathsf{X}_i = \mathsf{x}_i; \theta^{(t)}\right) = \frac{\mathbb{P}\left(\mathsf{X}_i = \mathsf{x}_i | Z_i = j; \theta^{(t)}\right) \mathbb{P}\left(Z_i = j; \theta^{(t)}\right)}{\mathbb{P}\left(\mathsf{X}_i = \mathsf{x}_i; \theta^{(t)}\right)}$$

$$= \frac{f\left(\mathsf{x}_i; \mu_j^{(t)}, \Sigma_j^{(t)}\right) \tau_j^{(t)}}{\sum_{j=1}^k \tau_j^{(t)} f\left(\mathsf{x}_i; \mu_j^{(t)}, \Sigma_j^{(t)}\right)}$$

Il valore atteso di $\log L$, calcolato rispetto alla legge di Z dato che $\mathrm{X} = \mathrm{x}$ con stima dei parametri $\theta^{(t)}$, è

$$Q\left(\theta|\theta^{(t)}\right) = \mathbb{E}_{\mathrm{Z}|\mathrm{X}=\mathrm{x};\theta^{(t)}}\left(\log L\left(\theta;\mathrm{x},\mathrm{Z}\right)\right) =$$

$$\sum_{i=1}^{n}\sum_{j=1}^{k} T_{ji}^{(t)}\left[\log\tau_j - \frac{d}{2}\log 2\pi - \frac{1}{2}\log\det\Sigma_j - \frac{1}{2}\left\langle \mathrm{x}_i - \mu_j, \Sigma_j^{-1}\left(\mathrm{x}_i - \mu_j\right)\right\rangle\right]$$

M-step

Dobbiamo massimizzare $Q\left(\theta|\theta^{(t)}\right)$ in θ.
Massimizziamo separatamente in τ e (μ_j, Σ_j):

$\tau^{(t+1)} = \arg\max_\tau \left\{\sum_{j=1}^k \left(\sum_{i=1}^n T_{ji}^{(t)}\right) \log \tau_j\right\}$, perciò

$$\tau_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^n T_{ji}^{(t)}$$

Analogamente $\left(\mu_j^{(t+1)}, \Sigma_j^{(t+1)}\right) =$
$\arg\max_{(\mu_j, \Sigma_j)} \sum_{i=1}^n T_{ji}^{(t)} \left[-\frac{1}{2} \log\det \Sigma_j - \frac{1}{2} \left\langle \mathrm{x}_i - \mu_j, \Sigma_j^{-1} \left(\mathrm{x}_i - \mu_j \right) \right\rangle \right]$,
perciò

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^n T_{ji}^{(t)} \mathrm{x}_i}{\sum_{i=1}^n T_{ji}^{(t)}}$$

$$\Sigma_j^{(t+1)} = \frac{\sum_{i=1}^n T_{ji}^{(t)} \left(\mathrm{x}_i - \mu_j^{(t+1)} \right) \left(\mathrm{x}_i - \mu_j^{(t+1)} \right)^t}{\sum_{i=1}^n T_{ji}^{(t)}}$$

Poiché si dimostra che, con le stime $\theta^{(t)}$ fornite da questo algoritmo, a ogni passo $\max Q$ aumenta, il processo iterativo termina quando

$$Q\left(\theta^{(t+1)}|\theta^{(t+1)}\right) \leq Q\left(\theta^{(t)}|\theta^{(t)}\right) + \varepsilon$$

con $\varepsilon > 0$ soglia prefissata.

Questo algoritmo risolve i due problemi di K-Means:

- indicazione del grado di sicurezza nell'assegnamento ai cluster tramite la matrice T: T_{ji} indica la probabilità che l'osservazione i appartenga al cluster j;
- clustering di dati non circolari corretto.

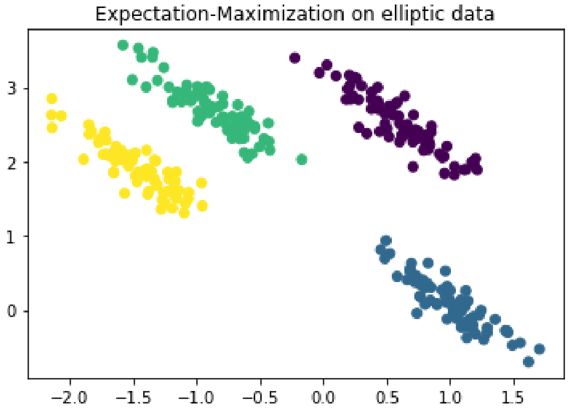
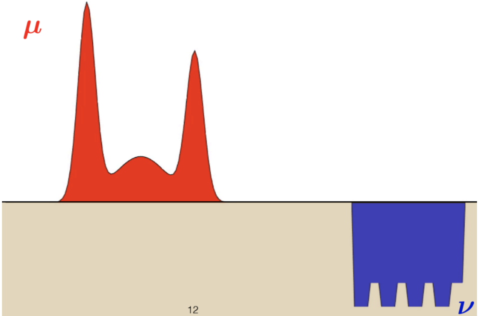


Figura: E-M on elliptic data, 4 clusters

Come usare questi algoritmi per fare clustering quando i dati sono misure di probabilità?
Ci serve una definizione di metrica e qualche nozione di statistica per quando i dati sono misure di probabilità, strumenti forniti dal trasporto ottimo.

Trasporto ottimo

Il trasporto ottimo è un settore dell'analisi matematica che studia il modo ottimale di trasportare risorse. Nel nostro caso, il trasporto ottimo permetterà di definire una nozione di distanza tra misure di probabilità.



Indichiamo con $P(\mathbb{X})$ l'insieme delle misure di probabilità su $(\mathbb{X}, \sigma_{\mathbb{X}})$.

Siano $\mathbb{X}, \mathbb{Y}$ spazi metrici completi e separabili,
$c : \mathbb{X} \times \mathbb{Y} \to [0, +\infty)$, $\mu \in P(\mathbb{X})$, $\nu \in P(\mathbb{Y})$.

$\mathbb{X}, \mathbb{Y}$ rappresentano, rispettivamente, lo spazio della sabbia e la buca; c ha il significato di funzione costo; μ e ν rappresentano, rispettivamente, le distribuzione della sabbia e la forma della buca.

Modello di Kantorovich

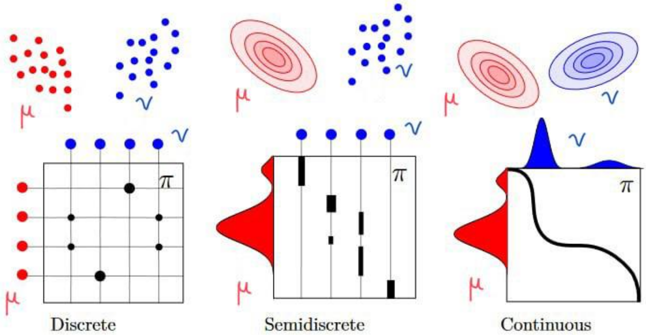
Kantorovich stabilisce che la scelta di come effettuare il trasporto sia descritta da una misura di probabilità $\pi \in P(\mathbb{X} \times \mathbb{Y})$, dove $\pi\left(A \times B\right)$ rappresenta la quantità di sabbia trasportata da $A \in \sigma_{\mathbb{X}}$ a $B \in \sigma_{\mathbb{Y}}$.

I vincoli naturali di conservazione della massa sono

$$\pi\left(A \times \mathbb{Y}\right) = \mu\left(A\right) \ \forall \ A \in \sigma_{\mathbb{X}}$$

$$\pi\left(\mathbb{X} \times B\right) = \nu\left(B\right) \ \forall \ B \in \sigma_{\mathbb{Y}}$$

L'insieme delle misure di probabilità che soddisfano questi vincoli si indica con $\Pi\left(\mu, \nu\right)$, e ogni suo elemento si dice piano di trasferimento da μ a ν.



Il costo totale del trasferimento secondo un piano di trasferimento π è

$$C(\pi) = \int_{\mathbb{X}\times\mathbb{Y}} c(x,y)\, d\pi(x,y)$$

Problema di Kantorovich

Vogliamo trovare

$$\inf_{\pi \in \Pi(\mu,\nu)} C(\pi) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{X} \times \mathbb{Y}} c(x,y)\, d\pi(x,y)$$

L'insieme ammissibile non è mai vuoto: la misura prodotto $\mu \otimes \nu$ soddisfa sempre i vincoli di Kantorovich.

Spazio di Wasserstein

Definition

Sia $(\mathbb{X}, ||\cdot||)$ uno spazio di Banach separabile. Dato $p \geq 1$, si dice **spazio di Wasserstein di ordine** p **su** $\mathbb{X}$ l'insieme

$$\mathbb{W}_p(\mathbb{X}) = \left\{ \mu \in P(\mathbb{X}) : \int_{\mathbb{X}} ||x||^p \, d\mu(x) < +\infty \right\}$$

Distanza di Wasserstein

Definition

Date $\mu, \nu \in \mathbb{W}_p(\mathbb{X})$, si dice **distanza di Wasserstein di ordine** p **tra** μ **e** ν

$$W_p(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathbb{X}^2} ||x - y||^p \, d\pi(x, y) \right)^{\frac{1}{p}}$$

Si dimostra che $\mathbb{W}_p(\mathbb{X})$ è uno spazio metrico rispetto alla distanza W_p.

Poiché il nostro obiettivo sarà fare clustering di misure in $P(\mathbb{R})$, concentriamoci sul caso $\mathbb{X} = \mathbb{Y} = \mathbb{R}$.

Definition

Data una funzione di ripartizione $F : \mathbb{R} \to [0, 1]$, si dice **pseudo-inversa di** F, e si indica con F^{-1}, la funzione $F^{-1} : (0, 1) \to \mathbb{R}$,

$$F^{-1}(u) = \inf \{x \in \mathbb{R} : F(x) \geq u\}$$

F^{-1} si dice anche funzione quantile; è continua da sinistra e non decrescente.

Soluzione del problema di Kantorovich nel caso $\mathbb{X} = \mathbb{Y} = \mathbb{R}$

Theorem

Siano F, G le funzioni di ripartizione di $\mu, \nu \in P(\mathbb{R})$ rispettivamente. Se $h : \mathbb{R} \to [0, +\infty)$ è convessa e $c(x, y) = h(|x - y|)$, allora

$$\inf_{\pi \in \Pi(\mu,\nu)} C(\pi) = \int_0^1 h\left(G^{-1}(u) - F^{-1}(u)\right) du$$

Nel caso $h(z) = z^2$, la tesi del teorema sopra diventa

$$\inf_{\pi \in \Pi(\mu,\nu)} C(\pi) = \int_0^1 \left(G^{-1}(u) - F^{-1}(u)\right)^2 du$$

cioè

$$W_2(\mu,\nu) = \left|\left|F^{-1} - G^{-1}\right|\right|_{L^2(0,1)}$$

Si dimostra che c'è un *isomorfismo isometrico* tra $\mathbb{W}_2(\mathbb{R})$, munito della distanza di Wasserstein W_2, e l'insieme $L^2_{\uparrow}([0,1])$ (insieme delle funzioni L^2 non decrescenti continue a sinistra) con la norma L^2.

Medie di Fréchet

Definition

Si dice funzionale di Fréchet associato alle misure $\mu_1, \ldots, \mu_n \in \mathbb{W}_2(\mathbb{R})$ il funzionale $F : \mathbb{W}_2(\mathbb{R}) \to \mathbb{R}$, $F(\gamma) = \frac{1}{n}\sum_{i=1}^{n} W_2^2(\gamma, \mu_i)$. Il punto di minimo di F in $W_2(\mathbb{R})$ si dice **media di Fréchet di** $\mu_1, \ldots, \mu_n$.

La definizione è ben posta: si può dimostrare che esiste un unico punto di minimo γ^* di F, e ha funzione quantile $F_{\gamma^*}^{-1} = \frac{1}{n}\sum_{i=1}^{n} F_{\mu_i}^{-1}$.

Misure aleatorie

Definition

Dato uno spazio di probabilità $(\Omega, \mathcal{A}, \mathbb{P})$, si dice **misura aleatoria su** $\mathbb{W}_2(\mathbb{R})$ una funzione misurabile $\mathfrak{F} : \Omega \to \mathbb{W}_2(\mathbb{R})$, dove $\mathbb{W}_2(\mathbb{R})$ è dotato della sua σ-algebra di Borel[1].

[1] Si considera la topologia indotta in $\mathbb{W}_2(\mathbb{R})$ dalla distanza di Wasserstein.

Funzioni di ripartizione di misure aleatorie

Data una misura aleatoria $\mathfrak{F} : \Omega \to \mathbb{W}_2(\mathbb{R})$, si definisce funzione di ripartizione di $\mathfrak{F}$

$$\mathbb{F}[\omega](t) := \mathfrak{F}[\omega]((-\infty, t])$$

Media e varianza di misure aleatorie

Definition

Data una misura aleatoria $\mathfrak{F} : \Omega \to \mathbb{W}_2(\mathbb{R})$, si dice **media di Wasserstein-Fréchet di** $\mathfrak{F}$ la misura di probabilità

$$\gamma_m(\mathfrak{F}) = \arg \min_{\gamma \in \mathbb{W}_2(\mathbb{R})} \mathbb{E}\left(W_2^2(\mathfrak{F}, \gamma)\right)$$

Si dice **varianza di Wasserstein-Fréchet di** $\mathfrak{F}$ il numero

$$var(\mathfrak{F}) = \mathbb{E}\left(W_2^2(\mathfrak{F}, \gamma_m)\right)$$

Si dimostra che, per ogni misura aleatoria $\mathfrak{F} : \Omega \to \mathbb{W}_2(\mathbb{R})$ con funzionale di Fréchet $F(\gamma) = \mathbb{E}\left(W_2^2(\mathfrak{F}, \gamma)\right)$ finito, esiste un unico punto di minimo di F, e ha funzione quantile $F_{\gamma_m}^{-1}(t) = \mathbb{E}(\mathbb{F}^{-1}[\omega](t))\ \forall\ t \in (0, 1)$.

Clustering di misure di probabilità

Supponiamo ora che le osservazioni siano $\mu_1, ..., \mu_n$ misure di probabilità su $\mathbb{R}$: vorremmo fare clustering di tali osservazioni, cioè raggrupparle in modo che si trovino nello stesso gruppo misure "simili" nel senso della distanza di Wasserstein.

Per fare ciò vorremmo sviluppare algoritmi analoghi a quelli visti nel caso euclideo.

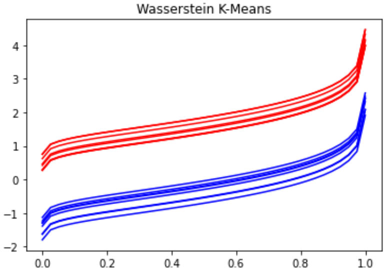
Algoritmo K-Means nello spazio di Wasserstein

L'algoritmo K-Means si estende da $\mathbb{R}^d$ allo spazio di Wasserstein senza difficoltà[1]:

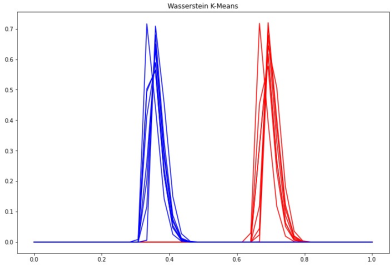
- ogni osservazione μ_i è assegnata al cluster $S_j^{(t)}$ più vicino secondo d_W
- il centroide per ogni cluster si aggiorna secondo la regola $\gamma_j^{(t+1)} = \arg\min_{\gamma \in \mathbb{W}_2(\mathbb{R})} \frac{1}{|S_j^{(t)}|} \sum_{i \in S_j^{(t)}} W_2^2 (\mu_i, \gamma)$

[1] Zhuang et al. (2022)

L'algoritmo K-Means nello spazio di Wasserstein presenta gli stessi problemi osservati in $\mathbb{R}^d$?



K-Means on Gaussian measures, quantile functions



K-Means on Gaussian measures, probability density functions

Qual è, nello spazio di Wasserstein, l'analogo dei dati ellittici in $\mathbb{R}^d$?

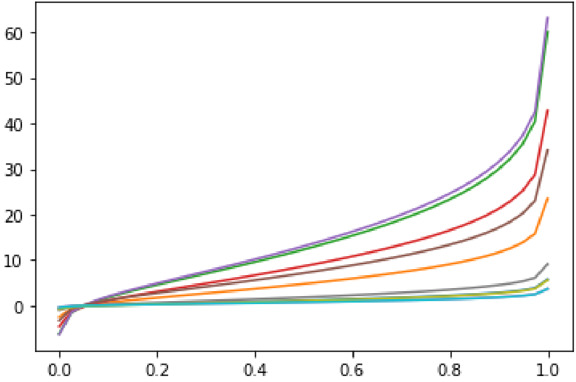
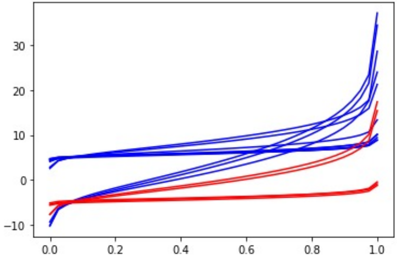


Figura: Skew-normal distributions, quantile functions

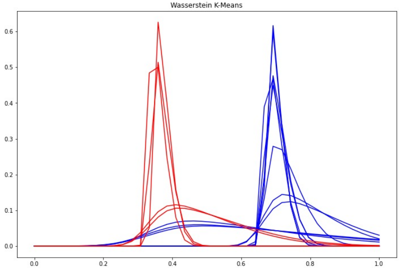
Generiamo funzioni quantile di distribuzioni skew normal:

- campioniamone le medie da due distribuzioni normali con medie uguali a 5 e -5 rispettivamente e varianza 0.2;
- campioniamone le varianze da una distribuzione log-normale.

K-Means non riesce a riconoscere i due cluster:



K-Means on elliptic measures, quantile functions



K-Means on elliptic measures, probability density functions

Cerchiamo allora di estendere anche l'algoritmo Expectation-Maximization allo spazio di Wasserstein, sfruttando, oltre alla distanza di Wasserstein, anche le nozioni di statistica nello spazio di Wasserstein.

Modello

Supponiamo che le osservazioni $\mu_1, \ldots, \mu_n$ siano realizzazioni di misure aleatorie $\mathfrak{F}_1, \ldots, \mathfrak{F}_n : \Omega \to \mathbb{W}_2(\mathbb{R})$, e supponiamo che queste misure siano una "mistura" di $k < n$ misure con media di Fréchet $\gamma_{m,j}$ e varianza di Fréchet $\mathfrak{S}_j^2$, cioè:
se $Z : \Omega \to \{1, \ldots, k\}^n$ è un vettore aleatorio discreto tale che $\mathbb{P}(Z_i = j) = \tau_j$ per $i = 1, \ldots, n, j = 1, \ldots, k$, allora $\mathfrak{F}_i | Z_i = j$ ha media di Fréchet $\gamma_{m,j}$ e varianza di Fréchet $\mathfrak{S}_j^2$.

Nella pratica, grazie all'isomorfismo isometrico menzionato, lavoriamo con le funzioni quantile e non con le misure.

Notazione:

- $F_1^{-1}, \ldots, F_n^{-1}$ sono le funzioni quantile dei dati;
- $\mathbb{F}_1^{-1}[\omega], \ldots, \mathbb{F}_n^{-1}[\omega]$ sono le funzioni quantile delle misure aleatorie;
- $\Gamma_{m,1}^{-1}, \ldots, \Gamma_{m,k}^{-1}$ sono le funzioni quantile delle medie di Fréchet $\gamma_{m,j}$ delle misure aleatorie.

Avendo i dati, l'obiettivo è stimare i parametri delle k misure aleatorie e il vettore $\boldsymbol{\tau}$ legge di Z:

$$\theta = \left(\tau, \Gamma^{-1}, \mathfrak{S}^2\right)$$

dove $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_k) \in \mathbb{R}^k$, Γ^{-1} raccoglie le funzioni quantile $\left\{\Gamma^{-1}_{m,1}, \ldots, \Gamma^{-1}_{m,k}\right\}$ delle medie di Fréchet e $\mathfrak{S}^2 \in \mathbb{R}^k$ raccoglie le varianze di Fréchet delle misure.

Per scrivere l'algoritmo E-M cerchiamo di scrivere, per ogni termine che appare nell'algoritmo E-M per dati euclidei, l'analogo nello spazio di Wasserstein.

In $\mathbb{R}^d$ la funzione da massimizzare è

$$Q\left(\theta|\theta^{(t)}\right) =$$

$$\sum_{i=1}^{n}\sum_{j=1}^{k} T_{ji}^{(t)}\left[\ln\tau_j - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln\det\Sigma_j - \frac{1}{2}\left\langle \mathrm{x}_i-\mu_j, \Sigma_j^{-1}\left(\mathrm{x}_i-\mu_j\right)\right\rangle\right]$$

Osserviamo i termini uno a uno.

$$Q\left(\theta|\theta^{(t)}\right) =$$

$$\sum_{i=1}^{n}\sum_{j=1}^{k} T_{ji}^{(t)}\left[\underline{\ln\tau_j - \frac{d}{2}\ln 2\pi} - \frac{1}{2}\ln\det\Sigma_j - \frac{1}{2}\left\langle \mathrm{x}_i - \mu_j, \Sigma_j^{-1}\left(\mathrm{x}_i - \mu_j\right)\right\rangle\right]$$

$$Q\left(\theta|\theta^{(t)}\right) =$$

$$\sum_{i=1}^{n}\sum_{j=1}^{k} T_{ji}^{(t)}\left[\ln\tau_j - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln\det\Sigma_j - \frac{1}{2}\underline{\left\langle \mathrm{x}_i - \mu_j, \Sigma_j^{-1}\left(\mathrm{x}_i - \mu_j\right)\right\rangle}\right]$$

$\left\langle \mathrm{x}_i - \mu_j, \Sigma_j^{-1} \left(\mathrm{x}_i - \mu_j\right) \right\rangle$ è il quadrato della distanza di Mahalanobis tra x_i e una distribuzione di probabilità con valore atteso μ_j e matrice di covarianza Σ_j.

Perciò l'analogo naturale di $\left\langle \mathrm{x}_i - \mu_j, \Sigma_j^{-1} \left(\mathrm{x}_i - \mu_j\right) \right\rangle$ è la distanza di Wasserstein normalizzata

$$\frac{W_2^2\left(\mu_i, \gamma_{m,j}\right)}{\mathfrak{S}_j^2} = \frac{\int_0^1 \left(F_i^{-1}(u) - \Gamma_{m,j}^{-1}(u)\right)^2 du}{\mathbb{E}\left(\int_0^1 \left(\mathbb{F}_j^{-1}[\omega](t) - \Gamma_{m,j}^{-1}(t)\right)^2 dt\right)}$$

$$Q\left(\theta|\theta^{(t)}\right) =$$

$$\sum_{i=1}^{n}\sum_{j=1}^{k} T_{ji}^{(t)}\left[\ln \tau_j - \frac{d}{2}\ln 2\pi - \frac{1}{2}\underline{\ln\det\Sigma_j} - \frac{1}{2}\left\langle \mathrm{x}_i - \mu_j, \Sigma_j^{-1}\left(\mathrm{x}_i - \mu_j\right)\right\rangle\right]$$

Il termine $\det \Sigma_j$ diventa semplicemente $\mathfrak{S}_j^2$.

Ricordiamo che $T_{ji}^{(t)} = \frac{f\left(x_i;\mu_j^{(t)},\Sigma_j^{(t)}\right)\tau_j^{(t)}}{\sum_{j=1}^k \tau_j^{(t)} f\left(x_i;\mu_j^{(t)},\Sigma_j^{(t)}\right)}$: troviamo un analogo

di ogni termine che appare nella densità gaussiana

$$f\left(x_i;\mu_j,\Sigma_j\right) = \frac{1}{(2\pi)^{n/2}\sqrt{\det\Sigma_j}} \exp\left(-\frac{1}{2}\left\langle x_i-\mu_j, \Sigma_j^{-1}\left(x_i-\mu_j\right)\right\rangle\right)$$

In base a quanto già detto,

$$T_{ji}^{(t)} = \frac{\tau_j^{(t)} \frac{1}{\mathfrak{S}_j^{(t)}} \exp\left(-\frac{1}{2}\frac{W_2^2\left(\mu_i,\gamma_{m,j}^{(t)}\right)}{\mathfrak{S}_j^{2,(t)}}\right)}{\sum_{j=1}^k \tau_j^{(t)} \frac{1}{\mathfrak{S}_j^{(t)}} \exp\left(-\frac{1}{2}\frac{W_2^2\left(\mu_i,\gamma_{m,j}^{(t)}\right)}{\mathfrak{S}_j^{2,(t)}}\right)}$$

Occupiamoci dei passi di aggiornamento.

Niente cambia per τ_j:

$$\tau_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} T_{ji}^{(t)}$$

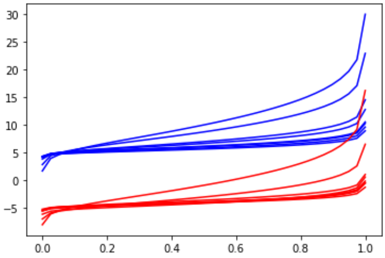
La stima successiva di $\Gamma_{m,j}^{-1}$, ricordando che $\mu_j^{(t+1)} = \frac{\sum_{i=1}^{n} T_{ji}^{(t)} x_i}{\sum_{i=1}^{n} T_{ji}^{(t)}}$, sarà

$$\Gamma_{m,j}^{-1,(t+1)} = \frac{\sum_{i=1}^{n} T_{ji}^{(t)} F_i^{-1}}{\sum_{i=1}^{n} T_{ji}^{(t)}}$$

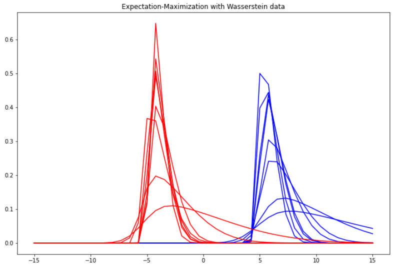
La stima successiva di $\mathfrak{S}_j^2$, ricordando che $\Sigma_j^{(t+1)} = \frac{\sum_{i=1}^n T_{ji}^{(t)}\left(x_i - \mu_j^{(t+1)}\right)\left(x_i - \mu_j^{(t+1)}\right)^t}{\sum_{i=1}^n T_{ji}^{(t)}}$, sarà

$$\mathfrak{S}_j^{2,(t+1)} = \frac{\sum_{i=1}^n T_{ji}^{(t)} W_2^2\left(\mu_i, \gamma_{m,j}^{(t+1)}\right)}{\sum_{i=1}^n T_{ji}^{(t)}}$$

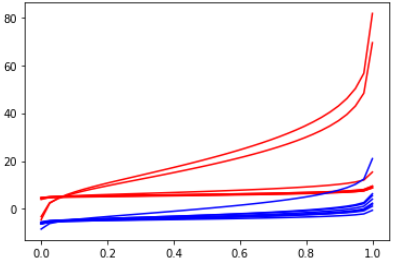
Testiamo l'algoritmo su misure non circolari: il clustering è fatto correttamente.



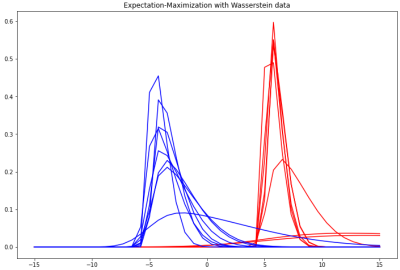
EM on elliptic measures, quantile functions; trial 1



EM on elliptic measures, probability density functions; trial 1



EM on elliptic measures, quantile functions; trial 2



EM on elliptic measures, probability density functions; trial 2

Conclusioni

Abbiamo osservato, sia nel caso euclideo che nel caso delle misure, le criticità dell'algoritmo K-Means e abbiamo visto come l'algoritmo EM sia in grado di risolverli col suo approccio probabilistico.

La nostra tecnica per estendere l'algoritmo EM è però intrinsecamente limitata al caso di misure di probabilità su $\mathbb{R}$: l'isomorfismo isometrico tra misure di probabilità e le loro funzioni quantile esiste solo per misure in $\mathbb{W}_2(\mathbb{R})$.

Bibliografia

[1] L. Ambrosio, A. Bressan, D. Helbing, A. Klar, E. Zuazua, and N. Gigli. *A user's guide to optimal transport*, in *Modelling and Optimisation of Flows on Networks*, SpringerLink, 2013.
[2] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics, 2009.
[3] V. M. Panaretos and Y. Zemel. *An Invitation to Statistics in Wasserstein Space*. SpringerLink, 2020.
[4] F. Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Birkhauser, 2015.
[5] Y. Zhuang, X. Chen, and Y. Yang. Wasserstein K-means for clustering probability distributions. NeurIPS, 2022.

Grazie per l'attenzione.