

## CLUSTERING

Il CLUSTERING È UNA TECNICA CHE CI PERMETTE DI APPRENDERE GRUPPI DI DATI CORRELATI TRA DI LORO E DI RAGGRUPPARLI QUINDI IN CLUSTER SULLA BASE DI CARATTERISTICHE COMUNI.

### K-MEANS

È UN ALGORITMO CHE PERMETTE DI ASSOCIARE CIASCUA DATA POINT CON UNO DI K CENTRI (MEANS) NEI DIVERSI SPAG.

DATO QUINDI UN INSERITO INIZIALE DI K-MEANS  $\bar{m}_1^1, \dots, \bar{m}_n^1$  SI HANNO DUE POSSIBILI CSE SI AGGIORNANO:

1. **ASSIGNMENT:**  $\xrightarrow{\text{ASSEGNA CLUSTERS}}$

$$S_i^t = \left\{ \bar{x}_p \mid \|\bar{x}_p - \bar{m}_i^t\| \leq \|\bar{x}_p - \bar{m}_j^t\| \forall j \right\}$$

2. **UPDATING:**  $\xrightarrow{\text{AGGIORNA IL MEDIO}}$

$$\bar{m}_i^{t+1} = \frac{1}{|S_i^t|} \sum_{\bar{x}_j \in S_i^t} \bar{x}_j$$

$\hookrightarrow$  SI ASSEGNA IL PUNTO AL CLUSTER  $m_i$  POICHE' È IL PIU' vicino

SI HA CHÉ L'ALGORITMO CONVERGE FINO AI ASSEGNAZIONI DEI CLUSTERS NON CAMBIANO PIÙ. LA CONVERGENZA È QUINDI GARANTITA, MA NON È L'OTTIMALITÀ!

ASSUMONO DI AVERE UN CERTO DATASET  $D = \{\bar{x}_1, \dots, \bar{x}_n\}$  DI N ELEMENTI. L'OGGETTIVO È QUINDI QUELLO DI PARTIZIONARE D IN K CLUSTER.

UN CLUSTER È QUINDI UN SOTTOSIEME DI D NEI PUÒ LI SUMMA DELLE DISTANZE INTER-POINT AL SUO INTERNO SONO MINIMIZZATE PRECEDENTEMENTE. //

INTRODUCIAMO, A QUESTO PUNTO, UN INSIEME DI  $K$  VETTORI PROTOTIPO  $\bar{\mu}_k$  PER  $k=1, \dots, K$ .  
QUINDI, PER OGNI CAMPIONE  $\bar{x}_m$  SI INTRODUCE UN INDICATORE BINARIO DEFINITO DA  $v_{mk}$ :

$$v_{mk} = \begin{cases} 1 & \text{SE } \bar{x}_m \text{ È ASSEGNAZO AL CLUSTER } k \\ 0 & \text{ALTRIMENTI} \end{cases}$$

POSSIAMO QUINDI DEFINIRE UNA FUNZIONE OBBIETTIVO  $J$ :

$$J = \sum_{m=1}^N \sum_{k=1}^K v_{mk} \| \bar{x}_m - \bar{\mu}_k \|^2$$

ABBIAMO QUINDI OTTENUTO UN PROBLEMA DI OTTIMIZZAZIONE IN  $v_{mk}$  E  $\bar{\mu}_k$ . SI RISOLVE  
ATTRAVERSO UN ALGORITMO ITERATIVO DIVISO IN DUE STEP:

- ① **E-STEP**: SI MINIMIZZA  $J$  RISPETTO A  $v_{mk}$  TENENDO FISSATO  $\bar{\mu}_k$
- ② **M-STEP**: SI MINIMIZZA  $J$  RISPETTO A  $\bar{\mu}_k$  TENENDO FISSATO  $v_{mk}$

1)  
E-STEP:

IN QUESTO CASO  $J$  È UNA FUNZIONE LINEARE DI  $v_{mk}$  E POSSIAMO QUINDI OTTIMIZZARLA  
PER OGNI  $m$  INDIPENDENTEMENTE.

LA RECA SUL PUNTO QUELLO DI ASSIGNARE  $\bar{x}_m$  AL  $\bar{\mu}_k$  PIÙ VICINO

$$V_{mn} = \begin{cases} 1 & \text{se } K = \underset{j}{\operatorname{arg\min}} \| \bar{x} - \bar{\mu}_j \|^2 \\ 0 & \text{ALTRIMENTI} \end{cases}$$

2)

### M-STEP:

IN QUESTO CASO SI HA CHE  $J$  È UNA FUNZIONE QUADRATICA IN  $\bar{\mu}_K$ ; ALLORA OCCORRE IMPOSTARE IL SUO GRADIENTE A ZERO:

$$2 \sum_{m=1}^N V_{mk} (\bar{x}_m - \bar{\mu}_k) = 0 \quad \forall k$$

E, risolvendo per  $\bar{\mu}_k$ , ottieniamo:

$$\bar{\mu}_k = \frac{\sum_{m=1}^N V_{mk} \bar{x}_m}{\sum_{m=1}^N V_{mk}}$$

Somma di tutti i punti associati al cluster  $k$

numero di punti presenti nel cluster  $k$

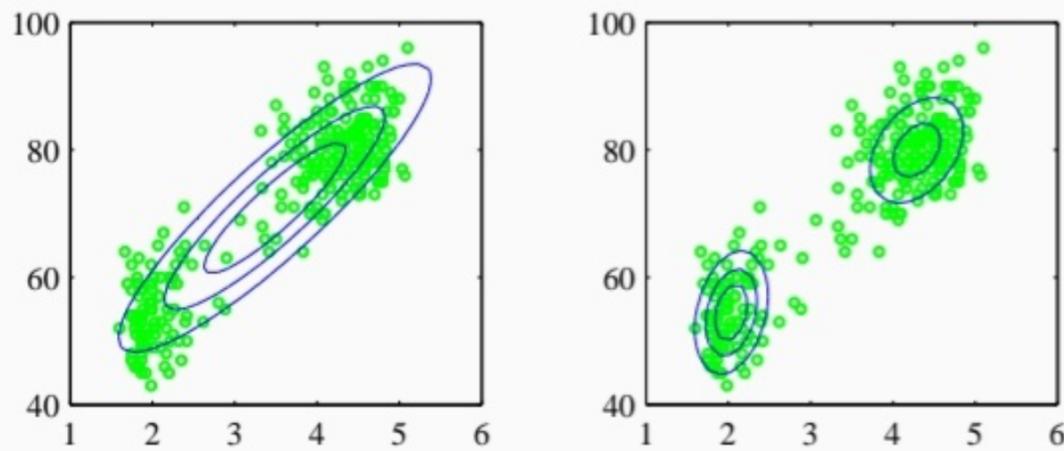
Allora si vanno ad iterare i due step fino a quando l'assegnamento non viene più modificato nell' $G$ -step.

PROPOGGIAMO CHIESA ITERAZIONE DELL'ALGORITMO VO A RIDURRE IL VALORE DI  $J$ . SI HA CHE LA CONVERGENZA È GARANTITA.

INFINE, A SECONDA DELL'INIZIALIZZAZIONE,  $K$ -MEANS PIÙ CONVERGERÀ O MENO AD UN PESSIMO MINIMO LOCALE!

## GAUSSIAN MIXTURE MODELS

MODELLARE DALLE DISTRIBUZIONI TRAMITE UNA SUCCESSIONE DI GAUSSIANE PUÒ FUNZIONARE BENE SULLE PARTI MA A VOLTE PUÒ RISULTARE IN UNA PESSIMA APPROSSIMAZIONE DELLA MOLTA



OBBIAMO GIÀ VISTO CHE TRAMITE L'APPROSSIMAZIONE CON KERNEL DENSITY ESTIMATION POSSIAMO AVVOLGERE UNA GAUSSIANA A TUTTO IL PUNTO E normalizzarla.

NEL CASO IN CUI I PUNTI SONO GENERATI A PARTIRE DA UN NUMERO FINITO E FISSATO DI DISTRIBUZIONI CONGREGATIVE, POSSIAMO INVECE FITTAZIONE UNA MIXTURE OF GAUSSIANS, ovvero un GAUSSIAN MIXTURE MODEL.

$$p(\bar{x}; \bar{\theta}) = \sum_{k=1}^K \pi_k N(\bar{x} | \bar{\mu}_k, \Sigma_k)$$

L → LIKELIHOOD  
K → COMPONENTI  
π\_k → COEFFICIENTI

CHE CONSISTE IN UNA SUPERPOSIZIONE DI K GAUSSIANE, ONSCIAZ DELLE quali è DEFINITA COMPONENTE DELLA MIXTURE CON UN MIXING COEFFICIENT  $\pi_k$ . È INOLTRE NECESSARIO, AI FINI DELLA NORMALIZZAZIONE DELL'ESPRESSIONE, CHE:

$$0 \leq \pi_k \leq 1$$

$$\sum_k \pi_k = 1$$

ADDESSO DOBBIAMO QUINDI STIMARE I VALORI CARATTERISTICI DI TUTTE QUESTE

DISTRIBUZIONI NORALI; SI ESEGUE QUINDI UN APPROCCIO DI MAXIMUM LIKELIHOOD:

$$p(D; \bar{\theta}) = \prod_{n=1}^N p(\bar{x}_n | \bar{\theta})$$

! LIKELIHOOD

$$= \prod_{n=1}^N \sum_{k=1}^K \pi_k N(\bar{x}_n | \bar{\mu}_k, \bar{\Sigma}_k)$$

DONC

$$\bar{\theta} = \left\{ \mu_1, \dots, \mu_K, \Sigma_1, \dots, \Sigma_K, \pi_1, \dots, \pi_K \right\}$$

↗ Sono tutti i vari parametri

Ora consideriamo il LUGARITO DEL LIKELIHOOD:

$$\ln p(D; \bar{\theta}) = \sum_{n=1}^N \sum_{k=1}^K \pi_k N(\bar{x}_n | \bar{\mu}_k, \bar{\Sigma}_k)$$

TUTTAVIA IN QUESTA FORMA SI HA DIFFICOLTÀ AD INSERIRE NUOVI DATI SENZA ANDARE AVERGHE INFORMAZIONI SU QUELLE COMPONENTI SONO RESPONSABILI PER SPECIFICI PUNTI DEL DATASET, OVVERO QUELLI SONO GLI GAUSSIANI PIÙ INFLUENTI NEL MODELLO.



PER FARLO QUESTO SI INTRODUCE UNA VARIABLE CASUALE LATENTE Z A QUALE NON VIENE EFFETTUAMENTE OSSERVATA, MA CHE È SERVITA PER SEMPLIFICARE E QUINDI RIDURRE IL PROBLEMA DI OBBESERVAZIONI.

DEFINIAMO QUINDI Z COME UNA VARIABLE LATENTE DI DIMENSIONE N TALE CHE:

$$\cdot Z_{H6} (0,1)$$

$$\cdot \sum_H z_H = 1$$

siamo adesso in curire questa variabile con  $\bar{x}$  definendo la DISTRIBUZIONE

CONDUTTA  $p(\bar{x}, \bar{z})$  in termini della MARGINALE  $p(\bar{z})$  è detta DISTRIBUZIONE CONDIZIONALE  $p(\bar{x}|\bar{z})$ :

$$\cdot P(z_{H=1}) = \pi_H$$

↓

$$\cdot P(\bar{z}) = \prod_{H=1}^K \pi_H^{z_H} \rightarrow \text{marginali}$$

$$\cdot P(\bar{x}|\bar{z}) = \prod_{H=1}^K N(\bar{x}|\bar{\mu}_H, \bar{\Sigma}_H)^{z_H}$$

→ condizionale

possiamo allora riscontrare la condutta come

$$P(\bar{x}, \bar{z}) = P(\bar{x}|\bar{z}) P(\bar{z})$$

E, intervallando via  $\bar{z}$ , otteniamo proprio la mixture:

$$\begin{aligned} P(\bar{x}) &= \sum_{\bar{z}} P(\bar{z}) P(\bar{x}|\bar{z}) \\ &= \sum_{H=1}^K \pi_H N(\bar{x}|\bar{\mu}_H, \bar{\Sigma}_H) \end{aligned}$$

→ si è marginalizzato via  $\bar{z}$

→ indica il peso della gaussina, che dipende dal numero di punti nello stesso intervallo.

Dove  $\bar{z} = \left\{ [1 \ 0 \ 0], [0 \ 1 \ 0], [0 \ 0 \ 1] \right\}$  per  $H=3$  sono le possibili classi.

QUESTO PROCEDIMENTO CI PERMETTE QUINDI DI LAVORARE CON LA DISTRIBUZIONE CONDIZIONATA AL POSTO DELLA MARGINALE CHE NON HA UNA STRUTTURA SODDA.  
PUNTO, SE ABBIANO  $N$  OSSERVAZIONI  $\bar{x}_1, \dots, \bar{x}_N$  AVREMO ANCHE UNA VARIABILE LATENTE  $\bar{z}_m$  PER OGNI  $\bar{x}_m$

PER QUANTO RIGUARDA LA POSTERIORI OTTENISMO CHE:

$$\begin{aligned} \delta(z_h) &= P(z_h=1|\bar{x}) = \frac{P(z_h=1)P(\bar{x}|z_h=1)}{\sum_{h=1}^K P(z_h=1)P(\bar{x}|z_h=1)} \\ &= \frac{\pi_h N(\bar{x}|\bar{\mu}_h, \Sigma_h)}{\sum_{j=1}^K \pi_j N(\bar{x}|\bar{\mu}_j, \Sigma_j)} \end{aligned}$$

Dove  $\pi_h$  è la prior di  $z_h=1$  e  $\delta(z_h)$  è la corrispondente posteriore.

FITTAZIONE DEL MODELLO GMF:

SI PARTE DA UN DATASET  $D = \{\bar{x}_1, \dots, \bar{x}_N\}$  IL QUALE È RAPPRESENTABILE TRAMITE UNA MATRICE  $\times$  DI DIMENSIONI  $N \times D$

LE CORRISPONDENTI VARIABILI LATENTI SARANNO RAPPRESENTATE DA UNA MATRICE  $Z$  DI DIMENSIONI  $N \times H$ .

SE ASSUMO CHE LE VARIABILI  $\bar{x}_m$  SONO INDEPENDENTI POSSIAMO SCRIVERE AD OTTIMEZZARE

## UNA SEGUENTE ESPRESSIONE:

$$\ln \text{IP}(\bar{x} | \bar{\pi}, \bar{\mu}, \Sigma) = \sum_{m=1}^N \ln \left[ \sum_{h=1}^H \pi_h N(\bar{x}_m | \bar{\mu}_h, \Sigma_h) \right]$$

osservo a parte il costante = 0, ottengendo:

$$\cdot \bar{\mu}_h = \frac{1}{N_h} \sum_{m=1}^N \underbrace{\delta(z_{mh})}_{\text{posterior}} \bar{x}_m$$

$$\cdot \pi_h = \frac{N_h}{N}$$

## ALGORITMO PER GMM:

1)

si iniziano  $\bar{\pi}, \bar{\mu}, \Sigma$

2)

(E-step) si calcola:

$$\delta(z_{mh}) = \frac{\pi_h N(\bar{x}_m | \bar{\mu}_h, \Sigma_h)}{\sum_{j=1}^H \pi_j N(\bar{x}_m | \bar{\mu}_j, \Sigma_j)}$$

$$N_h = \sum_{m=1}^N \delta(z_{mh})$$

3) (M-STEP) SI STIMA I PARAMETRI CON  $\delta(z_{mu})$  CORRENTI IN MODO DA  
MINIMIZZARE LA LIKELIHOOD.

$$\bar{\mu}_H^{\text{new}} = \frac{1}{N_H} \sum_{m=1}^N \delta(z_{mu}) \bar{x}_m$$

$$\Sigma_H^{\text{new}} = \frac{1}{N_H} \sum_{m=1}^N (\bar{x}_m - \bar{\mu}_H^{\text{new}})(\bar{x}_m - \bar{\mu}_H^{\text{new}})^T$$

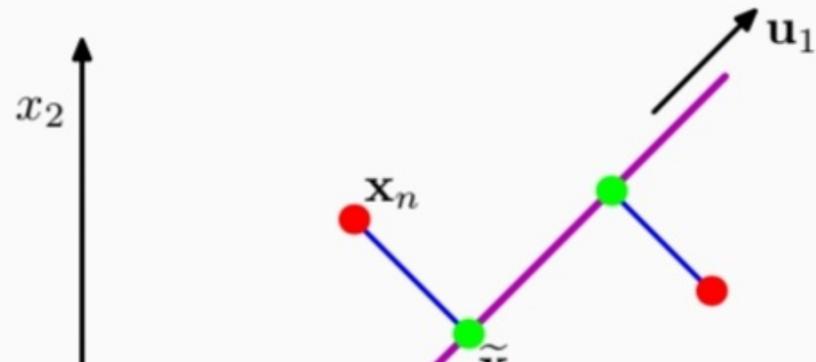
$$\bar{\mu}_H^{\text{new}} = \frac{N_H}{N}$$

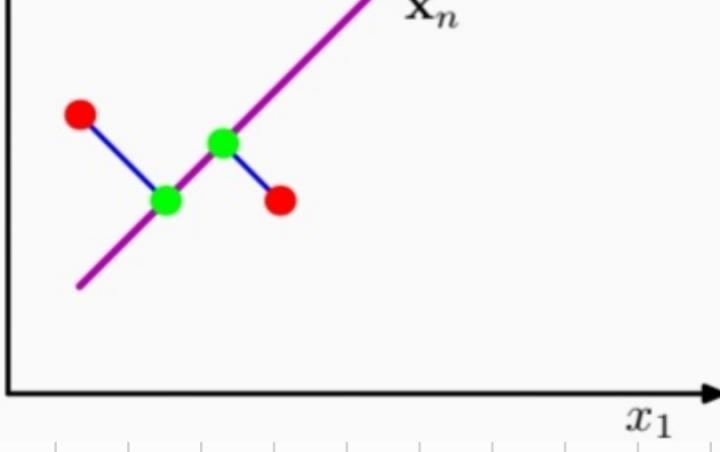
4)

SI VERIFICA L'EVENTUALE CONVERGENZA E SI RIPIEGA

PRINCIPAL COMPONENT ANALYSIS  
PCA

L'OBBIETTIVO DELLA PCA È QUELLO DI TROVARE UNA PROIEZIONE ORTOGONALE IN UN  
SOTTOSPAZIO A DIMENSIONE MINORE. UNA PRIMA IDEA È QUELLA DI MINIMIZZARE LA  
VARIANZA DEI PUNTI PROGETTATI.





$\bar{x}_n \in \mathbb{R}^d$

ASSUMEMO DI avere un certo dataset non etichettato  $D = \{x_1, \dots, x_N\}$   
 CERCHIAMO UN SOTOSPazio di dimensione  $M < D$  nel quale i punti proiettati  
hanno la varianza massima

SE consideriamo ad esempio  $M=1$ , cerchiamo trovare una direzione  $\bar{u}_1$  con norma unitaria, ovvero  $\bar{u}_1^T \bar{u}_1 = 1$

Allora la media dei punti proiettati sarà:

$$\frac{1}{N} \sum_{m=1}^N \bar{u}_1^T \bar{x}_m = \bar{u}_1^T \left\{ \frac{1}{N} \sum_{m=1}^N \bar{x}_m \right\} = \bar{u}_1^T \bar{x}$$

E la varianza proiettata

$$\frac{1}{N} \sum_{m=1}^N (\bar{u}_1^T \bar{x}_m - \bar{u}_1^T \bar{x})^2 = \bar{u}_1^T S \bar{u}_1$$

Matrice di covarianza che misura la relazione statistica tra i dati

con

$$S = \frac{1}{N} \sum_{m=1}^N (\bar{x}_m - \bar{x})(\bar{x}_m - \bar{x})^T$$

AL FINE DI MASSIMIZZARE LA VARIANZA DELL'INSERIMENTO DI UNO

PER FARLO SI COSTRUISCE LA LEGGIAMA:

$$\bar{u}_1^T S \bar{u}_1 + \lambda_1 (1 - \bar{u}_1^T \bar{u}_1)$$

E, PONENDONE IL GRADIENTE RISPETTO A  $\bar{u}_1$  PAR A ZERO:

$$\nabla_{\bar{u}_1} \left( \bar{u}_1^T S \bar{u}_1 + \lambda_1 (1 - \bar{u}_1^T \bar{u}_1) \right) = 0$$

$$\Leftrightarrow S \bar{u}_1 = \lambda_1 \bar{u}_1$$

→ AUTOGRAVE PER  $S \rightarrow$  NUOVA DIREZIONE DI PROIEZIONE  
→ AUTOGRAVE PER  $S \rightarrow$  NUOVA DIREZIONE DI PROIEZIONE

mentre la varianza nel sottospazio proiettato vale:

$$\bar{u}_1^T S \bar{u}_1 = \lambda_1$$

quindi per trovare il sottospazio ottimale dobbiamo ricavare l'autograve corrispondente al massimo autovalore  $\lambda_1$ .

Se poi vogliamo un sottospazio ridimensionale si tratta di usare  $\bar{u}_2$  per calcolare massima la varianza fra tutti i suoi vettori su  $\bar{u}_1$ .

In generale, per trovare il sottospazio  $n$ -dimensionale:

1. Si cercano  $\bar{x} \in S$  su  $D$

2. Si trova  $U \in \mathbb{R}^{D \times M}$  le cui colonne sono gli  $M$  autovetori corrispondenti agli  $M$  autovetori più grandi di  $S$

3. Si proietta i punti usando  $U^T(X - \bar{x})$

POSSIAMO quindi scrivere l'approssimazione tramite PCA di  $\tilde{x}_m$

$$\begin{aligned}\tilde{x}_m &= \sum_{i=1}^M (\tilde{x}_m^\top \tilde{u}_i) \tilde{u}_i + \sum_{i=M+1}^D (\tilde{x}_m^\top \tilde{u}_i) \tilde{u}_i \\ &= \bar{x} + \sum_{i=1}^M (\tilde{x}_m^\top \tilde{u}_i - \bar{x}^\top \tilde{u}_i) \tilde{u}_i\end{aligned}$$

CHE RAPPRESENTA UNA COMPRESIONE DELLE INFORMAZIONI PRINCIPALI  
POICHE' SI RAPPRESENTA UTILIZZANDO MOLTI VALORI

UNA PRATICA CONCRETA DI DATI PREPROCESSING E' LA STANDARDIZZAZIONE:  
SI CENTRA IL PUNTO SOTTRAENDO LA MEDIA E DIVIDENDO OGNICAUS DIMENSIONE  
ORIGINALI PER LA SUA DEVIAZIONE STANDARDO.

LA MATEMATICA DI CAVVOLTA INSERITA E' DATO DA:

$$p_{ij} = \frac{1}{N} \sum_{n=1}^N \frac{(x_{ni} - \bar{x}_i)(x_{nj} - \bar{x}_j)}{\sigma_i \cdot \sigma_j}$$

