Elisa Merigo e-mail: e.merigo1@campus.unimib.it / merigo.elisa@gmail.com
Matricola: 869436 - 8 aprile 2024

ANALISI DELLA GAP STATISTIC

(in rosso gli argomenti non presenti nell'articolo).

La gap statistic è un metodo, introdotto a Stanford nel 2000 da R. Tibshirani, per stimare il numero di cluster in un insieme di dati.
Questo metodo è progettato per poter essere usato in qualsiasi algoritmo di clustering, ma per semplicità viene analizzato utilizzando l'algoritmo K-Means.
Il centro di un cluster viene chiamato centroide, che tipicamente è la media dei punti del cluster, e, tramite il K-Means, ogni punto è assegnato ad un cluster con il centroide più vicino a tale punto. La "prossimità" può essere calcolata tramite la distanza euclidea, il cosine similarity, la correlazione, etc.
La misura più usata per valutare la bontà dei clustering K-Means è lo scarto quadratico medio (SSE, Sum of Squared Error). Il centroide che minimizza SSE quando si utilizza come misura la distanza euclidea è la media dei punti del cluster.
http://bias.csr.unibo.it/golfarelli/DataMining/MaterialeDidattico/DMISI-Clustering.pdf

Formula della gap statistic

Dati:

- $\{x_{ij}\}$ dove $i = 1, 2, \ldots, n$ e $j = 1, 2, \ldots, p$ con n che simboleggia le osservazioni indipendenti e p le caratteristiche misurate su di esse.
- $d_{ii'}$ è la distanza quadratica Euclidea misurata tra le osservazioni i e i'.

$$d_{ii'} = \sum_j (x_{ij} - x_{i'j})^2$$

- suddivido i dati in k clusters $C_1, C_2, \ldots, C_k$ e C_r denota gli indici delle osservazioni nel cluster r.
- $n_r = |C_r|$ ovvero è uguale al numero di elementi di C_r.

Da qui definisco:

$D_r = \sum_{i,i' \in C_r} d_{ii'}$ come somma delle distanze a coppie per tutti i punti nel cluster r.

e

$W_k = \sum_{r=1}^{k} \frac{D_r}{2n_r}$ come somma dei quadrati all'interno del cluster attorno alle medie del cluster (= centroidi).

La formula finale della gap statistic dunque è:

$Gap_n(k) = E_n^*\{ln(W_k)\} - ln(W_k)$ dove:

- E_n^* è il valore atteso (= la media) sotto una distribuzione di probabilità specificata.
- k è il numero di cluster che massimizzano il valore del Gap.
- Se si considerano cluster formati da n punti uniformi in p dimensioni, con k centri e si assume che i centri siano allineati in modo equispaziato, la previsione di $ln(W_k)$ è approssimativamente:

$$ln(pn/12) - (2/p)ln(k) + constant$$

https://thesis.unipd.it/retrieve/78857f63-62f1-43b3-b989-54caff174b37/Riotto_Luca.pdf

Per sviluppare la statistica del Gap bisogna:
- trovare una distribuzione di riferimento
- valutare la distribuzione campionaria della statistica del Gap (= la distribuzione campionaria è la distribuzione di tutti i possibili valori di una statistica ottenuta da campioni della stessa ampiezza estratti dalla popolazione)

https://labdisia.disia.unifi.it/rampi/Statistica2010_2011_cap15_camp.pdf

La statistica del gap confronta la variazione totale dell'intracluster per diversi valori di k con i loro valori attesi sotto distribuzione di riferimento nulla dei dati (cioè una distribuzione senza clustering evidenti).
Per cercare una distribuzione di riferimento appropriata, si considera per un momento la popolazione corrispondente alla statistica Gap nel caso del clustering K-Means:

$$g(k) = ln\{\frac{MSE_{X^*}(k)}{MSE_{X^*}(1)}\} - ln\{\frac{MSE_X(k)}{MSE_X(1)}\}$$

dove $MSE_X(k) = E(min_{\mu \in A_k} ||X - \mu||^2)$ = popolazione corrispondente a W_k

Si indica con:
- S^p l'insieme delle distribuzioni a singolo componente (o, variabili casuali = utilizzo l'ipotesi nulla)
- tutto è definito in R^p
- $A_k \subset R^p$ è il set di k-punti scelti per minimizzare $||X - \mu||^2$
- MSE è l'errore quadratico medio che corrisponde alla discrepanza quadratica media tra i valori dei dati osservati e i valori dei dati stimati. Per uno stimatore imparziale (= stimatore in cui non ci sono bias. Bias di uno stimatore = differenza tra valore atteso dello stimatore e valore reale del parametro stimato) MSE è la varianza dello stimatore.

Si sottrae $ln\{\frac{MSE_X(k)}{MSE_X(1)}\}$, che sono i logaritmi delle varianze, per avere $g(1) = 0$.

Ora si deve cercare la minima distribuzione favorevole a singola componente su X^* tale che $g(k) \leq 0 \; \forall X \in S^p e \; \forall k \geq 1$

Il *Teorema 1* dimostra che come distribuzione di riferimento, nel caso di distribuzioni univariate, è utile prendere la distribuzione uniforme:

Elisa Merigo e-mail: e.merigo1@campus.unimib.it / merigo.elisa@gmail.com
Matricola: 869436 - 8 aprile 2024

Teorema 1: Sia $p = 1$, allora $\forall k \geq 1$ si ha

$$inf_{X \in S^p}\{\frac{MSE_X(k)}{MSE_X(1)}\} = \frac{MSE_U(k)}{MSE_U(1)} \text{ dove } \frac{MSE_U(k)}{MSE_U(1)} = \frac{1}{k^2}$$

Analizzando, però, le distribuzioni multivariate questo approccio generalmente fallisce. Infatti il *Teorema 2* dice:

Teorema 2: Se $p > 1$ allora nessuna distribuzione $U \in S^p$ può soddisfare l'equazione sopra citata a meno che il suo supporto sia degenere al sottoinsieme di una linea

- supporto: in ambito probabilistico corrisponde ai casi possibili. https://www.quora.com/How-can-a-probability-distribution-work-with-the-subset-of-the-support
- sottoinsieme di una linea: punto. Una linea è un insieme di punti.

Tramite dei semplici calcoli si può dimostrare che una distribuzione di riferimento con un supporto degenere risulta una procedura inefficace.
Nella distribuzione multivariata non si sarà in grado di scegliere una distribuzione di riferimento applicabile in modo generale e utile.
Una soluzione a questo problema potrebbe essere quella di generare dei dati di riferimento a partire dalla stima della massima verosimiglianza (= maximum likelihood estimate = MLE) in S^p.

La MLE è un metodo per stimare i parametri di una distribuzione di probabilità presunta che deriva da alcuni dati osservati. Questo metodo massimizza una funzione di verosimiglianza, secondo il modello statistico presunto, in modo tale che i dati osservati siano più probabili. (si basa su un criterio di somiglianza).

In una dimensione, MLE può essere calcolata con l'aiuto dell'algoritmo iterativo del minimo convesso (Walther, 2000). In dimensioni maggiori, non può essere calcolato.

Come implementare la statistica gap:

- il teorema 2 riguarda la varianza strutturale.
- per la distribuzione di riferimento considero due metodi:
    1. genero ciascuna caratteristica di riferimento nell'intervallo dei valori osservati per tale caratteristica. Questo metodo è chiamato Gap/unif.
    2. genero la caratteristica di riferimento dalla distribuzione uniforme su una casella allineata con componenti principali dei dati. Siamo nella matrice X formata da n righe e p colonne e supponiamo che le colonne delle medie siano 0.

Si calcola:

- la decomposizione a valori singolari $X = UDV^T$ con U e V matrici ortogonali e D matrice dei valori singolari.
- $X' = XV$
- si genera Z' nell'intervallo delle colonne di X' utilizzando il metodo 1. Si trasforma $Z = Z'V^T$

Questo metodo è chiamato Gap/pc.

Il primo metodo risulta essere il più semplice, mentre nel secondo si tiene conto della distribuzione dei dati.
https://it.wikipedia.org/wiki/Decomposizione_ai_valori_singolari

In entrambi i casi, stimo $E_n^*\{ln(W_k)\}$ generando B copie di $ln(W_k{}^*)$ ognuna delle quali è calcolata a partire dal campione Monte Carlo $X_1^*, \ldots, X_n^*$ prelevato dalla distribuzione di riferimento.

Dunque: $Gap_n(k) = E_n^*\{ln(W_k)\} - ln(W_k) = \frac{1}{B}\sum_{b=1}^{B}\{ln(W_k^b)\} - ln(W_k)$

La simulazione Monte Carlo è un modello di previsione che prevede un insieme di possibili risultati sfruttando le distribuzioni probabilistiche, quali l'uniforme e la normale per esempio. Il suo compito è quello di ricalcolare ancora e ancora i risultati, utilizzando ogni volta una serie diversa di numeri casuali compresi tra il valore massimo e il valore minimo del set di dati su cui stiamo applicando tale simulazione.

Ora si può assemblare la distribuzione campionaria della statistica del Gap.
Si denota con:

- $sd(k)$ la deviazione standard delle B copie.
- $s_k = \sqrt{(1 + 1/B)}\, sd(k)$ l'errore di simulazione in $E_n^*\{ln(W_k)\}$. Questo errore è stato utilizzato in vari studi e si è dimostrato molto utile.

Usando queste formule, si sceglie come dimensione del cluster $\hat{k}$ la più piccola k in modo tale che:

$$Gap(k) \geq Gap(k + 1) - s_{k+1}$$

Come calcolare la Statistica Gap:

- passo 1: prendere i dati osservati, clusterizzarli variando i numeri dei clusters $k = 1, \ldots, K$ dando come misure di intra-dispersione $W_k, k = 1, \ldots, K$.
- passo 2: generare set di dati B di riferimento utilizzando uno dei due metodi sopra discussi, e cluster ognuno dei quali con misure di intra-dispersione $W_{kb}^*, b = 1, \ldots, B, k = 1, \ldots, K$.
  Calcolare la stima della statistica del Gap:

$$Gap(k) = (1/B)\sum_b ln(W_{kb}^*) - ln(W_k)$$

- passo 3: sia $\bar{l} = (1/B)\sum_b ln(W_{kb}^*)$ e si calcoli la deviazione standard

Elisa Merigo e-mail: e.merigo1@campus.unimib.it / merigo.elisa@gmail.com
Matricola: 869436 - 8 aprile 2024

$$sd_k = [(1/B)\sum_b \{ln(W^*_{kb}) - \bar{l}\}^2]^{1/2}$$

si definisca $s_k = \sqrt{(1 + 1/B)}\, sd(k)$.

Ora si può determinare il numero dei cluster con:

$$\hat{k} = il\ più\ piccolo\ k\ tale\ che\ Gap(k) \geq Gap(k+1) - s_{k+1}$$

Il valore minimo che la statistica Gap può avere è 0. Quando Gap = 0 significa che i cluster non hanno distanza tra di loro, dunque è considerato come valore non buono.
Il valore massimo che la statistica Gap può avere è sicuramente un valore positivo, ma non si è ancora scoperto quanto è.

Tramite degli studi di simulazioni si è venuto a conoscenza del fatto che la stima della statistica Gap è buona quando i cluster sono ben separati.
Nella situazione in cui i dati non sono ben separati, non si sa precisamente come agire tramite i cluster, poichè tale circostanza non è ancora definita in letteratura.

[La distanza intra-cluster non può essere maggiore della distanza inter-cluster.
Il modo più semplice per calcolare l'inter-cluster è trovare la distanza tra i due centroidi.]

Considerazioni finali:
La Gap statistic viene utilizzata per determinare il numero ottimale di cluster. Essa analizza la distribuzione dei dati all'interno dei clusters: se la gap ha un valore alto, allora i dati sono stati raggruppati in modo significativo. Se ha un valore basso vuol dire che i dati non sono ben raggruppati e quindi servirebbero più clusters.
Importante è anche la distanza tra i cluster; nell'analisi del cluster bisogna massimizzare la distanza inter-cluster e minimizzare quella intra-cluster. Se la gap ha un valore alto vuol dire che i dati sono raggruppati in modo significativo, che non ci sono outliers e che quindi i cluster sono ben separati tra di loro.

Se i dati sono distribuiti in modo simil uniforme avremo una gap alta. Quando sono distribuiti uniformemente avremo una gap bassa.

SCOPERTA:

$$Gap_n(k) = E^*_n\{ln(W_k)\} - ln(W_k)$$

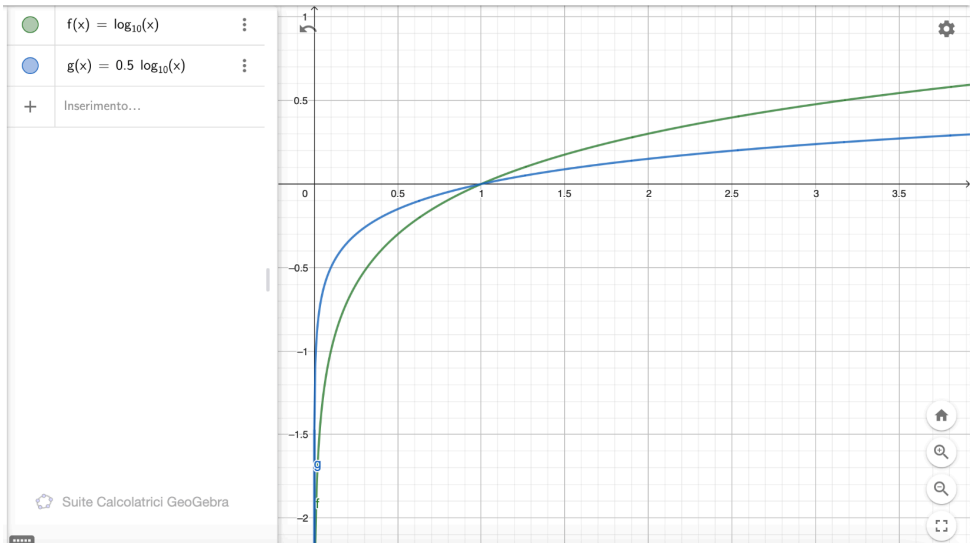
La Gap statistic risulta NEGATIVA quando:

- $E < 1\ e\ ln(W_k) > 0$
- $E < 1\ e\ ln(W_k) < 0$

La Gap statistic risulta POSITIVA quando:

- $E \geq 1\ e\ ln(W_k) > 0$
- $E \geq 1\ e\ ln(W_k) < 0$

https://www.mathworks.com/matlabcentral/answers/518864-why-are-negative-gap-statistic-values-provided-by-evalclusters-allowed-as-solution
→ Se il valore di Gap è negativo, la curva di $ln(W_k)$ è sopra la curva di riferimento $E_n^*\{ln(W_k)\}$. Ma Tibshirani dice che: " il valore ottimale di cluster è il valore di k per il quale $ln(W_k)$ cade più al di sotto della curva di riferimento $E_n^*\{ln(W_k)\}$".



Da quando $x = 1$ la curva $ln(W_k)$ si trova al di sopra della curva $E_n^*\{ln(W_k)\}$, contraddicendo ciò che Tibshirani afferma.
In questo articolo viene suggerito di mettere nel codice la gap maggiore di zero, in modo tale da rispettare la teoria di T.
Suggeriscono, inoltre, che quando tutti i valori della gap sono negativi bisogna prendere come numero di cluster ottimale 1, indicando nessun clustering. (In https://stackoverflow.com/questions/65799784/find-the-number-of-clusters-using-clusgap-function-in-r dicono ancora che quando gap negativa devo prendere k=1).

https://stats.stackexchange.com/questions/600195/negative-gap-statistic-interpretation-for-cluster-analysis → in questo articolo si parla del fatto che la gap statistic si basa sul logaritmo ed esiste un significato inerente al suo segno. Quindi se W_k è compreso tra $[0, 1)$, $ln(W_k)$ è negativo. Se W_k è compreso tra $[1, +\infty)$, $ln(W_k)$ è positivo.

La gap statistica può essere negativa, ma ciò significa che i dati non sono divisi in cluster, poiché i dati sono tutti dispersi e non raggruppati. A pagina 415, T. dice che in questo caso il k stimato è pari a 1.

Il range negativo è $(-\infty, 0)$. Dunque tale metrica può assumere valori tra (- infinito, + infinito), ma i valori ottimali sono in (0, + infinito). 0 è un caso limite.

Ulteriori considerazioni:
Il fatto che la gap possa essere sia positiva che negativa dipende dalla struttura del cluster:

- Se prendessimo un cluster patatoide con dati reali e un cluster sempre patatoide formato da dati presi da una generazione randomica, avremmo che la distanza media tra i dati reali (W_k) nel primo caso è inferiore a quella dei dati randomici ($E^*(W_k)$) $\rightarrow E^*(W_k) > W_k$. In questo caso dunque la gap risulterebbe positiva.
- Se prendessimo un cluster con i dati reali sparsi sulla cornice del cluster stesso e un cluster formato da dati presi da una generazione randomica sparsi all'interno del cluster stesso, avremmo che la distanza media dei dati reali (W_k) è maggiore a quella dei dati randomici ($E^*(W_k)$) $\rightarrow W_k > E^*(W_k)$. In questo caso dunque la gap è negativa.