Risposte Laboratorio 6

Membri del gruppo:

- Francesca Del Nin (1179732)
- Stefano Lia (1177743)
- Eugen Saraci (1171697)

Domanda 3

L'algoritmo kmeans ha una complessità O(q*n*k) mentre hierarchical clustering, che avrebbe complessità O(n^3) usando la funzione *SlowClosestPair*, ha complessità O((n-k)(n*log(n))) utilizzando *FastClosestPair* dal momento in cui quest'ultima ha complessità O(n*log(n)). Assumendo che k-means usi un piccolo numero di iterazioni q ed un numero di cluster k piccolo rispetto al numero dei punti n allora entrambi i numeri possono essere considerati come costanti e di conseguenza il tempo asintotico di k-means può essere considerato approsimativamente lineare sul numero dei punti n. Per quanto riguarda invece hierarchical clustering se il numero di cluster k è piccolo allora il numero di iterazioni è (n-k) che però può essere considerato approssimativamente n, di conseguenza il tempo asintotico totale dell'algoritmo diventa O(n^2(log(n))).

Quindi il più veloce è chiaramente **kmeans**.

Domanda 6

Algoritmi	Distorsione
K-means	2.814 x 10^11
Hierarchical Clustering	2.251 x 10^11

Domanda 7

La differenza che si crea tra i cluster prodotti è determinata dal modo con cui i due algoritmi procedono nella creazione dei cluster. Per quanto riguarda l'algoritmo **k-means**, esso sceglie come centroidi iniziali le città più popolose degli Stati Uniti. Nella costa occidentale però esse sono molto vicine tra di loro e sono concentrate tutte nella parte sud. Ciò causa dei cluster più allungati all'inizio per via di alcuni punti situati abbastanza lontani (più nello specifico nella parte nord-occidentale) dal centroide più vicino.

L'algoritmo poi procede eseguendo (più volte) due step fondamentali che sono: l'assegnamento dei punti al centroide più vicino e il ricalcolo del centroide. Nella costa occidentale quindi la situazione che si crea alla prima iterazione è quella in cui bisogna assegnare alcuni punti, quelli nella parte nord-occidentale della mappa, a centroidi che in realtà si trovano relativamente molto lontani, questo provoca quindi un elevato valore di distorsione. Nel secondo step però i centroidi vengono ricalcolati causando un loro spostamento verso i punti più a nord e man mano che le iterazioni avanzano i centroidi si distribuiscono sempre più, decrementando la distorsione, e i cluster assumono forme un po' più tondeggianti. Dopo cinque iterazioni però la situazione ancora non è ideale: si nota, infatti, che alcuni cluster (evidenti nella zona occidentale) hanno una forma "allungata" con ancora una distorsione elevata.

Per quanto riguarda invece l'algoritmo **hierarchical clustering** la situazione è differente. Esso procede creando inizialmente un cluster per ogni punto e man mano "unisce" quelli che sono più vicini tra di loro fino ad arrivare al numero di cluster k desiderato. In questo caso, sin dall'inizio, essendo i cluster distribuiti in maniera più uniforme, l'algoritmo procede individuando ed unendo i punti più vicini tra di loro causando una distorsione inferiore rispetto all'algoritmo k-means. Ciò risulta evidente nell'immagine prodotta in cui i cluster sono di forma meno allungata (e quindi con una minore distorsione).

Domanda 8

In base a quanto detto nella domanda 7 è chiaro che l'algoritmo **kmeans** richiede una maggiore supervisione umana. È necessario infatti controllare che il numero di iterazioni sia sufficiente per ottenere cluster con una distorsione sufficientemente bassa. Quindi è necessario trovare il valore migliore per l'iperparametro q della funzione k-means.

Domanda 10

Le performance danno risposte contrastanti. Sebbene nel cluster con 111 contee l'algoritmo hierarchical clustering sembri avere delle performance migliori, negli altri due casi non è evidente quale dei due sia preferibile, infatti i risultati prodotti sembrano essere migliori per un algoritmo o per l'altro a seconda del numero di cluser considerato. Questo caso sembra rientrare nel teorema denominato **no free lunch theorem**, dal momento in cui nessuno dei due algoritmi produce in modo coerente risultati con distorsione inferiore rispetto all'altro.