Big Data Homework 2

April 14, 2021

Exercise 1

Si consideri un sistema di raccomandazione basato su NBI con $U_1, \cdots, U_n$ utenti e $O_1, \cdots .O_m$ oggetti. Sia $a_{x,j}$ la utility matrix (contente il rating dell'oggetto x fatto dall'utente i).

$$w_{i,j} = \frac{1}{\Gamma(i,j)} \sum_{l=1}^{m} \frac{a_{i,l} \times a_{j_l}}{D(u_l)}$$

La matrice W indica quanto piacerà l'oggetto i ad un utente a cui piace l'ogetto j. Con $\Gamma(i,j)$ definito come $D(o_i)$ o $D(o_j)$. La matrice W conterrà m righe ed m colonne. Supponiamo di effetturare la raccomandazione usando il prodotto matriciale.

$$R = W \cdot A$$

Supponiamo di considerare il rating predetto per l'oggetto x da parte dell' utente i. Questo puo' essere rappresentato come:

$$\hat{r}_{x,i} = \sum_{l=1}^{m} w_{x,l} \times a_{l,i} = \sum_{l=1}^{m} \frac{1}{\Gamma(x,l)} \left[\sum_{h=1}^{n} \frac{a_{x,h} \times a_{l,h}}{D(u_h)} \right] \times a_{l,i}$$

Definiamo:

$$v_{x,l} = \sum_{h=1}^{n} \frac{a_{x,h} \times a_{l,h}}{D(u_h)}$$

Si noti che $v_{x,l}$ può essere ottentuto direttamente dal prodotto per A e A^t. Inoltre rimuoviamo dall'equazione la funzione arbitraria Γ e introduciamo una incognita γ. L'equazione diventerà:

$$\hat{r}_{x,i} = \sum_{l=1}^{m} \gamma_{x,l} \times v_{x,l} \times a_{l,i}$$

Misuriamo la qualità della predizione in funzione del' Root Mean Square Error (RMSE). Sia R l'insieme delle coppie di rating utente oggetto che desideriamo predire. L'errore sarà definito come:

$$RMSE = \frac{1}{|R|} \sqrt{\sum_{(x,i) \in R} \left(\hat{r}_{x,i} - r_{x,i}\right)^2}$$

quindi:

$$RMSE = \frac{1}{|R|}\sqrt{\sum_{(x,i)\in R}\left(\sum_{l=1}^{m} \gamma_{x,l} \times \upsilon_{x,l} \times a_{l,i} - r_{x,i}\right)^2}$$

1. Fornire le equazioni (incluso le derivate) di un algoritmo basato su Gradiente Discentente per stimare γ che minimizzi RMSE.
2. Implementare l'algoritmo GD o SGD e valutarne le presetazioni sui dati presenti in `http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/`. Usare i dati Enzyme: Adjacency matrix of the gold standard drug-target interaction data (text file):

Exercise 2

Supponiamo di voler risparmiare tempo quando effettuiamo la permutazione random nel min-hashing. Possiamo focalizzare la nostra attenzione a su k righe delle n, invece di fare l'hashing su tutte n. L'aspetto negativo di tale scelta è che se tra le k righe non ne troviamo nessuna che contiene un 1 in una colonna il risultato del min-hashing sarà "sconosciuto". Per quella colonna non otteniamo il valore di min-hash. Sarebbe un errore assumere che due colonne, entrambe con valore "sconosciuto" siano simili. In ogni caso se il valore di probabilità di ottenere "sconosciuto" è basso, possiamo tollerare tale situazione e ignorare questo valore di min-hash quando calcoliamo la similarità.

1. Supponiamo che una colonna contiene m 1 e quindi $n - m$ 0. Selezioniamo in modo random k righe per effettuare il min-hashing. Dimostrare che la probabilità di ottenere un valore "sconosciuto" è al piu' $\left(\frac{n-k}{n}\right)^m$
2. Supponiamo di voler portare la probabilità di "sconosciuto" al piu' a e^{-10}. Assumiamo che n e m siano molto grandi (con n molto piu' grande di m o k). Dai una semplice approssimazione del più piccolo k che assicura che al probabilità sia al più e^{-10}. L'espressione dovrebbe essere in funzione di n ed m. Suggerimenti usa $\left(\frac{n-k}{n}\right)^m$ e ricorda che per valori di x molto grandi $\left(1 - \frac{1}{x}\right)^x \approx \frac{1}{e}$.

Exercise 3

Siano $P = \{x_0, x_1, x_2, \dots x_{n-1}\}$ e $S = \{y_0, y_1, \dots, y_{K-1}\}$ due insieme di punti in uno spazio metrico (M, d) con $K \leq \sqrt{n}$. Per ogni $y \in S$ si definisca la distanza da P:

$$d(y, P) = \min_{x \in P} d(y, x)$$

Progettare un algoritmo Map-Reduce che calcola la distanza da P per ogni punto di S. Assumere che P sia rappresentato dalle coppie (i, x_i) con $0 \leq i \leq N$ e che S sia rappresentato dalle coppie $(N + i, y_i)$ con $0 \leq i \leq k$.

Rappresentare l'insieme dei punti in due file txt (P.txt e S.txt). Fornire il codice Hadoop con i dati usati per il test.