Singular Value Decomposition for Latent Semantic Index in Information Retrieval

Anno Accademico 2016/2017

A cura di
Gemma Martini

11 marzo 2017

Indice

1 Introduzione **2**

2 Latent Semantic Indexing **2**
2.1 I dati 2
2.2 Il modello 2
2.3 I concetti nascosti 3

3 Conclusioni **3**

A Cenni di SVD **4**

1 Introduzione

Un essere umano è capace di classificare una lista di documenti in base alla loro rilevanza su un certo argomento, definito mediante parole chiave. Questo comportamento può essere replicato dai computer mediante molti algoritmi.

In questo documento viene approfondito l'utilizzo del *latent semantic indexing* (in seguito LSI) per assegnare un punteggio ad un insieme di documenti mediante una *query*, assumendo nel lettore la conoscenza della *singular value decomposition* (da qui SVD). Per i lettori meno esperti, nell'appendice A si trovano le nozioni teoriche necessarie per spiegare LSI.

2 Latent Semantic Indexing

L'obiettivo di questo algoritmo è l'assegnamento di un punteggio ad un insieme di documenti, sulla base della loro pertinenza rispetto ad un insieme di parole, detto *query* (o richiesta).

Più in dettaglio, nella prima parte vengono gettate le fondamenta per introdurre lo *sketch*, che rappresenta sinteticamente le informazioni sui documenti e le parole. Nella seconda parte, in base a questo *sketch*, viene definita la pertinenza di un documento in relazione ad una *query*.

2.1 I dati

Sia D un insieme di documenti di cardinalità d.

Sia P l'insieme di tutte le parole che compongono i d documenti, con $|P| = p$.

Sia Q l'insieme delle parole che formano la *query*, di cardinalità l.

Sia $M \in M(p, d, \mathbb{N})$ la matrice delle occorrenze dell'insieme P nell'insieme D, in particolare (a_{ij}) rappresenta quante volte il termine p_i occorre nel documento d_j.

Sia $B = M^t M \in S(d, \mathbb{N})$ la matrice di elementi (b_{ij}), che rappresentano il numero di coppie di parole uguali tra il documento d_i ed il documento d_j.

Sia $C = MM^t \in S(p, \mathbb{N})$ la matrice di elementi $(c_{i,j})$, che rappresentano il numero di coppie di parole (p_i, p_j) in ogni documento.

2.2 Il modello

Con i dati definiti come sopra è possibile decomporre M mediante SVD, ottendendo tre matrici:

Matrice sinistra dei vettori singolari $S \in M(p, r, \mathbb{R})$

Matrice destra dei vettori singolari $U \in M(d, r, \mathbb{R})$

Matrice diagonale dei valori singolari $\Sigma \in D(r, \mathbb{R})$

tali che $A = S\Sigma U^t$. La matrice Σ ha, sulla diagonale, elementi decrescenti; di conseguenza gli ultimi elementi possono essere trascurati. Si può decidere di ridurre la matrice Σ ad una matrice in $M(k, \mathbb{R})$, ottenendo Σ_k, S_k e U_k, per calcoli più veloci e per ridurre il rumore dovuto a dati non significativi.

$A_k = S_k\Sigma_k U_k$ è di dimensione $p \times d$, come A, e la approssima.

2.3 I concetti nascosti

In modo un po' informale è possibile definire che cos'è un "concetto nascosto": come la diagonalizzazione di una matrice descrive, tramite gli autovettori, delle direzioni privilegiate dalle quali guardare la trasformazione, così gli autovettori in S ed in U rappresentano l'informazione sui documenti e sui dati in modo più comodo.

Lo *sketch* di questo algoritmo risiede nell'interpretazione di $S_k\Sigma_k$ e $\Sigma_k U_k^t$ come rappresentazione essenziale dei termini e dei documenti in termini di combinazione dei concetti.

Di conseguenza la *query* è un concetto modellato come $q = \frac{\sum_{i=1}^{l} (S_k\Sigma_k)_i}{l}$.

Concludendo, la pertinenza di un documento è espressa dalla distanza del coseno tra i due vettori d_i e q, ossia $\frac{d_i q}{|d_i||q|}$.

3 Conclusioni

Si conclude che, una volta calcolata la SVD della matrice termini-documenti e scelto il numero di elementi da considerare, è pressochè immediata una gerarchia di documenti ordinati in base alla loro pertinenza rispetto alle richieste.

A Cenni di SVD

Si supponga di avere i dati della sezione 2.1, senza interesse all'interpretazione che è stata data nell'ambito dell'*information retrieval*. Ossia i seguenti

- $M \in M(p, d, \mathbb{N})$, con $p > d$
- $B = M^t M \in S(p, \mathbb{N})$
- $C = MM^t \in S(d, \mathbb{N})$

Valgono i seguenti lemmi:

Lemma A.1
B e C sono simmetriche.

Dimostrazione.
$B^t = (M^t M)^t = M^t M = B$
$C^t = (MM^t)^t = MM^t = C$ □

Lemma A.2
Se $N = R^t R$ allora N è semi-definita positiva.

Dimostrazione. La tesi equivale a $x^t R^t R x \geq 0$, ma $x^t R^t R x = (Rx)^t (Rx) \geq 0$, perchè il prodotto scalare standard è definito positivo. □

Valgono inoltre le ipotesi del seguente teorema:

Teorema A.3 (Teorema spettrale)
Se $P \in S(n, \mathbb{R})$ esistono $x_1, x_2, \cdots, x_n$ autovettori ortonormali di P, con autovalori $\lambda_1, \lambda_2, \cdots, \lambda_n$ reali.

Corollario A.4
Sia B che C hanno autovalori reali non negativi.

Tali autovalori sono dunque quadrati di numeri reali non negativi, ordinati in senso decrescente $\sigma_1^2 \geq \sigma_1^2 \geq \cdots \geq \sigma_d^2$, tali che $\sigma_1, \sigma_2, \cdots, \sigma_r \in \mathbb{R}^+$ e $\sigma_{r+1}, \cdots, \sigma_d = 0$.

Quindi sia

$$U = \left(\begin{array}{c|c|c|c} & & & \\ x_1 & x_2 & \cdots & x_r \\ & & & \end{array} \right) \in M(d, r, \mathbb{R})$$

la matrice che ha per colonne gli autovettori ortonormali di B relativi ad autovalori positivi.

Siano $y_i = \frac{1}{\sigma_i} A x_i \ \forall i = 1, \cdots, r$, allora vale il seguente lemma:

Lemma A.5
Gli $y_i \ \forall i \in \{1, \cdots, r\}$ sono autovettori ortonormali per C.

Dimostrazione.

AUTOVETTORI È possibile riscrivere la tesi come $Cy_i = \lambda_i y_i$, ovvero $MM^t y_i = \lambda_i y_i$.

Si ha

$$MM^t y_i = MM^t \left(\frac{1}{\sigma_i} M x_i\right) = M \left(\frac{1}{\sigma_i} M^t M x_i\right) = M \left(\frac{1}{\sigma_i} \sigma_i^2 x_i\right) = \sigma_i^2 \frac{1}{\sigma_i} M x_i = \sigma_i^2 y_i$$

che corrisponde alla prima parte della tesi scegliendo $\lambda_i = \sigma_i^2$.

ORTONORMALI Vale la seguente catena di uguaglianze

$$\begin{aligned} y_i^t y_j &= \left(\frac{1}{\sigma_i} A x_i\right)^t \frac{1}{\sigma_j} A x_j \\ &= \frac{1}{\sigma_i \sigma_j} x_i^t A^t A x_j \\ &= \frac{1}{\sigma_i \sigma_j} x_i^t B x_j \\ &= \frac{1}{\sigma_i \sigma_j} x_i^t \sigma_j^2 x_j \\ &= \frac{\sigma_j}{\sigma_i} x_i^t x_j \end{aligned}$$

Quindi, poichè x_i e x_j sono ortonormali, si ha la tesi.

□

Sia

$$S = \left(\begin{array}{c|c|c|c} & & & \\ y_1 & y_2 & \cdots & y_r \\ & & & \end{array}\right) \in M(p, r, \mathbb{R})$$

la matrice che ha per colonne gli autovettori ortonormali relativi ad autovalori non nulli di C e si consideri la matrice $\Sigma = S^t A U$. Un suo generico elemento (i, j) vale $(S^t A U)_{ij} = y_j^t A x_i = y_j^t \sigma_i y_i = \sigma_i y_j^t y_i$, quindi, poichè gli y_i sono ortonormali, tale matrice è diagonale con elementi $\sigma_1 \dots \sigma_r$.

Inoltre, poichè S e U hanno per colonne vettori ortonormali, $SS^t = I_p$ e $UU^t = I_d$, quindi è possibile moltiplicare l'uguaglianza $S^t A U = \Sigma$ a sinistra per S e a destra per U^t, ottenendo il seguente teorema:

Teorema A.6
Sia $M \in M(p, d, \mathbb{R})$ e siano $B = M^t M$, $C = MM^t$, $U \in M(d, r, \mathbb{R})$ matrice che ha per colonne gli autovettori ortonormali relativi ad autovalori non nulli di B e $S \in M(p, r, \mathbb{R})$ matrice che ha per colonne gli autovettori ortonormali relativi ad autovalori non nulli di

C. Allora la matrice $\Sigma = S^t AU$ è diagonale e ha per elementi le radici quadrate positive degli autovalori della matrice B, ossia

$$S^t AU = \Sigma = \begin{pmatrix} \sigma_1 & & & & 0 \\ & \sigma_2 & & & \\ & & \ddots & & \\ & & & \sigma_{r-1} & \\ 0 & & & & \sigma_r \end{pmatrix}$$

Inoltre vale che $A = S\Sigma U^t$.