Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE

Corso di Laurea in Matematica

Rappresentazione in dimensioni ridotte di documenti mediante i centroidi e i minimi quadrati

Tesi di Laurea in Analisi Numerica

Relatore:
Chiar.ma Prof.ssa
Valeria Simoncini

Presentata da:
Francesca Bevilacqua

Sessione Unica
Anno Accademico 2016/2017

"Meravigliarsi di tutto è il primo passo
della ragione verso la scoperta."
(Louis Pasteur)

Introduzione

Questo elaborato ha l'obiettivo di presentare alcuni metodi di riduzione di dimensione mediante fattorizzazione di matrici. In particolare ci si soffermerà sulla riduzione di dimensione di documenti: essi vengono rappresentati come vettori le cui componenti contano le occorrenze di un termine all'interno del documento. Oltre ai metodi classici, si introdurranno due metodi recenti di riduzione e un algoritmo di classificazione che tengono conto della conoscenza a priori dei dati.
L'importanza di una buona riduzione di dimensione è legata all'estrapolazione di informazioni da grandi moli di dati; in essa il maggior costo computazionale è dato dal confronto tra due documenti/vettori. Si vogliono quindi ridurre le dimensioni dei vettori senza perdere troppe informazioni su di essi.
L'estrazione di informazioni consiste in:

- estrazione di documenti rilevanti da un database in relazione ad una domanda (*query*);
- assegnazione di un nuovo documento al gruppo di appartenenza;
- formazione di gruppi omogenei (*cluster*) in base ai valori dei vettori e ad una distanza assegnata.

Si considereranno set di dati con una struttura di cluster e ci si soffermerà sulla loro riduzione ai fini di una corretta classificazione.
Il primo capitolo introduce il concetto di matrice termini-documenti ed i metodi classici di riduzione di dimensione, con particolare attenzione alla SVD troncata.

Nel secondo capitolo vengono presentati un algoritmo di classificazione e due recenti algoritmi di riduzione (l'algoritmo dei centroidi e l'algoritmo dei centroidi ortogonali), che tengono conto della struttura di cluster dei dati.
Nel terzo capitolo sono illustrati i risultati sperimentali sugli algoritmi presentati a sostegno di quanto mostrato nei capitoli precedenti; in particolare si confronta l'accuratezza della classificazione nello spazio pieno e in quello in dimensione ridotta tramite i due metodi basati sui centroidi.

Indice

Elenco delle tabelle

Notazione

Introduciamo alcune definizioni che verranno usate nell'elaborato.

Definizione. Sia $A \in \mathbb{R}^{n\times n}$. Diciamo che A è **ortogonale** se $A^TA = AA^T = I_n$; in particolare se $A \in \mathbb{C}^{n\times n}$, si dice che A è **unitaria** se

$$A^*A = AA^* = I_n,$$

dove A^* indica la trasposta coniugata di A.

Definizione. Sia $x = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n$, si definisce la **norma-2 di** x (o norma Euclidea) come:

$$\|x\|_2 := \left(\sum_{i=1}^{n} x_i^2\right)^{1/2}.$$

Definizione. Sia $A \in \mathbb{R}^{m\times n}$, si definisce la **norma-2 indotta di** A come:

$$\|A\|_2 := \max_{0\neq x\in\mathbb{R}^n} \frac{\|Ax\|_2}{\|x\|_2} = \max_{x\in\mathbb{R}^n, \|x\|_2=1} \|Ax\|_2.$$

Questa è una norma matriciale, indotta dalla norma vettoriale Euclidea.

Definizione. Sia $A \in \mathbb{R}^{m\times n}$, si definisce la **norma di Frobenius di** A come:

$$\|A\|_F := \left(\sum_{i=1}^{m}\sum_{j=1}^{n} |A_{i,j}|^2\right)^{1/2}.$$

Capitolo 1

Rappresentazione di documenti in dimensioni ridotte

1.1 Matrici termini-documenti

Definizione 1.1. Una **matrice termini-documenti**

$$A = [a_1, a_2, \ldots, a_n] \in \mathbb{R}^{m \times n}, \quad a_j \in \mathbb{R}^m$$

è formata da una collezione di documenti, dove m è il numero dei termini (sostantivi, verbi, etc.) nella collezione di documenti e n è il numero dei documenti. Ogni colonna di A rappresenta un documento e nella matrice $A = (a_{i,j})$, $a_{i,j}$ rappresenta la frequenza del termine i-esimo nel documento j-esimo.

La più semplice forma di matrice termini-documenti è data da $A = (a_{i,j})$ con $a_{i,j} \in \{0, 1\}$ dove $1, 0$ indicano rispettivamente la presenza e assenza del termine all'interno del documento.

Osservazione 1. Generalmente il vocabolario di termini viene costituito effettuando delle semplificazioni:

- processo di stoplisting: esclusione di termini la cui frequenza è alta (es: articoli determinativi/indeterminativi, preposizioni, congiunzioni, etc...);

- processo di stemming: identificazione di termini mediante la loro radice.

Data una matrice termini-documenti $A \in \mathbb{R}^{m\times n}$ e un intero $k \ll \min(m,n)$, si vuole trovare una trasformazione $G^T \in \mathbb{R}^{k\times m}$ che manda ogni colonna a_i di A in un vettore k-dimensionale y_i , $i = 1,\ldots,n$

$$G^T : a_i \in \mathbb{R}^m \longmapsto y_i \in \mathbb{R}^k \quad 1 \leq i \leq n. \tag{1.1}$$

Trovata la trasformazione G^T, ogni vettore $q \in \mathbb{R}^m$ può essere rappresentato in k dimensioni come $\hat{q} = Q^T q \in \mathbb{R}^k$.
Invece che cercare esplicitamente la mappa G^T, si cerca di risolvere il problema di approssimazione

$$A \approx BY \tag{1.2}$$

dove $B \in \mathbb{R}^{m\times k}$, $Y \in \mathbb{R}^{k\times n}$, $\text{rank}(B) = \text{rank}(Y) = k$ e le colonne $\{y_i\}_{i=1,\ldots,n}$ di Y sono le rappresentazioni dei documenti della matrice A in dimensione ridotta. Una volta scelta la matrice B, un nuovo vettore $q \in \mathbb{R}^m$ può essere rappresentato in k dimensioni come $\hat{q} \in \mathbb{R}^k$ risolvendo il problema di minimo

$$\min_{\hat{q}\in\mathbb{R}^k} \|B\hat{q} - q\|_2. \tag{1.3}$$

Osservazione 2. Tale fattorizzazione non è unica, in quanto $\forall Z \in \mathbb{R}^{k\times k}$ non singolare

$$A = BY = \underbrace{(BZ)}_{=\hat{B}}\underbrace{(Z^{-1}Y)}_{=\hat{Y}}$$

e $rank(BZ) = k$, $rank(Z^{-1}Y) = k$.

Questo problema di fattorizzazione "approssimata"può essere affrontato in due modi diversi ma collegati: il primo in termini della formula di riduzione del rango di una matrice e il secondo come soluzione di un problema di minimo.

Osservazione 3. Osserviamo che, quando un documento viene ridotto, generalmente non c'è una connessione tra le componenti del vettore ridotto e una specifica parola o ambito semantico. Ci sono algoritmi specifici che hanno questa funzione, ma non ne tratteremo.

1.2 Teorema di riduzione del rango di una matrice

Teorema 1.2.1 (Teorema di riduzione del rango di una matrice)**.**
È data una matrice $A \in \mathbb{R}^{m\times n}$ di rango r. Allora la matrice

$$E = A - (AS)(PAS)^{-1}(PA) \tag{1.4}$$

dove $P \in \mathbb{R}^{k\times m}$ e $S \in \mathbb{R}^{n\times k}$,$k < r$, soddisfa:

$$rank(E) = rank(A) - rank((AS)(PAS)^{-1}(PA))$$

se e solo se $PAS \in \mathbb{R}^{k\times k}$ è non singolare.
Chiameremo la (1.4) *"Formula di riduzione del rango di una matrice".*

Osservazione 4. Notiamo che $T = (AS)(PAS)^{-1}P$ e $D = S(PAS)^{-1}(PA)$ sono proiettori obliqui poiché $T^2 = T$, $D^2 = D$ e non sono ortogonali.
Inoltre

$$P \cdot (AS)(PAS)^{-1} = P(AS)(PAS)^{-1} = (PAS)(PAS)^{-1} = I$$

$$(PA) \cdot S(PAS)^{-1} = (PA)S(PAS)^{-1} = (PAS)(PAS)^{-1} = I.$$

Quindi i vettori colonna di P^T e $(AS)(PAS)^{-1}$ sono biortogonali (analogamente lo sono anche i vettori colonna di $(PA)^T$ e $S(PAS)^{-1}$).

Osservazione 5. Le uniche restrizioni su P ed S riguardano le loro dimensioni e il fatto che PAS debba essere non singolare. È questa scelta di P ed S che rende flessibile la riduzione di dimensioni, e quindi ci permette di incorporare una conoscenza a priori dei dati nell'algoritmo di riduzione.
Si può fare vedere come molte fattorizzazioni possano essere ottenute dalla formula di riduzione di rango e proveremo nel prossimo paragrafo che l'approssimazione di rango k tramite SVD troncata minimizza $\|E\|_2$ o $\|E\|_{\mathrm{F}}$.

Minimizzare la matrice dell'errore E in una norma l equivale a risolvere il problema di minimo

$$\min_{B,Y} \|A - BY\|_l \tag{1.5}$$

con $B \in \mathbb{R}^{m\times k}$, $Y \in \mathbb{R}^{k\times n}$ e $rank(B) = k = rank(Y)$.
Notiamo che Eq. (1.2) ed Eq. (1.4) sono collegate, infatti:

$$BY = \underbrace{(AS)(PAS)^{-1}}_{=B}\underbrace{(PA)}_{Y}.$$

1.3 Decomposizione in valori singolari

In questa sezione introdurremo la decomposizione in valori singolari (*Singular Values Decomposition*, SVD), una decomposizione matriciale che può essere vista come un'estensione della diagonalizzazione a qualunque matrice. Ne daremo la definizione e proveremo alcune proprietà fondamentali. Alcune volte, senza specificare, considereremo $A \in \mathbb{R}^{m\times n}$ come una matrice alta ($m \geq n$).

Teorema 1.3.1. *Sia $A \in \mathbb{R}^{m\times n}$ e $q = \min\{m, n\}$. Allora esistono $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$ unitarie e $\Sigma = diag(\sigma_1, \sigma_2, \ldots, \sigma_q, 0, \ldots, 0) \in \mathbb{R}^{m\times n}$ con $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_q \geq 0$ tali che*

$$A = U\Sigma V^T. \tag{1.6}$$

Tale fattorizzazione è detta ***decomposizione in valori singolari di*** *A.*

Le colonne $\{u_i\}_{i=1,\ldots,m}$ della matrice U sono dette *vettori singolari sinistri*, le colonne $\{v_j\}_{j=1,\ldots,n}$ di V *vettori singolari destri* e gli scalari $\{\sigma_k\}_{k=1,\ldots,q}$, *valori singolari*. Indicheremo inoltre $\Sigma^* = diag(\sigma_1, \ldots, \sigma_q)$. In particolare, la decomposizione mostra che

$$Av_i = u_i\sigma_i, \quad A^T u_i = v_i\sigma_i, \quad i = 1, \ldots, n.$$

Inoltre, poiché U e V sono unitarie,

$$\|A\|_{\mathrm{F}}^2 = \|U\Sigma V^T\|_F^2 = \|\Sigma V^T\|_F^2 = \|\Sigma\|_F^2 = \sum_{i=1}^{q} \sigma_i^2. \tag{1.7}$$

Osservazione 6. Prima di iniziare la dimostrazione del teorema, osserviamo una proprietà fondamentale:

$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2 = \sigma_1$$

facilmente verificabile poiché

$$\frac{\|Ax\|_2^2}{\|x\|_2^2} = \frac{\|U\Sigma V^T x\|_2^2}{\|x\|_2^2} = \frac{\|\Sigma V^T x\|_2^2}{\|V^T x\|_2^2} = \frac{\|\Sigma^* y\|_2^2}{\|y\|_2^2} =$$
$$= \frac{\sum \sigma_i^2 y_i^2}{\|y\|_2^2} \leq \frac{\sum \sigma_1^2 y_i^2}{\|y\|_2^2} = \sigma_1^2 \frac{\|y\|_2^2}{\|y\|_2^2} = \sigma_1^2.$$

Inoltre tale valore è raggiunto per $x = v_1$, infatti si ha

$$\|Av_1\|_2 = \|u_1\sigma_1\|_2 = \sigma_1.$$

Dimostrazione.
Consideriamo $\max_{\|x\|_2=1}\|Ax\|_2$ e x il vettore soluzione. Sia $Ax = \sigma_1 y$ con $\|y\|_2 = 1$.
Definiamo $X_1 = [x, \hat{X_2}] \in \mathbb{R}^{n\times n}$ e $Y_1 = [y, \hat{Y_2}] \in \mathbb{R}^{m\times m}$ in modo che siano entrambe unitarie. Quindi

$$A_1 := Y_1^T A X_1 = \begin{bmatrix} \sigma_1 & d^T \\ 0 & B \end{bmatrix}$$

(poiché $y^T Ax = \sigma_1$ e $\hat{Y_2}^T Ax = \sigma_1 \hat{Y_2}^T y = 0$).
Si osserva che

$$\frac{\|Ax\|_2^2}{\|x\|_2^2} = \frac{\|A_1 x\|_2^2}{\|x\|_2^2}$$

in quanto A_1 è ottenuta da A mediante trasformazioni ortogonali.
Sia $x = (\sigma_1, d)^T \in \mathbb{R}^n$ con $d \in \mathbb{R}^{n-1}$:

$$\frac{\|Ax\|_2^2}{\|x\|_2^2} = \frac{\|A_1 \cdot (\sigma_1, d)^T\|_2^2}{\|(\sigma_1, d)^T\|_2^2} = \frac{1}{\sigma_1^2 + d^T d}\|A_1 \cdot (\sigma_1, d)^T\|_2^2 =$$
$$= \frac{1}{\sigma_1^2 + d^T d}\|[\sigma_1^2 + d^T d; Bd]\|_2^2 \geq \sigma_1^2 + d^T d.$$

Poiché $\sigma_1 = \max \frac{\|Ax\|}{\|x\|}$, l'unica possibilità è che $d = 0$. Quindi sia la prima riga che la prima colonna di A_1 sono zero, eccetto l'elemento diagonale. La procedura prosegue in modo iterativo con B.
Alla fine si avrà $U = Y_1 \dots Y_{n-1}$ e $V = X_1 \dots X_{m-1}$. □

Definizione 1.2. Sia $A \in \mathbb{R}^{m\times n}$ e $A = U\Sigma V^T$ la sua decomposizione in valori singolari. Se per $k \in \{1, \dots, m\}$ $\sigma_{k+1} \ll \sigma_k$, allora definiamo la **SVD troncata di A** come:

$$A \approx A_k = U \begin{pmatrix} \Sigma_k & 0 \\ 0 & 0 \end{pmatrix} V^T = \begin{pmatrix} U_k & \hat{U} \end{pmatrix} \begin{pmatrix} \Sigma_k & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V_k^T \\ \hat{V}^T \end{pmatrix} = U_k \Sigma_k V_k^T$$

con $U_k \in \mathbb{R}^{m\times k}$, $\Sigma_k = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_k) \in \mathbb{R}^{k\times k}$ e $V_k \in \mathbb{R}^{n\times k}$.

Teorema 1.3.2. *Sia $A \in \mathbb{R}^{m\times n}$ e $A_k = U_k\Sigma_k V_k^T$ la SVD troncata di A di rango k, allora:*

$$A_k = \arg \min_{B\in\mathbb{R}^{m\times n}, rank(B)=k} \|A - B\|_2$$

e

$$\|A - A_k\|_2 = \sigma_{k+1}.$$

Il teorema mostra che, tra tutte le matrici di rango k, quella che minimizza la distanza Euclidea rispetto alla matrice A è quella data dalla SVD troncata al termine k-esimo.

Dimostrazione.
Si ha k=rank(A_k) e $\|A - A_k\|_2 = \|\sum_{i>k} u_i\sigma_i v_i^T\|_2 = \sigma_{k+1}$ (analogamente a come abbiamo provato $\|A\|_2 = \sigma_1$).
Consideriamo B di rango k e $\{x_1, \dots, x_{n-k}\}$ base dello spazio nullo di B. Allora deve essere $Z := \{x_1, \dots, x_{n-k}\} \cap \{v_1, \dots, v_{k+1}\} \neq \{0\}$. Sia $z \in Z, \|z\| = 1$, allora $Bz = 0$ e $Az = \sum_{i=1}^{k+1} u_i\sigma_i v_i^T z$. Quindi:

$$\|A - B\|_2^2 \geq \|(A - B)z\|_2^2 = \|\sum_{i=1}^{k+1} u_i\sigma_i v_i^T z\|_2^2 = \sum_{i=1}^{k+1} \sigma_i^2 (v_i^T z)^2 \geq \sigma_{k+1}^2.$$

Inoltre tale valore è raggiunto per $B = A_k$, quindi

$$\sigma_{k+1} = \min_{B \in \mathbb{R}^{m \times n}, rank(B)=k} \|A - B\|_2.$$

□

Si ha un teorema analogo per la norma di Frobenius.

Teorema 1.3.3. *Sia $A \in \mathbb{R}^{m \times n}$ e $A_k = U_k \Sigma_k V_k^T$ la SVD troncata di A di rango k, allora:*

$$A_k = \arg \min_{Z \in \mathbb{R}^{m \times n}, rank(Z)=k} \|A - Z\|_F$$

e

$$\|A - A_k\|_F = \left(\sum_{i=k+1}^{r} \sigma_i^2 \right)^{1/2}$$

con $r = rank(A)$.

Dimostrazione.
Sia dato il seguente prodotto scalare definito mediante la norma di Frobenius:

$$< A, B > = \mathrm{tr}(A^T B) = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{i,j} b_{i,j}$$

con $A, B \in \mathbb{R}^{m \times n}$. Si nota che le matrici

$$u_i v_j^T \quad \forall i = 1, \ldots, m \ \forall j = 1, \ldots, n$$

costruiscono una base ortonormale di $\mathbb{R}^{m \times n}$ rispetto a questo prodotto scalare. Infatti

$$< u_i v_j^T, u_k v_l^T > = \mathrm{tr}(v_j u_i^T u_k v_l^T) = \mathrm{tr}(v_l^T v_j u_i^T u_k) = (v_l^T v_j)(u_i^T u_k).$$

Quindi le matrici $u_i v_j^T$ sono ortonormali e, essendo mn in numero, costituiscono una base di $\mathbb{R}^{m \times n}$. Scriviamo allora la matrice $Z \in \mathbb{R}^{m \times n}$ in termini di questa base:

$$Z = \sum_{i,j} \rho_{i,j} u_i v_j^T$$

dove i coefficienti $\rho_{i,j}$ devono essere determinati. Per l'ortogonalità della base considerata si ha

$$\|A - Z\|_F^2 = \sum_{i,j}(\sigma_{ij} - \rho_{ij})^2 = \sum_i(\sigma_{ii} - \rho_{ii})^2 + \sum_{i \neq j}\rho_{ij}^2 \tag{1.8}$$

dove con σ_{ij}si indicano i valori della matrice Σ. Possiamo scegliere la matrice Z in modo che essa sia diagonale e quindi il secondo addendo dell'Eq. (1.8) sia zero. Si ha quindi

$$Z = \sum_i \rho_{ii} u_i v_i^T.$$

Poiché il rango di Z è uguale al numero di termini non nulli in questa somma, imporre $\text{rank}(Z) = k$ implica che all'interno della somma ci debbano essere esattamente k termini non nulli. Indichiamo con σ_i e ρ_i gli elementi della diagonale rispettivamente di Σ e Z.
Per raggiungere il minimo in

$$\|A - Z\|_F^2 = \sum_i(\sigma_i - \rho_i)^2$$

si deve scegliere $\rho_i = \sigma_i \forall i = 1, \ldots, k$, il che dimostra la tesi, infatti:

$$Z = \sum_{i=1}^{k} \rho_i u_i v_i^T = \sum_{i=1}^{k} \sigma_i u_i v_i^T = A_k$$

e

$$\|A - A_k\|_F^2 = \sum_{i=1}^{r}(\sigma_i - \rho_i)^2 = \sum_{i=1}^{k}(\sigma_i - \sigma_i)^2 + \sum_{i=k+1}^{r}\sigma_i^2 = \sum_{i=k+1}^{r}\sigma_i^2.$$

□

Osservazione 7. Abbiamo visto che la SVD troncata di rango k di una matrice A è la migliore approssimazione di rango k per la norma-2 e la norma di Frobenius. Tale approssimazione può essere trovata anche minimizzando la norma della matrice E nella formula di riduzione del rango di una matrice. Il minimo errore è ottenuto con $P = U_k^T$ e $S = V_k$

$$\begin{aligned}(AS)(PAS)^{-1}(PA) &= (AV_k)(U_k^T AV_k)^{-1}(U_k^T A) = \\ &= (U_k\Sigma_k)(\Sigma_k)^{-1}(\Sigma_k V_k^T) = U_k\Sigma_k V_k^T\end{aligned}$$

(ricordando che $A \approx U_k\Sigma_k V_k^T$, $U_k\Sigma_k \approx AV_k$ e $\Sigma_k V_k^T \approx U_k^T A$).

Nelle procedure di classificazione, clustering e document retrieval l'operazione fondamentale è il confronto tra due documenti, pertanto è importante la scelta della misura di similarità. Nei metodi basati sugli spazi vettoriali vengono spesso usate la norma L_2, il prodotto scalare e il coseno.
Per il prodotto scalare

$$\begin{aligned} A^T A \approx A_k^T A_k &= (U_k\Sigma_k V_k^T)^T (U_k\Sigma_k V_k^T) = \\ &= V_k\Sigma_k^T U_k^T U_k\Sigma_k V_k^T = (V_k\Sigma_k^T)(\Sigma_k V_k^T). \end{aligned} \tag{1.9}$$

Quindi il prodotto scalare tra due colonne di A può essere approssimato con il prodotto scalare tra due colonne di $\Sigma_k V_k^T$, e $\Sigma_k V_k^T \in \mathbb{R}^{k\times n}$ è la rappresentazione della matrice A nello spazio di dimensioni ridotte, dato che $A \approx U_k(\Sigma_k V_k^T)$.

Osservazione 8. Questo ragionamento funziona bene anche con la norma del coseno se i vettori sono normalizzati; in generale essa è valida solo per il prodotto scalare.

Quindi, dato $q \in \mathbb{R}^m$

$$q^T A \approx q^T A_k = q^T(U_k\Sigma_k V_k^T) = (q^T U_k)(\Sigma_k V_k^T). \tag{1.10}$$

Eq. (1.10) mostra che un nuovo vettore $q \in \mathbb{R}^m$ può essere rappresentato come

$$\hat{q} = U_k^T q \tag{1.11}$$

nello spazio ridotto, poiché le colonne di $\Sigma_k V_k^T$ rappresentano le colonne di A in k dimensioni.
La rappresentazione k-dimensionale di $q \in \mathbb{R}^m$ di Eq. (1.11) può essere ottenuta anche risolvendo un problema di minimo. Abbiamo già visto che una soluzione ottima per il problema (1.2) è ottenuta dalla SVD di A con $B = U_k$ e $Y = \Sigma_k V_k^T$ nella norma 2 o di Frobenius.
Allora, risolvendo il problema di minimo (1.3) per $q \in \mathbb{R}^m$ e $B = U_k$

$$\min_{\hat{q}} \|U_k q - \hat{q}\|_2, \tag{1.12}$$

otteniamo

$$\hat{q} = U_k^T q \tag{1.13}$$

che è lo stesso risultato di Eq. (1.11).

Osservazione 9. Nell'estrazione di informazioni $\hat{q}$ deve essere confrontato con i vettori dei documenti nello spazio ridotto, ovvero con le colonne di $\Sigma_k V_k^T$. Notiamo che le colonne di $\Sigma_k V_k^T$ sono i vettori soluzione $\hat{y}_i$ che otteniamo risolvendo

$$\min_y \|U_k y - a_i\|_2$$

per $1 \leq i \leq n$.

Analogamente possiamo anche considerare $B = U_k \Sigma_k$ e $Y = V_k^T$ e risolvere il problema di minimo

$$\min_{\hat{q}} \|U_k \Sigma_k \hat{q} - q\|_2 \tag{1.14}$$

e ottenere

$$\hat{q} = \Sigma_k^{-1} U_k^T q. \tag{1.15}$$

Abbiamo quindi presentato un metodo generale per la riduzione di dimensioni basato sulla fattorizzazione $A \approx BY$ di rango k in cui il vettore ridotto $\hat{q} \in \mathbb{R}^k$ di $q \in \mathbb{R}^m$ è ottenuto risolvendo

$$\min_{\hat{q}} \|B\hat{q} - q\|.$$

In particolare abbiamo analizzato la fattorizzazione SVD troncata di rango k e il metodo di riduzione ad essa associato.

Capitolo 2

Riduzione di dimensioni di dati con struttura di cluster

Sebbene la SVD dia la migliore approssimazione BY di A in termini della minima distanza in norma-2 o di Frobenius di Eq.(1.2), la SVD non tiene conto che la matrice dei dati A abbia una struttura di cluster: le sue colonne possono essere raggruppate in un numero di cluster. Faremo vedere che ci sono altri modi per approssimare A, che sono spesso superiori alla SVD nel produrre una migliore rappresentazione in dimensioni ridotte di un data set in funzione di una classificazione successiva quando i dati hanno già una struttura di cluster.

2.1 Rappresentazione dei cluster

Per iniziare, assumiamo che il data set abbia già una struttura di cluster e sia diviso in un numero di cluster. Questa non è una restrizione in quanto, se il data set non ha una struttura di cluster, possiamo raggruppare i documenti usando uno dei tanti algoritmi di clustering come le k-medie.
Supponiamo di avere una matrice termini-documenti A, le cui colonne siano divise in k cluster. Vogliamo trovare le matrici B e Y rispettivamente con k colonne e k righe, con $A \approx BY$, in modo che i k cluster siano ben rappresen-

tati nello spazio ridotto.
Iniziamo definendo il rappresentante di ogni cluster.

Definizione 2.1. Siano $\alpha_1, \ldots, \alpha_s \in \mathbb{R}^m$ gli elementi di un cluster. Definiamo il rappresentante del cluster $c \in \mathbb{R}^m$, che chiameremo **centroide**, come:

$$c = \frac{1}{s}\sum_{i=1}^{s} \alpha_i = \frac{1}{s} Ae \in \mathbb{R}^m \tag{2.1}$$

dove $A = [\alpha_1, \ldots, \alpha_s] \in \mathbb{R}^{m\times s}$ ed $e = (1, \ldots, 1)^T \in \mathbb{R}^m$.

Osservazione 10. Il centroide è il vettore per cui si ha la minima varianza nel seguente senso: siano $\alpha_1, \ldots, \alpha_s \in \mathbb{R}^m$ i vettori del cluster, allora

$$\sum_{i=1}^{s} \|\alpha_i - c\|_2^2 = \min_{x\in\mathbb{R}^m} \sum_{i=1}^{s} \|\alpha_i - x\|_2^2 = \min_{x\in\mathbb{R}^m} \|A - xe^T\|_F^2. \tag{2.2}$$

L'equazione mostra che il centroide c minimizza la distanza, in norma di Frobenius, tra la matrice A e la sua rappresentazione di rango uno xe^T dove x deve essere determinato.

Definizione 2.2. Supponiamo che il data set rappresentato dalla matrice termini-documenti $A \in \mathbb{R}^{m\times n}$ sia diviso in k cluster. Definiamo la **matrice dei centroidi** C

$$C = [c_1, \ldots, c_k] \in \mathbb{R}^{m\times k} \tag{2.3}$$

dove $c_i \in \mathbb{R}^m$ è il centroide dell'i-esimo cluster.

2.2 Algoritmo dei centroidi per la riduzione di dimensioni

Sia $A \in \mathbb{R}^{m\times n}$ la matrice dei dati. In questo algoritmo vogliamo considerare l'approssimazione $A \approx BY$ scegliendo come B la matrice dei centroidi: $B := C$, e risolvere il problema di minimi quadrati

$$\min_{Y\in\mathbb{R}^{k\times n}} \|CY - A\|_F. \tag{2.4}$$

La soluzione $Y \in \mathbb{R}^{k\times n}$ dà una rappresentazione di A in k dimensioni.
Ogni nuovo dato $q \in \mathbb{R}^m$ è ridotto in k dimensioni risolvendo

$$\min_{\hat{q}\in\mathbb{R}^{k\times 1}} \|C\hat{q} - q\|_2. \tag{2.5}$$

Definizione 2.3. Sia N_j l'insieme di indici delle colonne di A che appartengono al j-esimo cluster. Definiamo la **matrice di raggruppamento** $H \in \mathbb{R}^{n\times k}$ come

$$H = F \cdot (F^T F)^{-1} \tag{2.6}$$

$$F(i,j) = \begin{cases} 1 & \text{se } i \in N_{\text{j}} \\ 0 & \text{altrimenti} \end{cases} \qquad F \in \mathbb{R}^{n\times k}$$

dove $F^T F \in \mathbb{R}^{k\times k}$ è diagonale con elementi della diagonale positivi, in particolare è non singolare.

Esempio 2.1. Se $n = 5$, $N_1 = (1,3,4), N_2 = (2,5)$

$$F = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad F^T F = \begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix}$$

$$H = F \cdot \begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{2} \\ \frac{1}{3} & 0 \\ \frac{1}{3} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}$$

Quindi la j-esima colonna di H ha elementi $\frac{1}{|N_j|}$ in corrispondenza degli indici dei documenti che appartengono al j-esimo cluster, e 0 altrimenti (dove $|N_j|$ indica la cardinalità dell'insieme N_j).
Per come abbiamo definito H segue che

$$C = AH. \tag{2.7}$$

Poiché $Y = arg\min_Y \|CY - A\|_F$, usando l'equazione normale si ottiene

$$Y = (C^T C)^{-1} C^T A. \tag{2.8}$$

Osservazione 11. Dalle Eq. (2.7) e (2.8), possiamo scrivere la matrice di errore $E = A - CY$ come:

$$\begin{aligned} E = A - CY &= A - (AH)(C^{\mathrm{T}}C)^{-1}C^{\mathrm{T}}A \\ &= A - (AH)(H^{\mathrm{T}}A^{\mathrm{T}}AH)^{-1}(H^{\mathrm{T}}A^{\mathrm{T}}A) \end{aligned} \tag{2.9}$$

ovvero la Formula di riduzione del rango di una matrice con $P = H^T A^T$ e $S = H$.

Questo metodo di risoluzione è sintetizzato nell'algoritmo 1.

ALGORITMO 1: (Algoritmo dei centroidi per la riduzione di dimensioni)
Dato un data set $A \in \mathbb{R}^{\mathrm{m}\times\mathrm{n}}$ con k cluster e un vettore $q \in \mathbb{R}^{\mathrm{m}}$, questo algoritmo calcola la rappresentazione k-dimensionale $\hat{q} \in \mathbb{R}^{\mathrm{k}}$ di q.

- Calcola il centroide c_{i} dell' i-esimo cluster $1 \leq i \leq k$;
- posto $C = [c_1 c_2 \dots c_{\mathrm{k}}]$;
- risolve $\min_{\hat{q}} \|C\hat{q} - q\|_2$.

2.3 Algoritmo di classificazione basato sui centroidi

Presentiamo ora un algoritmo di classificazione basato sui centroidi. Sia $A \in \mathbb{R}^{m\times n}$ la matrice dei dati con k cluster e k centroidi corrispondenti c_i, $1 \leq i \leq k$. L'algoritmo vuole assegnare ad un vettore $q \in \mathbb{R}^m$ il cluster corrispondente. Si calcolano i coefficienti di similarità $sim(q, c_i)$ tra q e ogni centroide c_i, $1 \leq i \leq k$, si trova quindi l'indice j del cluster per cui $sim(q, c_i)$ è minima (o massima).

La scelta di prendere l'indice per cui il coefficiente di similarità è minimo o massimo dipende dalla scelta di $sim(q, c_i)$:

- Se $sim(q, c_i) = \|q - c_i\|_2$ usando la norma L_2, considereremo l'indice j per cui si avrà il minimo, in quanto la distanza tra q e c_i sarà minore;
- se $sim(q, c_i) = cos(q, c_i) = \frac{q^T c_i}{\|q\|_2 \|c_i\|_2}$ considereremo l'indice j per cui si avrà il massimo.

Riassumiamo qui l'algoritmo.

ALGORITMO 2: (Algoritmo di classificazione basato sui centroidi)
Dato un data set $A \in \mathbb{R}^{m \times n}$ con k cluster e k centroidi corrispondenti $c_i \in \mathbb{R}^m$, $1 \leq i \leq k$, e un vettore $q \in \mathbb{R}^m$, l'algoritmo trova l'indice j del cluster a cui il vettore q appartiene.

- Calcola $sim(q, c_i)$ per $1 \leq i \leq k$;
- trova l'indice j per cui $sim(q, c_j)$ è minima (o massima, a seconda del coefficiente di similarità).

Abbiamo visto l'algoritmo generale; vogliamo ora applicarlo dopo aver usato l'algoritmo dei centroidi per ridurre le dimensioni del data set e del vettore q. Riduciamo le dimensioni prima per i centroidi c_i, $1 \leq i \leq k$ risolvendo

$$\min_{y \in \mathbb{R}^k} \|Cy - c_i\|_l. \tag{2.10}$$

La soluzione di questo problema di minimo è la i-esima colonna della matrice identità $e_i \in \mathbb{R}^k$, quindi i centroidi sono rappresentati come vettori unitari.

Osservazione 12. Osserviamo che i centroidi nello spazio ridotto sono ortogonali e che, con la proiezione nello spazio generato dalle colonne di C, viene massimizzata l'indipendenza dei vettori che rappresentano i centroidi.

Sia $\hat{q}$ la rappresentazione k-dimensionale di q ottenuta tramite l'algoritmo dei centroidi, nell'algoritmo di classificazione $\hat{q}$ è confrontato con e_i, $1 \leq i \leq k$ (che sono i centroidi dei cluster nello spazio ridotto). Usando la norma L_2 si cerca

$$\arg \min_{1 \leq i \leq k} \|\hat{q} - e_i\|_2; \tag{2.11}$$

esso è raggiunto per l'indice i per cui si ha la più grande componente di $\hat{q}$. Analogamente, usando la norma del coseno, poiché $\|e_i\|_2 = 1$, $1 \leq i \leq k$ e $\|\hat{q}\|_2$ è costante,

$$\arg \max_{1 \leq i \leq k} \frac{\hat{q}^T e_i}{\|\hat{q}\|_2 \|e_i\|_2} = \arg \max_{1 \leq i \leq k} \frac{\hat{q}^T e_i}{\|\hat{q}\|_2} = \arg \max_{1 \leq i \leq k} \hat{q}^T e_i \tag{2.12}$$

è raggiunto per l'indice i per cui si ha la più grande componente di $\hat{q}$.
Quindi il risultato della classificazione sarà lo stesso sia in norma L_2 che con la norma del coseno; inoltre l'algoritmo si riduce a cercare la più grande componente di $\hat{q}$, per cui il costo computazionale sarà nettamente più basso.

2.4 Algoritmo dei centroidi ortogonali per la riduzione di dimensioni

Se nell'approssimazione $A \approx BY$ di rango k la matrice B ha colonne ortogonali, allora la correlazione in A è ben approssimata dalla correlazione in Y

$$A^T A \approx Y^T B^T B Y = Y^T Y. \tag{2.13}$$

Avere una matrice B con colonne ortogonali non è così restrittivo, in quanto si può ovviare al problema facendo la fattorizzazione QR di B.

Definizione 2.4. Data una generica matrice $D \in \mathbb{R}^{s \times t}$, $s \geq t$, allora D può essere scritta come prodotto di due matrici in questo modo:

$$D = Q\hat{R} = \begin{bmatrix} Q_t & Q_r \end{bmatrix} \begin{bmatrix} R \\ 0 \end{bmatrix} = Q_t R, \tag{2.14}$$

dove $R \in \mathbb{R}^{t\times t}$, Q è una matrice $s \times s$ reale ortogonale, $Q = \begin{bmatrix} Q_t & Q_r \end{bmatrix}$ con $Q_t \in \mathbb{R}^{s\times t}$ e $Q_r \in \mathbb{R}^{s\times(s-t)}$, e le colonne di Q_t, Q_r sono ortonormali. Inoltre R è triangolare superiore e $\hat{R}$ ha le stesse dimensioni di A. La fattorizzazione $A = Q_t R$ è detta **fattorizzazione QR ridotta** di D.

Applicando la fattorizzazione QR ridotta alla matrice $B \in \mathbb{R}^{m\times k}$ si ha

$$B = Q \begin{bmatrix} R \\ 0 \end{bmatrix} = Q_k R \tag{2.15}$$

dove $Q = \begin{bmatrix} Q_k & Q_r \end{bmatrix} \in \mathbb{R}^{m\times m}$, $Q_k \in \mathbb{R}^{m\times k}$, $Q_r \in \mathbb{R}^{m\times(m-k)}$, Q con colonne ortogonali ed $R \in \mathbb{R}^{k\times k}$ triangolare superiore.

Osservazione 13. Osserviamo che moltiplicando a sinistra per $Q^T = [Q_k, Q_r]^T$ l'Eq. (2.15) si ha:

$$Q_k^T B = R \qquad Q_r^T B = 0. \tag{2.16}$$

Nell'algoritmo dei centroidi ortogonali per la riduzione di dimensioni consideriamo $A \approx BY$ con $B = C$ matrice dei centroidi, applichiamo la fattorizzazione QR ridotta a C e risolviamo il problema di minimo

$$\min_z \|Q_k z - q\|_F. \tag{2.17}$$

Il vettore soluzione $z \in \mathbb{R}^k$ sarà la rappresentazione k-dimensionale di $q \in \mathbb{R}^m$.
Riassumiamo qui l'algoritmo.

ALGORITMO 3: (Algoritmo dei centroidi ortogonali per la riduzione di dimensioni)
Dato un data set $A \in \mathbb{R}^{\mathrm{m}\times\mathrm{n}}$ con k cluster e un vettore $q \in \mathbb{R}^{\mathrm{m}}$, questo algoritmo calcola la rappresentazione k-dimensionale $\hat{q} \in \mathbb{R}^{\mathrm{k}}$ di q.

- Calcola il centroide c_{i} dell' i-esimo cluster $1 \le i \le k$;
- posto $C = [c_1 c_2 \dots c_{\mathrm{k}}]$;

- calcola la fattorizzazione QR ridotta di C: $C = Q_k R$;
- $\hat{q} = Q_k^T q$.

Osservazione 14. Notiamo che le rappresentazioni k dimensionali di un vettore $q \in \mathbb{R}^m$ con l'algoritmo 1 e con l'algoritmo 3 sono collegate. Siano y tale che si ha $\min_y \|Cy - q\|_F$ e $z = Q_k^T q$, allora

$$z = Ry.$$

2.4.1 Ordine di similarità

Mostriamo ora che la trasformazione attraverso Q_k dell'algoritmo dei centroidi ortogonali ha alcune proprietà di invarianza. In particolare i risultati di classificazione attraverso l'algoritmo di classificazione basato sui centroidi sono gli stessi nello spazio pieno e in quello ridotto con l'algoritmo 3. Introduciamo ora una definizione importante al fine di tali risultati.

Definizione 2.5. Siano $q \in \mathbb{R}^m$ e $B \in \mathbb{R}^{m\times k}$. L'**ordine di similarità** $S(q, B)$ è una k-pla di indici delle colonne di B ordinati in base alla vicinanza di q da ogni colonna di B, secondo la misura di similarità scelta.

Esempio 2.2. Se $B = \begin{bmatrix} b_1 & b_2 & b_3 \end{bmatrix} \in \mathbb{R}^{m\times 3}$ e $\|q-b_1\|_2 \leq \|q-b_3\|_2 \leq \|q-b_2\|_2$, allora $S(q, B) = (1, 3, 2)$ è l'ordine di similarità nella norma L_2.

I seguenti due teoremi mostrano che l'ordine di similarità tra un dato $q \in \mathbb{R}^m$ e la matrice dei centroidi C viene preservato dopo la riduzione di dimensione con l'algoritmo dei centroidi ortogonali.

Teorema 2.4.1. ***(La norma L_2 preserva l'ordine di similarità)*** *L'ordine di similarità $S(q, C)$ rispetto alla norma L_2 nello spazio pieno tra ogni vettore $q \in \mathbb{R}^m$ e la matrice dei centroidi $C \in \mathbb{R}^{m\times k}$ è uguale a quello nello spazio ridotto attraverso l'algoritmo dei centroidi ortogonali. Ovvero:*

$$S_{L_2}(q, C) = S_{L_2}(\hat{q}, \hat{C}), \tag{2.18}$$

dove $C = Q_k R$ è la fattorizzazione QR ridotta di C, $\hat{q} = Q_k^T q$ e $\hat{C} = Q_k^T C$.

Dimostrazione.
Poiché $Q_r^T c_j = 0$ dall'Eq. (2.16) (con $c_j = j$-esima colonna della matrice C) e Q è ortogonale, si ha:

$$\begin{aligned} \|q - c_j\|_2^2 = \|Q^T(q - c_j)\|_2^2 &= \|Q_k^T(q - c_j)\|_2^2 + \|Q_r^T(q - c_j)\|_2^2 \\ &= \|Q_k^T(q - c_j)\|_2^2 + \|Q_r^T q\|_2^2. \end{aligned} \tag{2.19}$$

Se $\|q - c_i\|_2^2 \leq \|q - c_h\|_2^2$, allora:

$$\|Q_k^T(q - c_j)\|_2^2 + \|Q_r^T q\|_2^2 \leq \|Q_k^T(q - c_h)\|_2^2 + \|Q_r^T q\|_2^2.$$

Poiché $\|Q_r^T q\|_2^2$ è costante:

$$\|Q_k^T(q - c_j)\|_2^2 \leq \|Q_k^T(q - c_h)\|_2^2, \tag{2.20}$$

ovvero

$$\|\hat{q} - \hat{c_j}\|_2^2 \leq \|\hat{q} - \hat{c_h}\|_2^2. \tag{2.21}$$

□

Teorema 2.4.2. ***(La norma del coseno preserva l'ordine di similarità)***
L'ordine di similarità $S(q, C)$ rispetto alla norma del coseno nello spazio pieno tra ogni vettore $q \in \mathbb{R}^m$ e la matrice dei centroidi $C \in \mathbb{R}^{m \times k}$ è uguale a quello nello spazio ridotto attraverso l'algoritmo dei centroidi ortogonali. Ovvero

$$S_{cos}(q, C) = S_{cos}(\hat{q}, \hat{C}) \tag{2.22}$$

dove $C = Q_k R$ è la fattorizzazione QR ridotta di C, $\hat{q} = Q_k^T q$ e $\hat{C} = Q_k^T C$.

Dimostrazione.
Sia $cos(q, c_j)$ la norma del coseno dei vettori q e c_j. Poiché $Q_r^T c_j = 0$ dall'Eq. (2.16),

$$cos(q, c_j) = cos(Q^T q, Q^T c_j) = \frac{(Q^T q)^T Q^T c_j}{\|Q^T q\|_2 \|Q^T c_j\|_2} = \frac{q^T Q_k Q_k^T c_j}{\|Q^T q\|_2 \|Q_k^T c_j\|_2}$$

e

$$cos(\hat{q}, \hat{c_j}) = cos(Q_k^T q, Q_k^T c_j) = \frac{q^T Q_k Q_k^T c_j}{\|Q_k^T q\|_2 \|Q_k^T c_j\|_2}. \tag{2.23}$$

Se $cos(q, c_j) \leq cos(q, c_h)$, allora:

$$\frac{q^T Q_k Q_k^T c_j}{\|Q^T q\|_2 \|Q_k^T c_j\|_2} \leq \frac{q^T Q_k Q_k^T c_h}{\|Q^T q\|_2 \|Q_k^T c_h\|_2}.$$

Poiché $\|Q^T q\|_2$ e $\|Q_k^T q\|_2$ sono costanti e positive, moltiplicando entrambi i membri per $\frac{\|Q^T q\|_2}{\|Q_k^T q\|_2}$ si ottiene:

$$\frac{q^T Q_k Q_k^T c_j}{\|Q_k^T q\|_2 \|Q_k^T c_j\|_2} \leq \frac{q^T Q_k Q_k^T c_h}{\|Q_k^T q\|_2 \|Q_k^T c_h\|_2} \tag{2.24}$$

ovvero

$$cos(\hat{q}, \hat{c_j}) \leq cos(\hat{q}, \hat{c_h}). \tag{2.25}$$

□

Osservazione 15. Nei teoremi precedenti abbiamo discusso l'ordine di similarità tra un vettore $q \in \mathbb{R}^m$ e la matrice dei centroidi $C \in \mathbb{R}^{m \times k}$ in quanto nell'algoritmo di classificazione basato sui centroidi ci interessano i coefficienti di similarità con i centroidi. Si può analogamente provare il teorema per una generica matrice $B \in \mathbb{R}^{m \times s}$, con fattorizzazione QR ridotta $B = Q_B R$, e la trasformazione $\hat{q} = Q_B^T q$; l'ordine di similarità, rispetto alla norma L_2 o alla norma del coseno, tra il vettore q e le colonne di B verrà mantenuto dopo la trasformazione Q_B.

Notiamo quindi che la classificazione basata sui centroidi (algoritmo 3) è la stessa sia nello spazio pieno che nello spazio ridotto usando l'algoritmo 3; è ovvio quanto si abbassi il costo computazionale della classificazione in k dimensioni mantenendo l'accuratezza della classificazione.
Nel prossimo capitolo vedremo dei risultati che mostrano l'applicazione pratica dei teoremi appena dimostrati.

Capitolo 3

Risultati sperimentali

In questo capitolo vengono presentati i risultati di alcuni test sperimentali che illustrano l'efficienza dei metodi di riduzione descritti nei capitoli precedenti. Confrontiamo i risultati della classificazione usando i dati nello spazio pieno e la loro rappresentazione ridotta mediante i nostri metodi, algoritmo dei centroidi e algoritmo dei centroidi ortogonali; per la classificazione usiamo l'algoritmo basato sui centroidi. Inoltre confrontiamo i risultati della classificazione ottenuta nello spazio ridotto con i due algoritmi presentati e con la SVD, che sappiamo essere la migliore approssimazione di rango k nel senso visto.

Nella pre-elaborazione dei dati è già stato effettuato un processo di stoplisting, per cui sono stati eliminati dall'elenco dei termini quelli più frequenti quali articoli, preposizioni, congiunzioni, etc. Non è stato invece fatto il processo di stemming, in cui si accorpano termini con la stessa radice, in quanto troviamo parole quali *communicating, communication, communications* e *communicative*, oppure *predominant, predominantly* e *predominate* come parole distinte. Gli elementi delle matrici termini-documenti usate negli esperimenti corrispondono alla frequenza del termine nel documento, e non solo alla sua presenza/assenza.
Nello specifico sono stati usati due set di dati: MEDLINE, riguardante il

settore medico, con 5735 termini e 1033 documenti, e CRANFIELD, sull'ingegneria aerospaziale, con 4563 termini e 1398 documenti. Si sono quindi considerate le matrici termini-documenti $A\_med \in \mathbb{R}^{5735\times1033}$ e $A\_cran \in \mathbb{R}^{4563\times1398}$.

3.1 Test I: confronto dei coefficienti di similarità

Nel primo test l'intento è quello di esaminare la relazione tra i dati (documenti) e i centroidi nello spazio pieno ed in quello ridotto; per questo mettiamo a confronto i coefficienti di similarità, sia in norma L_2 che in norma del coseno.
Per entrambi i data set prendiamo in considerazione solo i primi 200 documenti e li classifichiamo artificialmente in 3 gruppi di, rispettivamente, 63, 78 e 58 elementi (i primi 63 dati formano il primo cluster, i successivi 78 il secondo ed i restanti 58 il terzo); formiamo così due matrici $A\_med\_red \in \mathbb{R}^{5735\times200}$ e $A\_cran\_red \in \mathbb{R}^{4563\times200}$ ed indichiamo con MEDLINE-RED e CRANFIELD-RED i due data set ridotti. Riduciamo le dimensioni dei dati con il metodo dei centroidi e dei centroidi ortogonali, poi classifichiamo con l'algoritmo basato sui centroidi sia nello spazio pieno che in quello ridotto con i due metodi.
Le Tabelle 3.1, 3.2, 3.3 e 3.4 mostrano la distanza tra ogni dato ed i centroidi nello spazio pieno ed in quello ridotto secondo la norma L_2 e la norma del coseno per entrambi i data set; per motivi di spazio si sono presentati i risultati solo per i primi cinque documenti di ogni cluster. I numeri colorati rappresentano il miglior coefficiente di similarità (il più piccolo per la norma L_2 ed il più grande per quella del coseno): se la classificazione risulta corretta essi sono in blu, altrimenti in rosso.

In queste tabelle ritroviamo i risultati provati nel capitolo precedente. Notiamo come l'ordine di similarità nello spazio pieno venga mantenuto nel-

Tabella 3.1: Test I, coefficienti di similarità in norma L_2 per alcuni elementi del data set MEDLINE-RED nello spazio pieno ed in quello ridotto con i due metodi. Sia $q \in \mathbb{R}^{5735}$, indichiamo con $\hat{q}$ e q^* rispettivamente le sue rappresentazioni in dimensioni ridotte tramite l'algoritmo dei centroidi e dei centroidi ortogonali.

Data set MEDLINE-RED Norma L_2									
Dati	$\|a_i - c_j\|_2$			$\|\hat{a}_i - \hat{c}_j\|_2$			$\|Q_3^T a_i - c_j^*\|_2$		
	c_1	c_2	c_3	$\hat{c}_1$	$\hat{c}_2$	$\hat{c}_3$	c_1^*	c_2^*	c_3^*
a_1	12.95	13.51	13.44	1.10	2.49	2.29	1.85	4.28	4.05
a_2	12.36	12.72	12.27	0.90	1.87	1.04	2.26	3.75	1.74
a_3	10.97	11.30	11.39	0.12	1.51	1.49	0.40	2.76	3.09
a_4	11.72	11.98	12.08	0.08	1.41	1.46	0.22	2.46	2.94
a_5	11.55	11.98	12.09	0.91	2.18	2.20	1.90	3.70	4.03
a_{64}	13.44	13.09	13.62	1.65	0.35	1.74	3.17	0.91	3.87
a_{65}	15.73	15.65	15.75	0.95	0.96	1.28	3.06	2.67	3.20
a_{66}	13.17	13.01	13.41	1.11	0.42	1.49	2.20	0.76	3.35
a_{67}	8.57	8.42	8.49	1.28	0.64	0.79	2.17	1.45	1.82
a_{68}	9.72	9.90	9.95	0.44	1.07	1.04	0.75	2.06	2.27
a_{143}	15.55	15.65	14.70	2.88	2.81	1.53	6.10	6.35	3.36
a_{144}	19.23	19.28	18.10	4.07	3.95	2.73	8.99	9.10	6.24
a_{145}	7.08	7.30	6.56	1.53	1.51	0.35	2.97	3.48	1.36
a_{146}	12.49	12.57	11.79	2.15	2.06	0.77	4.43	4.66	1.64
a_{147}	7.98	8.05	7.70	1.29	1.14	0.34	2.35	2.59	1.08

Tabella 3.2: Test I, coefficienti di similarità in norma del coseno per alcuni elementi del data set MEDLINE-RED nello spazio pieno ed in quello ridotto con i due metodi. Sia $q \in \mathbb{R}^{5735}$, indichiamo con $\hat{q}$ e q^* rispettivamente le sue rappresentazioni in dimensioni ridotte tramite l'algoritmo dei centroidi e dei centroidi ortogonali.

Data set MEDLINE-RED Norma del coseno									
	$cos(a_i, c_j)$			$cos(\hat{a}, \hat{c}_j)$			$cos(Q_3^T a_i, c_j^*)$		
Dati	c_1	c_2	c_3	$\hat{c}_1$	$\hat{c}_2$	$\hat{c}_3$	c_1^*	c_2^*	c_3^*
a_1	0.26	0.06	0.10	0.96	-0.27	-0.03	0.92	0.22	0.35
a_2	0.22	0.10	0.25	0.72	-0.30	0.61	0.79	0.36	0.87
a_3	0.19	0.08	0.07	0.99	-0.09	-0.06	0.98	0.44	0.40
a_4	0.21	0.12	0.10	0.99	0.05	-0.02	0.99	0.60	0.50
a_5	0.34	0.18	0.15	0.99	-0.03	-0.05	0.99	0.51	0.44
a_{64}	0.15	0.27	0.11	0.02	0.99	-0.08	0.55	0.99	0.40
a_{65}	0.250	0.257	0.22	0.67	0.65	0.35	0.84	0.86	0.74
a_{66}	0.17	0.23	0.11	0.39	0.91	-0.09	0.72	0.97	0.45
a_{67}	0.14	0.23	0.22	-0.15	0.77	0.61	0.56	0.87	0.83
a_{68}	0.19	0.14	0.14	0.94	0.19	0.25	0.95	0.71	0.71
a_{143}	0.06	0.04	0.35	-0.31	-0.23	0.91	0.15	0.12	0.90
a_{144}	0.07	0.06	0.42	-0.33	-0.20	0.91	0.15	0.14	0.91
a_{145}	0.07	0.05	0.32	-0.26	-0.23	0.93	0.21	0.15	0.92
a_{146}	0.09	0.08	0.33	-0.26	-0.15	0.95	0.26	0.24	0.95
a_{147}	0.12	0.14	0.27	-0.14	0.11	0.98	0.46	0.51	0.99

Tabella 3.3: Test I, coefficienti di similarità in norma L_2 per alcuni elementi del data set CRANFIELD-RED nello spazio pieno ed in quello ridotto con i due metodi. Sia $q \in \mathbb{R}^{4563}$, indichiamo con $\hat{q}$ e q^* rispettivamente le sue rappresentazioni in dimensioni ridotte tramite l'algoritmo dei centroidi e dei centroidi ortogonali.

Data set CRANFIELD-RED Norma L_2									
	$\|a_i - c_j\|_2$			$\|\hat{a}_i - \hat{c}_j\|_2$			$\|Q_3^T a_i - c_j^*\|_2$		
Dati	c_1	c_2	c_3	$\hat{c_1}$	$\hat{c_2}$	$\hat{c_3}$	c_1^*	c_2^*	c_3^*
a_1	12.29	12.40	12.52	0.52	1.46	1.04	2.29	2.97	3.44
a_2	16.02	16.32	16.64	0.78	1.36	1.88	4.55	5.51	6.39
a_3	6.89	7.16	7.74	0.22	1.27	1.33	1.54	2.48	3.85
a_4	10.31	10.74	11.42	0.68	1.47	1.98	2.22	3.73	5.38
a_5	10.47	10.54	11.18	0.52	1.10	1.50	2.98	3.20	4.91
a_{64}	19.95	19.29	19.34	3.34	2.00	2.58	6.67	4.30	4.52
a_{65}	11.46	10.75	11.13	2.08	0.80	1.57	5.41	3.67	4.67
a_{66}	21.37	21.10	21.63	2.30	1.05	2.44	4.11	2.27	3.27
a_{67}	10.77	10.35	10.84	1.30	0.57	1.18	5.04	4.09	5.19
a_{68}	12.42	12.17	12.38	1.09	0.79	0.83	4.05	3.21	3.92
a_{143}	8.74	4.43	8.68	1.03	1.09	0.79	5.11	4.56	5.01
a_{144}	14.28	14.18	14,43	0.86	0.62	1.05	2.16	1.39	3.02
a_{145}	10.47	10.74	11.57	0.96	1.23	2.05	2.28	3.28	5.42
a_{146}	17.62	17.74	16.97	2.28	2.95	1.59	5.60	5.98	2.98
a_{147}	14.63	14.53	14.06	1.56	1.80	0.41	4.18	3.82	1.07

Tabella 3.4: Test I, coefficienti di similarità in norma del coseno per alcuni elementi del data set CRANFIELD-RED nello spazio pieno ed in quello ridotto con i due metodi. Sia $q \in \mathbb{R}^{4563}$, indichiamo con $\hat{q}$ e q^* rispettivamente le sue rappresentazioni in dimensioni ridotte tramite l'algoritmo dei centroidi e dei centroidi ortogonali.

Data set CRANFIELD-RED Norma del coseno									
	$cos(a_i, c_j)$			$cos(\hat{a}_i, \hat{c}_j)$			$cos(Q_3^T a_i, c_j^*)$		
Dati	c_1	c_2	c_3	$\hat{c_1}$	$\hat{c_2}$	$\hat{c_3}$	c_1^*	c_2^*	c_3^*
a_1	0.25	0.22	0.22	0.86	-0.40	0.30	0.97	0.82	0.85
a_2	0.54	0.51	0.43	0.87	0.46	-0.09	0.98	0.94	0.82
a_3	0.51	0.45	0.39	0.99	-0.001	-0.09	0.99	0.87	0.75
a_4	0.58	0.53	0.41	0.90	0.31	-0.29	0.98	0.89	0.70
a_5	0.27	0.23	0.15	0.85	0.23	-0.45	0.93	0.80	0.53
a_{64}	0.12	0.24	0.23	-0.62	0.74	0.23	0.39	0.74	0.73
a_{65}	0.08	0.18	0.16	-0.60	0.79	0.09	0.31	0.70	0.61
a_{66}	0.16	0.21	0.12	-0.14	0.93	-0.32	0.70	0.90	0.53
a_{67}	0.06	0.10	0.07	-0.40	0.90	-0.11	0.55	0.86	0.60
a_{68}	0.14	0.161	0.163	-0.15	0.75	0.63	0.84	0.95	0.96
a_{143}	0.07	0.07	0.11	-0.03	-0.36	0.93	0.63	0.66	0.96
a_{144}	0.25	0.26	0.23	0.50	0.85	0.14	0.95	0.98	0.87
a_{145}	0.54	0.51	0.37	0.74	0.54	-0.40	0.96	0.90	0.66
a_{146}	0.28	0.25	0.38	0.20	-0.56	0.79	0.67	0.61	0.93
a_{147}	0.25	0.26	0.35	0.02	-0.30	0.95	0.71	0.73	0.98

lo spazio ridotto con i centroidi ortogonali per entrambe le norme, nel caso sia di corretta classificazione, ad esempio per $a_4, a_{65}, a_{146} \in A\_med\_red$ e $a_2, a_{167} \in A\_cran\_red$, sia di una errata classificazione come $a_2 \in A\_med\_red$ e $a_{143}, a_{144} \in A\_cran\_red$. Invece esso non viene necessariamente mantenuto attraverso la riduzione con l'algoritmo dei centroidi, sia in norma L_2 che del coseno, come vediamo per $a_2, a_{65} \in A\_med\_red$ e $a_{143} \in A\_cran\_red$. Infatti nel caso di $a_2 \in A\_med\_red$ e $a_{143} \in A\_cran\_red$ la classificazione nello spazio pieno (e quindi anche in quello ridotto con i centroidi ortogonali) è sbagliata, mentre nello spazio ridotto con l'algoritmo dei centroidi è corretta; per $a_{65} \in A\_med\_red$ vale il contrario.
Inoltre si vede come la classificazione, dopo aver ridotto con l'algoritmo dei centroidi, sia la stessa sia con la norma L_2 che con quella del coseno (anche nei casi di errata classificazione), in quanto abbiamo provato che il numero del cluster a cui viene associato il documento corrisponde all'indice della più grande componente del vettore (nello spazio ridotto).
Ricordiamo come, in questo test, i cluster siano costruiti artificialmente senza seguire una relazione tra i documenti; nei test successivi aggiungeremo anche questo fattore per vedere come influenza i risultati.

3.1.1 Interpretazione del documento 68 di MEDLINE-RED

Vogliamo ora cercare di interpretare i risultati guardando nello specifico i casi difficili. Analizziamo il documento $a_{68} \in A\_med\_red$ che viene classificato in maniera errata in tutti i casi visti. Le parole presenti all'interno del documento sono: *accumulating, accumulation, acid, active, administered, appears, appreciable, biosynthesis, blood, bound, chains, components, compound, concluded, conversion, converted, derivatives, fed, fraction, glycoproteins, incorporation, intravenously, liver, microsomes, mitochondria, organ, peptide, plasma, protein, proteins, rapidly, rat, rats, recovered, released, removed, shown, soluble, stage, stream, tissues, transferred.* Guardando le

Tabella 3.5: Frequenza dei termini significativi in $a_{68} \in A\_med\_red$ e nei tre centroidi nello spazio pieno.

Documento a_{68} di MEDLINE-RED				
Termini	Frequenza			
	a_{68}	c_1	c_2	c_3
fegato	5	0.2857	0.2152	0.1552
proteine	3	0.1429	0.0633	0.1379
somministrato	2	0.1429	0.0380	0.0345
frazione	2	0.0159	0.0506	0.1034
glicoproteine	2	0	0.0253	0
plasma	2	0.2222	0.0633	0.1552
rapidamente	2	0.0317	0.1013	0.0862
rilasciato	2	0	0.0380	0
solubile	2	0	0.0253	0.1724
tessuti	2	0.0952	0.1266	0.2241

frequenze dei termini nel documento 68 si nota che la più alta è data dalla parola *liver* (fegato), seguita da *proteins* (proteine), *administered* (somministrato), *fraction* (frazione), *glycoproteins* (glicoproteine), *plasma*, *rapidly* (rapidamente), *released* (rilasciato), *soluble* (solubile) e *tissues* (tessuti). La Tabella 3.5 contiene la frequenza dei termini appena elencati nel documento a_{68} e nei tre centroidi; sono segnati in blu i numeri per cui si ha la minima distanza rispetto alla frequenza nel documento in esame. Si è deciso di riportare i risultati solamente per i termini con frequenza ≥ 2 nel documento.

Ricordiamo che, per il modo in cui sono stati costruiti i tre cluster, il documento a_{68} appartiene al secondo cluster. Dalle Tabelle 3.1 e 3.2 si vede che invece esso viene classificato come elemento del primo cluster, anche se per pochi decimi, in quanto i coefficienti di similarità rispetto ai tre centroidi risultano molto vicini tra loro per entrambe le norme. Notiamo come nella Tabella 3.5 i numeri in blu siano abbastanza equidistribuiti tra i centroidi, con

Tabella 3.6: Termini più frequenti in $a_{68} \in A\_med\_red$ e nei tre cluster

Termini più frequenti MEDLINE-RED				
	a_{68}	c_1	c_2	c_3
1	fegato	pazienti	pazienti	dna
2	proteine	nickel	sangue	lente
3	somministrato	renale	ventricolare	cellule
4	frazione	fetale	cellule	subtilis
5	glicoproteine	amiloidosi	aortico	fago
6	plasma	selenium	cardiaco	acido
7	rapidamente	cellula	percentuale	cellula
8	rilasciato	malattia	casi	pazienti
9	solubile	grasso	malattia	rna
10	tessuti	fegato	normale	tumore
11	raccogliendo	acido	corpo	umano
12	accumulazione	sangue	trovato	normale
13	acido	sindrome	cuore	bacillo
14	attivo	ratti	sinistra	tessuto
15	appare	amiloide	difetto	effetto

una leggera inclinazione verso il primo per cui viene minimizzata la distanza rispetto ai termini del documento più significativi, come *fegato* e *proteine*. Mettiamo a confronto le 15 parole più frequenti di ogni cluster con quelle del documento; per quelle con uguale frequenza si segue l'ordine alfabetico.

Dalla Tabella 3.6 si vede come il termine *fegato* sia frequente nel primo cluster; inoltre è l'unica parola (tra le 15 più numerose) del documento che troviamo anche in uno dei tre centroidi. Per come abbiamo ordinato i termini, la parola *ratti*, che nel documento ha frequenza uguale ad *acido*, non è presente nella lista, ma possiamo osservare che nella tabella troviamo, sempre nel primo gruppo, l'elemento *ratti* (in verde) tra i più frequenti: ciò ci può dare un'idea del perché il documento venga classificato come appartenente

al primo cluster. Sebbene la parola *tessuto* (in rosso) sia diversa da *tessuti* del documento, il fatto che *tessuto* sia frequente nel terzo gruppo porta a pensare che in esso si trovi spesso anche il termine *tessuti*: questo potrebbe dare una spiegazione del perché, nella Tabella 3.5, il numero in blu relativo a tale parola sia quello del terzo cluster.

3.1.2 Interpretazione del documento 145 di CRANFIELD-RED

Consideriamo il documento $a_{145} \in A\_cran\_red$: come nel caso precedente, questo documento viene classificato in modo erroneo con entrambe le norme sia nello spazio pieno che in quello ridotto (vedi Tabelle 3.3 e 3.4). Le parole presenti nel documento sono: *adverse, aero, analysis, approximate, boundary, case, compressible, constant, crocco, cylinder, derived, development, discussed, effect, empirical, equation, flat, flow, formula, formulae, forward, friction, general, gradient, heat, incidence, increase, laminar, layer, light, mach, marked, method, momentum, movement, number, plate, presence, problem, quart, radiation, semi, separation, skin, solving, transfer, velocity, young.* La Tabella 3.7 mostra i termini più importanti di $a_{145} \in A\_cran\_red$ e la loro frequenza nel documento stesso e nei tre centroidi; sono segnati in blu i numeri per cui si ha la minima distanza rispetto alla frequenza nel documento. Ricordiamo che il documento 145 appartiene al terzo cluster, mentre nella classificazione nelle Tabelle 3.3 e 3.4 esso viene assegnato al primo gruppo. Dalla Tabella 3.7 vediamo come i numeri che minimizzano la distanza (in blu) appartengano quasi tutti al primo centroide, pochi al secondo e solo uno al terzo. In particolare, i termini *confine, calore, strato* e *trasferire* sono poco più frequenti nel primo rispetto al secondo cluster, ma nettamente più frequenti nel primo rispetto al terzo. Poiché ciò avviene anche per la maggior parte dei termini nel documento, abbiamo trovato un spiegazione del perché i coefficienti di similarità associati ad a_{145} nella Tabella 3.3 siano, nell'ordine, 10.47, 10.74 e 11,57: i primi sono abbastanza vicini, mentre il terzo è maggiore.

Tabella 3.7: Frequenza dei termini significativi in $a_{145} \in A\_cran\_red$ e nei tre centroidi nello spazio pieno.

Documento a_{145} di CRANFIELD-RED				
Termini	Frequenza			
	a_{145}	c_1	c_2	c_3
confine	4	1.7619	1.3418	1.0517
calore	4	1.000	0.7468	0.4655
strato	4	1.6032	1.3418	0.9138
trasferire	4	0.7619	0.6456	0.2759
fluttuazioni	3	0	0.0127	0.0690
attrito	3	0.3016	0.3797	0.2414
laminare	3	0.5873	0.6076	0.3103
rivestimento	3	0.2857	0.4810	0.2414
comprimibile	2	0.3968	0.3038	0.1207
piatto/piano	2	0.4603	0.2025	0.1724
flusso	2	2.0952	0.8354	2.2931
metodo	2	0.5714	0.4177	0.4310
lastra	2	0.7302	0.3671	0.2069

Tabella 3.8: Termini più frequenti in $a_{145} \in A\_cran\_red$ e nei tre cluster.

Termini più frequenti CRANFIELD-RED				
	a_{145}	c_1	c_2	c_3
1	confine	flusso	flusso	flusso
2	calore	confine	confine	pressione
3	strato	strato	strato	numero
4	trasferire	calore	pressione	base
5	fluttuazioni	pressione	superficie	confine
6	attrito	numero	calore	urto
7	laminare	temperatura	velocità	mach
8	rivestimento	teoria	numero	strato
9	comprimibile	trasferire	trasferire	supersonico
10	piano/piatto	piano/piatto	urto	velocità
11	flusso	risultati	temperatura	jet
12	metodo	mach	laminare	gas
13	lastra	superficie	turbolento	corpo
14	avverso	ipersonico	teoria	risultati
15	aero	scs	risultati	reynolds

La Tabella 3.8 mostra i 15 termini più frequenti del documento e dei tre centroidi. Sono stati evidenziati con colori diversi i 4 termini più importanti del documento a_{145} ed i loro corrispondenti nei centroidi. Notiamo che, rispetto ai tre centroidi, il primo ha il maggior numero di parole (più frequenti) in comune con il documento in esame; segue poi il secondo ed infine il terzo, con solo un termine comune. Ciò dà un'ulteriore spiegazione alle frequenze della Tabella 3.8 e dell'errata classificazione tramite l'algoritmo.

Tabella 3.9: Test II, accuratezza della classificazione per MEDLINE-RED nello spazio pieno ed in quello ridotto con i metodi dei centroidi e dei centroidi ortogonali, per entrambe le norme.

Data set MEDLINE-RED			
Norma	Accuratezza della classificazione (in %) dell'algoritmo basato sui centroidi		
	Spazio pieno	Centroidi	Centroidi ortogonali
L_2	81.5	84	81.5
cos	83	84	83

3.2 Test II: accuratezza della classificazione

Nel secondo test vogliamo analizzare l'accuratezza della classificazione, perciò confrontiamo i risultati ottenuti nello spazio pieno ed in quello ridotto sia con la norma L_2 che con la norma del coseno.
Consideriamo gli stessi dati del test precedente (anche con la stessa suddivisione in cluster) e applichiamo l'algoritmo di classificazione basato sui centroidi, per entrambe le norme, sia nello spazio pieno che dopo aver ridotto le dimensioni con i metodi dei centroidi e dei centroidi ortogonali. Infine, per ogni documento, confrontiamo l'indice del cluster a cui il documento viene assegnato tramite l'algoritmo di classificazione con l'indice del cluster di appartenenza. Consideriamo quindi l'accuratezza della classificazione, ossia il numero di volte (in percentuale) in cui tali due indici coincidono. Facciamo questa operazione dopo aver classificato i dati, sia nello spazio pieno che in quello ridotto con i due metodi. Inoltre, poiché l'indice del cluster a cui viene assegnato un documento dipende dalla misura di similarità usata, riportiamo l'accuratezza della classificazione per entrambe le norme. Le Tabelle 3.9 e 3.10 riportano i risultati nei termini di accuratezza della classificazione (in %) per entrambi i data set.

Tabella 3.10: Test II, accuratezza della classificazione per CRANFIELD-RED nello spazio pieno ed in quello ridotto con i metodi dei centroidi e dei centroidi ortogonali, per entrambe le norme.

Data set CRANFIELD-RED			
Norma	Accuratezza della classificazione (in %) dell'algoritmo basato sui centroidi		
	Spazio pieno	Centroidi	Centroidi ortogonali
L_2	75.5	81.5	75.5
cos	78.5	81.5	78.5

Notiamo che, per entrambi i set di dati, l'accuratezza della classificazione sia abbastanza buona; inoltre, in accordo con i due teoremi sull'ordine di similarità, essa è uguale nello spazio pieno e nello spazio ridotto con i centroidi ortogonali sia per la norma L_2 che per la norma del coseno. Quindi è conveniente ridurre le dimensioni dei documenti con questo algoritmo, in quanto si abbassa di molto il costo computazionale per la classificazione, mantenendone però i risultati. Inoltre, come già visto, la classificazione è la stessa per entrambe le norme dopo aver ridotto con l'algoritmo dei centroidi: ritroviamo nelle due tabelle questo risultato, in quanto la percentuale di accuratezza è uguale per la norma L_2 e la norma del coseno (nello spazio ridotto l'algoritmo dei centroidi).

3.3 Test III: importanza della strategia di clustering

Nel terzo test si vuole analizzare quanto sia influente una buona divisione dei documenti nei cluster, al fine della classificazione tramite l'algoritmo basato sui centroidi. Nei test precedenti sono stati creati i cluster senza tener conto delle similarità tra i vari documenti; introduciamo ora un algoritmo di

clustering molto usato: l'algoritmo delle k-medie.

3.3.1 L'algoritmo delle k-medie

L'algoritmo delle k-medie è un algoritmo di clustering che considera inizialmente k gruppi e assegna ogni oggetto al cluster avente il più vicino centroide. Riassumiamo velocemente il procedimento.

ALGORITMO 4: (k-medie)
Dato un data set $A \in \mathbb{R}^{m \times n}$ con k cluster, assegna ogni documento al cluster avente il più vicino centroide.

1. Suddivide i documenti in k cluster e calcola il centroide di ogni cluster.

2. Per ogni documento: calcola la distanza dal centroide di ogni cluster; riposiziona il documento nel cluster con il centroide più vicino; ricalcola il centroide del cluster a cui è stato aggiunto il documento e del cluster che lo ha perso.

3. Riprende dal punto 2 finché ogni documento non cambia più cluster.

3.3.2 Il test

Nel terzo test consideriamo i set di dati completi MEDLINE e CRANFIELD e creiamo i cluster in due modi differenti:

- prendiamo cluster con numero simile di documenti tra loro;

- creiamo i cluster con l'algoritmo delle k-medie.

Poi applichiamo l'algoritmo di classificazione basato sui centroidi, per entrambe le norme, nello spazio pieno e nello spazio ridotto con i metodi dei centroidi e dei centroidi ortogonali. Infine confrontiamo l'accuratezza della

Tabella 3.11: Test III, accuratezza della classificazione per MEDLINE nello spazio pieno ed in quello ridotto con i due metodi.

Data set MEDLINE	
Cluster	Numero di documenti
1	345
2	344
3	344

Norma	Accuratezza della classificazione (in %) dell'algoritmo basato sui centroidi		
	Spazio pieno	Centroidi	Centroidi ortogonali
L_2	67.7	68.1	67.7
cos	70.1	68.1	70.1

classificazione come spiegato nel test precedente. Indichiamo con MEDLINE-K-MEDIE e CRANFIELD-K-MEDIE i due data set dopo la divisione in cluster attraverso l'algoritmo delle k-medie. Nelle Tabelle 3.11, 3.12, 3.13 e 3.14 sono riportati i risultati ottenuti. La parte superiore delle tabelle riporta come sono divisi i documenti nei cluster. Nel caso di divisione manuale, se il primo cluster ha h documenti, essi sono i primi h del data set, e a seguire per i cluster successivi. Per i cluster con le k-medie, il numero in corrispondenza del cluster ne rappresenta soltanto il numero di documenti. Nella parte inferiore delle tabelle sono riportati i risultati in termini di accuratezza della classificazione, per entrambe le norme e sia nello spazio pieno che nello spazio ridotto con i due metodi, per la strategia di clustering considerata.

Oltre alle osservazioni già fatte nel test precedente sul fatto che l'algoritmo dei centroidi ortogonali mantenga l'ordine di similarità, possiamo vedere quanto sia determinante il metodo di clustering per la buona riuscita della classificazione. Guardando le Tabelle 3.11 e 3.12, notiamo come, nello spazio pieno e usando la norma L_2, si passi da un'accuratezza della classificazione

Tabella 3.12: Test III, accuratezza della classificazione per MEDLINE-K-MEDIE nello spazio pieno ed in quello ridotto con i due metodi.

Data set MEDLINE-K-MEDIE	
Cluster	Numero di documenti
1	309
2	640
3	264

Norma	Accuratezza della classificazione (in %) dell'algoritmo basato sui centroidi		
	Spazio pieno	Centroidi	Centroidi ortogonali
L_2	98.4	98.6	98.4
cos	100	98.6	100

Tabella 3.13: Test III, accuratezza della classificazione per CRANFIELD nello spazio pieno ed in quello ridotto con i due metodi.

Data set CRANFIELD	
Cluster	Numero di documenti
1	466
2	466
3	466

Norma	Accuratezza della classificazione (in %) dell'algoritmo basato sui centroidi		
	Spazio pieno	Centroidi	Centroidi ortogonali
L_2	52.4	56.8	52.4
cos	56.7	56.8	56.7

Tabella 3.14: Test III, accuratezza della classificazione per CRANFIELD-K-MEDIE nello spazio pieno ed in quello ridotto con i due metodi.

Data set CRANFIELD-K-MEDIE	
Cluster	Numero di documenti
1	700
2	297
3	401

Norma	Accuratezza della classificazione (in %) dell'algoritmo basato sui centroidi		
	Spazio pieno	Centroidi	Centroidi ortogonali
L_2	91.4	97.8	91.4
cos	100	97.8	100

del 67.7% ad una del 98.4%; risultati ancora migliori si hanno con la norma del coseno, per cui dal 70.1% si arriva ad un'accuratezza perfetta. Nelle Tabelle 3.13 e 3.14 vediamo come la differenza di accuratezza, con e senza le k-medie, sia elevata: dal 56.6%, per la norma del coseno nello spazio pieno, al 100% e dal 56.8% al 97.8% nello spazio ridotto con l'algoritmo dei centroidi. Osserviamo infine come, in tutti i casi presentati fino ad ora, la migliore classificazione si ha con la norma del coseno nello spazio ridotto con i centroidi ortogonali; considerando invece solo la norma L_2, la percentuale massima di accuratezza si ha dopo aver ridotto con l'algoritmo dei centroidi.

3.4 Test IV: confronto con la SVD

Nel quarto test si vuole confrontare l'accuratezza della classificazione dopo aver ridotto le dimensioni dei documenti con i metodi dei centroidi e dei centroidi ortogonali con quella applicata nello spazio ridotto attraverso la

SVD troncata, che abbiamo provato essere la migliore approssimazione di rango k in norma 2 e di Frobenius.
Consideriamo entrambi i data set ridotti MEDLINE-RED e CRANFIELD-RED, divisi in 3 cluster con le k-medie (li indichiamo con MEDLINE-RED-KMEDIE e CRANFIELD-RED-KMEDIE), e riduciamo con l'algoritmo dei centroidi, con l'algoritmo dei centroidi ortogonali e con la SVD; classifichiamo infine con l'algoritmo basato sui centroidi. Inoltre, mantenendo la stessa divisione in 3 cluster, rappresentiamo i documenti nello spazio ridotto d-dimensionale attraverso la SVD troncata di rango d per $d = 5, 10, 20, 50, 100$ e 200; poi classifichiamo con l'algoritmo basato sui centroidi e riportiamo l'accuratezza della classificazione.
Le Tabelle 3.15 e 3.16 mostrano i risultati ottenuti. Nella parte superiore delle tabelle troviamo la suddivisione in cluster, mentre in quella inferiore è riportata l'accuratezza della classificazione per entrambe le norme. Essa è riportata per lo spazio ridotto in dimensione 3 mediante i due metodi basati sui centroidi e la SVD, mentre per gli spazi ridotti di dimensioni maggiori sono presenti i risultati solo attraverso il metodo della SVD.
Notiamo infatti come nei due algoritmi di riduzione presentati la dimensione dello spazio ridotto sia la stessa del numero di cluster (nel nostro test 3), mentre per la riduzione con la SVD non ci sono vincoli di questo genere.

Consideriamo il caso di 3 centroidi e la riduzione 3-dimensionale dei documenti: da entrambe le tabelle notiamo come la percentuale di accuratezza dopo aver ridotto con la SVD sia nettamente minore rispetto a quella degli altri due metodi. Per raggiungere tali percentuali si deve considerare una riduzione d-dimensionale tramite la SVD con $d \approx 200$, il che comporta un maggior costo computazionale, sia per il calcolo effettivo della SVD troncata che per il confronto tra due documenti.
Questi risultati mostrano che, sebbene la SVD troncata A_k di una matrice A sia la migliore approssimazione possibile di rango k della matrice, i due algoritmi di riduzione presentati lavorano meglio per quanto riguarda la classificazione su set di dati con una struttura di cluster.

Tabella 3.15: Test IV, accuratezza della classificazione per MEDLINE-RED-K-MEDIE (diviso in 3 cluster) nello spazio ridotto con i due metodi sui centroidi, e con la SVD nello spazio ridotto di dimensione $d = 3, 5, 10, 20, 50, 100$ e 200.

Data set MEDLINE-RED-K-MEDIE	
Cluster	Numero di documenti
1	61
2	78
3	61

Dimensione spazio ridotto	Accuratezza della classificazione (in %) dell'algoritmo basato sui centroidi					
	Centroidi ortogonali		Centroidi		SVD	
	L_2	cos	L_2	cos	L_2	cos
3	98	100	99	99	67	63
5					68.5	66.5
10					85.5	87
20					85.5	85
50					92.5	92
100					93	95
200					98	100

Tabella 3.16: Test IV, accuratezza della classificazione per CRANFIELD-RED-K-MEDIE (diviso in 3 cluster) nello spazio ridotto con i due metodi sui centroidi, e con la SVD nello spazio ridotto di dimensione $d = 3, 5, 10, 20, 50, 100$ e 200.

Data set CRANFIELD-RED-K-MEDIE	
Cluster	Numero di documenti
1	82
2	46
3	72

Dimensione spazio ridotto	Accuratezza della classificazione (in %) dell'algoritmo basato sui centroidi					
	Centroidi ortogonali		Centroidi		SVD	
	L_2	cos	L_2	cos	L_2	cos
3	91.5	100	98	98	62	76.5
5					79	87
10					81	90
20					83.5	90.5
50					86.5	95
100					88.5	97
200					91.5	100

3.5 Test V: classificazione di nuovi documenti

In questo test vogliamo considerare un data set unico formato da documenti di MEDLINE e di CRANFIELD, e analizzare la correttezza della classificazione di nuovi documenti.
Unificando i due data set si è creata una nuova matrice termini-documenti $A\_tot \in \mathbb{R}^{8347\times 2431}$; notiamo come il numero di parole del dizionario di questo nuovo set di dati (8347 parole) sia inferiore alla somma dei termini dei singoli data set, questo perché vi sono 1951 termini comuni che, nella creazione di A_tot, sono stati considerati una sola volta. Inoltre, in A_tot, non troviamo un blocco iniziale di righe corrispondenti a termini del primo data set, in quanto si sono volute intervallare parole del primo data set con parole del secondo.
Nel primo test consideriamo i due data set come due cluster di A_tot e guardiamo la correttezza della classificazione tramite l'algoritmo basato sui centroidi sia nello spazio pieno che in quello ridotto a dimensione 2 con i metodi dei centroidi e dei centroidi ortogonali. La Tabella 3.17 mostra i risultati ottenuti.
Notiamo come, essendo i due cluster divisi molto bene, si ha una classificazione nella maggior parte dei casi corretta, soprattutto con la norma del coseno o con il metodo di riduzione di dimensione dei centroidi.
Lo scopo di questa classificazione non è però quello di classificare documenti per i quali ci è noto il gruppo a cui appartengono, ma si vuole assegnare un cluster a dei documenti nuovi; ad esempio per poter consultare articoli correlati con quello che si sta studiando o scrivendo. Vogliamo quindi prima considerare un training set formato da 1000 documenti: 500 di MEDLINE e 500 di CRANFIELD, che formano rispettivamente il primo ed il secondo cluster; calcoliamo la matrice C e la sua fattorizzazione QR per applicare gli algoritmi di riduzione e di classificazione ai nuovi documenti. Per il test set consideriamo altri 1000 documenti, diversi da quelli del training test: 500 da MEDLINE e 500 da CRANFIELD. La Tabella 3.18 mostra i risultati dell'accuratezza della classificazione sia per i dati del training set che per quelli del

Tabella 3.17: Test V, accuratezza della classificazione per l'unico data set nello spazio pieno ed in quello ridotto con i due metodi.

Data set unico: *A_tot*	
Cluster	Numero di documenti
Medicina	1033
Ingegneria aerospaziale	1398

Norma	Accuratezza della classificazione (in %) dell'algoritmo basato sui centroidi		
	Spazio pieno	Centroidi	Centroidi ortogonali
L_2	89.1	99	89.1
cos	99.2	99	99.2

test set. Notiamo che i documenti del test set vengono classificati in modo corretto nella maggior parte dei casi, soprattutto nello spazio ridotto con l'algoritmo dei centroidi ed in generale con la norma del coseno.
La classificazione dei nuovi documenti (sia nello spazio pieno che in quello ridotto) è influenzata dagli elementi del training set con i quali si è costruita la matrice dei centroidi C: nel nostro caso si è considerato un numero esiguo di documenti (circa la metà dei documenti) per ogni cluster, il che ha portato ad una buona classificazione. Non sono stati effettuati altri test per osservare come la classificazione cambi a seconda del training set usato, ma risulta chiaro come, per una buona accuratezza della classificazione, i documenti del training set debbano essere rappresentativi degli ambiti considerati.
Infine confrontiamo l'accuratezza della classificazione dei nuovi documenti nello spazio ridotto con i due metodi e con la SVD; usiamo un training set di 200 documenti (100 di MEDLINE e 100 di CRANFIELD) ed un test set di 200 documenti diversi dai precedenti (sempre 100 di ambito medico e 100 di ambito aerospaziale). La Tabella 3.19 riporta i risultati ottenuti per la norma del coseno. Notiamo come, affinché l'accuratezza della classificazione

Tabella 3.18: Test V, accuratezza della classificazione di nuovi documenti nello spazio pieno ed in quello ridotto con i due metodi.

Data set: *A_tot*		
Cluster	Numero di documenti	
	Training set	Test set
MEDLINE	500	500
CRANFIELD	500	500

Norma	Accuratezza della classificazione (in %) dell'algoritmo basato sui centroidi					
	Spazio pieno		Centroidi		Centroidi ortogonali	
	Training	Test	Training	Test	Training	Test
L_2	91.5	81.9	99.2	96	91.5	81.9
cos	99.4	98.9	99.2	96	99.4	98.9

nello spazio ridotto tramite la SVD sia simile a quella ottenuta con i due metodi, la dimensione dello spazio ridotto deve essere vicina a 50.

Tabella 3.19: Test V, accuratezza della classificazione di nuovi documenti nello spazio ridotto con i due metodi sui centroidi, e con la SVD nello spazio ridotto di dimensione $d = 3, 5, 10, 20$ e 50 per la norma del coseno.

Data set: *A_tot*		
Cluster	Numero di documenti	
	Training set	Test set
MEDLINE	100	100
CRANFIELD	100	100

Dimensione spazio ridotto	Accuratezza della classificazione (in %) dell'algoritmo basato sui centroidi					
	Centroidi ortogonali		Centroidi		SVD	
	Training	Test	Training	Test	Training	Test
2	99.5	99	98.5	96	63	59.5
5					96	99
10					97.5	98.5
20					98.5	99
50					99	99.5

Conclusioni

In questo elaborato abbiamo presentato due algoritmi di riduzione di dimensione: l'algoritmo dei centroidi è basato sulla proiezione tramite la matrice dei centroidi mentre l'algoritmo dei centroidi ortogonali è una proiezione attraverso una base ortonormale per lo spazio generato dai centroidi. Questi due metodi hanno il vantaggio di richiedere un costo computazionale nettamente inferiore a quello richiesto dalla SVD; inoltre i risultati mostrano che essi sono più efficienti nella classificazione per dati con struttura di cluster.
D'altra parte, i nuovi algoritmi suppongono che i dati siano già divisi in gruppi, e che la dimensione dello spazio ridotto sia la stessa del numero di cluster. Se i dati non hanno questa struttura si può sempre applicare un algoritmo di clustering, ma abbiamo osservato quanto sia rilevante la scelta di tale metodo. Rimane in sospeso quale sia il miglior numero k di cluster in cui dividerlo, poiché non si devono perdere troppe informazioni nella riduzione, ma nello stesso tempo si ha che più piccolo è k e minore è il costo computazionale.
Infine, se il set di dati è già diviso in k cluster, e lo si vuole ridurre in una dimensione diversa da k, si possono accorpare o dividere a loro volta i cluster per ottenerne un numero uguale alla dimensione desiderata.
Abbiamo quindi mostrato come sia possibile incorporare una conoscenza a priori dei dati nella riduzione di dimensioni e come essa sia importante ai fini di una corretta classificazione.

Appendice A

Prima Appendice

A.1 Listati dei programmi

A.1.1 Algoritmo per calcolare i centroidi con $C = AH$

```
function [C]=centroids(A,k,N)
% Calcola i centroidi attraverso la matrice di raggruppamento
%
% INPUT:
% A: matrice termini-documenti m x n
% k: numero di cluster
% N: matrice con k colonne, in cui la colonna i-esima N(i) è il
%    vettore con elementi gli indici dei documenti che
%    appartengono all'i-esimo cluster
%
% OUTPUT:
% C: matrice dei centroidi m x k


[m,n]=size(A);
F=zeros(n,k);
```

```
for j=1:k
    for r=1:length(N(:,j))
        if N(r,j)>0;
            F(N(r,j),j)=1;
        end
    end
end
H=F*inv(diag(diag(F'*F)));
C=A*H;
```

A.1.2 Algoritmo dei centroidi per la riduzione di dimensioni

```
function[qcap]=centroid_algorithm_dim_reduction(C,q)

% Algoritmo dei centroidi per la riduzione di dimensioni
% si risolve il problema di minimo min ||C*qcap-q||
%
% INPUT:
% C: matrice dei centroidi m x k
% q: vettore di cui si vogliono ridurre le dimensioni
%
% OUTPUT:
% qcap = vettore q in dimensione ridotta

qcap=C\q;
```

A.1.3 Algoritmo dei centroidi ortogonali per la riduzione di dimensioni

```
function[qcap,Ccap]=orthogonal_centroid_algorithm_dim_red(C,q)

% Algoritmo dei centroidi ortogonali per la riduzione di dimensioni
%
% INPUT:
% C: matrice dei centroidi m x k (calcolata da C=centroids(A,k,N) )
% q: vettore di cui si vogliono ridurre le dimensioni
%
% OUTPUT:
% Ccap: matrice dei centroidi in dimensione ridotta
% qcap: vettore q in dimensione ridotta

[Q,R]=qr(C,0);
qcap=Q'*q;
Ccap=Q'*C;
```

A.1.4 Algoritmo di classificazione basato sui centroidi

```
function[ind_l2,ind_cos,sim_cos,sim_l2]=centroid_based_classification(C,q)

% Algoritmo di classificazione basato sui centroidi
%
% INPUT:
% C: matrice dei centroidi s x k
% q: vettore da classificare
%
% OUTPUT:
% sim_cos: coefficiente di similarità con la norma del coseno
```

```
% sim_l2:  coefficiente di similarità con la norma L2
% ind_cos: indice del cluster a cui viene assegnato il
%           vettore per la norma del coseno
% ind_l2:  indice del cluster a cui viene assegnato il
%          vettore per la norma L2

[m,k]=size(C);
norm_q=norm(q,2);
sim_cos=ones(k,1);
sim_l2=ones(k,1);
for j=1:k
    sim_cos(j)=q'*C(:,j)*1/(norm_q*norm(C(:,j),2));
    sim_l2(j)=norm(q-C(:,j),2);
end
[simcos,indcos]=sort(sim_cos,'descend');
indice_cos=indcos(1);
[siml2,indl2]=sort(sim_l2,'ascend');
indice_l2=indl2(1);
```

A.1.5 Algoritmo di classificazione basato sui centroidi dopo aver ridotto con il metodo dei centroidi

```
function[indice]=centroid_based_class_semplice(qred)

% Algoritmo di classificazione basato sui centroidi da usare
% dopo aver ridotto le dimensioni con il metodo dei centroidi.
%
% INPUT:
% qred: vettore da classificare
%
```

```
% OUTPUT:
% indice: indice del cluster a cui viene assegnato il vettore
%         qred per entrambe le norme

[elementi,indici]=sort(qred,'descend');
indice=indici(1);
```

Bibliografia

[1] Haesun Park, Moongu Jeon and J. Ben Rosen, *Lower dimensional representation of test data based on centroids and least squares*, BIT Numerical Mathematics 43: 427-448, 2003.

[2] Lars Elden, *Matrix Methods in Data Mining and Pattern Recognition*, SIAM, April 2007.

[3] Davide Palitta e Valeria Simoncini, *Dispense del corso di Calcolo Numerico. Modulo di Algebra Lineare*, 2016.

[4] Daniel B. Szyld, *The many proofs of an identity on the norm of oblique projections*, Numerical Algorithms 42: 309-323, 2006.

Ringraziamenti

Un ringraziamento particolare va alla mia relatrice, la Prof.ssa Valeria Simoncini, per la fiducia e la pazienza dimostrate nei miei confronti durante la scrittura della tesi, e soprattutto per il grande entusiasmo che mi ha trasmesso.

Poi vorrei rivolgermi a te:

a te che hai visto in me qualcosa in cui ancora non credevo,
a te che mi hai insegnato a mettermi in gioco,
a te che mi sei stato sempre accanto,
a te che hai letto formule e corretto accenti,
a te che hai condiviso con me questo viaggio,
a te che oggi sei con me.

Grazie