Paolo Speziali

Tesina di
Signal Processing and Optimization for Big Data
del Prof. Paolo Banelli

Low-Rank Matrix Completion con implementazione e verifica sperimentale

Perugia, Anno Accademico 2022/2023
Università degli Studi di Perugia
Corso di laurea magistrale in Ingegneria Informatica e Robotica
Curriculum Data Science
Dipartimento di Ingegneria

A.D. 1308
unipg
DIPARTIMENTO
DI INGEGNERIA

0. Indice

1. Introduzione

Lo scopo di questa tesina è l'implementazione in ambiente MATLAB di tre algoritmi atti a risolvere il problema dello stimare i valori mancanti di una matrice di cui abbiamo a disoposizione solo un sottoinsieme limtato di entry.

2. Concetti teorici

2.1 Formulazione del problema

Sia $M \in \mathbb{R}^{n_1 \times n_2}$ con rango r, e sia data la sua scomposizione ai valori singolari (SVD):

$$M = U\Sigma V^T \quad \text{con} \quad U \in \mathbb{R}^{n_1 \times r},\ \Sigma \in \mathbb{R}^{r \times r},\ V \in \mathbb{R}^{n_2 \times r}$$

dove U e V sono composte da colonne ortonormali, e Σ è una matrice diagonale con i valori singolari ordinati in modo non crescente ($\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$).

I **gradi di libertà** di M sono $(n_1+n_2-r)r$, che è il numero totale di parametri necessari per specificare univocamente la matrice M.

Supponiamo di avere delle osservazioni parziali di M su un insieme di indici

$$\Omega \subset \{1, 2, \ldots, n_1\} \times \{1, 2, \ldots, n_2\}$$

e definiamo l'**operatore di osservazione** $\mathcal{P}_\Omega : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$ come segue:

$$[\mathcal{P}_\Omega(M)]_{ij} = \begin{cases} M_{ij}, & \text{se } (i,j) \in X \\ 0, & \text{altrimenti} \end{cases}$$

Il nostro obiettivo è recuperare M da $\mathcal{P}_\Omega(M)$ quando il numero di osservazioni $m = |\Omega| \ll n_1 n_2$, ovvero quando è molto più piccolo del numero di elementi in M, e sotto l'assunzione che M sia a basso rango, ovvero $r \ll \min(n_1, n_2)$. Per semplicità notazionale, poniamo $n = \max(n_1, n_2)$.

2.2 Soluzione del problema

Quali tipi di matrici a basso rango possiamo completare? Consideriamo le matrici M_1 e M_2 di rango 1 e di dimensione 4×4:

$$M_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \qquad M_2 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

La matrice M_1 è più difficile da completare poiché la maggior parte delle sue voci sono nulle e quindi abbiamo bisogno di raccogliere più misure per assicurarsi che abbastanza "massa" venga dalle sue voci non nulle. Al contrario, la massa di M_2 è distribuita più uniformemente su tutte le voci, rendendolo più facile da propagare da una voce all'altra.

In altre parole, una matrice a basso rango è più facile da completare se la sua energia si distribuisce uniformemente su diverse coordinate. Questa proprietà è catturata dalla **coerenza**, che misura l'allineamento tra lo spazio delle colonne/righe della matrice a basso rango con i vettori della base standard.

Per una matrice $U \in \mathbb{R}^{n_1 \times r}$ con colonne ortonormali, P_U rappresenta la proiezione ortogonale sullo spazio delle colonne di U. La **coerenza** di U è definita come segue:

$$\mu(U) = \frac{n_1}{r} \max_{1 \leq i \leq n_1} \left\| P_U \, \underline{e}_i \right\|_2^2 = \frac{n_1}{r} \max_{1 \leq i \leq n_1} \left\| U^T \, \underline{e}_i \right\|_2^2$$

dove $\underline{e}_i$ è l'i-esimo vettore della base canonica.

Per una matrice a basso rango M il cui SVD è data da $M = U\Sigma V^T$, la coerenza di M è definita come:

$$\mu = max\{\mu(U), \mu(V)\}$$

Si noti che la coerenza μ è determinata dai vettori singolari di M ed è indipendente dai suoi valori singolari.

Poiché $1 \leq \mu(U) \leq \frac{n_1}{r}$ e $1 \leq \mu(V) \leq \frac{n_2}{r}$, abbiamo $1 \leq \mu \leq \frac{n}{r}$. Nell'esempio precedente, la coerenza di M_1 coincide con il limite superiore $\frac{n}{r}$, mentre quella di M_2 coincide con il limite inferiore 1. Più μ è piccolo, più è facile completare la matrice.

Possiamo incontrare alcune matrici la cui ricostruzione non è possibile, un esempio sarebbe una matrice con tutti i valori eccetto quelli di una colonna completamente mancante, essa non potrà essere recuperata in quanto potrebbe giacere ovunque nello spazio delle colonne della matrice. Ci servono quindi almeno r osservazioni per colonna/riga.

Per evitare di incorrere in questi casi sfavorevoli, supponiamo di star utilizzando un pattern di osservazione causale che segua un modello di distribuzione di probabilità noto come **Bernoulli**, per cui ogni valore viene osservato indipendentemente e con probabilità uguale a $p := \frac{m}{n_1 \cdot n_2}$.

Non è possibile recuperare una matrice a basso rango con un numero di osservazioni ad uno dell'ordine di $O(\mu n r \log n)$ utilizzando un qualsiasi algoritmo, questo è noto come l'**information-theoretic lower bound**. Rispetto ai gradi di libertà, che sono dell'ordine di nr, paghiamo un prezzo in complessità di campionamento di un fattore $\mu \log n$, mettendo ancora una volta in evidenza il ruolo della coerenza nel completamento di matrici a basso rango.

2.2.1 Completamento di matrici tramite ottimizzazione convessa

Cercando di sfruttare la struttura a basso rango della soluzione, un'euristica naturale è trovare la matrice con rango minore che permette tali osservazioni:

$$\min_{\Phi \in \mathbb{R}^{n_1 \times n_2}} \quad rank(\Phi)$$

$$s.t. \;\; \mathcal{P}_\Omega(\Phi) = \mathcal{P}_\Omega(M)$$

Tuttavia, essendo la minimizzazione del rango un problema NP-arduo, tale formulazione non è intrattabile, possiamo tuttavia pensare a un possibile rilassamento di questa euristica.

Notando che il rango di Φ è uguale al numero dei suoi valori singolari non nulli, sostituiamo $rank(\Phi)$ con la somma dei suoi valori singolari, indicata come **nuclear norm**:

$$\|\Phi\|_* \triangleq \sum_{i=1}^{n} \sigma_i(\Phi)$$

Quindi, invece di risolvere direttamente il problema visto precedentemente, risolviamo la minimizzazione della nuclear norm, che cerca una matrice con la nuclear norm minima che soddisfa tutte le misurazioni:

$$\min_{\Phi \in \mathbb{R}^{n_1 \times n_2}} \quad \|\Phi\|_*$$

$$s.t. \;\; \mathcal{P}_\Omega(\Phi) = \mathcal{P}_\Omega(M)$$

Si ottiene così un programma convesso che può essere risolto in modo efficiente in tempo polinomiale. Inoltre, non richiede la conoscenza del rango a priori.

La minimizzazione della nuclear norm può recuperare esattamente una matrice di basso rango non appena il numero di misurazioni è leggermente più grande dell'information-theoretic lower bound di un fattore logaritmico. Supponiamo che ogni valore della matrice M venga osservato indipendentemente con una probabilità $p \in (0, 1)$. Se:

$$p \leq C \, \frac{\mu r \log^2 n}{n}$$

per una qualche $C > 0$ abbastanza grande, allora con grande probabilità l'algoritmo recupera esattamente la matrice M come soluzione ottima.

2.2.2 Completamento di matrici tramite ottimizzazione non convessa

L'algoritmo appena visto può essere particolarmente costoso in termini di tempo e memoria per problemi su larga scala a causa del dover ottimizzare e memorizzare la variabile Φ. Pertanto, è necessario considerare approcci alternativi che scalino in modo più favorevole con n. Ciò porta al secondo algoritmo basato su gradient descent utilizzando un'inizializzazione adeguata.

Se il rango della matrice M è noto, è naturale incorporare questa conoscenza e considerare un problema least-square vincolato al rango:

$$\min_{\Phi \in \mathbb{R}^{n_1 \times n_2}} \quad \|\mathcal{P}_\Omega(\Phi - M)\|_F^2$$

$$s.t. \;\; rank(\Phi) \leq r$$

dove $\|\cdot\|_F$ è la **Frobenius norm** di una matrice. Utilizzando la fattorizzazione a basso rango $\Phi = XY^T$ dove $X \in \mathbb{R}^{n_1 \times r}$ e $Y \in \mathbb{R}^{n_2 \times r}$, riscriviamo il problema qui sopra come un problema d'ottimizzazione non vincolato e non convesso:

$$\min_{X,Y} \; f(X,Y) := \left\|\mathcal{P}_\Omega(XY^T - M)\right\|_F^2$$

Le complessità a livello di memoria di X e Y sono lineari in n. Introduciamo una loss function modificata per sistemare alcuni problemi di scalabilità e avere norme bilanciate:

$$F(X,Y) = \frac{1}{4p} f(X,Y) + \frac{1}{16} \left\|X^T X - Y^T Y\right\|_F^2$$

La probabilità p delle osservazioni può essere stimata com $p = \frac{|\Omega|}{n_1 \cdot n_2}$.

Ma come facciamo a ottimizzare la loss non convessa $F(X,Y)$?

1. Troviamo un'inizializzazione "spettrale" che sia vicina alla verità di base. Consideriamo la matrice parzialmente osservata $\frac{1}{p}\mathcal{P}_\Omega(M)$, che è una stima non polarizzata di M con valore atteso pari a $E[\frac{1}{p}\mathcal{P}_\Omega(M)] = M$. Perciò, un'approssimazione best rank-r produce una stima iniziale adeguata.

Sia tale approssimazione $U_0\Sigma_0 V_0^T$, inizializzeremo con:

$$X_0 = U_0\Sigma_0^{1/2} \quad \text{e} \quad Y_0 = V_0\Sigma_0^{1/2}$$

2. Raffiniamo la stima iniziale con semplici metodi iterativi secondo la seguente regola d'aggiornamento:
$$\begin{bmatrix} X_{t+1} \\ Y_{t+1} \end{bmatrix} = \begin{bmatrix} X_t \\ Y_t \end{bmatrix} - \eta_t \begin{bmatrix} \nabla_X F(X_t, Y_t) \\ \nabla_Y F(X_t, Y_t) \end{bmatrix}$$
dove η_t è la step-size.

Il gradient descent converge ad una velocità geometrica se il numero di osservazioni è dell'ordine di $\mu^3 r^3 n \log^3 n$. Il numero di iterazioni è indipendente dalla grandezza del problema e quindi il costo computazionale è molto più basso (unendolo al basso costo di un'iterazione).

Ricapitolando il tutto con una tabella che mette a confronto i tre algoritmi:

Algoritmo	**Complessità campionaria**	**Complessità computazionale**
Information-theoretic lower bound	$\mu nr \log n$	NP-arduo
Nuclear norm minimization	$\mu nr \log^2 n$	Tempo polinomiale
Gradient descent con inizializzazione spettrale	$\mu^3 nr^3 \log^3 n$	Tempo lineare

3. Implementazione degli algoritmi

4. Verifica sperimentale