Algorithmique des Réseaux Sociaux

10 Décembre

Décomposition en valeurs singulières et complétion de matrices

1. Montrer que $\|A\|_2 \leq \|A\|_F \leq \sqrt{r}\|A\|_2$ où r est le rang de A.
2. Calculer les valeurs propres de la matrice symétrique :
$$H = \begin{pmatrix} 0 & A^T \\ A & 0 \end{pmatrix}.$$
3. Soit A, B des matrices réelles de dimensions $m \times n$. Montrer que pour tout $i + j \leq 1 + m \wedge n$, on a :
$$\sigma_{i+j-1}(A) \leq \sigma_i(B) + \sigma_j(A - B).$$
En déduire que
$$\max_{1 \leq k \leq m \wedge n} |\sigma_k(A) - \sigma_k(B)| \leq \|A - B\|_2.$$
4. Étant donné une matrice A, on note A_k la meilleur approximation de A de rang k. Montrer que
$$\|B - B_k\|_2 \leq \|A - A_k\|_2 + \|A - B\|_2.$$
En déduire que $\|A - B_k\|_2 \leq \|A - A_k\|_2 + 2\|A - B\|_2$.

On se place dans le cadre suivant : soit M une matrice de dimensions $m \times n$ avec $m \leq n$ (modélisant les notes des utilisateurs) de rang r.

Soit $p \in [0, 1]$. On suppose que chaque entrée de M est observée avec probabilité p et non observée avec probabilité $1 - p$ indépendemment des autres entrées.

Pour simplifier, on supposera $M_{ij} \in [0, 1]$ et que les paramètres p et r sont connus.

Nous construisons une estimation $\hat{M}$ de M à partir des entées observées de la manière suivante :

a- Soit X la matrice avec $x_{ij} = m_{ij}$ si l'entrée est observée et $x_{ij} = 0$ sinon. Soit $X = \sum_{i=1}^m s_i \mathbf{u}_i {\mathbf{v}_i}^T$ sa décomposition en valeurs singulières.

b- On définit
$$W = \frac{1}{p} \sum_{i \leq r} s_i \mathbf{u}_i {\mathbf{v}_i}^T.$$

On définit alors la matrice $\hat{M}$ par $\hat{m}_{ij} = m_{ij}$ si l'entrée (ij) est observée et sinon par :
$$\hat{m}_{ij} = \begin{cases} w_{ij} & \text{si } 0 \leq w_{ij} \leq 1, \\ 1 & \text{si } w_{ij} > 1, \\ 0 & \text{si } w_{ij} < 0. \end{cases}$$

On définit alors la mesure d'erreur :
$$MSE(\hat{M}) = \mathbb{E}\left[\frac{1}{mn}\|M - \hat{M}\|_F^2\right].$$

5. Montrer que $\|M-W\|_F^2 \leq 8r\|M-\frac{1}{p}X\|_2^2$. En déduire

$$MSE(\hat{M}) \leq \frac{8r}{mnp^2}\mathbb{E}\left[\|pM-X\|_2^2\right]$$

Pour borner le dernier terme, nous allons utiliser l'inégalité matricielle de Bernstein : Soit $Z_1,\ldots,Z_k \in \mathbb{R}^{m\times n}$ des matrices aléatoires symétriques telles que $\mathbb{E}[Z_t]=0$, $\|Z_t\|_2 \leq 1$ et $\max\left\{\|\sum_{t=1}^k \mathbb{E}[Z_t Z_t^T]\|_2 ; \|\sum_{t=1}^k \mathbb{E}[Z_t^T Z_t]\|_2\right\} \leq \sigma^2$. On a

$$\mathbb{P}\left(\|\sum_{t=1}^k Z_t\|_2 \geq s\right) \leq (m+n)\exp\left(-\frac{s^2}{2(\sigma^2+s/3)}\right).$$

6. On définit $Y=pM-X$. Calculer $\mathbb{E}[Y]$. Montrer que $\|Y\| \leq n$.
7. Montrer que

$$\mathbb{P}\left(\|Y\| \geq s\right) \leq (m+n)\exp\left(-\frac{s^2}{2(np+s/3)}\right)$$

8. En déduire que pour $np > \ln n$, il existe une constante C telle que :

$$MSE(\hat{M}) \leq \frac{Cr\log n}{mp}.$$