CS 229r: Algorithms for Big Data Fall 2013

Lecture 15 — Oct 22nd, 2013

Prof. Jelani Nelson *Scribe: Chiheon Kim*

1 Overview

- Low-rank approximation,
- Mention somethings not covered in class,
- Some more subspace embedding stuff.

2 Low-rank approximation

Given a matrix $A \in \mathbb{R}^{n\times d}$, we want to compute $A_k := \text{argmin}_{\text{rank}(B)\le k}\|A - B\|_X$.

Theorem 1 (Schmidt approximation)**.** *Let $A = U\Sigma V^T$ be a singular-value decomposition of A where* $\text{rank}(A) = r$ *and Σ is diagonal with entries $\sigma_1 \ge \sigma_2 \ge \ldots \ge \sigma_r > 0$, then under $\|\cdot\|_X = \|\cdot\|_F$, $A_k = U_k\Sigma_k V_k^T$ is the minimizer where U_k and V_k are the first k columns of U and V and $\Sigma_k = \text{diag}(\sigma_1, \ldots, \sigma_k)$.*

Mirsky proved more general statement:

Theorem 2 (Mirsky, 1960)**.** *$A_k = U_k\Sigma_k V_k^T$ is the best approximation under any unitarily invariant norm, i.e., $\|x\| = \|Ux\|$ for any x and unitary U.*

2.1 Applications

There are several applications of low-rank approximation like principal component analysis and latent semantic indexing (LSI) . We will talk about LSI first.

LSI is a technique used in information retrieval. We have n documents and a dictionary of size d, and a $n \times d$ matrix A such that A_{ij} represents the (weighted) number of occurrences of word j in document i. This setting might have some problems like

1. (Synonymy) Like "automobile" and "car", some authors may use different words for same topic.
2. (Multiple definitions) Same word "surfing" has different meanings, like web surfing and surfing at the beach.

In LSI, it first compute SVD of A and project each document onto span of the top k singular vectors. Under some modeling assumption, LSI "performs well" [4].

2.2 Algorithm

Today we are going to use a sketch which is used both in subspace embedding and approximate matrix multiplication to compute $\tilde{A_k}$ with rank at most k such that $\|A-\tilde{A_k}\|_F \leq (1+\epsilon)\|A-A_k\|_F$, following Sarlós' approach [5]. The first works which got some decent error (like $\varepsilon\|A\|_F$) was due to Papadimitriou [4] and Frieze, Kanna and Vempala [2].

Theorem 3. *As long as $\Pi \in \mathbb{R}^{m\times n}$ is a subspace embedding with error $\Theta(1)$ for a certain k-dimensional subspace and satisfies approximate matrix multiplication with error $\sqrt{\varepsilon/k}$, then*

$$\|A - \mathrm{Proj}_{A\Pi^T,k}(A)\|_F \leq (1+\varepsilon)\|A-A_k\|_F,$$

where $\mathrm{Proj}_{V,k}(A)$ is the best rank k approximation to $\mathrm{Proj}_V(A)$, i.e., projecting the columns of A to V.

It is useful to define the following:

Definition 4. *Let $A = U\Sigma V^T$ be a singular decomposition. $A^+ = V\Sigma^{-1}U^T$ is called* Moore-Penrose pseudoinverse *of A.*

In the following proof, we will denote first k columns of U and V by U_k and V_k, and remaining $r-k$ columns by U_{r-k} and V_{r-k}.

Proof. Let X be the column span of $\mathrm{Proj}_{A\Pi^T}(A_k)$. Denote the projection operator onto X by P. Then,

$$\|A - \mathrm{Proj}_{A\Pi^T,k}(A)\|_F^2 \leq \|A-PA\|_F^2$$

since PA has rank k and columns in X. And,

$$\|A-PA\|_F^2 = \|U\Sigma V^T - PU\Sigma V^T\|_F^2 = \|U\Sigma - PU\Sigma\|_F^2$$

which can be decomposed into $\|U_k\Sigma_k - PU_k\Sigma_k\|_F^2 + \|U_{r-k}\Sigma_{r-k} - PU_{r-k}\Sigma_{r-k}\|_F^2$ since U_k and U_{r-k} has orthogonal columns. Note that

$$\|U_{r-k}\Sigma_{r-k} - PU_{r-k}\Sigma_{r-k}\|_F^2 = \|(I-P)U_{r-k}\Sigma_{r-k}\|_F^2 \leq \|U_{r-k}\Sigma_{r-k}\|_F^2$$

since P is a projection. The right hand side is $\|U_{r-k}\Sigma_{r-k}\|_F^2 = \|A_{r-k}\|_F^2 = \|A-A_k\|_F^2$.

Now, it suffices to show that $\|U_k\Sigma_kV_k^T - PU_k\Sigma_kV_k^T\|_F^2 \leq 2\epsilon\|A-A_k\|_F^2$. Note that $PA_k = (A\Pi^T)(A\Pi^T)^+A_k$. Also, by definition PA_k is the best rank k approximation to A_k in the subspace spanned by columns of $A\Pi^T$. So,

$$\begin{aligned}
\|A_k - (A\Pi^T)(A\Pi^T)^+A_k\|_F^2 &\leq \|A_k - (A\Pi^T)(A_k\Pi^T)^+A_k\|_F^2 \\
&= \|A_k^T - A_k^T(\Pi A_k^T)^+(\Pi A^T)\|_F^2 \\
&= \sum_i \|(A_k^T)^{(i)} - A_k^T(\Pi A_k^T)^+(\Pi A^T)^{(i)}\|_2^2.
\end{aligned}$$

Here superscript (i) means the ith column. Now we have a bunch of different approximate regression problems which have the following form:

$$\min_x \|\Pi A_k^T x - \Pi(A^T)^{(i)}\|_2,$$

which has optimal value $\tilde{x}^* = (\Pi A_k^T)^+(\Pi A^T)^{(i)}$. Consider the problem $\min_x \|\Pi A_k^T x - (A^T)^{(i)}\|_2$ as original regression problem. In this case optimal x^* gives $A_k^T x^* = \text{Proj}_{A_k^T}((A^T)^{(i)}) = (A_k^T)^{(i)}$. Now we can use the analysis on the approximate least square from last week.

Recall the notations: for the least square problem $\min_x \|Sx - b\|_2$, optimal solution $x^* = S^+ b$, and approximate solution $\tilde{x}^* = \text{argmin}_x \|\Pi S x - \Pi x\|_2$, we let $x^* = U\alpha$, $w = Sx^* - b$, $U\beta = S\tilde{x}^* - Sx^*$ where $S = U\Sigma V^T$. We proved that $(\Pi U)^T(\Pi U)\beta = (\Pi U)^T \Pi w$ last time.

In our problem, we have a bunch of w_i, β_i, α_i with $S = A_k^T = V_k \Sigma_k U_k^T$ and $b_i = (A^T)^{(i)}$. Here, $\|w_i\|^2 = \|Sx^* - b\|^2 = \|(A_k^T)^{(i)} - (A^T)^{(i)}\|^2$. Hence $\sum_i \|w_i\|^2 = \|A - A_k\|_F^2$. On the other hand, $\sum_i \|\beta_i\|^2 = \|A_k^T - A_k^T(\Pi A_k^T)^+(\Pi A^T)\|_F^2$. Since $(\Pi V_k)^T(\Pi V_k)\beta_i = (\Pi V_k)^T \Pi w_i$, if all singlar values of ΠV_k are at least $1/2^{1/4}$, we have

$$\frac{\sum_i \|\beta_i\|^2}{2} \leq \sum_i \|(\Pi V_k)^T(\Pi V_k)\beta_i\|^2 = \sum_i \|(\Pi V_k)^T \Pi w_i\|^2 = \|(\Pi V_k)^T \Pi W\|_F^T$$

where W has w_i as ith column. What does it look like? $(\Pi V_k)^T \Pi W$ exactly look like approximate matrix multiplication of V_k and W. Since columns of W and V_k are orthogonal, we have $V_k^T W = 0$, hence if Π is a sketch for approximate matrix multiplication of error $\varepsilon' = \sqrt{\varepsilon/k}$, then

$$\mathop{\mathbb{P}}_{\Pi}(\|(\Pi V_k)^T(\Pi W)\|_F^2 > \varepsilon \|W\|_F^2) < \delta$$

since $\|V_k\|_F^2 = k$. Clearly $\|W\|_F^2 = \sum_i \|w_i\|^2 = \|A - A_k\|_F^2$, we get the desired result. □

2.3 Further results

What we just talked about gives a good low-rank approximation but every column of $\tilde{A}_k$ is a linear combination of potentially all columns of A. In applications (e.g. information retrieval), we want a few number of columns be spanning our low dimensional subspace. There has been work on finding fewer columns of A (call them C) such that $\|A - (CC^+A)_k\|_F^2$ is small, but we will not talk about it deeply.

- Boutsidis et al. [1] showed that we can take C with $\approx 2k/\varepsilon$ columns and error $\leq \epsilon \|A - A_k\|_F$.
- Guruswami and Sinop got C with $\leq \frac{k}{\varepsilon} + k - 1$ columns such that $\|A - CC^+A\|_F \leq (1 + \epsilon)\|A - A_k\|_F$.

3 Ways to get subspace embeddings

We will not talk about other ways of getting subspace embeddings. Here are the known ways:

1. Use Johnson-Lindenstrauss lemma.
2. Use approximate matrix multiplication.
3. Moment method.

We have looked at first two methods, and we will see the last one in the problem set. Note that Π is ε-subspace embedding for column span of U if and only if $\|(\Pi U)^T(\Pi U) - I\| \leq \varepsilon$. Define $S = (\Pi U)^T(\Pi U)$ then

$$\mathop{\mathbb{P}}_{\Pi}(\|S - I\| > \varepsilon) < \frac{1}{\epsilon^\ell} \mathbb{E} \|S - I\|^\ell \leq \frac{1}{\epsilon^\ell} \mathbb{E} \operatorname{tr}(S - I)^\ell,$$

for even integer ℓ. To calculate the trace of random matrices, we may use

$$(B^\ell)_{ij} = \sum_{i_1 = i, i_2, \ldots, i_{\ell+1} = j} \prod_{j=1}^{\ell} B_{i_j i_{j+1}}.$$

References

[1] Christos Boutsidis, Petros Drineas, Malik Magdon-Ismail. Near Optimal Column-based Matrix Reconstruction. *FOCS*, 305-314, 2011.

[2] Alan M. Frieze, Ravi Kannan, Santosh Vempala. Fast Monte-carlo Algorithms for Finding Low-rank Approximations. *J. ACM*, 51(6):1025-1041, 2004.

[3] Venkatesan Guruswami, Ali Kemal Sinop. Optimal Column-based Low-rank Matrix Reconstruction. *SODA*, 1207-1214, 2012.

[4] Christos H. Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, Santosh Vempala. Latent Semantic Indexing: A Probabilistic Analysis. *J. Comput. Syst. Sci.*, 61(2):217-235, 2000.

[5] Tamás Sarlós. Improved Approximation Algorithms for Large Matrices via Random Projections. *FOCS*, 143-152, 2006.