13 matrix sketching

Monday, October 19, 2020 12:44 PM

Last time we finished off by showing that the SVD provides a natural way to approximate matrices by low-rank matrices. Why is this important?

Recall that the SVD $A = VDU^T = \begin{bmatrix} V \end{bmatrix} \begin{bmatrix} D \end{bmatrix} \begin{bmatrix} U^T \end{bmatrix} = \sum_{i=1}^{r} \sigma_i v_i u_i^T$.

gives an optimal rank-k approximation $A_k = \sum_{i=1}^{k} \sigma_i v_i u_i^T = \begin{bmatrix} V_k \end{bmatrix} \begin{bmatrix} D_k \end{bmatrix} \begin{bmatrix} U_k^T \end{bmatrix}$,

which is formed by taking only the top-k singular values and singular vectors.

But computing an SVD requires computing singular vectors and values, which can be slow and memory intensive. Can we do something faster?

Thm 6.9 [Foundations of Data Science, 2020, Blum, Hopcroft, Kannan]

Let $A \in \mathbb{R}^{m \times n}$, and $r, s \in \mathbb{Z}^+$.

$$A = [A^1 \cdots A^n] = \begin{bmatrix} A_1 \\ \vdots \\ A_m \end{bmatrix},$$ where A^i are cols of A, A_i are rows of A.

Let $C \in \mathbb{R}^{m \times s} = [C^1 \cdots C^s]$ chosen by randomly sampling A^i as follows:

$$C^i = A^j \text{ with probability } \frac{\|A^j\|_2^2}{\|A\|_F^2}.$$

Let $R \in \mathbb{R}^{r \times n} = \begin{bmatrix} R^1 \\ \vdots \\ R^r \end{bmatrix}$ chose by randomly samply rows A_i as follows:

$$R^i = A_j \text{ with probability } \frac{\|A_j\|_2^2}{\|A\|^2}.$$

$$R^i = A_j \quad \text{with probability } \frac{\|A_j\|_2^2}{\|A\|_F^2}.$$

Then there exists $U \in \mathbb{R}^{s \times r}$ s.t.

$$\mathbb{E}\left(\|A - CUR\|_2^2\right) \leq \|A\|_F^2 \left(\frac{2}{\sqrt{r}} + \frac{2r}{s}\right).$$

If we fix s, we minimize error with $s^{2/3}$.

Choose $s = \frac{1}{\varepsilon^3}$ and $r = \frac{1}{\varepsilon^2}$. Then $\mathbb{E}\left(\|A - CUR\|_2^2\right) = O(\varepsilon)\|A\|_F^2$.

i.e. $A = \begin{bmatrix} A \\ \\ n \times m \end{bmatrix} \approx \underset{C}{\begin{bmatrix} \text{sample} \\ \text{columns} \\ n \times s \end{bmatrix}} \underset{U}{\begin{bmatrix} \text{multiplier} \\ s \times r \end{bmatrix}} \underset{R}{\begin{bmatrix} \text{sample rows} \\ r \times m \end{bmatrix}}.$

Matrix multiplication through sampling

Let $A \in \mathbb{R}^{m \times n}$

$B \in \mathbb{R}^{n \times p}$. We want to approximated AB is less than $O(mnp)$ time.

$$AB = \left[A^1 \cdots A^n\right] \begin{bmatrix} B_1 \\ \vdots \\ B_n \end{bmatrix} = \sum_{k=1}^{n} A^k B_k \quad \text{(sum of outer products)}.$$

Let's try to sample AB by taking components with prob. p_k.

i.e. let $z = k$ w.p. p_k for $k \in \{1, \ldots, n\}$ a random variable.

Define $X = \frac{1}{p_z} A^z B_z$, a matrix r.v.

Then the entry-wise expectation

$$\mathbb{E} X = \sum_{k=1}^{n} P(z=k) \cdot \frac{1}{p_k} A^k B_k = \sum_{k=1}^{n} A^k B_k = AB.$$

But when using an estimator, we care about both mean and variance.

Def. $\operatorname{Var}(X) = \mathbb{E}\left(\|AB - X\|_F^2\right)$, the entry-wise variance

Then $\operatorname{Var}(X) = \sum_{i=1}^{m} \sum_{j=1}^{p} \operatorname{Var}(x_{ij}) = \sum_{ij} \mathbb{E}(x_{ij}^2) - \mathbb{E}(x_{ij})^2 = \left(\sum_{ij} \sum_{k=1}^{n} p_k \cdot \frac{1}{p_k^2} \cdot a_{ik}^2 b_{kj}^2\right) - \underbrace{\|AB\|_F^2}_{\text{doesn't matter for optimizing } p_k}$

We want to choose p_k's to minimize variance.

Note: $\sum_{i,j,k} \frac{1}{p_k} a_{ik}^2 b_{kj}^2 = \sum_k \frac{1}{p_k} \|A^k\|_2^2 \|B_k\|_2^2$

Lemma: $\forall c_k \geq 0$, $f(p_1, \dots, p_n) = \sum_{k=1}^{n} \frac{c_k}{p_k}$, subject to the constraint $p_1 + \dots + p_n = 1$, is minimized by $p_k \sim \sqrt{c_k}$.

proof.

So $f(p_2, \dots, p_n) = \frac{c_1}{1 - (p_2 + \dots + p_n)} + \sum_{k=2}^{n} \frac{c_k}{p_k}$

$$\frac{\partial f}{\partial p_k} = \frac{c_1}{(1 - (p_2 + \dots + p_n))^2} - \frac{c_k}{p_k^2} = 0 \quad \text{at optimum}$$

$$\Rightarrow \frac{p_k}{1 - (p_2 + \dots + p_n)} = \sqrt{\frac{c_k}{c_1}}$$

$$\Rightarrow p_k = \sqrt{c_k} \cdot \frac{1 - (p_2 + \dots + p_n)}{\sqrt{c_1}} \quad \forall k \neq 1.$$

$\square$

Thus, we want to pick $p_k \sim \|A^k\|_2 \|B_k\|_2$.

Note, when $B = A^T$, $p_k \sim \|A^k\|_2^2$ (squared length of cols)

Even if $B \neq A^T$, this is still an upper bound on $Var(X)$.

So use $p_k = \dfrac{\|A^k\|_2^2}{\|A\|_F^2}$.

$$\Rightarrow \mathbb{E}\left(\|AB - X\|_F^2\right) = Var(X) \leq \|A\|_F^2 \sum_{k=1}^{n} \|B_k\|_2^2 = \|A\|_F^2 \|B\|_F^2$$

Repeat with s independent trials to get $X_1, \dots, X_s$. Let $\bar{X} = \frac{1}{s}\sum_{i=1}^{s} X_i$.

Then $Var(\bar{X}) = \frac{1}{s} Var(X) \leq \frac{1}{s}\|A\|_F^2 \|B\|_F^2$

Note: $\frac{1}{s}\sum_{i=1}^{s} X_i = \frac{1}{s}\left(\frac{A^{k_1} B_{k_1}}{p_{k_1}} + \dots + \frac{A^{k_s} B_{k_s}}{p_{k_s}}\right)$, where each k_i is an ind. choice of col.

$= CR$, where

$$C = \left[\frac{A^{k_1}}{\sqrt{s p_{k_1}}} \quad \cdots \quad \frac{A^{k_s}}{\sqrt{s p_{k_s}}}\right], \quad R = \begin{bmatrix} \frac{B_{k_1}}{\sqrt{s p_{k_1}}} \\ \vdots \\ \frac{B_{k_s}}{\sqrt{s p_{k_s}}} \end{bmatrix}.$$

Thm 6.5 Suppose $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times p}$. CR as given above is an estimator for AB, and the error is bounded

$$\mathbb{E}\left(\|AB - CR\|_F^2\right) \leq \frac{\|A\|_F^2 \|B\|_F^2}{s},$$

To ensure $\mathbb{E}\left(\|AB - CR\|_F^2\right) \leq \varepsilon^2 \|A\|_F^2 \|B_F\|^2$ for some $\varepsilon > 0$, it suffices to make $s \geq \frac{1}{\varepsilon^2}$. Thus, CR can be computed in $O(msp)$ time. (often $< O(mnp)$ time)

Lemma 6.6: Given $R = VDU^T$ an SVD, let $P = R^+ R = UD^+V^T R$. P is

Lemma 6.6: Given $R = VDU^T$ an SVD, let $P = R^+R = UD^+V^TR$. P is a projection operator satisfying:

(i) $Px = x$ for every $x = R^Ty$ (i.e. $x \in$ col span of R^T / row span of R)

(ii) If $x \perp R^Ty$ $\forall y$, then $Px = 0$.

(Note $U^TU = I$, $V^TV = I$)

(But $UU^T \neq I$, $VV^T \neq I$)

proof. (i) If $x = R^Ty$ for some y, then

$$\begin{aligned} Px &= R^+RR^Ty \\ &= UD^+\underbrace{V^TV}DU^T\underbrace{U^{T}U}DV^Ty \\ &= UD^+D^2V^Ty = UDV^Ty \\ &= R^Ty = x. \end{aligned}$$

(ii) If $x \perp R^Ty$ $\forall y$, then

$$Px = R^+Rx = UD^+V^TVDU^Tx = UU^Tx.$$

But each row of U^T is a col of U and an eigenvector of R^TR, and thus in the col span of R^T / row span of R.

But $x \perp R^Ty$ $\forall y$, so $U^Tx = 0 \Rightarrow Px = 0$.

Prop. 6.7 $A \approx AP$ and the error $\mathbb{E}\|A - AP\|_2^2 \leq \frac{1}{\sqrt{r}}\|A\|_F^2$.

proof. $\|A - AP\|_2^2 = \max_{\|x\|_2 = 1} \|(A - AP)x\|_2^2$

If $x \in$ row space(R), then $Px = x$, so $(A - AP)x = 0$.

By linearity $\max_{\|x\|_2=1} \|(A - AP)x\|_2^2 = \max_{\substack{\|x\|_2 = 1 \\ x \perp \text{row space}(R)}} \|(A - AP)x\|_2^2 = \max_{\substack{\|x\|_2 = 1 \\ x \perp \text{row space}(R)}} \|Ax\|_2^2$.

Note $\|Ax\|_2^2 = x^TA^TAx = x^T(A^TA - R^TR)x \leq \|A^TA - R^TR\|_2$

↑ Rayleigh-Ritz

$$\Rightarrow \|A - AP\|_2^2 \leq \|A^TA - R^TR\|_2 \leq \|A^TA - R^TR\|_F$$

Choose R by sampling r rows of A according to a length square distribution.
(cols of A^T)

[illegible] multiplication sampling $\mathbb{E}\|A - AP\|^2 \leq \frac{\|A\|_F^2}{\text{[illegible]}}$

Then by the matrix multiplication sampling, $\mathbb{E}\|A-AP\|_2^2 \le \frac{\|A\|_F^2}{\sqrt{r}}$

Prop. 6.8 $\|P\|_F^2 \le r$, if we choose R by sampling r rows of A and $P = R^+R$.

pf. $P = UD^+V^TVDU^T = UU^T = UI_qU^T$, where $q = \mathrm{rank}(R) \le r$.

Then $\|P\|_F^2 = q \le r$ because all the singular values are 1.

Proof of matrix sketch $\mathbb{E}(\|A-CUR\|_2^2) \le \|A\|_F^2\left(\frac{2}{\sqrt{r}} + \frac{2r}{s}\right)$

Let's approximate AP by matrix multiplication sampling.

Let C = s sampled cols of A.

P' = s sampled rows of $P = R^+R$

Then $CP' = C\underbrace{\begin{bmatrix} 1 & & & \\ & & 1 & \\ & 1 & & \\ & & & 1 \end{bmatrix}}_{\text{sampling matrix}}R^+R \equiv CUR$.

Then $\mathbb{E}\|AP-CUR\|_2^2 \le \mathbb{E}\|AP-CUR\|_F^2 \le \frac{\|A\|_F^2\|P\|_F^2}{s} \le \frac{r}{s}\|A\|_F^2$.

But $\|A-CUR\|_2 \le \|A-AP\|_2 + \|AP-CUR\|_2$

$\left(\begin{array}{l} c = a+b \\ c^2 \le 2a^2+2b^2 \end{array}\right)$

$\Rightarrow \|A-CUR\|_2^2 \le 2\|A-AP\|_2^2 + 2\|AP-CUR\|_2^2$

$\Rightarrow \mathbb{E}\|A-CUR\|_2^2 \le \frac{2}{\sqrt{r}}\|A\|_F^2 + \frac{2r}{s}\|A\|_F^2$