Singular Value Decomposition (SVD)

Given a matrix $A \in \mathbb{R}^{n \times d}$, the SVD factorization is given by

$$A = U \Sigma V^T$$  \hspace{1cm} (1)

where $U \in \mathbb{R}^{n \times r}$ and $V \in \mathbb{R}^{d \times r}$ contains the left and right singular vectors as columns and they are unitary matrices. $V \in \mathbb{R}^{r \times r}$ is a diagonal matrix with the non-zero singular values ranked from the largest to the smallest on the diagonal. $r$ is the rank of matrix $A$. Let $u_i$ and $v_i (1 \leq i \leq r)$ be the columns of matrices $U$ and $V$, then $A$ can be written as

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$$  \hspace{1cm} (2)

where $u_i v_i^T$ is a matrix with rank 1. Every matrix has an SVD decomposition, however it might not be unique. A common application of SVD is in dimensionality reduction. Suppose we want to find the rank 1 subspace that keeps the most information of $A$ which has $n$ points of $d$ dimensions. Then the loss function would be

$$l(v) = \sum_{i=1}^{n} \|a_i - (a_i^T v)v\|^2$$

$$= \sum_{i=1}^{n} (\|a_i\|^2 - (a_i^T v)^2)$$  \hspace{1cm} (3)

where $v$ is a unit vector. Since $\|a_i\|$ is constant, the loss minimization problem could change to the following maximization problem:

$$\arg \max_{v, \|v\|=1} \|Av\|^2$$  \hspace{1cm} (4)

The columns $v_i$ in matrix $V$ form orthogonal basis. So $v$ could be represented as

$$v = \sum_{i=1}^{r} \alpha_i v_i + v$$  \hspace{1cm} (5)

leading to

$$Av = \sum_{i=1}^{r} \alpha_i \sigma_i u_i$$  \hspace{1cm} (6)

Thus, we could get

$$\|Av\|^2 = \sum_{i=1}^{r} \alpha_i^2 \sigma_i^2$$  \hspace{1cm} (7)

Because $\|v\|=1$, then $\sum_{i=1}^{r} \alpha_i^2 = 1$. That means if we want to maximize $\sum_{i=1}^{r} \alpha_i^2 \sigma_i^2$, we need to put all the weights on the largest $\sigma_i^2$, i.e. $\sigma_1^2$. Therefore, $v_1$ is the optimal solution and the best rank 1 approximation of $A$ should be

$$A_1 = \sigma_1 u_1 v_1^T$$  \hspace{1cm} (8)
Similarly, the best-k approximation of $A$ will be

$$A_k = \sum_{i=1}^{k} \sigma_i \mu_i v_i^T$$  \hspace{1cm} (9)

Next we introduce a simple algorithm that computes the top singular vector. The algorithm is known as the Power Method. Let $B = A^T A$, then the algorithm is described in Algorithm 1.

Algorithm 1 Power Method

1: initialize unit vector $x_0 \sim \mathcal{N}(0, I)$
2: for $k = 0$ to $m$ do
3: $x_{k+1} = \frac{Bx_k}{\|Bx_k\|}$
4: return $v = x_{m+1}$

Theorem 1. Assume that $|v_1^T x_0| \geq \delta = \frac{1}{20\sqrt{d}}$. If for some $\epsilon > 0$, $\sigma_2 < (1-\epsilon)\sigma_1$, then after $m = \Omega(\frac{1}{\epsilon} \log \frac{d}{\epsilon^2})$ iterations, $x_{m+1}^T v_1 \geq 1 - \epsilon$.

Proof. Representing $x_0$ using $V$ as a basis, we get $x_0 = \sum_{i=1}^{r} \alpha_i v_i + v_1$. Similarly, $B$ could be written as $B = \sum_{i=1}^{r} \sigma_i^2 v_i v_i^T$. Then, $B^k x_0 = \sum_{i=1}^{r} \sigma_i^{2k} \alpha_i v_i = \sigma_1^{2k}[\alpha_1 v_1 + \sum_{i=2}^{r} (\frac{\sigma_i}{\sigma_1})^{2k} \alpha_i v_i]$. The norm squares of projections on $v_1$ and its orthogonal complements $v_1^\perp$ would be $\|B^k x_0 v_1\|^2 = \sigma_1^{4k}\alpha_1^2 \geq \sigma_1^{4k}\delta^2$ and $\|B^k x_0 v_1^\perp\|^2 = \sum_{i=2}^{r} \sigma_i^{4k}\alpha_i^2 \leq (1-\epsilon)^{4k}\sigma_1^{4k}\sum_{i=2}^{r} \alpha_i^2 \leq (1-\epsilon)^{4k} \sigma_1^{4k}$. In order to get $\|B^k x_0 v_1\|^2/\|B^k x_0 v_1^\perp\|^2 \geq 1 - \epsilon$, it is enough to run for $k \geq \frac{10}{\epsilon^2} \log \frac{1}{\epsilon} = \Omega(\frac{1}{\epsilon} \log (\frac{d}{\epsilon^2}))$ iterations. \hfill \Box

Next we show that a randomly chosen unit vector will have a decent correlation with $v_1$.

Lemma 1. For fixed vector $v$, if $x_0 \sim \mathcal{N}(0, I)$, then with probability at least $\frac{4}{5}$, $\frac{\|v^T x_0\|}{\|x_0\|} \geq \frac{1}{20\sqrt{d}}$.

Proof. Since $x_0$ is a random Gaussian vector, with probability at least $1 - 3e^{-d/64}$ we have that $\|x_0\| \leq 20\sqrt{d}$. Also, notice that $x^T v$ is a one dimensional Gaussian random variable with mean 0 and variance 1. Hence, with probability at least $\frac{9}{10}$, $|x_0^T v| \leq \frac{1}{10}$. Combining the two, we get the claim. \hfill \Box

Another way to understand the power method is as a special case of a more general strategy known as Alternate Minimization. Alternate minimization is a heuristic to solve the following optimization problem

$$\arg \min_{x, \|x\|=1} \|B - xx^T\|_F$$  \hspace{1cm} (10)

where $\|\cdot\|_F$ is the Frobenius norm of matrix. Since the objective is non-convex, the algorithm alternately optimizes the left and right $x$ in the objective function by fixing the other. The benefit of alternating minimization is that by fixing one $x$, the objective function becomes quadratic which has a closed form solution. If we fix $x_{right}$, and let the derivative equal to 0, we could get $x_{left} = Bx_{right}$. Likewise, when $x_{left}$ is fixed, then $x_{right} = Bx_{left}$. The details are given in Algorithm 2.
Algorithm 2 Alternate Minimization

1: Randomly initialize $x_{right} = x_0$;
2: for $k=1,2,\ldots,m$ do
3: \quad $x_{left}^k = \arg \min_x \|B - x(x_{right}^{k-1})^T\|_F^2$;
4: \quad $x_{right}^k = \arg \min_x \|B - x_{left}^k x^T\|_F^2$;
\textbf{return} $v = x_{left}^{m+1}$.

0.1 Additional Reading


3