zhuanlan.zhihu.com

矩阵的奇异值分解（singular value decomposition，简称 SVD）是线性代数中很重要的内容，并且奇异值分解过程也是线性代数中相似对角化分解（也被称为特征值分解，eigenvalue decomposition，简称 EVD）的延伸。因此，以下将从线性代数中最基础的矩阵分解开始讲起，引出奇异值分解的定义，并最终给出奇异值分解的低秩逼近问题相关的证明过程。

1 线性代数中的矩阵分解

我们在学习线性代数时，就已经接触了线性代数中的两个重要定理，即对角化定理和相似对角化定理，在这里，我们先简单地回顾一下这两个定理。另外，在接下来的篇幅里，我们所提到的矩阵都是指由实数构成的矩阵，即实矩阵。

给定一个大小为$m \times m$的矩阵A（是方阵），其对角化分解可以写成

$$A = U\Lambda U^{-1}$$

其中，U的每一列都是特征向量，Λ对角线上的元素是从大到小排列的特征值，若将U记作$U = (\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_m)$，则

$$AU = A(\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_m) = (\lambda_1\vec{u}_1, \lambda_2\vec{u}_2, \ldots, \lambda_m\vec{u}_m)$$

$$= (\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_m)\begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_m \end{bmatrix}$$

$$\Rightarrow AU = U\Lambda \Rightarrow A = U\Lambda U^{-1}$$

更为特殊的是，当矩阵A是一个对称矩阵时，则存在一个对称对角化分解，即

$$A = Q\Lambda Q^T$$

其中，Q的每一列都是相互正交的特征向量，且是单位向量，Λ对角线上的元素是从大到小排列的特征值。

当然，将矩阵Q记作$Q = (\vec{q}_1, \vec{q}_2, \ldots, \vec{q}_m)$，则矩阵$A$也可以写成如下形式：

$$A = \lambda_1\vec{q}_1\vec{q}_1^T + \lambda_2\vec{q}_2\vec{q}_2^T + \ldots + \lambda_m\vec{q}_m\vec{q}_m^T$$

举一个简单的例子，如给定一个大小为2×2的矩阵$A=\begin{bmatrix}2 & 1\\ 1 & 2\end{bmatrix}$，根据

$|\lambda I-A|=\begin{vmatrix}\lambda-2 & -1\\ -1 & \lambda-2\end{vmatrix}=0$求得特征值为$\lambda_1=3$，$\lambda_2=1$，相应地，

$\vec{q}_1=\left(\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2}\right)^T$，$\vec{q}_2=\left(-\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2}\right)^T$，则

$A=\lambda_1\vec{q}_1\vec{q}_1^T+\lambda_2\vec{q}_2\vec{q}_2^T=\begin{bmatrix}2 & 1\\ 1 & 2\end{bmatrix}$.

这样，我们就很容易地得到了矩阵A的对称对角化分解。

2 奇异值分解的定义

在上面，对于对称的方阵而言，我们能够进行对称对角化分解，试想：对称对角化分解与奇异值分解有什么本质关系呢？

当给定一个大小为$m\times n$的矩阵A，虽然矩阵A不一定是方阵，但大小为$m\times m$的AA^T和$n\times n$的A^TA却是对称矩阵，若$AA^T=P\Lambda_1P^T$，$A^TA=Q\Lambda_2Q^T$，则矩阵A的奇异值分解为

$$A=P\Sigma Q^T$$

其中，矩阵$P=(\vec{p}_1,\vec{p}_2,...,\vec{p}_m)$的大小为$m\times m$，列向量$\vec{p}_1,\vec{p}_2,...,\vec{p}_m$是$AA^T$的特征向量，也被称为矩阵$A$的**左奇异向量**（left singular vector）；矩阵$Q=(\vec{q}_1,\vec{q}_2,...,\vec{q}_n)$的大小为$n\times n$，列向量$\vec{q}_1,\vec{q}_2,...,\vec{q}_n$是$A^TA$的特征向量，也被称为矩阵$A$的**右奇异向量**（right singular vector）；矩阵Λ_1大小为$m\times m$，矩阵Λ_2大小为$n\times n$，两个矩阵对角线上的非零元素相同（即矩阵AA^T和矩阵A^TA的非零特征值相同，推导过程见附录 1）；矩阵Σ的大小为$m\times n$，位于对角线上的元素被称为**奇异值**（singular value）。

接下来，我们来看看矩阵Σ与矩阵AA^T和矩阵A^TA的关系。令常数k是矩阵A的秩，则$k\leq\min(m,n)$，当$m\neq n$时，很明显，矩阵Λ_1和矩阵Λ_2的大小不同，但矩阵Λ_1和矩阵Λ_2对角线上的非零元素却是相同的，若将矩阵Λ_1（或矩阵Λ_2）对角线上的非零元素分别为$\lambda_1,\lambda_2,...,\lambda_k$，其中，这些特征值也都是非负的，再令矩阵$\Sigma$对角线上的非零元素分别为$\sigma_1,\sigma_2,...,\sigma_k$，则

$$\sigma_1=\sqrt{\lambda_1},\sigma_2=\sqrt{\lambda_2},...,\sigma_k=\sqrt{\lambda_k}$$

即非零奇异值的平方对应着矩阵Λ_1（或矩阵Λ_2）的非零特征值，到这里，我们就不难看出奇异值分解与对称对角化分解的关系了，即我们可以由对称对角化分解得到我们想要的奇异值分解。

为了便于理解，在这里，给定一个大小为2×2的矩阵$A=\begin{bmatrix}4 & 4\\ -3 & 3\end{bmatrix}$，虽然这个矩阵是方阵，但却不是对称矩阵，我们来看看它的奇异值分解是怎样的。

由$AA^T=\begin{bmatrix}32&0\\0&18\end{bmatrix}$进行对称对角化分解，得到特征值为$\lambda_1=32$，$\lambda_2=18$，相应地，特征向量为$\vec{p}_1=(1,0)^T$，$\vec{p}_2=(0,1)^T$；由$A^TA=\begin{bmatrix}25&7\\7&25\end{bmatrix}$进行对称对角化分解，得到特征值为$\lambda_1=32$，$\lambda_2=18$，相应地，特征向量为$\vec{q}_1=\left(\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2}\right)^T$，$\vec{q}_2=\left(-\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2}\right)^T$。取$\Sigma=\begin{bmatrix}4\sqrt{2}&0\\0&3\sqrt{2}\end{bmatrix}$，则矩阵$A$的奇异值分解为

$$A=P\Sigma Q^T=(\vec{p}_1,\vec{p}_2)\,\Sigma(\vec{q}_1,\vec{q}_2)^T$$

$$=\begin{bmatrix}1&0\\0&1\end{bmatrix}\begin{bmatrix}4\sqrt{2}&0\\0&3\sqrt{2}\end{bmatrix}\begin{bmatrix}\frac{\sqrt{2}}{2}&\frac{\sqrt{2}}{2}\\-\frac{\sqrt{2}}{2}&\frac{\sqrt{2}}{2}\end{bmatrix}=\begin{bmatrix}4&4\\-3&3\end{bmatrix}.$$

若矩阵A不再是一个方阵，而是一个大小为3×2的$A=\begin{bmatrix}1&2\\0&0\\0&0\end{bmatrix}$，由

$AA^T=\begin{bmatrix}5&0&0\\0&0&0\\0&0&0\end{bmatrix}$得到特征值为$\lambda_1=5$，$\lambda_2=\lambda_3=0$，特征向量为$\vec{p}_1=(1,0,0)^T$，$\vec{p}_2=(0,1,0)^T$，$\vec{p}_3=(0,0,1)^T$；由$A^TA=\begin{bmatrix}1&2\\2&4\end{bmatrix}$得到特征值为$\lambda_1=5$，$\lambda_2=0$，特征向量为$\vec{q}_1=\left(\frac{\sqrt{5}}{5},\frac{2\sqrt{5}}{5}\right)^T$，$\vec{q}_2=\left(-\frac{2\sqrt{5}}{5},\frac{\sqrt{5}}{5}\right)^T$，令$\Sigma=\begin{bmatrix}\sqrt{5}&0\\0&0\\0&0\end{bmatrix}$（注意：矩阵$\Sigma$大小为$3\times 2$），此时，矩阵$A$的奇异值分解为

$$A=P\Sigma Q^T=(\vec{p}_1,\vec{p}_2)\,\Sigma(\vec{q}_1,\vec{q}_2)^T$$

$$=\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}\begin{bmatrix}\sqrt{5}&0\\0&0\\0&0\end{bmatrix}\begin{bmatrix}\frac{\sqrt{5}}{5}&\frac{2\sqrt{5}}{5}\\-\frac{2\sqrt{5}}{5}&\frac{\sqrt{5}}{5}\end{bmatrix}=\begin{bmatrix}1&2\\0&0\\0&0\end{bmatrix}.$$

比较有趣的是，假设给定一个对称矩阵$A=\begin{bmatrix}2&1\\1&2\end{bmatrix}$，它是对称矩阵，则其奇异值分解是怎么样的呢？

分别计算AA^T和A^TA，我们会发现，$AA^T=A^TA=\begin{bmatrix}2&1\\1&2\end{bmatrix}\begin{bmatrix}2&1\\1&2\end{bmatrix}$ $=\begin{bmatrix}5&4\\4&5\end{bmatrix}$，左奇异向量和右奇异向量构成的矩阵也是相等的，即

$P=Q=\begin{bmatrix}\frac{\sqrt{2}}{2}&-\frac{\sqrt{2}}{2}\\\frac{\sqrt{2}}{2}&\frac{\sqrt{2}}{2}\end{bmatrix}$，更为神奇的是，该矩阵的奇异值分解和对称对角化分解相同，都是$A=\begin{bmatrix}\frac{\sqrt{2}}{2}&-\frac{\sqrt{2}}{2}\\\frac{\sqrt{2}}{2}&\frac{\sqrt{2}}{2}\end{bmatrix}\begin{bmatrix}3&0\\0&1\end{bmatrix}\begin{bmatrix}\frac{\sqrt{2}}{2}&\frac{\sqrt{2}}{2}\\-\frac{\sqrt{2}}{2}&\frac{\sqrt{2}}{2}\end{bmatrix}$。这是由于对于正定对称矩阵而言，奇异值分解和对称对角化分解结果相同。

3 奇异值分解的低秩逼近

在对称对角化分解中，若给定一个大小为3×3的矩阵$A=\begin{bmatrix}30&0&0\\0&20&0\\0&0&1\end{bmatrix}$，很显然，矩阵$A$的秩为$rank(A)=3$，特征值为$\lambda_1=30$，$\lambda_2=20$，$\lambda_3=1$，对应的特征向量分别为$\vec{q}_1=(1,0,0)^T$，$\vec{q}_2=(0,1,0)^T$，$\vec{q}_3=(0,0,1)^T$，考虑任意一个向量$\vec{v}=(2,4,6)^T=2\vec{q}_1+4\vec{q}_2+6\vec{q}_3$，则

$$A\vec{v}=A(2\vec{q}_1+4\vec{q}_2+6\vec{q}_3)$$

$$=2\lambda_1\vec{q}_1+4\lambda_2\vec{q}_2+6\lambda_3\vec{q}_3=60\vec{q}_1+80\vec{q}_2+6\vec{q}_3$$

在这里，我们会发现，即使$\vec{v}$是一个任意向量，用矩阵A去乘以$\vec{v}$的效果取决于A较大的特征值及其特征向量，类似地，在奇异值分解中，较大的奇异值会决定原矩阵的“主要特征”，下面我们来看看奇异值分解的低秩逼近（有时也被称为截断奇异值分解）。需要说明的是，接下来的部分是从文献《A Singularly Valuable Decomposition: The SVD of a Matrix》整理而来的。

给定一个大小为$m\times n$的矩阵A，由于$A=P\Sigma Q^T$可以写成

$$A=\sum_{i=1}^{k}\sigma_i\vec{p}_i\vec{q}_i^T=\sigma_1\vec{p}_1\vec{q}_1^T+\sigma_2\vec{p}_2\vec{q}_2^T+\ldots+\sigma_k\vec{p}_k\vec{q}_k^T$$

其中，向量$\vec{p}_1,\vec{p}_2,\ldots,\vec{p}_k$之间相互正交，向量$\vec{q}_1,\vec{q}_2,\ldots,\vec{q}_k$之间也相互正交，由内积$\left\langle\sigma_i\vec{p}_i\vec{q}_i^T,\sigma_j\vec{p}_j\vec{q}_j^T\right\rangle=0,1\leq i\leq k,1\leq j\leq k,i\neq j$（有兴趣的读者可以自行推算）得到矩阵$A$的 F - 范数的平方为

$$||A||_F^2 = ||\sigma_1 \vec{p}_1 \vec{q}_1^T + \sigma_2 \vec{p}_2 \vec{q}_2^T + \ldots + \sigma_k \vec{p}_k \vec{q}_k^T||_F^2$$
$$= \sigma_1^2 ||\vec{p}_1 \vec{q}_1^T||_F^2 + \sigma_2^2 ||\vec{p}_2 \vec{q}_2^T||_F^2 + \ldots + \sigma_k^2 ||\vec{p}_k \vec{q}_k^T||_F^2$$
$$= \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_k^2 = \sum_{i=1}^{r} \sigma_i^2$$

知道了矩阵A的 F - 范数的平方等于其所有奇异值的平方和之后，假设$A_1 = \sigma_1 \vec{p}_1 \vec{q}_1^T$是矩阵$A$的一个秩一逼近（rank one approximation），那么，它所带来的误差则是$\sigma_2^2 + \sigma_3^2 + \ldots + \sigma_k^2$（$k$是矩阵$A$的秩），不过如何证明$A_1 = \sigma_1 \vec{p}_1 \vec{q}_1^T$是最好的秩一逼近呢？

由于$||A - A_1||_F^2 = ||P\Sigma Q^T - A_1||_F^2 = ||\Sigma - P^T A_1 Q||_F^2$（证明过程见附录 2），令$P^T A_1 Q = \alpha \vec{x} \vec{y}^T$，其中，$\alpha$是一个正常数，向量$\vec{x}$和$\vec{y}$分别是大小为$m \times 1$和$n \times 1$的单位向量，则

$$||\Sigma - P^T A_1 Q||_F^2 = ||\Sigma - \alpha \vec{x} \vec{y}^T||_F^2 = ||\Sigma||_F^2 + \alpha^2 - 2\alpha \left\langle \Sigma, \vec{x} \vec{y}^T \right\rangle$$

单独看大小为$m \times n$的矩阵Σ和$\vec{x} \vec{y}^T$的内积$\left\langle \Sigma, \vec{x} \vec{y}^T \right\rangle$，我们会发现，

$$\left\langle \Sigma, \vec{x} \vec{y}^T \right\rangle = \sum_{i=1}^{k} \sigma_i x_i y_i \leq \sum_{i=1}^{k} \sigma_i |x_i| |y_i|$$

$$\leq \sigma_1 \sum_{i=1}^{k} |x_i| |y_i| = \sigma_1 \langle \vec{x}^*, \vec{y}^* \rangle$$
$$\leq \sigma_1 ||\vec{x}^*|| \cdot ||\vec{y}^*|| \leq \sigma_1 ||\vec{x}|| \cdot ||\vec{y}|| = \sigma_1$$

其中，需要注意的是，x_i, y_i分别是向量$\vec{x}$和$\vec{y}$的第i个元素；向量$\vec{x}^* = (|x_1|, |x_2|, \ldots, |x_k|)^T$的大小为$k \times 1$，向量$\vec{y}^* = (|y_1|, |y_2|, \ldots, |y_k|)^T$的大小也为$k \times 1$，另外，以$\vec{x}^*$为例，$||\vec{x}^*|| = \sqrt{x_1^2 + x_2^2 + \ldots + x_k^2}$是向量的模，则$||A - A_1||_F^2$（残差矩阵的平方和）为

$$||\Sigma - \alpha \vec{x} \vec{y}^T||_F^2 \geq ||\Sigma||_F^2 + \alpha^2 - 2\alpha\sigma_1 = ||\Sigma||_F^2 + (\alpha - \sigma_1)^2 - \sigma_1^2$$

当且仅当$\alpha = \sigma_1$时，$||A - A_1||_F^2$取得最小值$\sigma_2^2 + \sigma_3^2 + \ldots + \sigma_k^2$，此时，矩阵$A$的秩一逼近恰好是$A_1 = \sigma_1 \vec{p}_1 \vec{q}_1^T$。

当然，我们也可以证明$A_2=\sigma_2\vec{p}_2\vec{q}_2^T$是矩阵$A-A_1$的最佳秩一逼近，以此类推，$A_r=\sigma_r\vec{p}_r\vec{q}_r^T, r<k$是矩阵$A-A_1-A_2-...-A_{r-1}$的最佳秩一逼近。由于矩阵$A_1+A_2+...+A_r$的秩为$r$，这样，我们可以得到矩阵$A$的最佳秩$r$逼近（rank-$r$ approximation），即

$$A\approx A_1+A_2+...+A_r=\sum_{i=1}^{r}A_i.$$

这里得到的矩阵P_r的大小为$m\times r$，矩阵Σ_r的大小为$r\times r$，矩阵Q_r的大小为$n\times r$，矩阵A可以用$P_r\Sigma_rQ_r^T$来做近似。

用低秩逼近去近似矩阵A有什么价值呢？给定一个很大的矩阵，大小为$m\times n$，我们需要存储的元素数量是mn个，当矩阵A的秩k远小于m和n，我们只需要存储$k(m+n+1)$个元素就能得到原矩阵A，即k个奇异值、km个左奇异向量的元素和kn个右奇异向量的元素；若采用一个秩r矩阵$A_1+A_2+...+A_r$去逼近，我们则只需要存储更少的$r(m+n+1)$个元素。因此，奇异值分解是一种重要的数据压缩方法。

另外，关于奇异值分解的应用将在该系列后续文章中进行详述。

附录 1：相关链接：[Largest eigenvalues of AA' equals to A'A]，截图如下：

For any $m\times n$ and $n\times m$ matrices A and B, the nonzero eigenvalues of AB and BA are the same. Namely, if $ABv=\lambda v$ with $\lambda\neq 0$ and $v\neq 0$, then $Bv\neq 0$ and $BA(Bv)=B(ABv)=\lambda Bv$.

share cite improve this answer

answered Apr 24 '15 at 6:32
Robert Israel
254k ● 15 ■ 167 ▲ 379

附录 2：求证：$||P\Sigma Q^T-A_1||_F^2=||\Sigma-P^TA_1Q||_F^2$，其中，$QQ^T=I$，$PP^T=I$.

证明：$||P\Sigma Q^T-A_1||_F^2$

$$=trace\left(\left(P\Sigma Q^T-A_1\right)\left(P\Sigma Q^T-A_1\right)^T\right)$$

$$=trace\left(\left(P\Sigma Q^T-A_1\right)QQ^T\left(P\Sigma Q^T-A_1\right)^T\right)$$

$$=trace\left(\left(P\Sigma-A_1Q\right)\left(\Sigma^TP^T-Q^TA_1^T\right)\right)$$

$$=trace\left(\left(\Sigma^TP^T-Q^TA_1^T\right)\left(P\Sigma-A_1Q\right)\right)$$

$$=trace\left(\left(\Sigma^TP^T-Q^TA_1^T\right)PP^T\left(P\Sigma-A_1Q\right)\right)$$

$$= trace\left(\left(\Sigma^T - Q^T A_1^T P\right)\left(\Sigma - P^T A_1 Q\right)\right)$$

$$= ||\Sigma - P^T A_1 Q||_F^2.$$