奇异值分解（SVD）原理总结

原创 石头 机器学习算法那些事 2019-03-07

前言

奇异值分解（SVD）在降维，数据压缩，推荐系统等有广泛的应用，任何矩阵都可以进行奇异值分解，本文通过正交变换不改变基向量间的夹角循序渐进的推导SVD算法，以及用协方差含义去理解行降维和列降维，最后介绍了SVD的数据压缩原理 。

目录

1. 正交变换

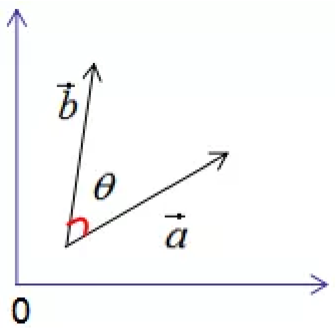
正交变换公式：

$$X = UY$$

上式表示：X是Y的正交变换 ，其中U是正交矩阵，X和Y为列向量 。

下面用一个例子说明正交变换的含义：

假设有两个单位列向量a和b，两向量的夹角为θ，如下图：



现对向量a，b进行正交变换：

$$\vec{a}' = U * \vec{a}$$

$$\vec{b}' = U * \vec{b}$$

$\vec{a}'$，$\vec{b}'$ 的模：

$$\| \vec{a}' \| = \| U * \vec{a} \| = \| U \| * \| \vec{a} \| = \| \vec{a} \| = 1$$

$$\| \vec{b}' \| = \| U * \vec{b} \| = \| U \| * \| \vec{b} \| = \| \vec{b} \| = 1$$

由上式可知 $\vec{a}'$ 和 $\vec{b}'$ 的模都为1。

$\vec{a}'$ 和 $\vec{b}'$ 的内积：

$$\vec{a}'^T * \vec{b}' = (U * \vec{a})^T * (U * \vec{b}) = \vec{a}^T U^T U \vec{b}$$

$$\Rightarrow \vec{a}'^T * \vec{b}' = \vec{a}^T * \vec{b} \qquad (1)$$

由上式可知，正交变换前后的内积相等。

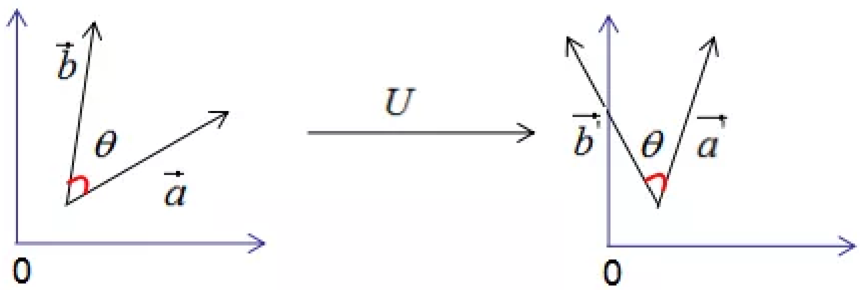
$\vec{a}'$ 和 $\vec{b}'$ 的夹角 θ'：

$$\cos\theta' = \frac{\vec{a}'^T * \vec{b}'}{\| \vec{a}' \| * \| \vec{b}' \|} \qquad (2)$$

$$\cos\theta = \frac{\vec{a}^T * \vec{b}}{\| \vec{a} \| * \| \vec{b} \|} \qquad (3)$$

比较（2）式和（3）式得：正交变换前后的夹角相等，即：$\theta = \theta'$

因此，正交变换的性质可用下图来表示：

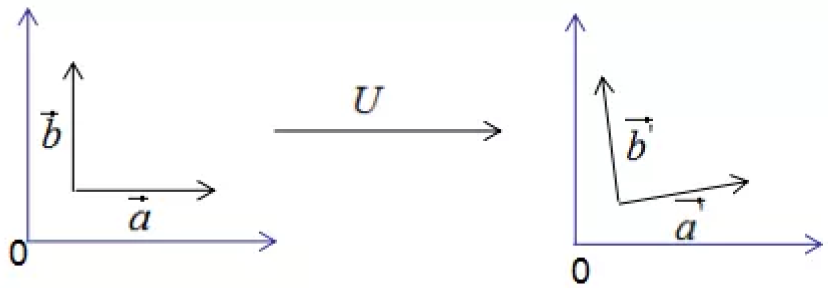


正交变换的两个重要性质：

1）正交变换不改变向量的模。

2）正交变换不改变向量的夹角。

如果向量 $\vec{a}$ 和 $\vec{b}$ 是基向量，那么正交变换的结果如下图：



上图可以得到重要结论：基向量正交变换后的结果仍是基向量 。基向量是表示向量最简洁的方法，向量在基向量的投影就是所在基向量的坐标，我们通过这种思想去理解特征值分解和推导SVD分解。

2. 特征值分解的含义

对称方阵A的特征值分解为：

$$A = U\Sigma U^{-1} \quad (2.1)$$

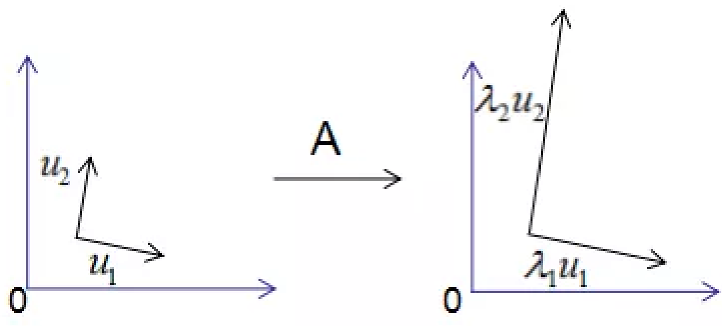
其中U是正交矩阵，Σ是对角矩阵。

为了可视化特征值分解，假设A是2×2的对称矩阵，$U=(u_1,u_2)$，$\Sigma=(\lambda_1,\lambda_2)$。（2.1）式展开为：

$$Au_1 = \lambda_1 u_1$$

$$Au_2 = \lambda_2 u_2$$

用图形表示为：



由上图可知，矩阵A没有旋转特征向量，它只是对特征向量进行了拉伸或缩短（取决于特征值的大小），因此，对称矩阵对其特征向量（基向量）的变换仍然是基向量（单位化）。

特征向量和特征值的几何意义：若向量经过矩阵变换后保持方向不变，只是进行长度上的伸缩，那么该向量是矩阵的特征向量，伸缩倍数是特征值。

3. SVD分解推导

我们考虑了当基向量是对称矩阵的特征向量时，矩阵变换后仍是基向量，但是，我们在实际项目中遇到的大都是行和列不相等的矩阵，如统计每个学生的科目乘积，行数为学生个数，列数为科目数，这种形成的矩阵很难是方阵，因此SVD分解是更普遍的矩阵分解方法 。

先回顾一下正交变换的思想：基向量正交变换后的结果仍是基向量 。

我们用正交变换的思想来推导SVD分解：

假设A是M*N的矩阵，秩为K，Rank(A)=k。

存在一组正交基V：

$$V = (v_1, v_2, \ldots, v_k)$$

矩阵对其变换后仍是正交基，记为U：

$$U = (Av_1, Av_2, \ldots, Av_k)$$

由正交基定义，得：

$$(Av_i)^T(Av_j)=0 \qquad (3.1)$$

上式展开：

$$v_i^T A^T A v_j = 0 \qquad (3.2)$$

当v_i是A^TA的特征向量时，有：

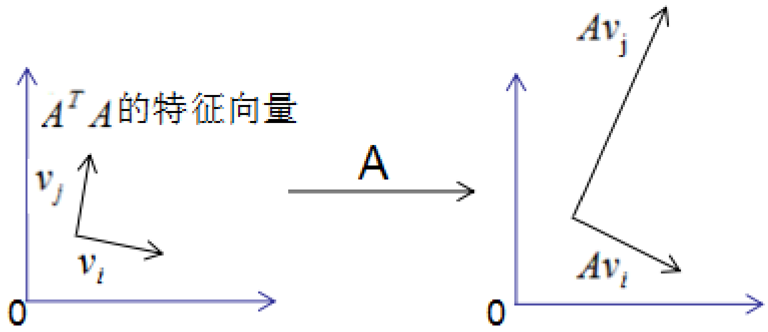
$$(A^TA)v_i = \lambda v_i$$

∴ （3.2）式得：

$$\lambda v_i^T v_j = 0$$

即假设成立 。

图形表示如下：



正交向量的模：

$$\| Av_i \|^2 = (Av_i)^T * (Av_i)$$
$$\Rightarrow \| Av_i \|^2 = v_i^T A^T A v$$
$$\Rightarrow \| Av_i \|^2 = \lambda_i v_i^T v = \lambda_i$$
$$\therefore \quad \| Av_i \| = \sqrt{\lambda_i}$$

单位化正交向量，得：

$$u_i = \frac{Av_i}{\| Av_i \|} = \frac{1}{\sqrt{\lambda_i}} Av_i$$
$$\Rightarrow Av_i = \sqrt{\lambda_i} * u_i \quad (3.3)$$

结论：当基向量是A^TA的特征向量时，矩阵A转换后的向量也是基向量 。

用矩阵的形式表示（3.3）式：

$$AV = U\Sigma \quad (3.4)$$

$$\text{其中}V=(v_1,v_2,...,v_k),\Sigma=\begin{pmatrix}\sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_k\end{pmatrix},U=(u_1,u_2,...,u_k)$$

$$\sigma_i=\sqrt{\lambda_i}$$

V是N*K矩阵，U是M*K矩阵，Σ 是M*K的矩阵，需要扩展成方阵形式：

将正交基$U=(u_1,u_2,...,u_k)$ 扩展$(u_1,u_2,...,u_m)R^m$ 空间的正交基，即U是M*M方阵 。

将正交基$V=(v_1,v_2,...,v_k)$ 扩展成$(v_1,v_2,...,v_n)R^n$ 空间的正交基，其中$(v_{k+1},v_{k+2},...,v_n)$ 是矩阵A的零空间，即：

$$Av_i=0\,,i>k$$

对应的特征值σ_i =0，Σ 是M*N对角矩阵,V是N*N方阵

因此（3.4）式写成向量形式为：

$$A(v_1,v_2,...,v_k\mid v_{k+1},v_{k+2},...,v_m)=(u_1,u_2,...,u_k\mid u_{k+1},u_{k+2},...,u_m)\begin{pmatrix}\sigma_1 & & & & & & \\ & \sigma_2 & & & & & \\ & & \ddots & & & & \\ & & & \sigma_k & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0\end{pmatrix}$$

得：

$$AV=U\Sigma$$

两式右乘V^T，可得矩阵的奇异值分解：

$$A=U\Sigma V^T \quad (3.5)$$

（3.5）式写成向量形式：

$$A=(u_1,u_2,\dots,u_k\mid v_{k+1},v_{k+2},\dots,v_m)\begin{pmatrix}\sigma_1 & & & & & \\ & \sigma_2 & & & & \\ & & \ddots & & & \\ & & & \sigma_k & & \\ & & & & 0 & \\ & & & & & \ddots \\ & & & & & & 0\end{pmatrix}\begin{pmatrix}v_1^T\\ \vdots\\ v_k^T\\ v_{k+1}^T\\ \vdots\\ v_n^T\end{pmatrix}$$

$$=(u_1,u_2,\dots,u_k)\begin{pmatrix}\sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_k\end{pmatrix}\begin{pmatrix}v_1^T\\ \vdots\\ v_k^T\end{pmatrix}$$

令:

$$X=(u_1,u_2,\dots,u_k)\begin{pmatrix}\sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_k\end{pmatrix}=(\sigma_1u_1,\sigma_2u_2,\dots,\sigma_ku_k)$$

$$Y=\begin{pmatrix}v_1^T\\ \vdots\\ v_k^T\end{pmatrix}$$

则:

$$A = XY$$

因为X和Y分别是列满秩和行满秩，所以上式是A的满秩分解。

（3.5）式的奇异矩阵Σ的值σ_i是A^TA特征值的平方根，下面推导奇异值分解的U和V:

$$\begin{aligned}A^TA&=(U\Sigma V^T)^T(U\Sigma V^T)\\&=V\Sigma U^TU\Sigma V^T\\&=V\Sigma^2V^T\end{aligned}$$

即V是A^TA的特征向量构成的矩阵，称为右奇异矩阵。

$$\begin{aligned}AA^T&=(U\Sigma V^T)(U\Sigma V^T)^T\\&=U\Sigma V^TV\Sigma U^T\\&=U\Sigma^2U^T\end{aligned}$$

即U是AA^T的特征向量构成的矩阵，称为左奇异矩阵 。

小结：矩阵A的奇异值分解：

$$A = U\Sigma V^T$$

其中U是AA^T的特征向量构成的矩阵，V是A^TA的特征向量构成的矩阵，奇异值矩阵Σ的值是A^TA特征值的平方根 。

3. 奇异值分解的例子

本节用一个简单的例子来说明矩阵是如何进行奇异值分解的。矩阵A定义为：

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 0 \end{pmatrix}$$

首先求出A^TA和AA^T：

$$A^TA = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

$$AA^T = \begin{pmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

求A^TA的特征向量V和特征值λ：

$$V = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}，对应的特征值：\lambda_1 = 3，\lambda_2 = 1$$

奇异值是特征值的平方根：$\sigma_1 = \sqrt{3}$， $\sigma_2 = 1$

求AA^T的特征向量U：

$$U = \begin{pmatrix} \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \\ \frac{2}{\sqrt{6}} & 0 & -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \end{pmatrix}$$

所以矩阵A的奇异值分解：

$$A=U\Sigma V^T=\begin{pmatrix}\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}}\\ \frac{2}{\sqrt{6}} & 0 & -\frac{1}{\sqrt{3}}\\ \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}}\end{pmatrix}\begin{pmatrix}\sqrt{3} & 0\\ 0 & 1\\ 0 & 0\end{pmatrix}\begin{pmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{pmatrix}$$

4. 行降维和列降维

本节通过协方差的角度去理解行降维和列降维，首先探讨下协方差的含义：

单个变量用方差描述，无偏方差公式：

$$D(x)=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

其中，n是样本数，$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$

两个变量用协方差描述，协方差公式：

$$\mathrm{cov}(x,y)=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

多个变量（如三个变量）之间的关系可以用协方差矩阵描述：

$$\mathrm{cov}(x,y,z)=\begin{pmatrix}\mathrm{cov}(x,x) & \mathrm{cov}(x,y) & \mathrm{cov}(x,z)\\ \mathrm{cov}(y,y) & \mathrm{cov}(y,x) & \mathrm{cov}(y,z)\\ \mathrm{cov}(z,x) & \mathrm{cov}(z,y) & \mathrm{cov}(z,z)\end{pmatrix}$$

相关系数公式：

$$\rho=\frac{\mathrm{cov}(x,y)}{\sqrt{D(x)}\sqrt{D(y)}}$$

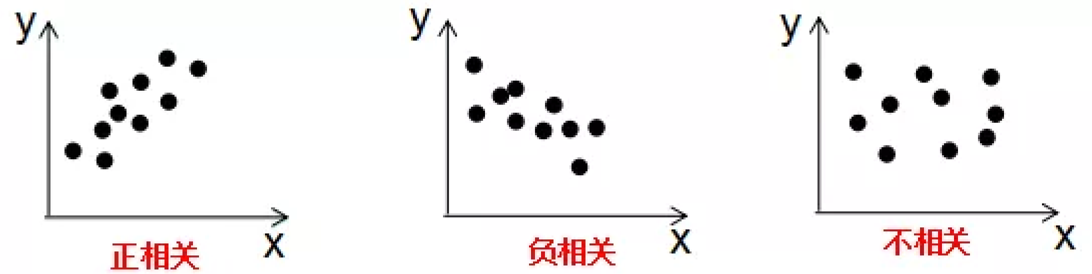
由上式可知，协方差是描述变量间的相关关系程度：

1）协方差cov(x,y) > 0时，变量x与y正相关；

2）协方差cov(x,y)<0时，变量x与y负相关；

3）协方差cov(x,y)=0时，变量x与y不相关；

变量与协方差关系的定性分析图：



现在开始讨论A^TA和AA^T的含义：

假设数据集是n维的，共有m个数据，每一行表示一例数据，即：

$$A=\begin{pmatrix}(x^{(1)})^T\\(x^{(2)})^T\\\vdots\\(x^{(m)})^T\end{pmatrix}$$

$x^{(i)}$表示第i个样本，x_j 表示第j维特征，$x_j^{(i)}$ 表示第i个样本的第j维特征 。

$$A^TA=(x^{(1)},x^{(2)},\dots,x^{(m)})\begin{pmatrix}(x^{(1)})^T\\(x^{(2)})^T\\\vdots\\(x^{(m)})^T\end{pmatrix}=x^{(1)}(x^{(1)})^T+x^{(1)}(x^{(1)})^T+\dots+x^{(m)}(x^{(m)})^T$$

$$\Rightarrow A^TA=\begin{pmatrix}\mathrm{cov}(x_1,x_1) & \mathrm{cov}(x_1,x_2) & \cdots & \mathrm{cov}(x_1,x_n)\\ \mathrm{cov}(x_2,x_1) & \mathrm{cov}(x_2,x_2) & \cdots & \mathrm{cov}(x_2,x_n)\\ \vdots & \vdots & & \vdots\\ \mathrm{cov}(x_n,x_1) & \mathrm{cov}(x_n,x_2) & \cdots & \mathrm{cov}(x_n,x_n)\end{pmatrix}$$

由上式可知，A^TA是描述各特征间相关关系的矩阵，所以A^TA的正交基V是以数据集的特征空间进行展开的。

数据集A在特征空间展开为：

$$X_{M*N}=A_{M*N}V_{N*N}\qquad(4.1)$$

由上一篇文章可知，特征值表示了A^TA在相应特征向量的信息分量。特征值越大，包含矩阵A^TA的信息分量亦越大。

若我们选择前r个特征值来表示原始数据集，数据集A在特征空间展开为：

$$X'_{M*r}=A_{M*N}V_{N*r}\qquad(4.2)$$

（4.2）式对列进行了降维，即右奇异矩阵V可以用于列数的压缩，与PCA降维算法一致。

行降维：

$$AA^T=\begin{pmatrix}(x^{(1)})^T\\(x^{(2)})^T\\\vdots\\(x^{(m)})^T\end{pmatrix}(x^{(1)},x^{(2)},\dots,x^{(m)})=\begin{pmatrix}(x^{(1)})^Tx^{(1)} & (x^{(1)})^Tx^{(2)} & \cdots & (x^{(1)})^Tx^{(m)}\\(x^{(2)})^Tx^{(1)} & (x^{(2)})^Tx^{(2)} & \cdots & (x^{(2)})^Tx^{(m)}\\\vdots & \vdots & & \vdots\\(x^{(m)})^Tx^{(1)} & (x^{(m)})^Tx^{(2)} & \cdots & (x^{(m)})^Tx^{(m)}\end{pmatrix}$$

$$\Rightarrow AA^T=\begin{pmatrix}\operatorname{cov}(x^{(1)},x^{(1)}) & \operatorname{cov}(x^{(1)},x^{(2)}) & \cdots & \operatorname{cov}(x^{(1)},x^{(m)})\\\operatorname{cov}(x^{(2)},x^{(1)}) & \operatorname{cov}(x^{(2)},x^{(2)}) & \cdots & \operatorname{cov}(x^{(2)},x^{(m)})\\\vdots & \vdots & \vdots & \\\operatorname{cov}(x^{(m)},x^{(1)}) & \operatorname{cov}(x^{(m)},x^{(2)}) & \cdots & \operatorname{cov}(x^{(m)},x^{(m)})\end{pmatrix}$$

由上式可知：AA^T 是描述样本数据间相关关系的矩阵，因此，左奇异矩阵U是以样本空间进行展开，原理与列降维一致，这里不详细介绍了 。

若我们选择前r个特征值来表示原始数据集，数据集A在样本空间展开为：

$$Y_{r*N}=U_{r*M}^T*A_{M*N}$$

因此，上式实现了行降维，即左奇异矩阵可以用于行数的压缩 。

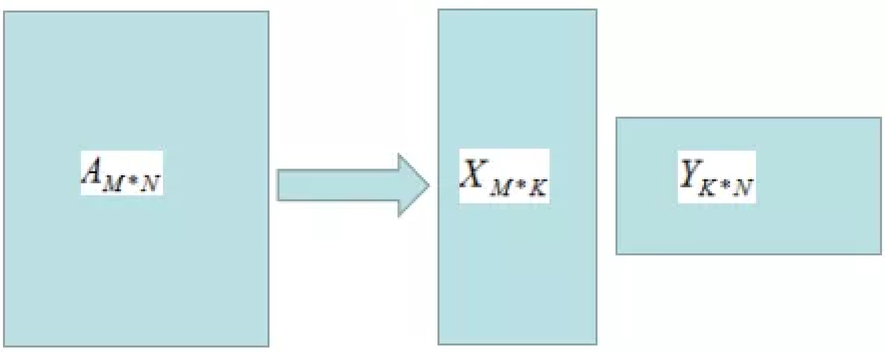
5. 数据压缩

本节介绍两种数据压缩方法：满秩分解和近似分解

矩阵A的秩为k，A的满秩分解：

$$A_{M*N}=X_{M*K}Y_{K*N}$$

满秩分解图形如下：



由上图可知，存储X和Y的矩阵比存储A矩阵占用的空间小，因此满秩分解起到了数据压缩作用。

若对数据再次进行压缩，需要用到矩阵的近似分解。

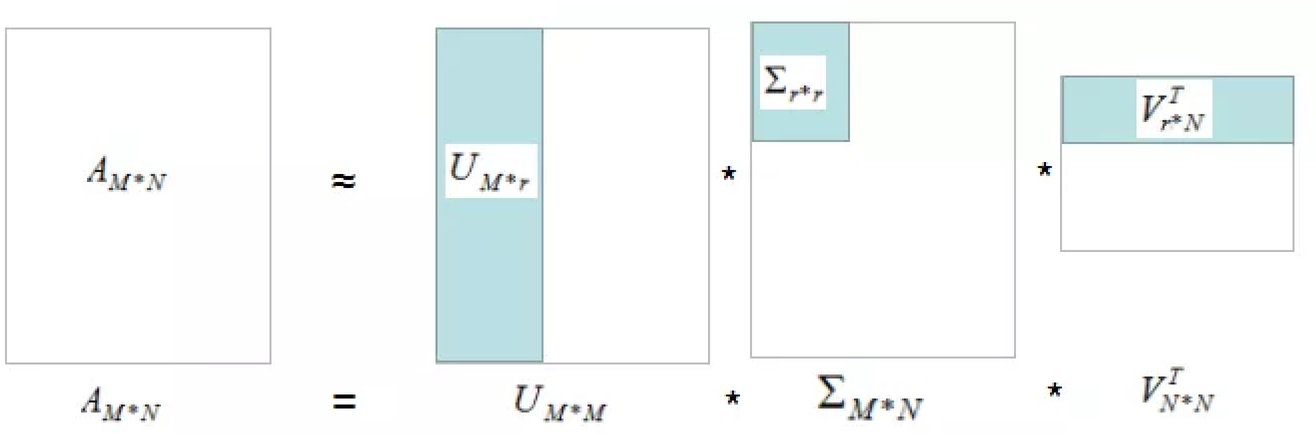
矩阵A的奇异值分解：

$$A_{M*N}=U_{M*M}\Sigma_{M*N}V^T_{N*N}$$

若我们选择前r个特征值近似矩阵A，得：

$$A_{M*N}\approx U_{M*r}\Sigma_{r*r}V^T_{r*N}$$

如下图：



我们用灰色部分的三个小矩阵近似表示矩阵A，存储空间大大的降低了。

6. SVD总结

任何矩阵都能进行SVD分解，SVD可以用于行降维和列降维，SVD在数据压缩、推荐系统和语义分析有广泛的应用，SVD与PCA的缺点一样，分解出的矩阵解释性不强 。

参考：

https://blog.csdn.net/zhongkejingwang/article/details/43053513

https://www.cnblogs.com/pinard/p/6251584.html

推荐阅读

主成分分析（PCA）原理总结