Singular Value Decomposition

本文的主要思路来源于Google研究院Jonathon Shlens的A Tutorial on Principle Component Analysis，值得一提的是，他是TensorFlow的作者之一，他在arxiv.org上的文章都很棒。

奇异值分解是线性代数中一个非常重要的技巧，本文从一个具体例子出发，介绍奇异值分解的推导、计算和应用，这个技巧在七日文后续系列中也很重要。

首先SVD很好记忆，对于任何矩阵A，都可以得到这样的形式：

$$A = U\Sigma V^T \quad (1)$$

其中Σ是一个对角阵，U和V都是正交矩阵，经过简单的变换，我们就可以得到这样的形式：

$$A^T = V\Sigma U^T \quad (2)$$

那么，下面的结论显而易见：

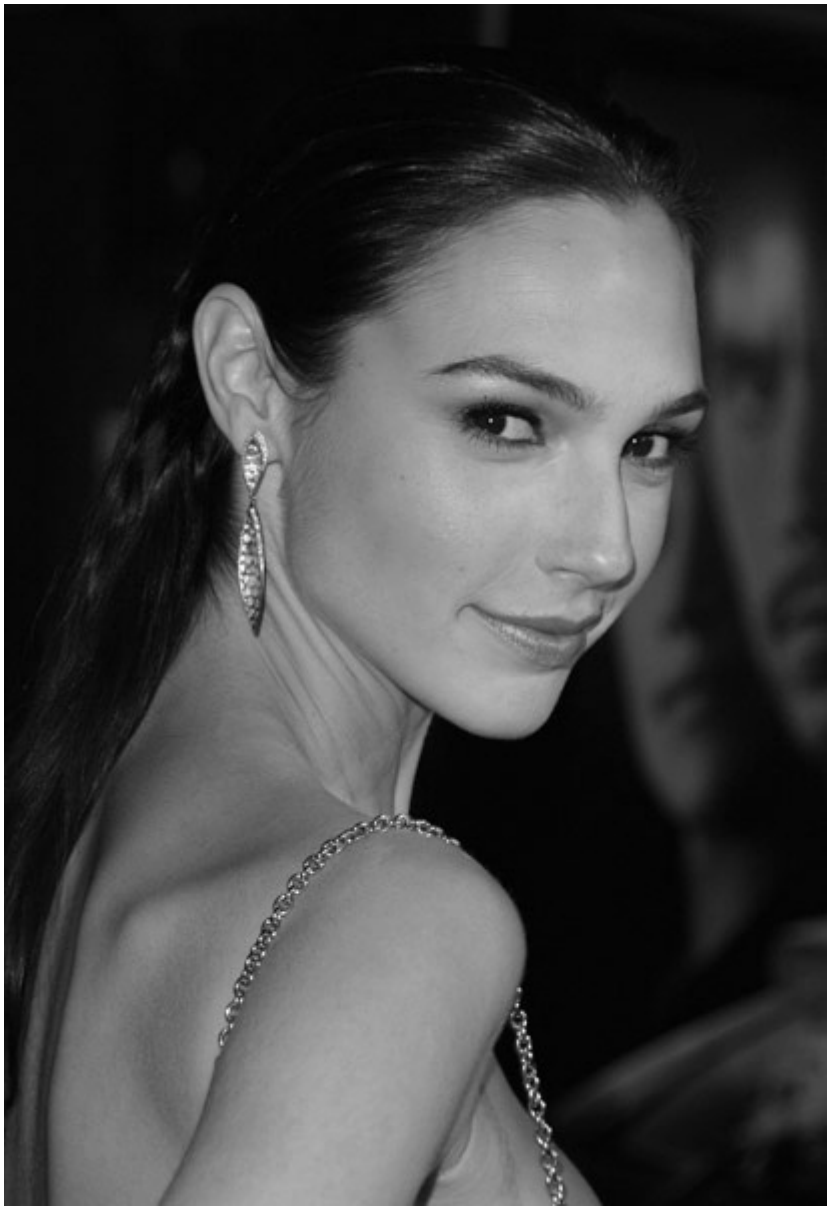
$$AA^T = U\Sigma^2 U^T \quad (3)$$
$$A^T A = V\Sigma^2 V^T \quad (4)$$

现在已知AA^T和$A^T A$都是对称矩阵，根据对称矩阵特征值分解的性质，我们就可以得到SVD的结论了：

> U是AA^T特征分解的结果，V是$A^T A$特征分解的结果。

SVD的具体应用

接下来，我们举例一个SVD的具体应用，有请神奇女侠—盖尔加朵：



我们知道，灰度图片可以被表示成矩阵的样子，我们将矩阵进行奇异值分解

$A = U\Sigma V^T$，进一步：

$$A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \ldots + \sigma_r u_r v_r^T$$

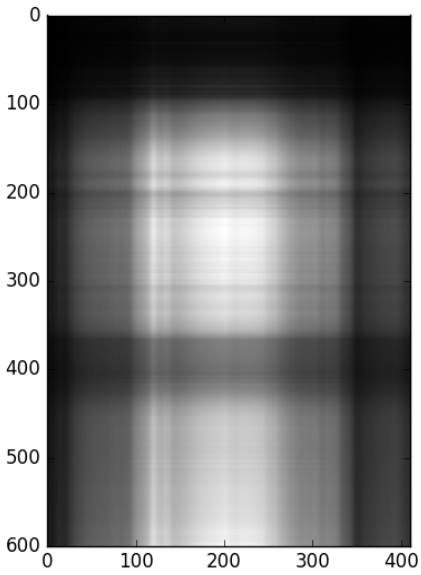
由于σ奇异值是从大到小排列的，而u和v是秩为1的向量，因此uv^T是秩为1的矩阵，我们只保留比较大的σ，试试将这样得到的图像和原图相比。

> u和v是秩为1的向量，因此uv^T是至少秩为1的矩阵。

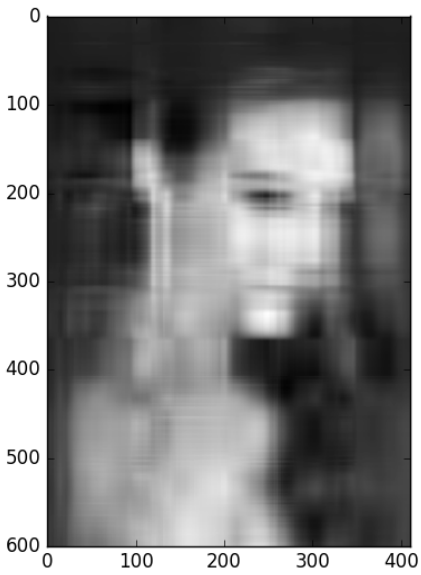
$$
\begin{gathered}
det(AB) = detA * detB \\
R(AB) \leq min(R(A), R(B)) \\
A = (a_{ij})_{n*m} \quad B = (b_{ij})_{m*s} \quad AB = C = (c_{ij})_{n*s} \\
a_{i1}B_1 + a_{i2}B_2 + \ldots + a_{im}B_m \\
= (a_{i1}b_{11} + a_{i2}b_{21} + \ldots + a_{im}b_{m1}, \ldots,) \\
= (c_{i1}, c_{i2,\ldots,}) = C_i
\end{gathered}
$$

因此C可以由B线性表示，因此$R(C) \leq R(B)$,同理可得$R(C) \leq R(A)$。

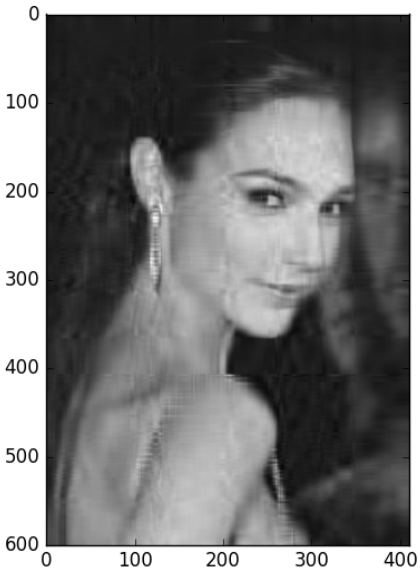
这是只取第一个特征值时的重建结果：



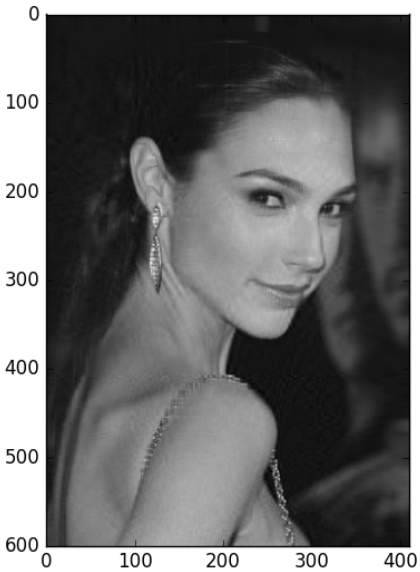
这是取前五个特征值时的重建结果：



这是取前二十个特征值时的重建结果：



这是取前五十个特征值时的重建结果：



可以看到，随着所保留特征值的增多，我们发现图像越来越清晰。

那么这里就可以发现SVD分解的一个良好的性质，对于这张$600*411$的图像，u是一个$600*1$的向量，v是一个$411*1$的向量，奇异值共有411个。也就是说存储一个σuv^T需要$600+411+1$的空间，假如我们可以在尽可能保持图像精度的情况下，减少奇异值的个数，那么我们就可以去掉较小的奇异值部分，节省空间。

SVD可以进行数据压缩，同时也可以进行数据去噪。如果一幅图像中包含噪声，那么我们有理由相信，较小的奇异值是由噪声引起的，通过设定这些奇异值为0，可以达到去噪的目的。

保留奇异值较大对应的特征向量部分，去掉奇异值较小对应的特征向量部分，是数据压缩和数据去噪的主要步骤。

> 奇异值往往对应着矩阵中隐藏的重要信息，而且重要性和奇异值大小正相关。每个矩阵A都可以表示为一系列秩为1的小矩阵之和，而奇异值则衡量了这些小矩阵对于A的权重。

上述程序所对应的代码是：

```
import numpy as np
import matplotlib.pyplot as plt
import matplotlib.image as mpimg

def rgb2gray(rgb):
    return np.dot(rgb[...,:3], [0.299, 0.587, 0.114])

#Convert color image into gray image
img = mpimg.imread('gal.jpg')
gray = rgb2gray(img)
#plt.imshow(gray, cmap = plt.get_cmap('gray'))
#plt.show()
plt.imsave('gray.jpg', gray, cmap =
plt.get_cmap('gray'))

#SVD
U, s, V = np.linalg.svd(gray, full_matrices = False)
print U.shape, V.shape, s.shape
#Change the number to test
s[50:] = 0
S = np.diag(s)
re_1 = np.dot(U, np.dot(S, V))
plt.imshow(re_1, cmap = plt.get_cmap('gray'))
plt.show()
```

那么SVD的轮子已经造好了，放在眼前等着我们用了，就是已知一个任意形状的矩阵A，都可以分解成三个矩阵。

这是经过先人们的努力得到的结果，我们应当想一下，这样的结果从何而来：

> *1*、实对称矩阵可以被它的特征向量对角化。（暗含：实对称矩阵有n个线性无关的实特征向量。）

> *2*、对于任何矩阵X，我们都可以得到$\boldsymbol{X^TX}$和$\boldsymbol{XX^T}$的特征值分解结果。

> 3、X可以分解成三个矩阵的形式。

后面是无聊的数学推导时间：

实对称矩阵可以被它的特征向量对角化

首先我们回顾下线性代数基础知识：

$$Av = \lambda v \quad (5)$$

那么我们就称v是矩阵A的特征向量，λ是矩阵A的特征值。这个等式说明了一件事：特征向量被施加线性变换A，只会使得向量伸长或者缩短，而其方向不发生变化。这也是特征向量和特征值在线性代数中这么重要的原因。

那么如何计算特征向量和特征值？

$$(A - \lambda I)v = 0 \quad (6)$$

上述等式成立，同时v不为零向量，说明$(A - \lambda I)x = 0$存在非零解，进而说明$A - \lambda I$为奇异矩阵(行或者列向量线性相关)，那么就可以得到：

（PS：行向量线性相关，会使得高斯行变换后的矩阵出现一行0，列向量线性相关，会使得高斯列变换后的矩阵出现一列0，这两个情况说明的是同一件事：化简为对角阵时，对角线元素上会出现0）

$$det(A - \lambda I) = 0 \quad (7)$$

求解上述方程可得λ，随后代入6式中即可以求解出对应的v。

> 另外，特征方程在复数范围内恒有解，其个数为方程的次数，因此n阶矩阵A有n个特征值。

> 假设特征方程求解得到重根，那么此m重根作为特征值最多对应了m个特征向量。即若λ的重数为k，如果是一般矩阵，那么特征向量的个数不大于特征值的重数，如果是可对角矩阵，那么特征向量的个数等于特征值的重数。也就是所谓的几何重数不超过代数重数。

> 不同特征值对应的特征向量不会相等，即一个特征向量只能属于一个特征值。

> 属于不同特征值的特征向量一定线性无关。

可以根据数学归纳法证明得到：

假设λ_m是矩阵A的不同特征值，而ξ_m是对应的特征向量，我们要证明ξ_m是线性无关的。

当m=1时，结果显然成立；

假设对于m-1时成立，那么对于m时，假设线性相关，那么就是$k_1, k_2, \ldots, k_m$不全为0，可得下式：

$$A\xi_i = \lambda_i \xi_i$$
$$k_1\xi_1 + k_2\xi_2 + \ldots + k_m\xi_m = 0$$

经过变换：

$$k_1\lambda_m\xi_1 + \ldots + k_m\lambda_m\xi_m = 0$$
$$A(k_1\xi_1 + \ldots + k_m\xi_m) = 0$$
$$k_1\lambda_1\xi_1 + \ldots + k_m\lambda_m\xi_m = 0$$
$$k_1(\lambda_m - \lambda_1)\xi_1 + \ldots + k_{m-1}(\lambda_m - \lambda_{m-1})\xi_{m-1} = 0$$

根据假设，$\xi_1, \ldots, \xi_{m-1}$线性无关：

则：

$$k_i(\lambda_m - \lambda_i) = 0$$

由于特征值不同，那么只能说明k_i为0，即$\xi_1, \ldots, \xi_m$线性无关。

> 如果存在满秩矩阵P，使得$B = P^{-1}AP$，则称A和B相似。相似矩阵具有相同的特征值。

> n阶矩阵A和对角矩阵$\Sigma = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$相似的充要条件是矩阵$A$有$n$个线性无关的特征向量（但是特征值可以存在相同的）。

对于任何矩阵$\mathbf{X}$，我们都可以得到X^TX和XX^T的特征值分解结果。

首先对应任意矩阵A，AA^T和A^TA都是对称矩阵，那么我们需要证明：

> 对称矩阵，可以被它的特征向量对角化，同时它的特征向量是正交向量。

证明如下：

首先假设一个大前提：对称矩阵具有n个线性无关的特征向量，同时具有n个实数根（可以存在重根）。

假设B是对称矩阵，E是B的特征向量构成的矩阵，D是一个对角阵，对角线上的元素是B矩阵的特征值，则有：

$$BE = ED \quad (8)$$

证明8式：

$$BE = [Be_1, Be_2, \ldots, Be_n] \quad (9)$$
$$ED = [\lambda_1 e_1, \lambda_2 e_2, \ldots, \lambda_n e_n] \quad (10)$$

而：

$$Be_m = \lambda_m e_m \quad (11)$$

而这正是特征向量和特征值的定义，因此得证！

如果E可逆，我们可以得到：$B = EDE^{-1}$

那么接下来我们要证明：对于对称矩阵来说，所有的特征向量不仅线性无关，而且正交，线性无关上面已经说明。

$$\begin{aligned}\lambda_1 e_1 \cdot e_2 &= (\lambda_1 e_1)^T e_2 \\ &= (Ae_1)^T e_2 \\ &= e_1^T A^T e_2 \\ &= e_1^T A e_2 \\ &= e_1^T (\lambda_2 e_2) \\ &= \lambda_2 e_1 \cdot e_2 \quad (12)\end{aligned}$$

进而我们得到：

$$\lambda_1 e_1 \cdot e_2 = \lambda_2 e_1 \cdot e_2 \quad (13)$$

对于不同的特征值，13式说明只能$e_1 \cdot e_2$为0，也就是说单位特征向量是正交的。

因此可以得到：$E^T = E^{-1}$

那么$B = EDE^T$

最后我们可以得到一个结论：

> 一个矩阵是对称的，当且仅当它可以被正交分解。

首先证明一个矩阵可以被正交分解，那么这个矩阵是对称的。

$$\begin{aligned}A &= EDE^T \\ A^T &= (EDE^T)^T = EDE^T = A\end{aligned}$$

得证！

一个矩阵是对称的，因此它可以被正交分解，上面已经证明。

当然以上的推导基于一个假设：即对称矩阵具有n个线性无关的特征向量和n个根（包括重根），接下来我们证明这个假设：

> 实对称矩阵的特征值恒为实数，从而它的特征向量都可以取为实向量。

$$A' = A \quad A_{\text{共轭}} = A \quad Aa = \lambda a \quad a \neq 0$$

$$(a^{共轭})'Aa=(a^{共轭})'A'a=(Aa^{共轭})'a=((Aa)^{共轭})'a'$$

$$\lambda(a^{共轭})'a=(\lambda^{共轭})(a^{共轭})'a$$

因为$a\neq 0$，则$\lambda=\lambda^{共轭}$，因此λ为实数。

> 设A为n阶对称矩阵，λ为A的特征方程的r重根，那么$(A-\lambda E)$的秩为$n-r$，从而特征值λ恰好有r个线性无关的特征向量，也就是说A有n个线性无关的特征向量。

仍然使用数学归纳法证明：

当A为一阶实对称矩阵时，$A=a,E=(1)$，$E^TAE=\lambda_1$，其中$\lambda_1=a$，结论成立。

假设对于A为n-1阶实对称矩阵成立。

那么当A为n阶实对称矩阵时：

第一步：构建一个正交矩阵M

假设a_1是属于A的特征值λ_1的一个单位特征向量，通过构造，选择n-1个非零向量$a_2,\dots,a_n$，设定正交矩阵$M=(a_1,\dots,a_n)$。

第二步：

$$\begin{aligned}AM&=A(a_1,\dots,a_n)\\&=(\lambda_1a_1,Aa_2,\dots,Aa_n)\end{aligned}$$

设$Aa_2=\beta_2$，则$AM=(\lambda_1a_1,\beta_2,\dots,\beta_n)$。

$$M^TAM=\begin{bmatrix}\lambda_1 & a_1^T\beta_2 & \dots & a_1^T\beta_n\\0 & a_2^T\beta_2 & \dots & a_2^T\beta_n\\\dots & \dots & \dots & \dots\end{bmatrix}$$

因为M^TAM是对称矩阵

$$(M^TAM)^T=M^TAM$$

因此$a_1^T\beta_2=0$，以此类推。

$$M^TAM=\begin{bmatrix}\lambda_1 & \\ & B_{n-1}\end{bmatrix}$$

同时B_{n-1}也应当是实对称矩阵。

由归纳假设可知，对于n-1阶实对称矩阵，存在n-1个线性无关的特征向量，因此得证，对于n阶实对称矩阵，存在n个线性无关的特征向量。

任意矩阵都可以被**SVD**分解

那么接下来我们要证明，任意矩阵确实可以表示为SVD分解的形式：

上述关于对称矩阵的结论引出了一个绝妙的性质，而这也是打开SVD大门的钥匙：

> 对于任何一个$m*n$的矩阵X，对称矩阵X^TX有一个正交向量集合$v_1, v_2, \dots, v_n$，以及对应的特征值$\lambda_1, \lambda_2, \dots, \lambda_n$，同时$Xv_1, Xv_2, \dots, Xv_n$形成了一组新的正交基底，且$Xv_i$的长度为$\sqrt{\lambda_i}$。

$$
\begin{aligned}
(Xv_i) \cdot (Xv_j) &= (Xv_i)^T (Xv_j) \\
&= v_i^T X^T X v_j \\
&= v_i^T (\lambda_j v_j) \\
&= \lambda_j v_i \cdot v_j
\end{aligned}
$$

由于v是正交的，因此

$$
(Xv_i) \cdot (Xv_j) = \begin{cases} \lambda_j & i = j \\ 0 & i \neq j \end{cases}
$$

这个证明引出了一个很棒的结果，我们将所有的特征值放在对角线组成一个主对角阵，由于V和U只包含r（矩阵的秩）个特征向量，我们再补齐m和n个正交向量基底，补充的基底对应特征值为0，所以事实上是无用的，就可以得到SVD分解：

$$
\begin{aligned}
Xv_i &= \sigma_i u_i \\
V &= [v_1, v_2, \dots, v_m] \\
U &= [u_1, u_2, \dots, u_n] \\
XV &= U\Sigma \\
X &= U\Sigma V^T
\end{aligned}
$$

最后我们还可以观察到一个性质：

X^TX和XX^T具有相同的特征值，这个可以直接推导出来。

$$
\begin{aligned}
X^T X v &= \lambda v \\
XX^T X v &= \lambda X v \\
XX^T (Xv) &= \lambda (Xv)
\end{aligned}
$$

我们最后得到结论，Xv就是XX^T的特征向量，这和SVD的结果相一致，这也解释了为什么XX^T和X^TX共享同样的特征值。

最后我们聊聊奇异值的几何性质：

$$
\begin{aligned}
Mv_1 &= \sigma_1 u_1 \\
Mv_2 &= \sigma_2 u_2
\end{aligned}
$$

v_1和v_2本身就是正交的，在经过M变化后，生成的u_1和u_2仍然是正交的，而奇异值代表这个向量的长度。

假设矩阵A的奇异值分解为：

$$A = [u_1 \quad u_2][\begin{matrix}3 & 0\\ 0 & 1\end{matrix}][\begin{matrix}v_1^T\\ v_2^T\end{matrix}]$$

假设：

$$x = \xi_1 v_1 + \xi_2 v_2$$
$$y = Ax = A[v_1 \quad v_2][\begin{matrix}\xi_1\\ \xi_2\end{matrix}] = 3\xi_1 u_1 + \xi_2 u_2$$
$$\eta_1 = 3\xi_1, \eta_2 = \xi_2$$

如果x是在单位圆$\xi_1^2 + \xi_2^2 = 1$，那么y就会在椭圆$\frac{\eta_1^2}{3^2} + \frac{\eta_2^2}{1^2} = 1$上，这表明矩阵A将单位圆变成了椭圆，而椭圆的半轴长正好是对应的奇异值。

推广到一般情况：一个矩阵A将单位球变换为超椭球面，那么矩阵A的每一个奇异值恰好就是超椭球的每条半轴长度。

> *SVD*分解的几何含义是：对于任何一个矩阵，我们都可以找到一组两两正交单位向量序列，使得矩阵作用在此向量序列上后得到新的向量序列依然保持正交关系，而奇异值的意义就是这组变换后的新的向量序列的长度。