奇异值分解(Singular Value Decompostion, SVD)是在机器学习领域广泛应用的算法，不光可以用于降维算法中的特征分解，还可以用于推荐系统，以及自然语言处理等领域，是很多机器学习算法的基石。本篇文章对SVD原理做主要讲解，在学习之前，确保你已经熟悉线性代数中的基本知识，包括特征值、特征向量、相似矩阵相关知识点。如果不太熟悉的话，推荐阅读如下两篇文章，[如何理解矩阵特征值？知乎马同学的回答]和[如何理解相似矩阵？马同学高等数学]，读完之后再看本篇文章会有很大帮助。

1. 回顾特征值和特征向量

我们首先回顾下特征值和特征向量的定义，如下所示。其中A是一个n×n的矩阵，x是一个n维向量，则我们说λ是矩阵A的一个特征值，x是矩阵A的特征值λ所对应的特征向量。但是求出特征值和特征向量有什么好处呢?

$$Ax = \lambda x$$

求出了矩阵A的n个特征值$\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$，以及这n个特征值所对应的特征向量$w_1, w_2, \ldots, w_n$，如果这n个特征值线性无关，那么矩阵A就可以用下式的特征分解表示

$$A = W\Sigma W^{-1}$$

其中Σ是以这n个特征值为主对角线的n×n维矩阵，W是这n个特征向量所组成的n×n维矩阵。一般我们会把W的这n个特征向量标准化，即满足$||w_i||_2 = 1$，或者说$w_i^T w_i = 1$，此时W的n个特征向量为标准正交基，满足$W^T W = I$，即$W^T = W^{-1}$，也就是说W为酉矩阵。这样我们的特征分解便可写作

$$A = W\Sigma W^T$$

上面矩阵能够进行特征分解，需要满足矩阵A必须为方阵。那么如果A不是方阵，即行和列不相同时，我们还可以进行矩阵分解吗?

2. 奇异值分解(SVD)

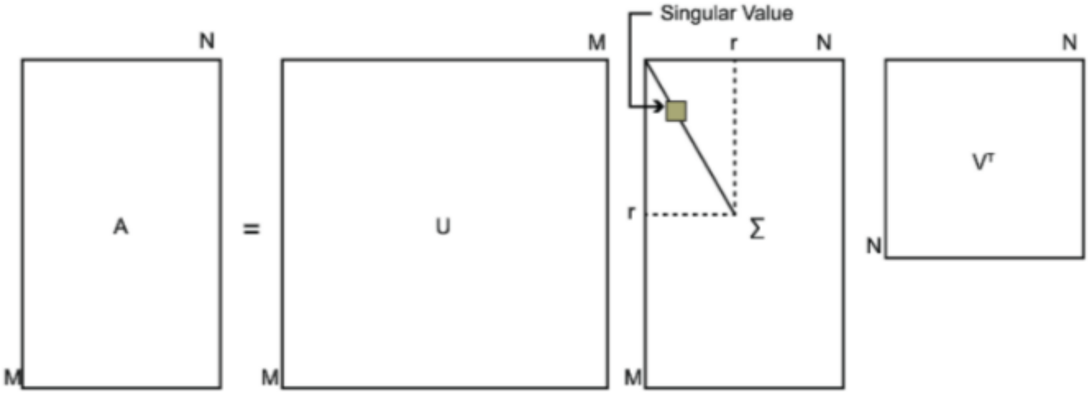
当矩阵A不是方阵时，可以用奇异值进行分解，假设我们的的矩阵A时一个m×n的矩阵，那么我们定义矩阵A的SVD为

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

其中U时一个m×m的矩阵，Σ是一个m×n的矩阵，除了主对角线上的元素以外全为0，主对角线上的每个元素都称为奇异值，V时一个n×n的矩阵。U和V都是酉矩阵，即满足

$$U^T U = I$$
$$V^T V = I$$

下图可以形象的表示出上述SVD的定义，但我们如何求出SVD分解后的U,Σ,V这三个矩阵呢?



如果我们将A的转置和A做矩阵乘法，那么会得到n×n的一个方阵A^TA。既然A^TA是方阵，那么我们就可以进行特征分解，得到的特征值和特征向量满足下式

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

这样我们就可以得到矩阵A^TA的n个特征值和对应的n个特征向量v。将A^TA的所有特征向量组成一个n×n的矩阵V，就是SVD公式里面的V矩阵，一般我们将V中的每个特征向量叫做A的右奇异向量。

同样，如果我们将A和A的转置做矩阵乘法，那么会得到m×m的一个方阵AA^T。因为AA^T是方阵，那么我们就可以进行特征分解，得到的特征值和特征向量满足下式

$$(AA^T)u_i = \lambda_i u_i$$

这样我们就可以得到矩阵AA^T的m个特征值和对应的m个特征向量u。将AA^T的所有特征向量组成一个m×m的矩阵U，就是SVD公式里面的U矩阵，一般我们将U中的每个特征向量叫做A的左奇异向量。

U和V都已经求出，现在只有奇异值矩阵Σ没有求出。由于Σ除了对角线上是奇异值，其他位置都是0，因此我们只需要求出每个奇异值σ就可以了。我们注意到

$$\begin{gathered} A = U\Sigma V^T \\ AV = U\Sigma V^TV \\ AV = U\Sigma \\ Av_i = u_i\sigma_i \\ \sigma_i = \frac{Av_i}{u_i} \end{gathered}$$

通过上式，我们便可以求出每个奇异值，进而求出奇异值矩阵Σ。

上面还有一个问题没有细讲，就是我们说A^TA的特征向量组成的就是SVD中的V矩阵，AA^T的特征向量就是SVD的U矩阵，为什么呢？下面我们做简短证明。

$$\begin{gathered} A = U\Sigma V^T \\ A^T = V\Sigma^T U^T \\ A^TA = V\Sigma^T U^T U\Sigma V^T = V\Sigma^2 V^T \end{gathered}$$

上式证明中使用到$U^TU = I, \Sigma^T\Sigma = \Sigma^2$，可以看出$A^TA$的特征向量，的确就是SVD中的V矩阵。类似的方法同样可以得到AA^T的特征向量，就是SVD中的U矩阵。进一步我们还可以看出特征值矩阵等于奇异值矩阵的平方，也就是特征值和奇异值满足如下关系

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

这也就是说，我们不仅可以通过用$\sigma_i = \frac{Av_i}{u_i}$来计算奇异值，也可以通过求出$A^TA$的特征值，然后取平方根得到奇异值。

3. SVD示例

下面我们通过一个简单例子来说明矩阵式如何进行奇异值分解的，假设矩阵A为

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

然后求出A^TA和AA^T

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

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

进而求出A^TA的特征值和特征向量

$$\lambda_1 = 3; v_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$$

$$\lambda_2 = 1; v_2 = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$$

求出AA^T的特征值和特征向量

$$\lambda_1 = 3; u_1 = \begin{pmatrix} \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{pmatrix}$$

$$\lambda_2 = 1; u_2 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ -\frac{1}{\sqrt{2}} \end{pmatrix}$$

$$\lambda_3 = 0; u_3 = \begin{pmatrix} \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{pmatrix}$$

利用$Av_i = \sigma_i u_i, i = 1, 2$求解奇异值

$$\begin{pmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} = \sigma_1 \begin{pmatrix} \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{pmatrix} \Rightarrow \sigma_1 = \sqrt{3}$$

$$\begin{pmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} = \sigma_2 \begin{pmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ -\frac{1}{\sqrt{2}} \end{pmatrix} \Rightarrow \sigma_2 = 1$$

当然，我们也可以利用$\sigma_i = \sqrt{\lambda_i}$直接求出奇异值$\sqrt{3}$和1。最终得到A的奇异值分解为

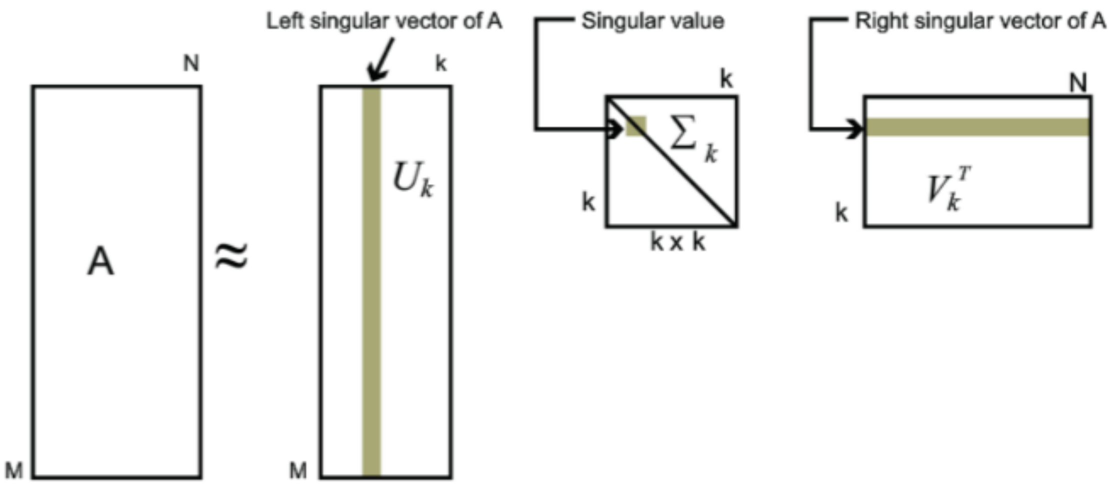
$$A = U\Sigma V^T = \begin{pmatrix} \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \\ \frac{2}{\sqrt{6}} & 0 & -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \end{pmatrix} \begin{pmatrix} \sqrt{3} & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix}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

4. SVD性质

对于SVD有哪些重要的性质值得我们注意呢? 对于奇异值，它跟特征分解中的特征值类似，在奇异值矩阵中也是按照从大到小排列，而且奇异值的减少特别的快，在很多情况下，前10%甚至1%的奇异值的和就占了全部奇异值之和的99%以上的比例。也就是说，我们可以用最大的k个奇异值和对应的左右奇异向量来近似描述矩阵，即

$$A_{m*n} = U_{m*m}\Sigma_{m*n}V_{n*n}^T \approx U_{m*k}\Sigma_{k*k}V_{k*n}^T$$

其中k要比n小很多，也就是一个大的矩阵A可以用三个小的矩阵$U_{m*k}, \Sigma_{k*k}, V_{k*n}^T$来表示。如下图所示，现在矩阵A只需要灰色部分的三个小矩阵就可以近似描述。



由于这个重要的性质，因此SVD可以用于PCA降维，用来做数据压缩和去噪。也可以用于推荐算法，将用户和喜好对应的矩阵做特征分解，进而得到隐含的用户需求来做推荐。同时也可以用于NLP之中的算法，比如潜在语义索引(LSI)。

5. SVD在PCA之中的应用

在[机器学习降维之主成分分析(PCA)](#)之中，我们讲到PCA降维时，需要找到样本协方差矩阵X^TX最大的d个特征向量，然后用着最大的d个特征向量组成的矩阵来做低维投影降维。可以看出，在这个过程中需要先求出协方差矩阵X^TX，当样本数多、样本特征数也多的时候，比如10000*10000的矩阵，这个计算量是很大的。

注意到SVD也可以求出协方差矩阵X^TX最大的d个特征向量组成的矩阵，但是SVD有个好处，就是可以不求出协方差矩阵X^TX，也能通过某些算法求出右奇异矩阵V，比如https://arxiv.org/abs/0909.4061。也就是说，PCA算法可以不用做特征分解，而是用SVD来进行完成。

另一方面，PCA仅仅使用了SVD的右奇异矩阵，没有使用左奇异矩阵，那么左奇异矩阵有什么用呢？假如我们的样本是m×n的矩阵X，如果通过SVD找到矩阵XX^T最大的d个特征向量组成的m×d的矩阵U，则我们进行如下处理

$$X'_{d*n} = U^T_{d*m} X_{m*n}$$

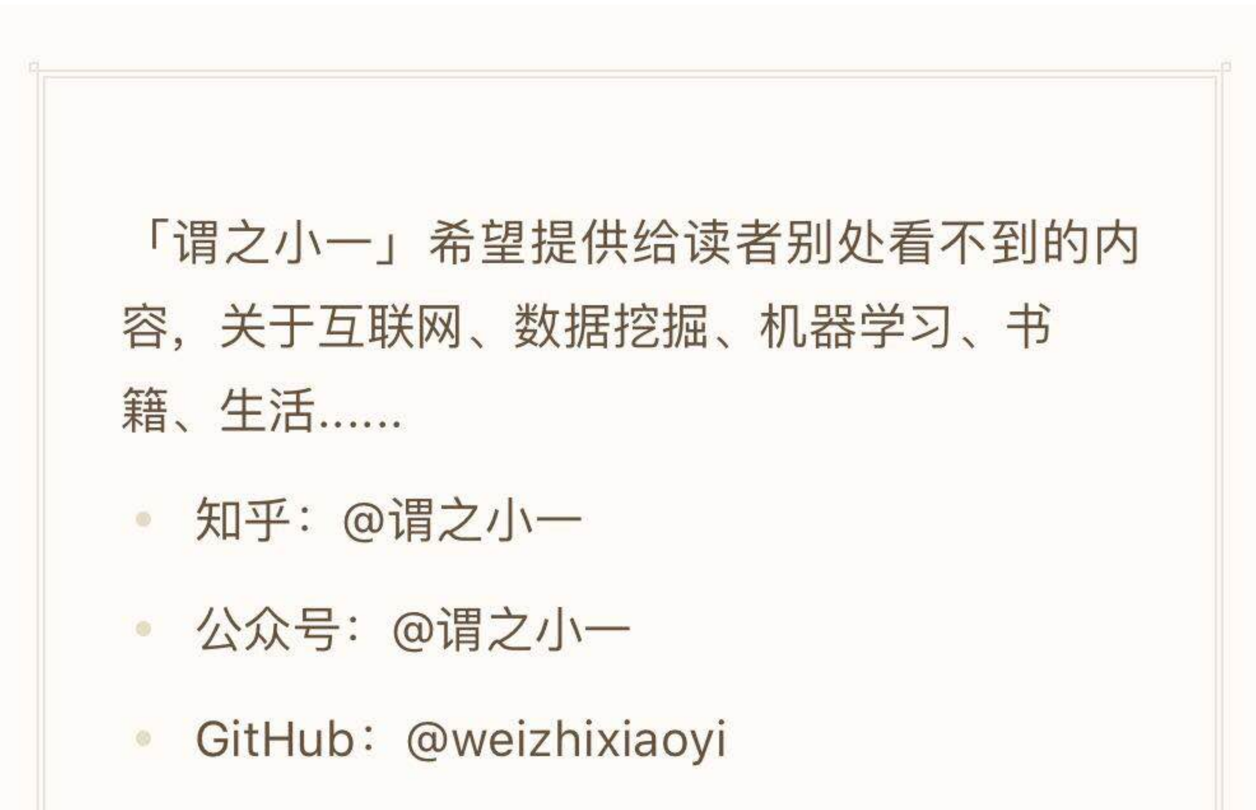
可以得到一个d×n的矩阵X'，这个矩阵和我们原来的m×n维样本矩阵X相比，行数从m减到了d，可见对行数进行了压缩。也就是说，左奇异矩阵可以用于行数的压缩，右奇异矩阵可以用于列数压缩，即特征降维。

6. SVD算法总结

SVD作为一个很基本的算法，在很多机器学习算法中都有它的身影，特别是在现在的大数据时代，由于SVD可以实现并行化，因此更是大展身手。当然，SVD的缺点是分解出的矩阵解释性往往不强，不过这不影响它的使用。

7. 推广

更多内容请关注公众号**谓之小一**，若有疑问可在公众号后台提问，随时回答，欢迎关注，内容转载请注明出处。



参考

> [刘建平Pinard-奇异值分解(SVD)原理转载降维中的应用]