奇異值分解(SVD)原理與在降維中的應用

其他 · 發表 2018-12-20

奇異值分解(Singular Value Decomposition，以下簡稱SVD)是在機器學習領域廣泛應用的演算法，它不光可以用於降維演算法中的特徵分解，還可以用於推薦系統，以及自然語言處理等領域。是很多機器學習演算法的基石。本文就對SVD的原理做一個總結，並討論在在PCA降維演算法中是如何運用運用SVD的。

1. 回顧特徵值和特徵向量

我們首先回顧下特徵值和特徵向量的定義如下：

$$Ax=\lambda x$$

其中A是一個n×n的矩陣，x是一個n維向量，則我們說λ是矩陣A的一個特徵值，而x是矩陣A的特徵值λ所對應的特徵向量。

求出特徵值和特徵向量有什麼好處呢？ 就是我們可以將矩陣A特徵分解。如果我們求出了矩陣A的n個特徵值$\lambda_1\leq\lambda_2\leq\ldots\leq\lambda_n$,以及這n個特徵值所對應的特徵向量$\{w_1,w_2,\ldots w_n\}$，，如果這n個特徵向量線性無關，那麼矩陣A就可以用下式的特徵分解表示：

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

其中W是這n個特徵向量所張成的n×n維矩陣，而Σ為這n個特徵值為主對角線的n×n維矩陣。

一般我們會把W的這n個特徵向量標準化，即滿足$||w_i||_2=1$, 或者說$w_i^Tw_i=1$，此時W的n個特徵向量為標準正交基，滿足$W^TW=I$，即$W^T=W^{-1}$, 也就是說W為酉矩陣。

這樣我們的特徵分解表示式可以寫成

A=WΣWT

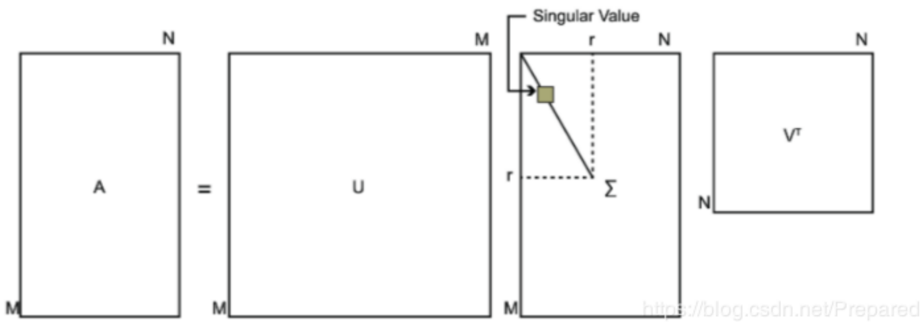
注意到要進行特徵分解，矩陣A必須為方陣。那麼如果A不是方陣，即行和列不相同時，我們還可以對矩陣進行分解嗎？答案是可以，此時我們的SVD登場了。

2. SVD的定義

SVD也是對矩陣進行分解，但是和特徵分解不同，SVD並不要求要分解的矩陣為方陣。假設我們的矩陣A是一個m×n的矩陣，那麼我們定義矩陣A的SVD為：

A=UΣVT

其中U是一個m×m的矩陣，Σ是一個m×n的矩陣，除了主對角線上的元素以外全為0，主對角線上的每個元素都稱為奇異值，V是一個n×n的矩陣。U和V都是酉矩陣，即滿足UTU=I,VTV=I。下圖可以很形象的看出上面SVD的定義：



那麼我們如何求出SVD分解後的U,Σ,V這三個矩陣呢？

如果我們將A的轉置和A做矩陣乘法，那麼會得到n×n的一個方陣ATA。既然ATA是方陣，那麼我們就可以進行特徵分解，得到的特徵值和特徵向量滿足下式：

(ATA)vi=λivi

這樣我們就可以得到矩陣ATA的n個特徵值和對應的n個特徵向量v了。將ATA的所有特徵向量張成一個n×n的矩陣V，就是我們SVD公式裡面的V矩陣了。一般我們將V中的每個特徵向量叫做A的右奇異向量。

如果我們將A和A的轉置做矩陣乘法，那麼會得到m×m的一個方陣AAT。既然AAT是方陣，那麼我們就可以進行特徵分解，得到的特徵值和特徵向量滿足下式：

(AAT)ui=λiui

這樣我們就可以得到矩陣AAT的m個特徵值和對應的m個特徵向量u了。將AAT的所有特徵向量張成一個m×m的矩陣U，就是我們SVD公式裡面的U矩陣了。一般我們將U中的每個特徵向量叫做A的左奇異向量。

U和V我們都求出來了，現在就剩下奇異值矩陣Σ沒有求出了。由於Σ除了對角線上是奇異值其他位置都是0，那我們只需要求出每個奇異值σ就可以了。

我們注意到:

A=UΣVT⇒AV=UΣVTV⇒AV=UΣ⇒Avi=σiui⇒σi=Avi/ui

這樣我們可以求出我們的每個奇異值，進而求出奇異值矩陣Σ。

上面還有一個問題沒有講，就是我們說ATA的特徵向量組成的就是我們SVD中的V矩陣，而AAT的特徵向量組成的就是我們SVD中的U矩陣，這有什麼根據嗎？這個其實很容易證明，我們以V矩陣的證明為例。

A=UΣVT⇒AT=VΣTUT⇒ATA=VΣTUTUΣVT=VΣ2VT

上式證明使用了:UTU=I,ΣTΣ=Σ2。可以看出ATA的特徵向量組成的的確就是我們SVD中的V矩陣。類似的方法可以得到AAT的特徵向量組成的就是我們SVD中的U矩陣。

進一步我們還可以看出我們的特徵值矩陣等於奇異值矩陣的平方，也就是說特徵值和奇異值滿足如下關係：

σi=λi−−√

這樣也就是說，我們可以不用σi=Avi/ui來計算奇異值，也可以通過求出ATA的特徵值取平方根來求奇異值。

3. SVD計算舉例

裡我們用一個簡單的例子來說明矩陣是如何進行奇異值分解的。我們的矩陣A定義為：

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

我們首先求出ATA和AAT

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

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

進而求出ATA的特徵值和特徵向量：

$$\lambda_1 = 3; v_1 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}; \lambda_2 = 1; v_2 = \begin{pmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$$

接著求AAT的特徵值和特徵向量：

$$\lambda_1 = 3; u_1 = \begin{pmatrix} 1/\sqrt{6} \\ 2/\sqrt{6} \\ 1/\sqrt{6} \end{pmatrix}; \lambda_2 = 1; u_2 = \begin{pmatrix} 1/\sqrt{2} \\ 0 \\ -1/\sqrt{2} \end{pmatrix}; \lambda_3 = 0; u_3 = \begin{pmatrix} 1/\sqrt{3} \\ -1/\sqrt{3} \\ 1/\sqrt{3} \end{pmatrix}$$

利用Avi=σiui,i=1,2求奇異值：

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

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

當然，我們也可以用σi=λi−−√直接求出奇異值為3–√和1.

最終得到A的奇異值分解為：

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

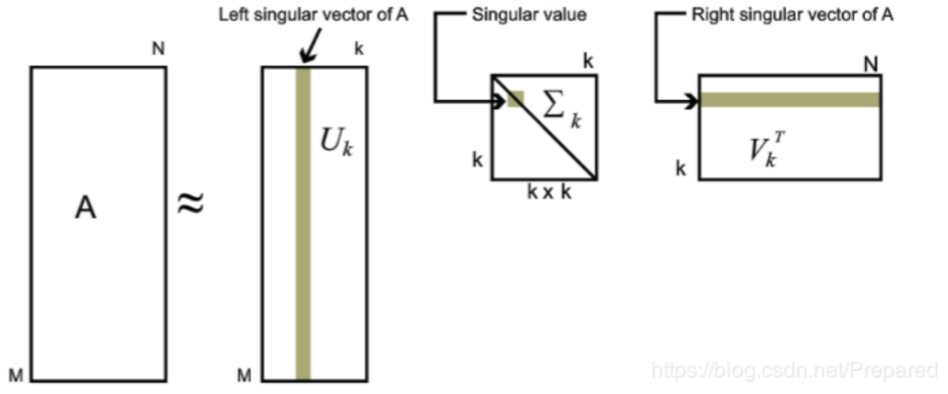
4. SVD的一些性質

上面幾節我們對SVD的定義和計算做了詳細的描述，似乎看不出我們費這麼大的力氣做SVD有什麼好處。那麼SVD有什麼重要的性質值得我們注意呢？

對於奇異值,它跟我們特徵分解中的特徵值類似，在奇異值矩陣中也是按照從大到小排列，而且奇異值的減少特別的快，在很多情況下，前10%甚至1%的奇異值的和就佔了全部的奇異值之和的99%以上的比例。也就是說，我們也可以用最大的k個的奇異值和對應的左右奇異向量來近似描述矩陣。也就是說：

Am×n=Um×mΣm×nVTn×n≈Um×kΣk×kVTk×n

其中k要比n小很多，也就是一個大的矩陣A可以用三個小的矩陣Um×k,Σk×k,VTk×n來表示。如下圖所示，現在我們的矩陣A只需要灰色的部分的三個小矩陣就可以近似描述了。



由於這個重要的性質，SVD可以用於PCA降維，來做資料壓縮和去噪。也可以用於推薦演算法，將使用者和喜好對應的矩陣做特徵分解，進而得到隱含的使用者需求來做推薦。同時也可以用於NLP中的演算法，比如潛在語義索引（LSI）。下面我們就對SVD用於PCA降維做一個介紹。

5. SVD用於PCA

在主成分分析（PCA）原理總結中，我們講到要用PCA降維，需要找到樣本協方差矩陣XTX的最大的d個特徵向量，然後用這最大的d個特徵向量張成的矩陣來做低維投影降維。可以看出，在這個過程中需要先求出協方差矩陣XTX，當樣本數多樣本特徵數也多的時候，這個計算量是很大的。

注意到我們的SVD也可以得到協方差矩陣XTX最大的d個特徵向量張成的矩陣，但是SVD有個好處，有一些SVD的實現演算法可

以不求先求出協方差矩陣XTX，也能求出我們的右奇異矩陣V。也就是說，我們的PCA演算法可以不用做特徵分解，而是做SVD來完成。這個方法在樣本量很大的時候很有效。實際上，scikit-learn的PCA演算法的背後真正的實現就是用的SVD，而不是我們我們認為的暴力特徵分解。

另一方面，注意到PCA僅僅使用了我們SVD的右奇異矩陣，沒有使用左奇異矩陣，那麼左奇異矩陣有什麼用呢？

假設我們的樣本是m×n的矩陣X，如果我們通過SVD找到了矩陣XXT最大的d個特徵向量張成的m×d維矩陣U，則我們如果進行如下處理：

X′d×n=UTd×mXm×n

可以得到一個d×n的矩陣X‘,這個矩陣和我們原來的m×n維樣本矩陣X相比，行數從m減到了k，可見對行數進行了壓縮。也就是說，左奇異矩陣可以用於行數的壓縮。相對的，右奇異矩陣可以用於列數即特徵維度的壓縮，也就是我們的PCA降維。

6. SVD小結

SVD作為一個很基本的演算法，在很多機器學習演算法中都有它的身影，特別是在現在的大資料時代，由於SVD可以實現並行化，因此更是大展身手。SVD的原理不難，只要有基本的線性代數知識就可以理解，實現也很簡單因此值得仔細的研究。當然，SVD的缺點是分解出的矩陣解釋性往往不強，有點黑盒子的味道，不過這不影響它的使用。

標籤：

您可能也會喜歡…

奇異值分解(SVD)原理與在降維中的應用
奇異值分解(SVD)原理詳解及推導（轉）
機器學習實戰（Machine Learning in Action）學習筆記————10.奇異值分解(SVD)原理、基於協同過濾的推薦引擎、數據降維
機器學習實戰（Machine Learning in Action）學習筆記————10.奇異值分解(SVD)原理、基於協同過濾的推薦引擎、資料降維
奇異值分解(SVD)原理詳解及推導
奇異值分解（SVD）原理
奇異值分解(SVD)原理詳解
奇異值分解(SVD)原理及應用
差分陣列原理與其字首和的應用
B樹和B+樹原理及在索引中的應用

奇異值分解SVD計算原理及JAVA程式碼

矩陣的奇異值分解（SVD）（理論）

奇異值的物理意義是什麼？強大的矩陣奇異值分解(SVD)及其應用

[數學] 奇異值分解SVD的理解與應用

數學基礎系列(六)----特徵值分解和奇異值分解(SVD)

首頁
Python教學