

刘建平Pinard

十年码农，对数学统计学，数据挖掘，机器学习，大数据平台，大数据平台应用开发，大数据可视化感兴趣。

博客园 首页 新随笔 联系 订阅 管理

奇异值分解(SVD)原理与在降维中的应用

奇异值分解(Singular Value Decomposition，以下简称SVD)是在机器学习领域广泛应用的算法，它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。本文就对SVD的原理做一个总结，并讨论在PCA降维算法中是如何运用运用SVD的。

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：

$Ax = \lambda x$

其中A是一个 $n \times n$ 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，而 x 是矩阵A的特征值 λ 所对应的特征向量。

求出特征值和特征向量有什么好处呢？就是我们可以将矩阵A特征分解。如果我们求出了矩阵A的 n 个特征值 $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$ ，以及这 n 个特征值所对应的特征向量 $\{w_1, w_2, \dots, w_n\}$ ，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W是这 n 个特征向量所张成的 $n \times n$ 维矩阵，而 Σ 为这 n 个特征值为主对角线的 $n \times n$ 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表达式可以写成

$A = W \Sigma W^T$

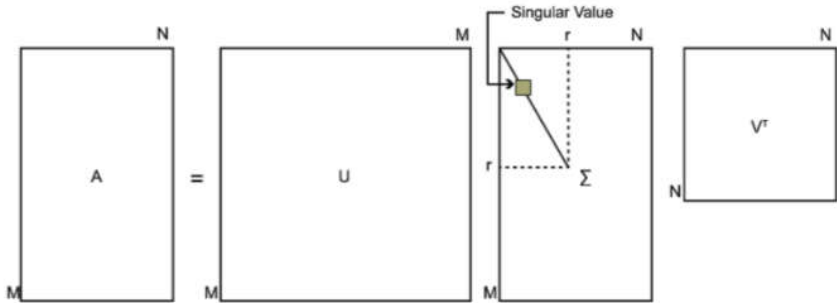
注意到要进行特征分解，矩阵A必须为方阵。那么如果A不是方阵，即行和列不相同，我们还可以对矩阵进行分解吗？答案是可以，此时我们的SVD登场了。

2. SVD的定义

SVD也是对矩阵进行分解，但是和特征分解不同，SVD并不要求要分解的矩阵为方阵。假设我们的矩阵A是一个 $m \times n$ 的矩阵，那么我们定义矩阵A的SVD为：

$A = U \Sigma V^T$

其中U是一个 $m \times m$ 的矩阵， Σ 是一个 $m \times n$ 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是一个 $n \times n$ 的矩阵。U和V都是酉矩阵，即满足 $U^T U = I, V^T V = I$ 。下图可以很形象的看出上面SVD的定义：



那么我们如何求出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 $n \times n$ 的一个方阵 $A^T A$ 。既然 $A^T A$ 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：

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

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

公告

★珠江追梦，饮岭南茶，恋鄂北家★
昵称：刘建平Pinard
园龄：7个月
粉丝：220
关注：12
+加关注

2017年6月						
<	日	一	二	三	四	五
	28	29	30	31	1	2
	4	5	6	7	8	9
	11	12	13	14	15	16
	18	19	20	21	22	23
	25	26	27	28	29	30
	2	3	4	5	6	7

常用链接

我的随笔
我的评论
我的参与
最新评论
我的标签

随笔分类(91)

- 0040. 数学统计学(4)
- 0081. 机器学习(62)
- 0082. 深度学习(10)
- 0083. 自然语言处理(13)
- 0121. 大数据挖掘(1)
- 0122. 大数据平台(1)
- 0123. 大数据可视化

随笔档案(91)

- 2017年6月 (2)
- 2017年5月 (7)
- 2017年4月 (5)
- 2017年3月 (10)
- 2017年2月 (7)
- 2017年1月 (13)
- 2016年12月 (17)
- 2016年11月 (22)
- 2016年10月 (8)

常去的机器学习网站

52 NLP
Analytics Vidhya
机器学习库
机器学习路线图
深度学习进阶书

如果我们将A和A的转置做矩阵乘法，那么会得到 $m \times m$ 的一个方阵 AA^T 。既然 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：

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

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 了。将 AA^T 的所有特征向量张成一个 $m \times m$ 的矩阵U，就是我们SVD公式里面的U矩阵了。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我们都求出来了，现在就剩下奇异值矩阵Σ没有求出了。由于Σ除了对角线上是奇异值其他位置都是0，那我们只需要求出每个奇异值σ就可以了。

我们注意到：

$$A = U\Sigma V^T \Rightarrow AV = U\Sigma V^T V \Rightarrow AV = U\Sigma \Rightarrow Av_i = \sigma_i u_i \Rightarrow \sigma_i = Av_i / u_i$$

这样我们可以求出我们的每个奇异值，进而求出奇异值矩阵Σ。

上面还有一个问题没有讲，就是我们说 $A^T A$ 的特征向量组成的就是我们SVD中的V矩阵，而 AA^T 的特征向量组成的就是我们SVD中的U矩阵，这有什么根据吗？这个其实很容易证明，我们以V矩阵的证明为例。

$$A = U\Sigma V^T \Rightarrow A^T = V\Sigma U^T \Rightarrow A^T A = V\Sigma U^T U \Sigma V^T = V\Sigma^2 V^T$$

上式证明使用了： $U^T U = I, \Sigma^T = \Sigma$ 。可以看出 $A^T A$ 的特征向量组成的就是我们SVD中的V矩阵。类似的方法可以得到 AA^T 的特征向量组成的就是我们SVD中的U矩阵。

进一步我们还可以看出我们的特征值矩阵等于奇异值矩阵的平方，也就是说特征值和奇异值满足如下关系：

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

这样也就是说，我们可以不用 $\sigma_i = Av_i / u_i$ 来计算奇异值，也可以通过求出 $A^T A$ 的特征值取平方根来求奇异值。

3. SVD计算举例

这里我们用简单的例子来说明矩阵是如何进行奇异值分解的。我们的矩阵A定义为：

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

我们首先求出 $A^T A$ 和 AA^T

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

进而求出 $A^T A$ 的特征值和特征向量：

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

接着求 AA^T 的特征值和特征向量：

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

利用 $Av_i = \sigma_i u_i, i = 1, 2$ 求奇异值：

$$\begin{aligned} \begin{pmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} &= \sigma_1 \begin{pmatrix} 1/\sqrt{6} \\ 2/\sqrt{6} \\ 1/\sqrt{6} \end{pmatrix} \Rightarrow \sigma_1 = \sqrt{3} \\ \begin{pmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} &= \sigma_2 \begin{pmatrix} 1/\sqrt{2} \\ 0 \\ -1/\sqrt{2} \end{pmatrix} \Rightarrow \sigma_2 = 1 \end{aligned}$$

当然，我们也可以用 $\sigma_i = \sqrt{\lambda_i}$ 直接求出奇异值为 $\sqrt{3}$ 和1。

最终得到A的奇异值分解为：

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

深度学习入门书

积分与排名

积分 - 119247
排名 - 2030

阅读排行榜

- 1. scikit-learn决策树算法类库使用小结(5162)
- 2. 梯度提升树(GBDT)原理小结(4641)
- 3. scikit-learn随机森林调参小结(4339)
- 4. 用scikit-learn和pandas学习线性回归(3648)
- 5. 用scikit-learn学习K-Means聚类(3370)

评论排行榜

- 1. 线性回归原理小结(18)
- 2. scikit-learn随机森林调参小结(18)
- 3. 谱聚类 (spectral clustering) 原理总结(12)
- 4. 集成学习之Adaboost算法原理小结(10)
- 5. BIRCH聚类算法原理(9)

推荐排行榜

- 1. 机器学习研究与开发平台的选择(6)
- 2. 支持向量机原理(五)线性支持回归(5)
- 3. scikit-learn决策树算法类库使用小结(5)
- 4. 协同过滤推荐算法总结(5)
- 5. 支持向量机高斯核调参小结(5)

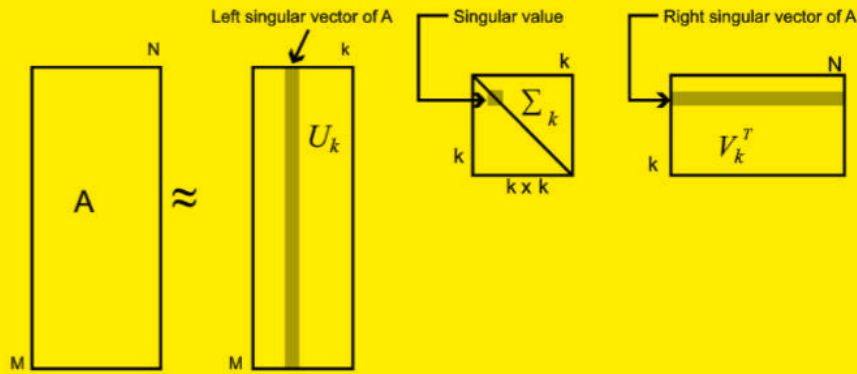
4. SVD的一些性质

上面几节我们对SVD的定义和计算做了详细的描述，似乎看不出我们费这么大的力气做SVD有什么好处。那么SVD有什么重要的性质值得我们注意呢？

对于奇异值,它跟我们特征分解中的特征值类似,在奇异值矩阵中也是按照大到小排列,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的奇异值之和的99%以上的比例。也就是说,我们也可以用最大的k个的奇异值和对应的左右奇异向量来近似描述矩阵。也就是说:

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

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 $U_{m \times k}$, $\Sigma_{k \times k}$, $V_{k \times n}^T$ 来表示。如下图所示,现在我们的矩阵A只需要灰色的部分的三个小矩阵就可以近似描述了。



由于这个重要的性质, SVD可以用于PCA降维,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中的算法,比如潜在语义索引(LSI)。下面我们就对SVD用于PCA降维做一个介绍。

5. SVD用于PCA

在主成分分析(PCA)原理总结中,我们讲到要用PCA降维,需要找到样本协方差矩阵 $X^T X$ 的最大的d个特征向量,然后用这最大的d个特征向量张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 $X^T X$,当样本数多样本特征数也多的时候,这个计算量是很大的。

注意到我们的SVD也可以得到协方差矩阵 $X^T X$ 最大的d个特征向量张成的矩阵,但是SVD有个好处,有一些SVD的实现算法可以不先求出协方差矩阵 $X^T X$,也能求出我们的右奇异矩阵 V 。也就是说,我们的PCA算法可以不用做特征分解,而是做SVD来完成。这个方法在样本量很大的时候很有效。实际上,scikit-learn的PCA算法的背后真正的实现就是用的SVD,而不是我们我们认为是暴力特征分解。

另一方面,注意到PCA仅仅使用了我们的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?

假设我们的样本是 $m \times n$ 的矩阵X,如果我们通过SVD找到了矩阵 $X X^T$ 最大的d个特征向量张成的 $m \times d$ 维矩阵U,那么我们如果进行如下处理:

$$X'_{d \times n} = U_{d \times m}^T X_{m \times n}$$

可以得到一个 $d \times n$ 的矩阵X',这个矩阵和我们原来的 $m \times n$ 维样本矩阵X相比,行数从m减到了k,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。相对的,右奇异矩阵可以用于列数即特征维度的压缩,也就是我们的PCA降维。

6. SVD小结

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。SVD的原理不难,只要有基本的线性代数知识就可以理解,实现也很简单因此值得仔细的研究。当然,SVD的缺点是分解出的矩阵解释性往往不强,有点黑盒子的味道,不过这不影响它的使用。

(欢迎转载,转载请注明出处。欢迎沟通交流: pinard.liu@ericsson.com)

分类: 0081. 机器学习

标签: 维度规约



刘建平Pinard
关注 - 12
粉丝 - 220
+加关注

计算过程:
1. 先对目标信息进行SVD分解,获得分解后的“左奇异矩阵,奇异值矩阵,右奇异矩阵”
2. 根据奇异值矩阵中的奇异值大小,找出最大的n个奇异值对应的n个行号
3. 到左、右奇异矩阵中找到n个行号对应的列与行,由这些左、右奇异矩阵,奇异值矩阵重建原始信息矩阵
4. 由于 $n < k$,因此i. 数据被压缩; ii. 对噪声进行了去噪

左、右奇异矩阵:
右奇异矩阵用于列信息(特征维度)的压缩,由k压缩到n
左奇异矩阵用于行数的压缩,由k压缩到n

« 上一篇: [用scikit-learn进行LDA降维](#)

» 下一篇: [局部线性嵌入\(LLE\)原理总结](#)

posted @ 2017-01-05 15:44 刘建平Pinard 阅读(1330) 评论(0) 编辑 收藏

[刷新评论](#) [刷新页面](#) [返回顶部](#)

注册用户登录后才能发表评论，请 [登录](#) 或 [注册](#)，[访问网站首页](#)。

【推荐】50万行VC++源码：大型组态工控、电力仿真CAD与GIS源码库



最新IT新闻:

- Apple Pay将推个人转账功能：信用卡需3%的手续费
 - 微软分享Xbox One向后兼容性新数据
 - 雅虎股东批准44.8亿美元出售核心互联网业务 股价大涨10%
 - 庆祝《Pokémon GO》一周年 7月22日芝加哥将举行真人线下活动
 - watchOS 4 beta首个上手视频公布
- » [更多新闻...](#)



最新知识库文章:

- 小printf的故事：真正的程序员？
 - 程序员的工作、学习与绩效
 - 软件开发为什么很难
 - 唱吧DevOps的落地，微服务CI/CD的范本技术解读
 - 程序员，如何从平庸走向理想？
- » [更多知识库文章...](#)

Copyright ©2017 刘建平Pinard