PCA 与 SVD 笔记

李向阳　d1142845997@gmail.com

目录

1 引入

PCA 与 SVD 是一种降维技术, 在很多领域都有重要应用, 因此五花八门的介绍也很多. 为了建立自己的体系, 这里稍微介绍一下, 也方便回顾.

SVD(Singular Value Decomposition), 也就是奇异值分解, 严格的数学介绍是在大四的矩阵分析课上, 而 PCA(Principal Component Analysis), 即主成分分析, 是在大三的数据分析课上学的, 是偏统计的角度, 对细节也都做了证明. 因此, 这篇笔记会略去一些证明, 着重对关系进行梳理.

2 SVD

SVD 本身就具有丰富的内容, 详细可见维基 https://en.wikipedia.org/wiki/Singular_value_decomposition.

3 PCA

3.1 统计中的 PCA

统计中经常用 n 表示样本数, 用 p 表示特征数 (维数), 这里我们仍延续机器学习系列笔记的记号, 用 m 表示样本数, 用 n 表示特征数.

设总体为 n 维随机向量 $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^T$, 主成分分析是构造原变量的一系列线性组合 $Y_1, Y_2, \cdots$, 使其方差达到最大 (具体可回顾数据分析方法课本). 实际中, 是从总体中抽样得到了 m 个样本, 其中 $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{in})^T, \boldsymbol{x}_i \in \mathbb{R}^n, i = 1, 2, \cdots, m$, 把样本按行并起来得到了 $m \times n$ 的观测数据矩阵

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_m^T \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}$$

为了后面讨论的方便, 我们要求 $\boldsymbol{X}$ 是去中心化后的矩阵 (即每一列已经减去了列均值), 即 $\sum_{i=1}^m x_{i,j} = 0, j = 1, 2, \cdots, n$, 事实上可写为 $\sum_{i=1}^m \boldsymbol{x}_i = \boldsymbol{0}$, 这样样本的协方差阵为 (为了推导形式的方便, 分母取为 m 也可, 本身二者都可以)

$$\boldsymbol{S} = \frac{1}{m-1} \sum_{i=1}^m \boldsymbol{x}_i \boldsymbol{x}_i^T = \frac{1}{m-1} \boldsymbol{X}^T \boldsymbol{X}$$

数据分析课上讲过, 设 $\boldsymbol{S}$ 为样本的协方差矩阵 (半正定矩阵), 其特征值按大小顺序排列为 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n \geqslant 0$, 相应的正交单位化特征向量为 $\boldsymbol{e}_1, \cdots, \boldsymbol{e}_n$, 则总体的第 j 个主成分可表示为

$$y_j = \boldsymbol{e}_j^T \boldsymbol{x} = e_{j1}x_1 + e_{j2}x_2 + \cdots + e_{jn}x_n, j = 1, 2, \cdots, n$$

其中 $\boldsymbol{e}_j = (e_{j1}, e_{j2}, \cdots, e_{jn})^T$. 把第 i 个样本值 $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{in})^T$ 代入, 便得到该样本在第 j 个主成分上的得分 y_{ij}, 如下表所示

表 1: 原始数据及主成分得分

样本序号	原始变量				主成分			
	X_1	X_2	$\cdots$	X_n	Y_1	Y_2	$\cdots$	Y_n
1	x_{11}	x_{12}	$\cdots$	x_{1n}	y_{11}	y_{12}	$\cdots$	y_{1n}
2	x_{21}	x_{22}	$\cdots$	x_{2n}	y_{21}	y_{22}	$\cdots$	y_{2n}
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$
m	x_{m1}	x_{m2}	$\cdots$	x_{mn}	y_{m1}	y_{m2}	$\cdots$	y_{mn}

一般我们会选取前 $k(k < n)$ 个主成分, 用前 k 个主成分的得分替代原始数据做分析, 这样便达到了数据降维的目的. 其中第 j 个主成分的贡献率为 $\lambda_j / \sum_{i=1}^{n} \lambda_i$, 前 k 个主成分的累计贡献率为 $\sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{n} \lambda_i$.

用矩阵形式表达, 即首先对于样本协方差阵 S 对角化, 有

$$\boldsymbol{S} = \boldsymbol{Q}\ \mathrm{diag}(\lambda_1, \cdots, \lambda_n)\ \boldsymbol{Q}^T$$

其中 $\boldsymbol{Q} = (\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_n)$. 然后可得新的数据矩阵 $\boldsymbol{Y} = (y_{ij})$ 为

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{Q}$$

若取前 k 个主成分, 即令 $\boldsymbol{Q}_k = (\boldsymbol{e}_1, \cdots, \boldsymbol{e}_k)$, 降维矩阵 $\boldsymbol{Y}_k$ 为

$$\boldsymbol{Y}_k = \boldsymbol{X}\boldsymbol{Q}_k$$

3.2 利用 SVD 做 PCA

关于 SVD 与 PCA 的关系, 可参考 http://stats.stackexchange.com/questions/134282/relationship-between-svd-and-pca-how-to-use-svd-to-perform-pca, 介绍的比较清楚.

这里说一下利用 SVD 做 PCA.

原始数据矩阵 $\boldsymbol{X}$ 是一个 $m \times n$ 矩阵, 对其做奇异值分解

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$$

其中 $\boldsymbol{U}=(\boldsymbol{u}_1,\cdots,\boldsymbol{u}_m)$ 为 $m\times m$ 阶正交阵, $\boldsymbol{V}=(\boldsymbol{v}_1,\cdots,\boldsymbol{v}_n)$ 为 $n\times n$ 阶正交阵, $\boldsymbol{D}$ 为一个对角阵 (实际上为 $m\times n$), 其对角元为矩阵 $\boldsymbol{X}$ 的奇异值 σ_i, 也就是 $\boldsymbol{X}^T\boldsymbol{X}$ 特征值的平方根. 于是可得

$$\boldsymbol{X}^T\boldsymbol{X}=\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^T$$

这里 $\boldsymbol{D}^2=\mathrm{diag}(\sigma_1^2,\cdots,\sigma_n^2)$ 表示 n 阶对角阵. 注意到样本协方差阵为 $\boldsymbol{S}=\frac{1}{m-1}\boldsymbol{X}^T\boldsymbol{X}$, 因此

$$\boldsymbol{S}=\boldsymbol{V}\frac{\boldsymbol{D}^2}{m-1}\boldsymbol{V}^T$$

这不正好是协方差阵 $\boldsymbol{S}$ 的对角化吗? 其中 $\boldsymbol{S}$ 的特征值为 $\lambda_i=\sigma_i^2/(m-1)$. 那么根据上面主成分得分的求法, 新的得分矩阵为

$$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{V}=\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T\boldsymbol{V}=\boldsymbol{U}\boldsymbol{D}$$

同样的, 若取前 $k(k<n)$ 个主成分, 即令 $\boldsymbol{V}_k$ 为矩阵 $\boldsymbol{V}$ 的前 k 列, $\boldsymbol{U}_k$ 为矩阵 $\boldsymbol{U}$ 的前 k 列, $\boldsymbol{D}_{mk}$ 为矩阵 $\boldsymbol{D}$ 的前 k 列, $\boldsymbol{D}_k$ 为矩阵 $\boldsymbol{D}$ 的左上 $k\times k$ 对角阵, 可得降维得分矩阵为

$$\boldsymbol{Y}_k=\boldsymbol{X}\boldsymbol{V}_k=\boldsymbol{U}\boldsymbol{D}_{mk}=\boldsymbol{U}_k\boldsymbol{D}_k$$

注意用分块矩阵可以去理解证明上式, $\boldsymbol{U}_k=(\boldsymbol{u}_1,\cdots,\boldsymbol{u}_k)$, $\boldsymbol{V}_k=(\boldsymbol{v}_1,\cdots,\boldsymbol{v}_k)$.

$$\begin{aligned}\boldsymbol{Y}_k&=\boldsymbol{X}\boldsymbol{V}_k=\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T\boldsymbol{V}_k\\&=\boldsymbol{U}\boldsymbol{D}\begin{pmatrix}\boldsymbol{v}_1^T\\\boldsymbol{v}_2^T\\\vdots\\\boldsymbol{v}_n^T\end{pmatrix}(\boldsymbol{v}_1,\cdots,\boldsymbol{v}_k)=\boldsymbol{U}\boldsymbol{D}\begin{pmatrix}\boldsymbol{I}_k&\boldsymbol{0}\\\boldsymbol{0}&\boldsymbol{0}\end{pmatrix}\\&=\boldsymbol{U}\boldsymbol{D}_{mk}=\boldsymbol{U}_k\boldsymbol{D}_k\end{aligned}$$

注 3.1. 若令 $\boldsymbol{X}_k=\boldsymbol{U}_k\boldsymbol{D}_k\boldsymbol{V}_k^T$, 即

$$\boldsymbol{X}_k=\sum_{j=1}^{k}\sigma_j\boldsymbol{u}_j\boldsymbol{v}_j^T$$

则 $\boldsymbol{X}_k$ 仍是一个 $m\times n$ 矩阵, 只不过它的秩为 $k<n$, 在矩阵分析的课上我们证明 $\boldsymbol{X}_k$ 是 $\boldsymbol{X}$ 的一个最佳低秩逼近, 证明过程也可见 *http://stats.stackexchange.com/questions/130721/what-norm-of-the-reconstruction-error-is-minimized*

$\boldsymbol{X}_k$ 也可看做是利用前 k 个主成分对矩阵 $\boldsymbol{X}$ 的重建.

4 Kernel PCA

4.1 PCA 重述

PCA 实际上是将样本点在几个互相正交的方向上进行投影来达到降维的目的. 投影方向就是协方差阵 $\boldsymbol{S}$ 的特征向量的方向.

设 $\boldsymbol{Sv} = \lambda \boldsymbol{v}$, 将特征向量 $\boldsymbol{v}$ 按照特征值降序排列, 然后将样本点投影在特征向量上. 协方差阵 $\boldsymbol{S}$ 的特征向量其实可表示为样本点的线性组合. 事实上

$$\lambda \boldsymbol{v} = \boldsymbol{S}\boldsymbol{v} = \frac{1}{m-1}\sum_{i=1}^{m} \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{v} = \frac{1}{m-1}\sum_{i=1}^{m} (\boldsymbol{x}_i^T \boldsymbol{v})\boldsymbol{x}_i = \frac{1}{m-1}\sum_{i=1}^{m} \langle \boldsymbol{x}_i, \boldsymbol{v}\rangle \boldsymbol{x}_i$$

其中 $\boldsymbol{x}_i^T \boldsymbol{v}$ 是一个数, 因此用了欧式内积表示, 于是可得

$$\boldsymbol{v} = \frac{1}{\lambda(m-1)}\sum_{i=1}^{m} \langle \boldsymbol{x}_i, \boldsymbol{v}\rangle \boldsymbol{x}_i = \sum_{i=1}^{m} \alpha_i \boldsymbol{x}_i = \boldsymbol{X}^T \boldsymbol{\alpha}$$

如何寻求系数 $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_m)^T$ 呢?

再次利用 $\lambda \boldsymbol{v} = \boldsymbol{S}\boldsymbol{v}$, 可得

$$(m-1)\lambda \boldsymbol{v} = \sum_{i=1}^{m} \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{v}, \Rightarrow \mu \boldsymbol{v} = \sum_{i=1}^{m} \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{v}$$

也即

$$\mu \sum_{i=1}^{m} \alpha_i \boldsymbol{x}_i = \sum_{i=1}^{m} \left(\boldsymbol{x}_i \boldsymbol{x}_i^T \cdot \sum_{j=1}^{m} \alpha_j \boldsymbol{x}_j \right)$$

变形可得

$$\sum_{i=1}^{m} \mu \alpha_i \cdot \boldsymbol{x}_i = \sum_{i=1}^{m} \left(\sum_{j=1}^{m} \alpha_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \right) \cdot \boldsymbol{x}_i$$

于是, 只需下式成立即可

$$\mu \alpha_i = \sum_{j=1}^{m} \alpha_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle, i = 1, 2, \cdots, m$$

因此 $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_m)^T$ 满足 $\boldsymbol{K}\boldsymbol{\alpha} = \mu \boldsymbol{\alpha}$, 即 $\boldsymbol{\alpha}$ 是矩阵 $\boldsymbol{K}$ 的特征向量, 其中 $K \in \mathbb{R}^{m \times m}$ 为样本的内积矩阵

$$\boldsymbol{K} = \begin{pmatrix} \langle \boldsymbol{x}_1, \boldsymbol{x}_1 \rangle & \langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle & \cdots & \langle \boldsymbol{x}_1, \boldsymbol{x}_m \rangle \\ \langle \boldsymbol{x}_2, \boldsymbol{x}_1 \rangle & \langle \boldsymbol{x}_2, \boldsymbol{x}_2 \rangle & \cdots & \langle \boldsymbol{x}_2, \boldsymbol{x}_m \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \boldsymbol{x}_m, \boldsymbol{x}_1 \rangle & \langle \boldsymbol{x}_m, \boldsymbol{x}_2 \rangle & \cdots & \langle \boldsymbol{x}_m, \boldsymbol{x}_m \rangle \end{pmatrix}$$

当然, 此时的矩阵 $\boldsymbol{K}$ 还可表示为 $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{X}^T$.

4.2 核方法

考虑先将样本数据映射到另外一个空间, $\phi : \mathcal{X} \to \mathcal{H}, \boldsymbol{x} \mapsto \phi(\boldsymbol{x})$, 然后在新的空间中对样本点 $\phi(\boldsymbol{x}_1), \cdots, \phi(\boldsymbol{x}_m)$ 做 PCA.

假设数据已经中心化, 即 $\sum_{i=1}^{m} \phi(\boldsymbol{x}_i) = \boldsymbol{0}$, 如何将映射后的数据中心化后面再提. 此时样本的协方差为

$$\boldsymbol{S} = \frac{1}{m} \sum_{i=1}^{m} \phi(\boldsymbol{x}_i) \phi(\boldsymbol{x}_i)^T$$

接下来按照 PCA 的步骤是一样的, 只需要把上面的 $\boldsymbol{x}_i$ 换为 $\phi(\boldsymbol{x}_i)$ 即可. 关键是计算样本的内积矩阵 $\boldsymbol{K}$, 其 (i, j) 元为 $\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j)$.

如同 SVM 一样, 在映射后的空间中计算内积是复杂的, 我们通过构造核函数 $\kappa(\cdot)$ 来计算这个内积 (相关回顾可看 SVM 笔记), 即有

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j)$$

核函数的选取跟 SVM 一样, 也有很多种选择, 比如多项式核、高斯核等等.

5 总结

5.1 参考资料

(1) 博客: http://www.cnblogs.com/LeftNotEasy/archive/2011/01/19/svd-and-applications.html, 对奇异值分解的强大应用做了通俗介绍.

(2) JerryLead 的博客: http://www.cnblogs.com/jerrylead/archive/2011/04/18/2020216.html, 对主成分分析做了不同解释.

(3) PRML: 关于 Kernel PCA 的推导介绍, 可见 12.3 节.

附录