小锅的机器学习笔记–降维

补一下数学

奇异值分解

设 A 为一个 $m \times n$ 的矩阵，则 A 一定可以进行奇异值分解:

$$A_{m\times n} = U_{m\times m}\Sigma_{m\times n}V^T_{n\times n} \tag{1}$$

公式中 $U_{m\times m}, V_{n\times n}$ 分别是 m 阶，n 阶的正交矩阵，$\Sigma_{m\times n}$ 是 $m \times n$ 的矩形对角矩阵，对角线的值均非负，且从左到右降序排列。

证明:

设 $m \geq n$，$m \leq n$ 同理: 因为 A^TA 是 n 阶实对称矩阵，所有 A^TA 一定可以相似对角化，且特征值均为实数，即存在一个 n 阶正交矩阵 V 使：

$$A^TA = V\Lambda V^T \qquad \Lambda = diag(\lambda_1, \ldots, \lambda_n) \tag{2}$$

λ_i 是 A^TA 的特征值，设 v_i 是 λ_i 对应的特征向量，那么有:

$$\begin{aligned} ||Av_i||^2 &= (Av_i)^T Av_i \\ &= v_i^T A^T A v_i \\ &= v_i^T \lambda_i v_i \\ &= \lambda_i v_i^T v_i \\ &= \lambda_i ||v_i||^2 \end{aligned}$$

所以有 $\lambda_i \geq 0$：

$$\lambda_i = \frac{||Av_i||^2}{||v_i||^2} \geq 0 \tag{3}$$

让 Λ 对角线及特征值按降序排列, 并且取 $\sigma_i = \sqrt{\lambda_i}$，也就是:

$$\lambda_1 \geq \lambda_2 \ldots\ldots \lambda_n \geq 0 \tag{4}$$

设 $r(A) = r$，由于 $r(A) = r(A^TA)$, 且 A^TA 的特征值全部为正，所以 $\lambda_{r+1}, \lambda_{r+2}, \ldots,$ 均为 0。所以有:

$$\begin{cases} \sigma_i \geq 0 \quad if \quad i \leq r \\ \sigma_i = 0 \quad if \quad i > r \end{cases} \tag{5}$$

我们取 $V = [V_1, V_2]$，V_1, V_2 满足下式：

$$V_1 = [v_1, v_2, \ldots, v_r] \qquad V_2 = [v_{r+1}, v_{r+2}, \ldots, v_n] \tag{6}$$

式中 v_i 为 λ_i 所对应的特征向量。我们令:

$$\Sigma_1 = \begin{bmatrix} \sigma_i & & & \\ & \sigma_2 & & \\ & & \cdots & \\ & & & \sigma_r \end{bmatrix}_{r \times r} \tag{7}$$

我们令奇异值分解公式中的 $m \times n$ 对角矩阵 Σ 为:

$$\Sigma_{m \times n} = \begin{bmatrix} \Sigma_1 & 0_{r \times (n-r)} \\ 0_{(m-r) \times r} & 0_{(m-r) \times (n-r)} \end{bmatrix} \tag{8}$$

因为 $u_{r+1}, u_{r+2}, \ldots, un$ 对应的特征值都是 0，所以有:

$$A^T A V_2 = 0 \tag{9}$$

又因为 $Ax = 0$ 与 $A^T Ax = 0$ 为同解方程，所以有:

$$AV_2 = 0 \tag{10}$$

所以有:

$$\begin{aligned} A &= AE \\ &= AVV^T \\ &= A\left[V_1, V_2\right] \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} \\ &= A\left[V_1 V_1^T + V_2 V_2^T\right] \\ &= AV_1 V_1^T \end{aligned}$$

我们令:

$$u_i = \frac{1}{\sigma_i} A v_i \qquad i = 1, 2, \ldots, r \tag{11}$$

$$U_1 = [u_1, u_2, \ldots, u_r] \tag{12}$$

则有：

$$AU_1 = \Sigma_1 V_1 \tag{13}$$

因为:

$$\begin{aligned} u_i^T u_j &= \frac{1}{\sigma_i \sigma_j} v_i^T A^T A v_j \\ &= \frac{\lambda_j}{\sigma_i \sigma_j} v_i^T v_j \\ &= \frac{\lambda_j}{\sigma_i \sigma_j} v_i^T v_j \\ &= \frac{\sigma_j}{\sigma_i} v_i^T v_j \\ &= \begin{cases} 0 & if \quad i \not= j \\ 1 & if \quad i == j \end{cases} \end{aligned}$$

所以 $(u_1, u_2, \ldots, u_r)$ 是相互正交。
观察奇次方程:

$$U_1^T x = 0 \tag{14}$$

我们可以知道这个奇次方程有 $m-r(U_1) = m-r$ 个基础解系，我们对着 $m-r$ 基础解系做史密斯正交化，记正交化的后的 $m-r$ 个基础解系为 $(u_{r+1}, u_{r+2}, \ldots, u_m)$，它们均与 U_1 中的每个列向量 $u_1, u_2, \ldots, u_r$ 正交，且它们之间相互正交，记 $U_2 = [u_{r+1}, u_{r+2}, \ldots, u_m]$。
令:

$$U = [U_1, U_2] \tag{15}$$

U 的列向量相互正交，所以 $U^T U = E$，U 也是正交矩阵。
最后证明 $U\Sigma V^T = A$：

$$\begin{aligned} U\Sigma V^T &= [U_1, U_2] \begin{bmatrix} \Sigma_1 & 0_{r\times(n-r)} \\ 0_{(m-r)\times r} & 0_{(m-r)\times(n-r)} \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} \\ &= U_1 \Sigma_1 V_1^T \\ &= A V_1 V_1^T \\ &= A \end{aligned}$$

矩阵求导:

设 $S_{p\times p}$ 是一个对称矩阵 ($S^T = S$)，$u_{p\times p}$ 是一个 p 维列向量，求证:

$$\frac{\partial \left[u^T S u\right]}{\partial u} = 2Su \tag{16}$$

第一种角度 (复杂点):

因为 S 是对称的，那么这个一定可以相识对角化，那么也就是有 $S = AA^T$, 那么 $u^T Su = \left[A^T u\right]^T \left[A^T u\right]$，那么可以写成:

$$\frac{\partial \left[u^T Su\right]}{\partial u} = \frac{\partial \left[A^T u\right]}{\partial u} \times \frac{\partial \left[A^T u\right]^T \left[A^T u\right]}{\partial A^T u} = A \times 2A^T u = 2AA^T u = 2Su \tag{17}$$

第二种角度:

$$\frac{\partial \left[u^T Su\right]}{\partial u} = \frac{\partial u^T}{\partial u} Su + \left(u^T S\right)^T \frac{\partial u}{\partial u} = 2Su \tag{18}$$

协方差矩阵的推导

我们记入数据矩阵为 $X_{N\times p}$，x_i 为每第 i 个样本, 是一个 p 维列向量，数据矩阵写成:

$$X = [x_1, x_2, ..., x_N]^T \tag{19}$$

首先记均值向量 $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$: 协方差矩阵可以写成:

$$\begin{aligned} S &= \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(x_i - \overline{x})^T \\ &= \frac{1}{N}\left[\ x_1 - \overline{x}, x_2 - \overline{x}, \ldots, x_N - \overline{x}\ \right] \begin{bmatrix} x_1 - \overline{x} \\ x_2 - \overline{x} \\ \vdots \\ x_N - \overline{x} \end{bmatrix} \end{aligned}$$

我们记：$L = (1, 1, \ldots, 1)^T$，E 为 p 维单位矩阵：所以有:

$$\begin{aligned} &\left[\ x_1 - \overline{x}, x_2 - \overline{x}, \ldots, x_N - \overline{x}\ \right] \\ &= X^T - \overline{x}L^T \\ &= X^T - \frac{1}{N}\sum_{i=1}^{N} x_i L^T \\ &= X^T - \frac{1}{N}X^T LL^T \\ &= X^T(E - \frac{1}{N}LL^T) \end{aligned}$$

所以有：

$$S = X^T(E - \frac{1}{N}LL^T)(E - \frac{1}{N}LL^T)^T X \tag{20}$$

我们记:$H = (E - \frac{1}{N}LL^T)$ 显然有 $H = H^T$, 且 $H^2 = H$, 所以协方差矩阵可以写成:

$$S = X^T H X \tag{21}$$

PCA:

最大投影方差

我们首先设置投影的一个方向向量为 u_j，是一个 p 维列向量且 $u^T u = 1$，那么样本 x_i 在这个方向上的投影为 $x_i^T u_j$: 我们的整个数据集往这上面投影，投影后的方差可以表示成:

$$\begin{aligned} S_u &= \frac{1}{N}\sum_{i=1}^{N}(x_i^T u_j - \overline{x}^T u_j)^2 \\ &= \frac{1}{N}\sum_{i=1}^{N} u_j^T (x_i - \overline{x})(x_i - \overline{x})^T u_j \\ &= u_j^T \left[\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(x_i - \overline{x})^T\right] u_j \\ &= u_j^T S u_j \end{aligned}$$

我们要在约束 $u^T u = 1$ 下求得 S_u 的最大值，依据拉格朗日乘子法只需要求解：

$$J(u_j, \lambda_j) = u_j^T S u_j + \lambda_j (1 - u_j^T u_j) \tag{22}$$

求导得到:

$$\frac{\partial J(u_j, \lambda_j)}{\partial u_j} = 2Su_j - 2\lambda_j u_j \tag{23}$$

我们令 $\frac{\partial J(u_j, \lambda_j)}{\partial u_j} = 0$, 可以得到:

$$Su_j = \lambda_j u_j \tag{24}$$

所以很显然 u_j 就只是协方差矩阵 S 的特征向量，但是 S 有 p 个线性无关的特征向量，比如我要降到 q 维，就只能在这 p 个特征向量中选择 q 个出来，那么该选择哪 q 个呢？很容易理解，我们要选择的是能让 $J(u_j, \lambda_j)$ 最大的那 q 个特征向量，我们把 $Su_j = \lambda_j u_j$ 带入到 $J(u_j, \lambda_j)$ 中，有:

$$J(u_j, \lambda_j) = \lambda_j u_j^T u_j + \lambda_j (1 - u_j^T u_j) = \lambda_j \tag{25}$$

根据上式，特征值越大的特征向量，它对应的 $J(u_j, \lambda_j)$ 的值就越大，所以我们要选择最大的 q 个特征值所对应的 q 个特征向量作为我们的投影向量即可。

最小重构代价

PCA 的本质是让我们降维度后的样本依旧能保持更多的信息，最大投影方差也是这个思想，那么另外一个思想就是让投影后的向量重新投影会原向量的代价最小。假设我们不做降维,就是我们利用 p 个线性无关的单位向量作为我们的投影向量,$(u_1, u_2, \ldots, u_p)$ 构成了原本样本的特征空间的一组新的基，此时对样本 x_i 的投影就是线性变化，不会损失任何的空间信息。

我们以所有的样本均值 $\overline{x}$ 为原点，那么一个样本 x_i 在第 j 个单位向量的投影为 $(x_i - \overline{x})^T u_j$，那么 $((x - \overline{x})^T u_j)u_j$ 是这个特征向量在 u_j 这个方向上的取值，那么有衡等式:

$$x_i - \overline{x} = \sum_{k=1}^{p}((x_i - \overline{x})^T u_k)u_k \tag{26}$$

我要降低维度，降低维度到 q 维，那么我就是在这 p 个基中选择 q 个基出来，当我们使用 q 个基的时候，这 q 个向量构成了集合 D_p，把样本 x_i 投影到 q 维度，在变换回来得到的向量为 x_i^*：

$$x_i^* - \overline{x}^* = \sum_{u_k \in D_p}((x_i - \overline{x})^T u_k)u_k \tag{27}$$

因为 $q < p$，所以我们降维后对于每个样本的信息是有损失的，我们希望选择的这 q 个基能使这份损失最小，那么就是让 J 最小:

$$\begin{aligned} J &= \frac{1}{N}\sum_{i=1}^{N}\|(x_i - \overline{x}) - (x_i^* - \overline{x}^*)\|_2 \\ &= \frac{1}{N}\sum_{i=1}^{N}\|\sum_{u_k \notin D_p}((x_i - \overline{x})^T u_k)u_k\|_2 \\ &= \frac{1}{N}\sum_{i=1}^{N}\sum_{u_k \notin D_p}\left[(x_i - \overline{x})^T u_k\right]^2 \end{aligned}$$

所以

$$\begin{aligned} J &= \frac{1}{N}\sum_{i=1}^{N}\sum_{u_k \notin D_p} u_k^T(x - \overline{x})(x_i - \overline{x})^T u_k \\ &= \sum_{u_k \notin D_p} u_k^T\left[\frac{1}{N}\sum_{i=1}^{N}(x - \overline{x})(x_i - \overline{x})^T\right]u_k \\ &= \sum_{u_k \notin D_p} u_k^T S u_k \end{aligned}$$

那么这个意思是，我要找出 $p - q$ 个单位向量，他们能使得 J 取得最小值，为了方便表示，我们分别记入这 $z = p - q$，且这 z 个向量为 $(u_1, u_2, \ldots, u_z)$。

我们只需要分别让 $u_1^T S u_1, u_2^T S u_2, \ldots, u_z^T S u_z$ 最小，并且当 $i \neq j$ 时 $u_i \neq u_j$，即可。我们利用拉格朗日方法构造我们的目标函数 (其实就是损失函数)，使目标函数取得最小值:

$$Loss = u_k^T S u_k + \lambda_k (1 - u_k^T u_k) \tag{28}$$

我们令 $\frac{\partial Loss}{\partial u_k} = 0$: 可以得到:

$$S u_k = \lambda_k u_k \tag{29}$$

带入 $Loss$，得到:

$$Loss = \lambda_k \tag{30}$$

这就很明显了，这 z 个向量 $(u_1, u_2, ..., u_z)$ 就是协方差矩阵 S 最小的那 z 个特征值所对应的特征向量，那么剩下的那 $q = p - z$ 个特征向量就是我们要用来降维的向量啦。

奇异值分解的角度

首先我们先对 $X_{N \times p}$ 做中心化:

$$\begin{aligned} X_{N\times p} - \begin{bmatrix} \overline{x} \\ \overline{x} \\ \vdots \\ \overline{x} \end{bmatrix}_{N\times p} &= X_{N\times p} - \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \overline{x} \\ &= X_{N\times p} - L\overline{x} \\ &= X_{N\times p} - L\frac{1}{N}L^T X_{N\times p} \\ &= \left(E - \frac{1}{N}LL^T\right) X_{N\times p} \\ &= H X_{N\times p} \end{aligned}$$

由奇异值分解定理，$H X_{N\times p}$ 可以写成:

$$(HX)_{N\times p} = U_{N\times N} \Sigma_{N\times p} V^T p \times p \tag{31}$$

斜方差矩阵 S 可以写成:

$$\begin{aligned} S &= \frac{1}{N} X^T H X \\ &= \frac{1}{N} X^T H^T H X \\ &= \frac{1}{N} (HX)^T H X \\ &= \frac{1}{N} V \Sigma U^T U \Sigma V^T \end{aligned}$$

由奇异值分解定理，所以 $U^TU = E$，所以:

$$S = V\frac{\Sigma^2}{N}V^T \tag{32}$$

其中 V 的列向量是 $(HX)^THX$ 的特征向量，也就是 S 的特征向量，而 $\frac{\Sigma^2}{N}$ 是 S 的特征值所组成的对角矩阵，所以如果我们要对 $X_{N\times p}$ 使用 PCA，我们其实不需要去对协方差矩阵 S 计算特征值，特征向量，甚至不需要计算协方差矩阵 S。只需要对中心化后的 HX 做特征值分解即可。