

偏最小二乘法回归 (Partial Least Squares Regression)

JerryLead

csxulijie@gmail.com

2011年8月20日星期六

1. 问题

这节我们请出最后的有关成分分析和回归的神器 PLSR。PLSR 感觉已经把成分分析和回归发挥到极致了，下面主要介绍其思想而非完整的教程。让我们回顾一下最早的 Linear Regression 的缺点：如果样例数 m 相比特征数 n 少 ($m < n$) 或者特征间线性相关时，由于 $X^T X$ ($n \times n$ 矩阵) 的秩小于特征个数 (即 $X^T X$ 不可逆)。因此最小二乘法 $\theta = (X^T X)^{-1} X^T \vec{y}$ 就会失效。

为了解决这个问题，我们会使用 PCA 对样本 X ($m \times n$ 矩阵) 进行降维，不妨称降维后的 X 为 X' ($m \times r$ 矩阵，一般加了'就表示转置，这里临时改变下)，那么 X' 的秩为 r (列不相关)。

2. PCA Revisited

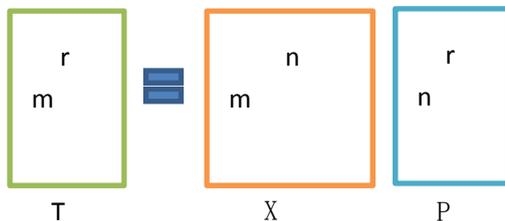
所谓磨刀不误砍柴工，这里先回顾下 PCA。

令 X 表示样本，含有 m 个样例 $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$ ，每个样例特征维度为 n ， $x^{(i)} = \{x_1^{(i)}, x_2^{(i)}, \dots, x_n^{(i)}\}$ 。假设我们已经做了每个特征均值为 0 处理。

如果 X 的秩小于 n ，那么 X 的协方差矩阵 $\frac{1}{m} X^T X$ 的秩小于 n ，因此直接使用线性回归的话不能使用最小二乘法来求解出唯一的 θ ，我们想使用 PCA 来使得 $X^T X$ 可逆，这样就可以用最小二乘法来进行回归了，这样的回归称为主元回归 (PCR)。

PCA 的一种表示形式：

$$T = XP$$



其中 X 是样本矩阵， P 是 X 的协方差矩阵的特征向量 (当然是按照特征值排序后选取的前 r 个特征向量)， T 是 X 在由 P 形成的新的正交子空间上的投影 (也是样本 X 降维后的新矩阵)。

在线性代数里面我们知道，实对称阵 A 一定存在正交阵 P ，使得 $P^{-1} A P$ 为对角阵。因此可以让 $X^T X$ 的特征向量矩阵 P 是正交的。

其实 T 的列向量也是正交的，不太严谨的证明如下：

$$T^T T = (XP)^T (XP) = P^T X^T X P = P^T (P \Lambda P^T) P = P^T P \Lambda P^T P = \Lambda$$

其中利用了 $X^T X = P \Lambda P^T$ ，这是求 P 的过程， Λ 是对角阵，对角线上元素就是特征值 λ 。这里对 P 做了单位化，即 $P^T P = I$ 。这就说明了 T 也是正交的， P 是 $X^T X$ 的特征向量矩阵，更进一步， T 是 XX^T 的特征向量矩阵 ($XX^T T = XX^T X P = X P \Lambda P^T P = T \Lambda$)。

这样经过 PCA 以后，我们新的样本矩阵 T ($m \times r$) 是满秩的，而且列向量正交，因此直接代入最小二乘法公式，就能得到回归系数 θ 。

PCA 的另一种表示：

$$X = M_1 + M_2 + M_3 + \dots + M_n = t_1 p_1^T + t_2 p_2^T + t_3 p_3^T + \dots + t_n p_n^T = T P^T \quad (\text{假设 } X \text{ 秩为 } n)$$

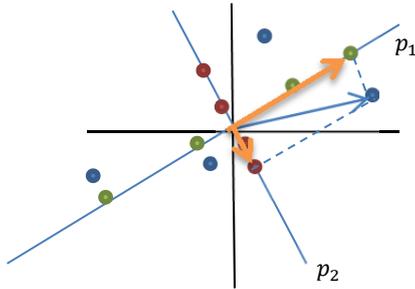
这个公式其实和上面的表示方式 $T = X P$ 没什么区别。

$$T = X P \rightarrow T P^T = X P P^T \rightarrow X = T P^T \quad (\text{当然我们认为 } P \text{ 是 } n \times n \text{ 的，因此 } P^T = P^{-1})$$

如果 P 是 $n \times r$ 的，也就是舍弃了特征值较小的特征向量，那么上面的加法式子就变成了

$$X = M_1 + M_2 + M_3 + \dots + M_r + E = t_1 p_1^T + t_2 p_2^T + t_3 p_3^T + \dots + t_r p_r^T + E = T P^T + E$$

这里的 E 是残差矩阵。其实这个式子有着很强的几何意义， p_i 是 $X^T X$ 第 i 大特征值对应的归一化后的特征向量， t_i 就是 X 在 p_i 上的投影。 $t_i p_i^T$ 就是 X 先投影到 p_i 上，再以原始坐标系得到的 X' 。下面这个图可以帮助理解：



黑色线条表示原始坐标系，蓝色的点是原始的 4 个 2 维的样本点，做完 PCA 后，得到两个正交的特征向量坐标 p_1 和 p_2 。绿色点是样本点在 p_1 上的投影（具有最大方差），红色点是在 p_2 上的投影。 t_1 的每个分量是绿色点在 p_1 上的截距， t_2 是红色点在 p_2 上的截距。 $t_i p_i^T$ 中的每个分量都可以看做是方向为 p_i ，截距为 t_i 相应分量大小的向量，如那个 p_1 上的橘色箭头。 $t_i p_i^T$ 就得到了 X 在 p_i 的所有投影向量，由于 p_1 和 p_2 正交，因此 $t_1 p_1^T + t_2 p_2^T$ 就相当于每个点的橘色箭头的加和，可想而知，得到了原始样本点。

如果舍弃了一些特征向量如 p_2 ，那么通过 $t_1 p_1^T$ 只能还原出原始点的部分信息（得到的绿色点，丢失了蓝色点在另一维度上的信息）。另外， P 有个名字叫做 loading 矩阵， T 叫做 score 矩阵。

3. PLSR 思想及步骤

我们还需要回味一下 CCA 来引出 PLSR。在 CCA 中，我们将 X 和 Y 分别投影到直线得到 u 和 v ，然后计算 u 和 v 的 Pearson 系数（也就是 $\text{Corr}(u, v)$ ），认为相关度越大越好。形式化表示：

$$\begin{aligned} & \text{Maximize } a^T \text{Cov}(x, y) b \\ & \text{Subject to: } a^T \text{Var}(x) a = 1, b^T \text{Var}(y) b = 1 \end{aligned}$$

其中 a 和 b 就是要求的投影方向。

想想 CCA 的缺点：对特征的处理方式比较粗糙，用的是线性回归来表示 u 和 x 的关系， u 也是 x 在某条线上的投影，因此会存在线性回归的一些缺点。我们想把 PCA 的成分提取技术引入 CCA，使得 u 和 v 尽可能携带样本的最主要信息。还有一个更重要的问题，CCA 是寻找 X 和 Y 投影后 u 和 v 的关系，显然不能通过该关系来还原出 X 和 Y ，也就是找不到 X 到 Y 的直接映射。这也是使用 CCA 预测时大多配上 KNN 的原因。

而 PLSR 更加聪明，同时兼顾 PCA 和 CCA，并且解决了 X 和 Y 的映射问题。看 PCA Revisited 的那张图，假设对于 CCA， X 的投影直线是 p_1 ，那么 CCA 只考虑了 X 的绿色点与 Y 在某条直线上投影结果的相关性，丢弃了 X 和 Y 在其他维度上的信息，因此不存在 X 和 Y 的映射。而 PLSR 会在 CCA 的基础上再做一步，由于原始蓝色点可以认为是绿色点和红色点的叠加，因此先使用 X 的绿色点 t_1 对 Y 做回归 ($Y = t_1 r_1^T + F$ ，样子有点怪，两边都乘以 r_1 就明白了，这里的 Y 类似于线性回归里的 X ， t_1 类似 y)，然后用 X 的红色点 t_2 对 Y 的剩余部分 F 做回归 (得到 r_2 ， $F = t_2 r_2^T + F'$)。这样 Y 就是两部分回归的叠加。当新来一个 x 时，投影一下得到其绿色点 t_1 和红色点 t_2 ，然后通过 r 就可以还原出 Y ，实现了 X 到 Y 的映射。当然这只是几何上的思想描述，跟下面的细节有些出入。

下面正式介绍 PLSR:

- 1) 设 X 和 Y 都已经过标准化 (包括减均值、除标准差等)。
- 2) 设 X 的第一个主成分为 p_1 ， Y 的第一个主成分为 q_1 ，两者都经过了单位化。(这里的主成分并不是通过 PCA 得出的主成分)
- 3) $u_1 = Xp_1$ ， $v_1 = Yq_1$ ，这一步看起来和 CCA 是一样的，但是这里的 p 和 q 都有主成分的性质，因此有下面 4) 和 5) 的期望条件。
- 4) $Var(u_1) \rightarrow max, Var(v_1) \rightarrow max$ ，即在主成分上的投影，我们期望是方差最大化。
- 5) $Corr(u_1, v_1) \rightarrow max$ ，这个跟 CCA 的思路一致。
- 6) 综合 4) 和 5)，得到优化目标 $Cov(u_1, v_1) = \sqrt{Var(u_1)Var(v_1)}Corr(u_1, v_1) \rightarrow max$ 。

形式化一点：

$$\text{Maximize } \langle Xp_1, Yq_1 \rangle$$

$$\text{Subject to: } \|p_1\| = 1, \|q_1\| = 1$$

看起来比 CCA 还要简单一些，其实不然，CCA 做完一次优化问题就完了。但这里的 p_1 和 q_1 对 PLSR 来说只是一个主成分，还有其他成分呢，那些信息也要计算的。

先看该优化问题的求解吧：

引入拉格朗日乘子

$$\mathcal{L} = p_1^T X^T Y q_1 - \frac{\lambda}{2} (p_1^T p_1 - 1) - \frac{\theta}{2} (q_1^T q_1 - 1)$$

分别对 p_1, q_1 求偏导，得

$$\frac{\partial \mathcal{L}}{\partial p_1} = X^T Y q_1 - \lambda p_1 = 0$$

$$\frac{\partial \mathcal{L}}{\partial q_1} = Y^T X p_1 - \theta q_1 = 0$$

从上面可以看出 $\lambda = \theta$ (两边都乘以 p 或 q ，再利用 $=1$ 的约束)

下式代入上式得到

$$X^T Y Y^T X p_1 = \lambda^2 p_1$$

上式代入下式得到

$$Y^T X X^T Y q_1 = \lambda^2 q_1$$

目标函数 $\langle X p_1, Y q_1 \rangle \rightarrow p_1^T X^T Y q_1 \rightarrow p_1^T (\lambda p_1) \rightarrow \lambda$, 要求最大。

因此 p_1 就是对称阵 $X^T Y Y^T X$ 的最大特征值对应的单位特征向量, q_1 就是 $Y^T X X^T Y$ 最大特征值对应的单位特征向量。

可见 p_1 和 q_1 是投影方差最大和两者相关性最大上的权衡, 而 CCA 只是相关性上最大化。求得了 p_1 和 q_1 , 即可得到

$$u_1 = X p_1$$

$$v_1 = Y q_1$$

这里得到的 u_1 和 v_1 类似于上图中的绿色点, 只是在绿色点上找到了 X 和 Y 的关系。如果就此结束, 会出现与 CCA 一样的不能由 X 到 Y 映射的问题。

利用我们在 PCA Revisited 里面的第二种表达形式, 我们可以继续做下去, 建立回归方程:

$$X = u_1 c_1^T + E$$

$$Y = v_1 d_1^T + G$$

这里的 c 和 d 不同于 p 和 q, 但是它们之间有一定联系, 待会证明。E 和 G 是残差矩阵。

我们进行 PLSR 的下面几个步骤:

- 1) $Y = u_1 r_1^T + F$, 使用 u_1 对 Y 进行回归, 原因已经解释过, 先利用 X 的主成分对 Y 进行回归。
- 2) 使用最小二乘法, 计算 c, d, r 分别为:

$$c_1 = \frac{X^T u_1}{\|u_1\|^2}$$

$$d_1 = \frac{Y^T v_1}{\|v_1\|^2}$$

$$r_1 = \frac{Y^T u_1}{\|u_1\|^2}$$

实际上这一步计算出了各个投影向量。

p_1 和 c_1 的关系如下:

$$p_1^T c_1 = p_1^T \frac{X^T u_1}{\|u_1\|^2} = \frac{u_1^T u_1}{\|u_1\|^2} = 1$$

再谈谈 p_1 和 c_1 的关系, 虽然这里将 c_1 替换成 p_1 可以满足等式要求和几何要求, 而且 p_1 就是 X 投影出 u_1 的方向向量。但这里我们想做的是回归 (让 E 尽可能小), 因此根据最小二乘法得到的 c_1 一般与 p_1 不同。

- 3) 将剩余的 E 当做新的 X, 剩余的 F 当做新的 Y, 然后按照前面的步骤求出 p_2 和 q_2 , 得到:

$$u_2 = E p_2$$

$$v_2 = F q_2$$

目标函数 $\langle E p_2, F q_2 \rangle \rightarrow p_2^T E^T F q_2 \rightarrow p_2^T (\lambda p_2) \rightarrow \lambda$, 这个与前面一样, p_2 和 q_2 分别是新的 $E^T F F^T E$ 和 $F^T E E^T F$ 的最大特征值对应的单位特征向量。

- 4) 计算得到第二组回归系数:

$$c_2 = \frac{E^T u_2}{\|u_2\|^2}$$

$$r_2 = \frac{F^T u_2}{\|u_2\|^2}$$

这里的 u_2 和之前的 u_1 是正交的，证明如下：

$$u_1^T u_2 = u_1^T E p_2 = u_1^T (X - u_1 c_1^T) p_2 = \left[u_1^T X - u_1^T u_1 \frac{u_1^T X}{\|u_1\|^2} \right] p_2 = 0$$

其实 u_i 和不同的 u_j 都是相互正交的。

同样 p_i 和不同的 p_j 也是正交的。

$$\begin{aligned} p_1^T p_2 &= p_1^T \frac{1}{\lambda} E^T F q_2 = p_1^T \frac{1}{\lambda} E^T v_2 = \frac{1}{\lambda} p_1^T (X - u_1 c_1^T)^T v_2 \\ &= \frac{1}{\lambda} (X p_1 - u_1 c_1^T p_1)^T v_2 = \frac{1}{\lambda} (u_1 - u_1)^T v_2 = 0 \end{aligned}$$

但 c_i 和不同的 c_j 一般不是正交的。

- 5) 从上一步得到回归方程：

$$\begin{aligned} E &= u_2 c_2^T + E' \\ F &= u_2 r_2^T + F' \end{aligned}$$

如果还有残差矩阵的话，可以继续计算下去。

- 6) 如此计算下去，最终得到：

$$\begin{aligned} X &= u_1 c_1^T + u_2 c_2^T + u_3 c_3^T + \dots + u_n c_n^T + E \\ Y &= u_1 r_1^T + u_2 r_2^T + u_3 r_3^T + \dots + u_n r_n^T + F \end{aligned}$$

与 PCA 中表达式不一样的是这里的 c_i 和不同的 c_j 之间一般不是正交的。

其实这里不必一直计算到 n ，可以采用类似于 PCA 的截尾技术，计算到合适的 r 即可。关于 r 数目的选取可以使用交叉验证方法，这与 PCA 里面的问题类似。

另外， p_i 和 c_j 的关系是 $p_i^T c_j = 1 (i = j)$, $p_i^T c_j = 0 (i \neq j)$

上面的公式如果写成矩阵形式如下：

$$\begin{aligned} X &= UC^T + E \\ Y &= UR^T + F = XPR^T + F = XB + F \end{aligned}$$

这就是 $X \rightarrow Y$ 的回归方程，其中 $B = PR^T$ 。

在计算过程中，收集一下 P 和 R 的值即可。

- 7) 使用 PLSR 来预测。

从 6) 中可以发现 Y 其实是多个回归的叠加(其实 $u_1 r_1^T$ 已经回归出 Y 的最主要信息)。

我们在计算模型的过程中，得到了 p 和 r 。那么新来一个 x ，首先计算 u (这里的 u 变成了实数，而不是向量了)，得到

$$u_1 = x^T p_1, u_2 = x^T p_2, u_3 = x^T p_3 \dots$$

然后代入 Y 的式子即可求出预测的 y 向量，或者直接代入 $y^T = x^T B$

- 8) 至此，PLSR 的主要步骤结束。

4. PLSR 相关问题

- 1) 其实不需要计算 v 和 q ，因为我们使用 u 去做 Y 的回归时认为了 $u_i = c v_i$ ，其中 c 是常数。之所以这样是因为前面提到过的 Y 可以首先在 X 的主要成分上做回归，然后将 Y 的残差矩阵在 X 的残差矩阵的主要成分上做回归。最后 X 的各个成分回归之和就是 Y 。

- 2) 一般使用的 PLSR 求解方法是迭代化的求解方法,称之为 NIPALS,还有简化方法 SIMPLS,这些方法在一般论文或参考文献中提供的网址里都有,这里就不再贴了。
- 3) PLSR 里面还有很多高级话题,比如非线性的 Kernel PLSR,异常值检测,带有缺失值的处理方法,参数选择,数据转换,扩展的层次化模型等等。可以参考更多的论文有针对性的研究。

5. 一些感悟

本文试图将 PCA、CCA、PLSR 综合起来对比、概述和讨论,不免对符号的使用稍微都点混乱,思路也有穿插混淆。还是以推导出的公式为主进行理解吧。另外,本文有很多个人理解在里面,难免有误,还望批评指正。提供 PDF 版本,只是为了格式好看些。

之前也陆陆续续地关注了一些概率图模型和时间序列分析,以后可能会转向介绍这两方面的内容,也会穿插一些其他的内容。说实话,自学挺吃力的,尤其对我这样一个不是专业搞 ML 的人来说,也需要花大量时间。感叹国外的资料多,lecture 多,视频多,可惜因为我这的网速和 GFW 原因,看不了教学视频,真是遗憾。

6. 参考文献:

1. PARTIAL LEAST-SQUARES REGRESSION: A TUTORIAL. Paul Geladi and Bruce R. Kowalski
2. 王惠文一偏最小二乘回归方法及应用
3. Partial Least Squares (PLS) Regression.
4. A Beginner's Guide to Partial Least Squares Analysis
5. Nonlinear Partial Least Squares: An Overview
6. <http://www.statsoft.com/textbook/partial-least-squares/>
7. Canonical Correlation a Tutorial
8. Pattern Recognition And Machine Learning