PCA

-PROMISE-

Thursday 21st April, 2022

参考：https://zhuanlan.zhihu.com/p/77151308

1 向量与基变换

1.1 基变换

- 向量 **a** 和 **b** 的内积:

$$(a_1, a_2, ..., a_n) \cdot (b_1, b_2, ..., b_n)^T = a_1b_1 + a_2b_2 + ... + a_nb_n$$

内积运算得到的是实数值，可以根据下图理解，其实就是向量 **A** 在向量 **B** 角度的一个映射乘上 **B** 的长度，当 |**B**|=1 的时候，就可以把 **B** 看成基向量。（例如向量（3，2）就是在（1，0）（0，1）这组基下的一个坐标

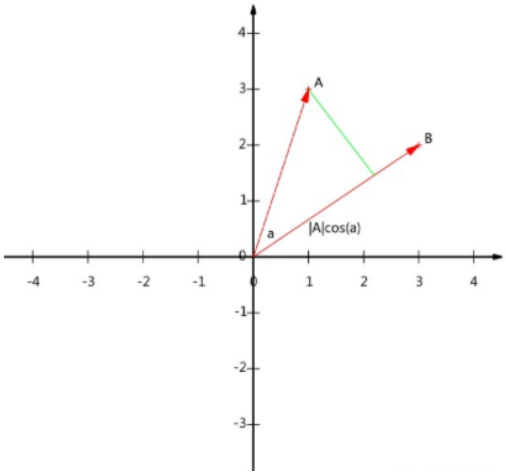

图 1:

• 所以我们可以得出一个结论，我们要准确描述向量，首先要确定一组基，然后给出基所在的各个直线上的投影就可以了。

1.2 基变换的矩阵表示

• 对于向量 a_1,a_2...$a_n(n \times 1)$，在基 x_1,x_2...$x_n(x_i\ n \times 1)$ 的基向量
• 坐标变换表现为

$$\begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} \cdot \begin{pmatrix} a_1^T & a_2^T & \dots & a_n^T \end{pmatrix}$$

2 降维

2.1 基本思想

• 假如基的数量小于向量本身的维数，就可以达到降维的效果。而对于基向量的选择，我们希望选择的 K 个基进行保留原有的信息。(希望投影后的数据尽可能分散，因为如果重叠就会有数据消失)

2.2 方差和协方差

2.2.1 方差

• 假设 μ 为向量的均值，那么方差就是 $Var(a) = \frac{1}{m}\sum_{i=1}^{m}(a_i-\mu)^2$，方便处理，将均值变成 0（所有向量减去均值）那么方差就是 $Var(a) = \frac{1}{m}\sum_{i=1}^{m}a_i^2$

总的来说就是找到一个基使得数据在这个基想变换后坐标的方差最大

• 至此，我们得到了降维问题的优化目标：将一组 N 维向量降为 K 维，其目标是选择 K 个单位正交基，使得原始数据变换到这组基上后，各变量两两间协方差为 0，而变量方差则尽可能大（在正交的约束下，取最大的 K 个方差）。

2.2.2 协方差

• 协方差表示的是两个变量之间的关系，协方差为 0 的时候说明两个变量之间没有线性关系，为了使结果包括更多有效信息，我们在选择之后的基的时候使得它和之前的基保持正交同样均值为 0 的时候，有：

• Cov(a,b)=$\sum_{i=1}^{m}a_ib_i$

• 协方差矩阵：

假设两个变量 a，b

组成矩阵 X:

$$X = \begin{pmatrix} a_1 & a_2 & ... & a_m \\ b_1 & b_2 & ... & b_m \end{pmatrix}$$

此时有：

$$\frac{1}{m}XX^T = \begin{pmatrix} \frac{1}{m}\sum_{i=1}^{m}a_i^2 & \frac{1}{m}\sum_{i=1}^{m}a_ib_i \\ \frac{1}{m}\sum_{i=1}^{m}a_ib_i & \frac{1}{m}\sum_{i=1}^{m}b_i^2 \end{pmatrix} = \begin{pmatrix} Cov(a,a) & Cov(a,b) \\ Cov(b,a) & Cov(b,b) \end{pmatrix}$$

综上，对于有 m 和 n 维的数据记录，将其排列成矩阵 $X_{n,m}$，设 $C = \frac{1}{m}XX^T$ 则 C 是以恶个对称矩阵，其对角线是每个变量的方差，而 x_{ij} 是数据 i 和数据 j 的协方差

2.2.3 矩阵对角化

• 设原始矩阵 X 对应的协方差矩阵为 C，P 是一组基按行构成的矩阵，设 Y=PX，则 Y 为 X 对 P 基变换后的数据，设 Y 的协方差为 D，推导出 D 和 C 的关系是：$D = PCP^T$

所以说我们需要找的 P 是一个能让原始矩阵对角化的矩阵，也就是说 P 会满足 PCP^T 为一个对焦矩阵，它的对角线值降序排列，其他位置是 0（根据前面的优化条件，方差需要找最大的但是数据之间没有线性关系）

对于原始的协方差矩阵 C，因为 C 是一个实对称矩阵，有以下性质

· 实对称矩阵对于的特征向量正交

· 特征向量 λ 重数为 r，则必然存在 r 个线性无关的特征向量对应于 λ, 因此可以把这 r 个特征向量正交化

由上面两条可知，一个 n 行 n 列的实对称矩阵一定可以找到 n 个单位正交特征向量，设这 n 个特征向量为 $e_1, e_2, ..., e_n$，我们将其按列组成矩阵：$E = (e_1, e_2, ..., e_n)$

对协方差矩阵有：

$$E^T C E = \Lambda = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \cdots & \\ & & & \lambda_n \end{pmatrix}$$

• 因此我们需要的矩阵 P 就是 E^T

• P 是协方差矩阵的特征向量单位化后按行排列出的矩阵，其中每一行都是 C 的一个特征向量。如果设 P 按照 Λ 中特征值的从大到小，将特征向量从上到下排列，则用 P 的前 K 行组成的矩阵乘以原始数据矩阵 X，就得到了我们需要的降维后的数据矩阵 Y。

• 拉格朗日乘子法:

设样本点为 x_i，在基 w 下的坐标为：$(x_i, w) = x_i^T w$, 有方差：

$D(x) = \frac{1}{m}\sum_{i=1}^{m}(x_i^T w)^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i^T w)^T(x_i^T w) = \frac{1}{m}\sum_{i=1}^{m} w^T x_i x_i^T w = w^T(\frac{1}{m}\sum_{i=1}^{m} x_i x_i^T)w$

我们看到 $\frac{1}{m}\sum_{i=1}^{m} x_i x_i^T$ 为原样本的协方差，设这个矩阵为 Λ 我们有

$$\max w^T \Lambda w s.t. w^T w = 1$$

构造拉格朗日函数：$L(w) = w^T \Lambda w + \lambda(1 - w^T w)$, 对 w 求导有 $\Lambda w = \lambda w$

得到方差为 $D(x) = w^t \Lambda w = \lambda w^T w = \lambda$

于是我们发现，x 投影后的方差就是协方差矩阵的特征值。我们要找到最大方差也就是协方差矩阵最大的特征值，最佳投影方向就是最大特征值所对应的特征向量，次佳就是第二大特征值对应的特征向量，以此类推。

3 求解步骤

- 设有 m 条 n 维数据。

  将原始数据按列组成 n 行 m 列矩阵 X；

  将 X 的每一行进行零均值化，即减去这一行的均值；

  求出协方差矩阵 $C = \frac{1}{m}XX^T$；

  求出协方差矩阵的特征值及对应的特征向量；

  将特征向量按对应特征值大小从上到下按行排列成矩阵，取前 k 行组成矩阵 P；

  $Y = PX$ 即为降维到 k 维后的数据。

4 性质

- 可以实现降维，缓解维度灾难
- 降噪，往往最小的特征值对应的向量与噪声有关，舍弃之后可以得到降噪的效果
- PCA 保留了主要信息，但是这只是对于训练集而言的主要信息，如果舍弃的一些看似无用的信息出现在测试集当中，PCA 可能加剧过拟合
- 降维后数据各特征独立

5 其他细节

5.1 中心化

- 各个数据都减去均值

5.2 与 SVD 对比

- 略