PCA 主成分分析概述

PCA（Principal Components Analysis）即主成分分析，是图像处理中经常用到的降维方法，大家知道，我们在处理有关数字图像处理方面的问题时，比如经常用的图像的查询问题，在一个几万或者几百万甚至更大的数据库中查询一幅相近的图像。这时，我们通常的方法是对图像库中的图片提取响应的特征，如颜色，纹理，sift，surf，vlad 等等特征，然后将其保存，建立响应的数据索引，然后对要查询的图像提取相应的特征，与数据库中的图像特征对比，找出与之最近的图片。这里，如果我们为了提高查询的准确率，通常会提取一些较为复杂的特征，如 sift，surf 等，一幅图像有很多个这种特征点，每个特征点又有一个相应的描述该特征点的 128 维的向量，设想如果一幅图像有 300 个这种特征点，那么该幅图像就有 300*vector（128 维）个，如果我们数据库中有一百万张图片，这个存储量是相当大的，建立索引也很耗时，如果我们对每个向量进行 PCA 处理，将其降维为 64 维，是不是很节约存储空间啊？

PCA 主成分分析数学原理

- Principal Component Analysis
  - 用途：降维中最常用的一种手段
  - 目标：提取最有价值的信息（基于方差）
  - 问题：降维后的数据的意义?

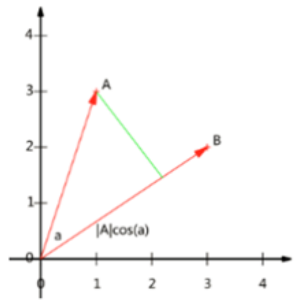
向量的表示及基变换

- 向量的表示及基变换

  - 内积：$(a_1,a_2,\cdots,a_n)^\mathsf{T}\cdot(b_1,b_2,\cdots,b_n) \;=a_1b_1+a_2b_2+\cdots+a_nb_n$

  - 解释：$A\cdot B=|A||B|cos(a)$

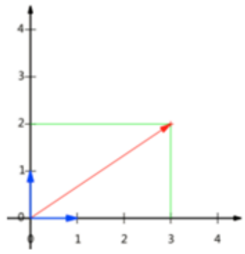
  - 设向量B的模为1，则A与B的内积值等于A向B所在直线投影的矢量长度



- 向量的表示及基变换

  - 向量可以表示为(3,2)
    实际上表示线性组合：$x(1,0)^\mathsf{T}+y(0,1)^\mathsf{T}$

  - 基：(1,0)和(0,1)叫做二维空间中的一组基

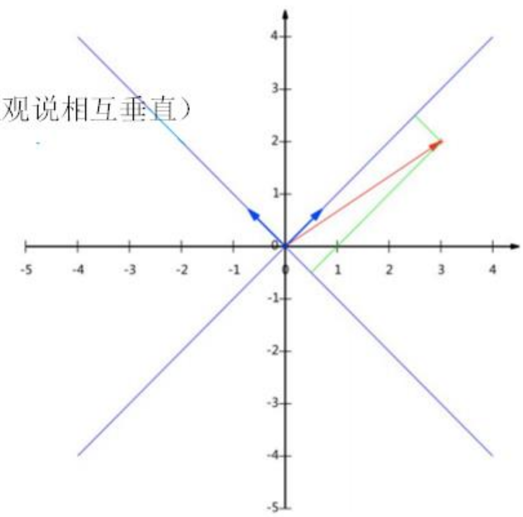


基变换

- 基变换

  - 基是正交的（即内积为0，或直观说相互垂直）

  - 要求：线性无关



- 基变换

  - 变换：数据与一个基做内积运算，结果作为第一个新的坐标分量，然后与第二个基做内积运算，结果作为第二个新坐标的分量

  - 数据（3，2）映射到基中坐标： $\begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 5/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$

- 基变换

$$\begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_R \end{pmatrix} \begin{pmatrix} a_1 & a_2 & \cdots & a_M \end{pmatrix} = \begin{pmatrix} p_1a_1 & p_1a_2 & \cdots & p_1a_M \\ p_2a_1 & p_2a_2 & \cdots & p_2a_M \\ \vdots & \vdots & \ddots & \vdots \\ p_Ra_1 & p_Ra_2 & \cdots & p_Ra_M \end{pmatrix}$$

  - 两个矩阵相乘的意义是将右边矩阵中的每一列列向量变换到左边矩阵中每一行行向量为基所表示的空间中去

协方差矩阵

- 协方差矩阵

  - 方向：如何选择这个方向（或者说基）才能尽量保留最多的原始信息呢？ 一种直观的看法是：希望投影后的投影值尽可能分散

  - 方差：$Var(a) = \frac{1}{m}\sum_{i=1}^{m}(a_i - \mu)^2$

  - 寻找一个一维基，使得所有数据变换为这个基上的坐标表示后，方差值最大

  - 协方差（假设均值为0时）： $Cov(a,b) = \frac{1}{m}\sum_{i=1}^{m}a_ib_i$

- 协方差

  - 如果单纯只选择方差最大的方向，后续方向应该会和方差最大的方向接近重合。

  - 解决方案：为了让两个字段尽可能表示更多的原始信息，我们是不希望它们之间存在（线性）相关性的

  - 协方差：可以用两个字段的协方差表示其相关性 $Cov(a,b)=\frac{1}{m}\sum_{i=1}^{m}a_ib_i$

  - 当协方差为0时，表示两个字段完全独立。为了让协方差为0，选择第二个基时 只能在与第一个基正交的方向上选择。因此最终选择的两个方向一定是正交的。

优化目标

- 优化目标

  - 将一组N维向量降为K维（K大于0，小于N），目标是选择K个单位正交基，使 原始数据变换到这组基上后，各字段两两间协方差为0，字段的方差则尽可能大

  - 协方差矩阵： $X=\begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \\ \cdots & \cdots \\ a_m & b_m \end{pmatrix}$ $\frac{1}{m}X^\mathsf{T}X=\begin{pmatrix} \frac{1}{m}\sum_{i=1}^{m}a_i^2 & \frac{1}{m}\sum_{i=1}^{m}a_ib_i \\ \frac{1}{m}\sum_{i=1}^{m}a_ib_i & \frac{1}{m}\sum_{i=1}^{m}b_i^2 \end{pmatrix}$

  - 矩阵对角线上的两个元素分别是两个字段的方差，而其它元素是a和b的协方差。

- 优化目标

  - 协方差矩阵对角化：即除对角线外的其它元素化为0，并且在对角线上将元素按大小从上到下排列

  - 协方差矩阵对角化： $PCP^\mathsf{T}=\Lambda=\begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix}$

PCA 实例

- PCA实例

  - 数据：$\begin{pmatrix} -1 & -2 \\ -1 & 0 \\ 0 & 0 \\ 2 & 1 \\ 0 & 1 \end{pmatrix}$

  - 协方差矩阵：$C = \frac{1}{5}\begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix}\begin{pmatrix} -1 & -2 \\ -1 & 0 \\ 0 & 0 \\ 2 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \frac{6}{5} & \frac{4}{5} \\ \frac{4}{5} & \frac{6}{5} \end{pmatrix}$

  - 特征值：$\lambda_1 = 2, \lambda_2 = 2/5$　　特征向量：$c_1\begin{pmatrix} 1 \\ 1 \end{pmatrix}, c_2\begin{pmatrix} -1 \\ 1 \end{pmatrix}$

  - 对角化：$PCP^{\mathsf{T}} = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}\begin{pmatrix} 6/5 & 4/5 \\ 4/5 & 6/5 \end{pmatrix}\begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2/5 \end{pmatrix}$

  - 降维：$Y = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}\begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix} = \begin{pmatrix} -3/\sqrt{2} & -1/\sqrt{2} & 0 & 3/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$