主成分分析

Principal Component Analysis

王栋

wdice@dlut.edu.cn

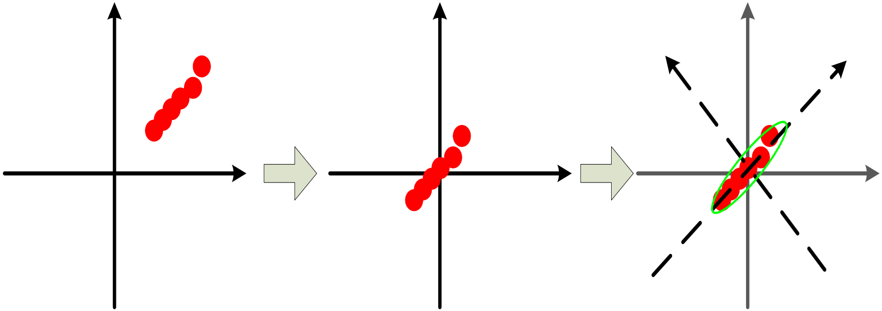
信息与通信工程学院

内容提要

预备知识

- 基（坐标）变换，空间和子空间
  参看第三讲最小二乘预备知识
- SVD分解和特征值分解
  参看最后一页PPT
- 拉格朗日乘子法
  google：维基百科，拉格朗日乘数
- KL (Karhunen-Loeve)变换
  参看PCA人脸识别
- 《Eigenfaces for Face Detection/Recognition》
  理解特征脸算法与PCA算法最大不同（“技巧加速”）
  理解特征脸的含义（“基向量的可视化”）和物理意义（“任何一个人脸都可以看做平均脸和特征脸的线性组合

从坐标变换看PCA



从坐标变换看PCA以及压缩

$$\begin{bmatrix}-1\\-1\end{bmatrix},\begin{bmatrix}0\\0\end{bmatrix},\cdots,\begin{bmatrix}1\\1\end{bmatrix}\leftarrow\begin{bmatrix}1&0\\0&1\end{bmatrix}$$

$$\begin{bmatrix}-\sqrt{2}\\0\end{bmatrix},\begin{bmatrix}0\\0\end{bmatrix},\cdots,\begin{bmatrix}\sqrt{2}\\0\end{bmatrix}\leftarrow\begin{bmatrix}\sqrt{2}/2&\sqrt{2}/2\\\sqrt{2}/2&-\sqrt{2}/2\end{bmatrix}$$

问题介绍

- 给定n个m维的样本$\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ $(\mathbf{x}_i \in \mathbb{R}^{m\times 1})$,主成分分析（PCA）的目的寻找一个均值$\mu$,和一组正交基向量$\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_d](d < \min(m, n))$来最优的表示给定的样本。

$$\min \sum_{i=1}^{n} \|\mathbf{x}_i - (\mu + \mathbf{U}\mathbf{y}_i)\|_2^2,$$

这里的$\mathbf{y}_i$称为表示系数。假设我们已经知道了U和μ,那么最优的$\mathbf{y}_i$就为：$\mathbf{y}_i = \mathbf{U}^\top(\mathbf{x}_i - \mu)$。注意：与最小二乘类比。

$$\min \sum_{i=1}^{n} \|\mathbf{x}_i - (\mu + \mathbf{U}\mathbf{y}_i)\|_2^2 \Rightarrow \min \|\mathbf{x}_i - (\mu + \mathbf{U}\mathbf{y}_i)\|_2^2$$

$$\Rightarrow \min \|(\mathbf{x}_i - \mu) - \mathbf{U}\mathbf{y}_i\|_2^2 \Rightarrow \mathbf{y}_i = \left(\mathbf{U}^\top\mathbf{U}\right)^{-1}\mathbf{U}^\top(\mathbf{x}_i - \mu)$$

$$\Rightarrow \mathbf{y}_i = \mathbf{U}^\top(\mathbf{x}_i - \mu)$$

原问题转化为：$\min \sum_{i=1}^{n} \left\|\mathbf{x}_i - \mu - \mathbf{U}\mathbf{U}^\top(\mathbf{x}_i - \mu)\right\|_2^2$

1. 预备知识
2. 问题介绍
3. 算法推导和解释
4. 自然图像的主成分
5. 总结

问题介绍

- 回顾原问题：
$$\min \sum_{i=1}^{n} \left\| \mathbf{x}_i - \mu - \mathbf{U}\mathbf{U}^{\top} (\mathbf{x}_i - \mu) \right\|_2^2$$
$$s.t.\ \mathbf{U}\mathbf{U}^{\top} = \mathbf{I}$$
- 注意：要把ℓ_2范数和i.i.d高斯噪声假设联系起来。所以PCA的假设也是噪声服从i.i.d高斯分布。所以PCA也不能处理Outliers的问题（回顾上一讲的最小二乘问题）。
- Robust Principal Component Analysis (Low Rank Matrix Recovery)
http://perception.csl.illinois.edu/matrix-rank/home.html
- Online Object Tracking with Sparse Prototypes
http://ice.dlut.edu.cn/lu/Project/TIP12-SP/TIP12-SP.htm

均值推导和解释

- 均值的推导：

$$\mu = \arg\min_{\mu} \sum_{i=1}^{n} (\mathbf{x}_i - \mu)_2^2$$

$$J(\mu) = \sum_{i=1}^{n} (\mathbf{x}_i - \mu)_2^2 = \sum_{i=1}^{n} (\mathbf{x}_i - \mu)^T (\mathbf{x}_i - \mu)$$

$$\frac{\partial J}{\partial \mu} = -2 \sum_{i=1}^{n} (\mathbf{x}_i - \mu) = 0 \Rightarrow \mu = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$$

- 均值的解释
  - 如果我们只用一个向量来近似所有样本，均值就是最佳的选择。
  - 类比信号的直流分量，均值可以看作样本数据集的零维表达。
  - 均值向量简单而且描述了样本数据集的最大共性，但是缺点是不能反映出样本数据集的不同。
  - 补充阅读：《模式分类》中文版，第94页，另一种方式的推导。

基向量推导和解释

- 最小误差V.S.最大方差：（令$\overline{\mathbf{x}}_i = \mathbf{x}_i - \mu$）

$$\mathbf{u} = \arg\min_{\mathbf{u}} \sum_{i=1}^{n} \left\| \overline{\mathbf{x}}_i - \mathbf{u}\mathbf{u}^\top \overline{\mathbf{x}}_i \right\|_2^2 \quad \text{等价于}$$

$$s.t.\ \ \|\mathbf{u}\|_2^2 = 1$$

$$\mathbf{u} = \arg\max_{\mathbf{u}} \mathbf{u}^\top \left(\sum_{i=1}^{n} \overline{\mathbf{x}}_i \overline{\mathbf{x}}_i^\top \right) \mathbf{u}$$

$$s.t.\ \ \|\mathbf{u}\|_2^2 = 1$$

  - 令$S = \sum_{i=1}^{n} \overline{\mathbf{x}}_i \overline{\mathbf{x}}_i^T$称为散布矩阵（Scatter Matrix）。我们可以看出它和协方差矩阵（Convariance Matrix）$C = \frac{1}{n-1}\left(\sum_{i=1}^{n} \overline{\mathbf{x}}_i \overline{\mathbf{x}}_i^\top \right)$只相差个倍数。
  - 所以，对于PCA有两种原则解释：（1）最小重构误差原则；（2）最大方差原则。两者之间是等价的。

最小误差v.s.最大方差细节推导

$$\sum_{i=1}^{n}\left\|\overline{\mathbf{x}}_i-\mathbf{u}\mathbf{u}^\top\overline{\mathbf{x}}_i\right\|_2^2$$

$$=\sum_{i=1}^{n}\left(\overline{\mathbf{x}}_i-\mathbf{u}\mathbf{u}^\top\overline{\mathbf{x}}_i\right)^\top\left(\overline{\mathbf{x}}_i-\mathbf{u}\mathbf{u}^\top\overline{\mathbf{x}}_i\right)$$

$$=\sum_{i=1}^{n}\overline{\mathbf{x}}_i^\top\overline{\mathbf{x}}_i-\overline{\mathbf{x}}_i^\top\mathbf{u}\mathbf{u}^\top\overline{\mathbf{x}}_i-\overline{\mathbf{x}}_i^\top\mathbf{u}\mathbf{u}^\top\overline{\mathbf{x}}_i+\overline{\mathbf{x}}_i^\top\mathbf{u}\mathbf{u}^\top\mathbf{u}\mathbf{u}^\top\overline{\mathbf{x}}_i$$

$$\left(\mathbf{u}^\top\mathbf{u}=\|\mathbf{u}\|_2^2=1\right)$$

$$=\sum_{i=1}^{n}\overline{\mathbf{x}}_i^\top\overline{\mathbf{x}}_i-\overline{\mathbf{x}}_i^\top\mathbf{u}\mathbf{u}^\top\overline{\mathbf{x}}_i-\overline{\mathbf{x}}_i^\top\mathbf{u}\mathbf{u}^\top\overline{\mathbf{x}}_i+\overline{\mathbf{x}}_i^\top\mathbf{u}\mathbf{u}^\top\overline{\mathbf{x}}_i$$

$$\left(\mathbf{u}^\top\overline{\mathbf{x}}_i=\overline{\mathbf{x}}_i^\top\mathbf{u},\overline{\mathbf{x}}_i^\top\mathbf{u}\mathbf{u}^\top\overline{\mathbf{x}}_i=\mathbf{u}^\top\overline{\mathbf{x}}_i\overline{\mathbf{x}}_i^\top\mathbf{u}\right)$$

$$=-\left(\sum_{i=1}^{n}\mathbf{u}^\top\overline{\mathbf{x}}_i\overline{\mathbf{x}}_i^\top\mathbf{u}\right)+\sum_{i=1}^{n}\overline{\mathbf{x}}_i^\top\overline{\mathbf{x}}_i$$

$$=-\mathbf{u}^\top\left(\sum_{i=1}^{n}\overline{\mathbf{x}}_i\overline{\mathbf{x}}_i^\top\right)\mathbf{u}+\sum_{i=1}^{n}\overline{\mathbf{x}}_i^\top\overline{\mathbf{x}}_i$$

补充阅读：《模式分类》中文版，第95页，另一种方式的推导。

利用特征值分解来求基向量

- 回顾目标函数：

$$\mathbf{u} = \arg\max_{\mathbf{u}} \mathbf{u}^\top \mathbf{S}\mathbf{u} \qquad \left(\mathbf{S} = \sum_{i=1}^{n} \overline{\mathbf{x}}_i \overline{\mathbf{x}}_i^\top\right)$$

$$s.t.\ \mathbf{u}^\top \mathbf{u} = 1$$

- 拉格朗日乘子法：

$$J(\mathbf{u}) = \mathbf{u}^\top \mathbf{S}\mathbf{u} - \lambda \mathbf{u}^\top \mathbf{u}$$

$$\Rightarrow \frac{\partial J}{\partial \mathbf{u}} = 2\mathbf{S}\mathbf{u} - 2\lambda \mathbf{u} = 0 \Rightarrow \mathbf{S}\mathbf{u} = \lambda \mathbf{u}$$

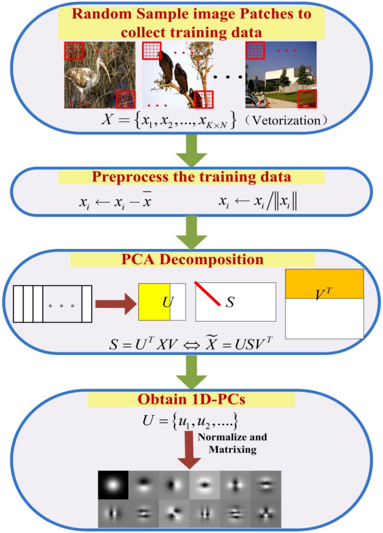
所以最优的$\mathbf{u}$为最大特征值对应的特征向量。

  - 这是选取一个基向量$\mathbf{u}$的证明。扩展到选取d个基向量，就是选取前d个特征值对应的特征向量。
  - 为什么要选取最大的特征值对应的特征向量？ 目标函数$\mathbf{u}^\top \mathbf{S}\mathbf{u} = \mathbf{u}^\top \lambda \mathbf{u} = \lambda$。关键看目标函数是最大还是最小。最大化目标函数当然选择最大特征值，如：PCA，LDA算法。最小化目标函数当然选择最小特征值，如：LPP，LLE，N-Cut算法。

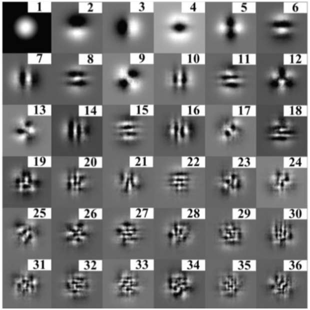
自然图像主成分分析步骤

(1)在自然图像数据库上采样 $S\times S$的小块组成训练样本。
(2)数据预处理：向量化，去均值，归一化。
(3)进行PCA分解。
最后，得到的特征向量被称为自然图像的主成分。

自然图像的主成分用途：
(1)加深我们对PCA的理解；
(2)用于纹理特征提取，如PPBTF。



自然图像主成分显示和分析



自然图像主成分显示(前36个)

第1个主成分相当于高斯滤波器，第2, 3个主成分相当于一阶导数滤波器，第4, 5个主成分相当于二阶导数滤波器。而后面的主成分越来越像杂乱的噪声模式。理解：PCA类似于低通滤波，越靠前的主成分越能用于表征图像的低频信息。

总结

- 需要从坐标变换（基变换）角度理解PCA。PCA的目的是找到**最能够表示**样本数据的主方向（或基函数，坐标轴）。
- PCA的**两个原则**：最小重构误差原则和最大方差原则，及等价性。
- 掌握利用（广义）特征值分解来求解基向量的推导和意义。**尤其要理解为什么选择最大特征值对应的特征向量。** 看目标函数是最大还是最小。（注：以后很多算法如LDA，LPP，LLE，NCut等会用到）。
- 从SVD分解或因子分解的角度看PCA（下页图解）。

SVD分解和PCA