机器学习白板推导系列笔记 (5.1~5.6)

Dexing Huang
dxhuang@bupt.edu.cn
Beijing University of Posts and Telecommunications

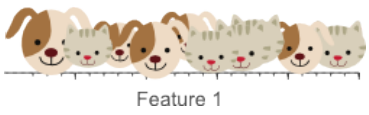
日期：2021 年 11 月 3 日

1 降维背景介绍

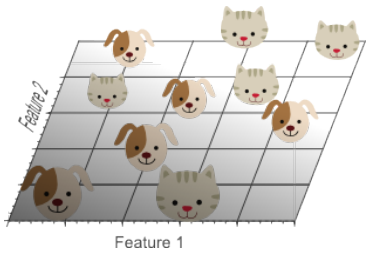
在机器学习问题中, 主要关系的是泛化误差, 而非训练误差. 在降低泛化误差的过程中, 遇到的问题之一就是过拟合 (overfit), 解决这一问题有三个思路, 1) 增加数据的数量, 2) 正则化, 3) 降维. 正则化在线性回归中讲述过 (Lasso 和 Ridge), 其实其本质也是降维, 通过增加惩罚项来消除一些特征.

高纬会带来什么问题呢? 从数学角度上看, 高维度有更大的特征空间, 需要更多的数据才可以进行较准确的估计, 每增加一个维度 (设该维度上取值只有两种选择), 那么如果想要覆盖这个维度, 所需要的数据的增量是以 2 的指数级上升的. 如数据只有一维时, 只需要两个数据点 $\{0,1\}$ 即可, 二维时则需要 $\{(0,0),(0,1),(1,0),(1,1)\}$ 四个数据点, 因此 n 维时需要 2^n 个数据点, 当 n 很大时, 如果没有足够多的数据, 很容易造成过拟合.

从几何的角度上看[1], 假设需要训练一个猫狗分类器, 训练数据集个数为 10 个, 目标是训练出的分类器要有较好的泛化能力. 先使用简单的线性分类器, 考虑只使用一个特征, 结果如下

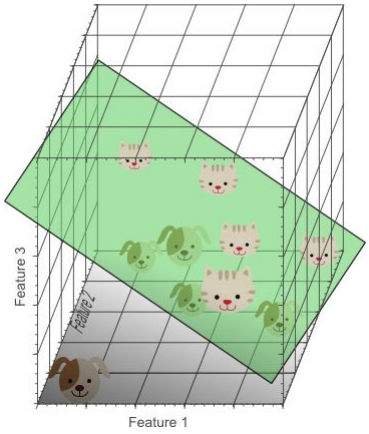


很显然, 仅凭一个特征无法进行分类. 因此一个自然的想法就是增加特征的维度到二维, 结果如下



[1]https://zhuanlan.zhihu.com/p/26945814

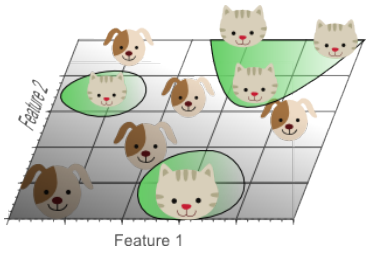
增加维度后, 还是无法使用一条直线将两类完全分隔开 (即线性不可分). 那么将数据增加到三维, 结果如下



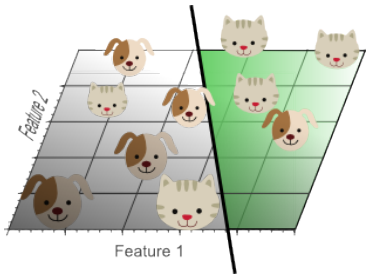
可以看到, 存在一个平面可以将两类完美的分开. 以上的例子似乎证明了不断增加特征数量, 直到获得最佳分类效果, 是构建一个分类器的最好方法. 但是, 随着特征维度的增加, 训练样本的在特征空间的密度是如何呈指数型下降的.

假设每一维的特征宽度都为 5, 在一维空间中, 样本的密度为 $\frac{10}{5}=2$, 在二维空间中, 样本密度变为 $\frac{10}{5^2}=0.4$, 在三维空间中， 样本密度为 $\frac{10}{5^3}=0.08$.

如果继续增加特征, 整个特征空间维度增加, 并变得越来越稀疏. 由于稀疏性, 很容易找到一个超平面来实现分类. 然而, 如果将高维的分类结果投影到低维空间中, 将会出现一个严重的问题



上图展示了三维的分类结果投影到二维特征空间的情况. 样本数据在三维是线性可分的, 但是在二维却并非如此, 事实上, 增加第三个维度来获得最佳的线性分类效果, 等同于在低维特征空间中使用非线性分类器, 其结果是, 分类器学习了训练数据的噪声和异常, 而对样本外的数据拟合效果并不理想, 甚至很差. 这种情况称为过拟合, 是维度灾难的一个直接后果.



上图简单的线性分类器比三维的分类器的效果差, 但是泛化能力强, 这是因为分类器没有把样本数据的噪声和异常也进行学习.

在上面的例子中, 展示了维度灾难会引起训练数据的稀疏化. 使用的特征越多, 数据就会变得越稀疏, 从而导致分类器的分类效果就会越差, 维度灾难还会造成搜索空间的数据稀疏程度分布不均, 围绕原点的数据 (在超立方体的中心) 比在搜索空间的角落处的数据要稀疏得多

想象一个单位正方形代表了二维的特征空间, 特征空间的平均值位于这个单位正方形的中心处, 距中心处单位距离的所有点构成了正方形的内接圆, 没有落在单位圆的训练样本距离搜索空间的角落处比距离中心处更近, 而这些样本由于特征值差异很大(样本分布在正方形角落处), 难以分类, 因此, 如果大部分样本落在单位内接圆里, 就会更容易分类.

当增加特征空间的维度时, 超立方体的体积都是 1, 而半径为 0.5 的超球体的体积随着维度的变化为

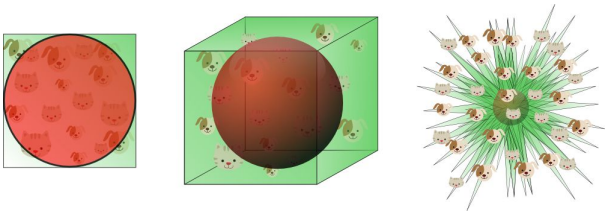
$$V_{ball} = K(0.5)^d \tag{1}$$

注. K 为常数

当 $d \to \infty$

$$\lim_{d\to\infty} V_{ball} = 0 \tag{2}$$

那么在高维空间中, 大部分的训练数据分布在定义为特征空间的超立方体的角落处, 这是非常不利于分类的



从上述例子可以看到, 降维是很有必要的, 常见的降维方法有

- 直接降维特征选择简单粗暴, 直接把不要的特征扔掉
- 线性降维 PCA, MDS(多维空间缩放)
- 非线性降维流形 (嵌入了高维空间的地维结构), 等度量映射 (ISOMAP), 局部线性嵌入 (LLE)

2 样本均值 & 样本方差的矩阵表示

设样本数据 $\mathbf{X}$

$$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)^T = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_N^T \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{pmatrix}_{N\times p} \tag{3}$$

其中 $\mathbf{x}_i \in \mathbb{R}^p$, 样本均值和方差表示如下

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i \tag{4}$$

$$\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T \tag{5}$$

接下来将式 (4), (5) 转换为矩阵表示

$$\begin{aligned} \bar{\mathbf{x}} &= \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i \\ &= \frac{1}{N}(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}_{N\times 1} \\ &= \frac{1}{N}\mathbf{X}^T\mathbf{1}_N \end{aligned} \tag{6}$$

$$\begin{aligned} \mathbf{S} &= \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T \\ &= \frac{1}{N}((\mathbf{x}_1 - \bar{\mathbf{x}}), (\mathbf{x}_2 - \bar{\mathbf{x}}), \cdots, (\mathbf{x}_N - \bar{\mathbf{x}}))\begin{pmatrix} (\mathbf{x}_1 - \bar{\mathbf{x}})^T \\ (\mathbf{x}_2 - \bar{\mathbf{x}})^T \\ \vdots \\ (\mathbf{x}_N - \bar{\mathbf{x}})^T \end{pmatrix} \\ &= \frac{1}{N}((\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N) - (\bar{\mathbf{x}}, \bar{\mathbf{x}}, \cdots, \bar{\mathbf{x}}))\begin{pmatrix} (\mathbf{x}_1 - \bar{\mathbf{x}})^T \\ (\mathbf{x}_2 - \bar{\mathbf{x}})^T \\ \vdots \\ (\mathbf{x}_N - \bar{\mathbf{x}})^T \end{pmatrix} \\ &= \frac{1}{N}(\mathbf{X}^T - \bar{\mathbf{x}}\mathbf{1}_N^T)\begin{pmatrix} (\mathbf{x}_1 - \bar{\mathbf{x}})^T \\ (\mathbf{x}_2 - \bar{\mathbf{x}})^T \\ \vdots \\ (\mathbf{x}_N - \bar{\mathbf{x}})^T \end{pmatrix} \\ &= \frac{1}{N}\mathbf{X}^T(\mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T)(\mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T)^T\mathbf{X} \end{aligned} \tag{7}$$

设 $\mathbf{H}=\mathbf{I}_N-\frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T$ 得, $\mathbf{H}$ 称为 centering matrix, 作用是将一组数据中化.

$$\mathbf{S}=\frac{1}{N}\mathbf{X}^T\mathbf{H}\mathbf{H}^T\mathbf{X} \tag{8}$$

$\mathbf{H}$ 具有如下的性质

- $\mathbf{H}=\mathbf{H}^T$

$$\begin{aligned}\mathbf{H}^T&=(\mathbf{I}_N-\frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T)^T\\&=(\mathbf{I}_N^T-\frac{1}{N}(\mathbf{1}_N^T)^T\mathbf{1}_N^T)\\&=(\mathbf{I}_N-\frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T)\\&=\mathbf{H}\end{aligned}$$

- $\mathbf{H}^n=\mathbf{H}$

$$\begin{aligned}\mathbf{H}^2&=(\mathbf{I}_N-\frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T)(\mathbf{I}_N-\frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T)\\&=\mathbf{I}_N\mathbf{I}_N-\frac{1}{N}\mathbf{I}_N\mathbf{1}_N\mathbf{1}_N^T-\frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T\mathbf{I}_N+\frac{1}{N^2}\mathbf{1}_N\mathbf{1}_N^T\mathbf{1}_N\mathbf{1}_N^T\\&=\mathbf{I}_N-\frac{2}{N}\mathbf{1}_N\mathbf{1}_N^T+\frac{1}{N^2}\mathbf{1}_N\mathbf{1}_N^T\mathbf{1}_N\mathbf{1}_N^T\end{aligned}$$

对于 $-\frac{2}{N}\mathbf{1}_N\mathbf{1}_N^T$ 有

$$-\frac{2}{N}\mathbf{1}_N\mathbf{1}_N^T=-\frac{2}{N}\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix}(1,1,\cdots,1)=\begin{pmatrix}-\frac{2}{N}&-\frac{2}{N}&\cdots&-\frac{2}{N}\\-\frac{2}{N}&-\frac{2}{N}&\cdots&-\frac{2}{N}\\\vdots&\vdots&\ddots&\vdots\\-\frac{2}{N}&-\frac{2}{N}&\cdots&-\frac{2}{N}\end{pmatrix}$$

对于 $\frac{1}{N^2}\mathbf{1}_N\mathbf{1}_N^T\mathbf{1}_N\mathbf{1}_N^T$ 有

$$\begin{aligned}\frac{1}{N^2}\mathbf{1}_N\mathbf{1}_N^T\mathbf{1}_N\mathbf{1}_N^T&=\frac{1}{N^2}\begin{pmatrix}1&1&\cdots&1\\1&1&\cdots&1\\\vdots&\vdots&\ddots&\vdots\\1&1&\cdots&1\end{pmatrix}\begin{pmatrix}1&1&\cdots&1\\1&1&\cdots&1\\\vdots&\vdots&\ddots&\vdots\\1&1&\cdots&1\end{pmatrix}\\&=\begin{pmatrix}\frac{1}{N}&\frac{1}{N}&\cdots&\frac{1}{N}\\\frac{1}{N}&\frac{1}{N}&\cdots&\frac{1}{N}\\\vdots&\vdots&\ddots&\vdots\\\frac{1}{N}&\frac{1}{N}&\cdots&\frac{1}{N}\end{pmatrix}\end{aligned}$$

那么

$$\mathbf{H}^2=\mathbf{I}_N-\frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T=\mathbf{H}$$

很容易推广到 $\mathbf{H}^n=\mathbf{H}$, 式 (8) 可写为

$$\mathbf{S}=\frac{1}{N}\mathbf{X}^T\mathbf{H}\mathbf{X} \tag{9}$$

综上, 原始数据均值和方差的矩阵表示如下

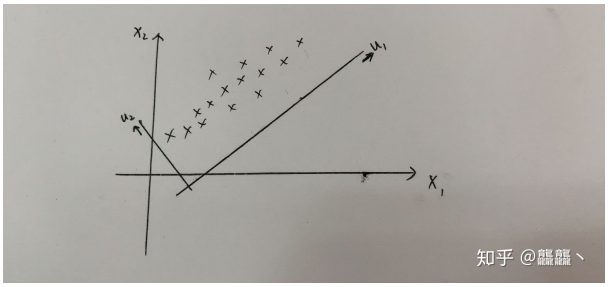
$$\bar{\mathbf{x}}=\frac{1}{N}\mathbf{X}^T\mathbf{1}_N \tag{10}$$

$$\mathbf{S}=\frac{1}{N}\mathbf{X}^T\mathbf{H}\mathbf{X} \tag{11}$$

3 主成分分析 (PCA)

主成分分析围绕着一个中心和两个基本点展开, 中心指的是对原始特征空间的重构, 将相关的特征转为无关的特征, 将特征空间变成一组相互正交的基. 两个基本点分别是最大投影方差和最小重构距离.

3.1 最大投影方差



主成分分析的目的是将线性相关的特征转化为线性无关, 即找到一个投影轴，使得其投影后方差最大, 如上图所示, 这一组数据点投影到 $\mathbf{u}_1$ 方向后方差更大, 数据更分散, 而投影到 $\mathbf{u}_2$ 方向会很密集, 因此称 $\mathbf{u}_1$ 方向为主成份, 主成份分析的意思是找到一组线性无关的基, 这组基就是主成份, 若想降到 p 维, 便选择其前 p 个基即可.

首先对数据进行中心化, 即

$$\mathbf{x}_i' = \mathbf{x}_i - \bar{\mathbf{x}} \tag{12}$$

把 $\mathbf{x}_i'$ 投影到 $\mathbf{u}_1$, 令 $\mathbf{u}_1^T\mathbf{u}_1 = 1$, 因此得

$$(\mathbf{x}_i')^T\mathbf{u}_1 = (\mathbf{x}_i - \bar{\mathbf{x}})^T\mathbf{u}_1 \tag{13}$$

又由于已经归一化了, 因此投影的方差即

$$((\mathbf{x}_i - \bar{\mathbf{x}})^T\mathbf{u}_1 - 0)((\mathbf{x}_i - \bar{\mathbf{x}})^T\mathbf{u}_1 - 0)^T \tag{14}$$

对整个数据集求和, 构造目标函数

$$\begin{aligned}\mathcal{J} &= \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{x}})^T\mathbf{u}_1)^2 = \frac{1}{N}\sum_{i=1}^{N}\mathbf{u}_1^T(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T\mathbf{u}_1 \\ &= \mathbf{u}_1^T(\frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T)\mathbf{u}_1 \\ &= \mathbf{u}_1^T\mathbf{S}\mathbf{u}_1\end{aligned} \tag{15}$$

有了目标函数就可以构造最优化问题了

$$\begin{aligned}&\hat{\mathbf{u}}_1 = \arg\max \mathbf{u}_1^T\mathbf{S}\mathbf{u}_1 \\ &s.t. \quad \mathbf{u}_1^T\mathbf{u}_1 = 1\end{aligned} \tag{16}$$

使用 Lagrange 乘数法就可以解决

$$\mathcal{L}(\mathbf{u}_1, \lambda) = \mathbf{u}_1^T\mathbf{S}\mathbf{u}_1 + \lambda(1 - \mathbf{u}_1^T\mathbf{u}_1) \tag{17}$$

将问题转化为无约束最优化问题

$$\hat{\mathbf{u}}_1 = \arg\max(\mathbf{u}_1^T \mathbf{S}\mathbf{u}_1 + \lambda(1 - \mathbf{u}_1^T \mathbf{u}_1)) \tag{18}$$

求导并令导数为零得

$$\frac{\partial}{\partial \mathbf{u}_1}\mathcal{L}(\mathbf{u}_1, \lambda) = 2\mathbf{S}\mathbf{u}_1 - 2\lambda\mathbf{u}_1 = 0 \tag{19}$$

得到如下等式

$$\mathbf{S}\mathbf{u}_1 = \lambda\mathbf{u}_1 \tag{20}$$

可以发现 $\mathbf{u}_1$ 和 λ 分别是 $\mathbf{S}$ 的特征向量和特征值. 解出此式后, 特征向量 $\mathbf{u}_1$ 便是投影方向, 特征值最大的特征向量是投影方差最大的主成份.

3.2 最小重构代价

其本质与最大投影方差一致, 以最小代价将投影后的数据重构回去, 若投影后数据越分散, 则重构越容易, 若数据越集中, 甚至重合到一个点, 便很难重构回去. 因此最小重构距离也需要寻找投影后数据最分散的方向.

PCA 的目标是将 p 维的数据降到 q 维去, 设将 p 维的数据投影到另一个 p 维的特征向量的基底 $\mathbf{u}_k$ 上, 则数据点重构为

$$\mathbf{x}_i'' = \sum_{k=1}^{p}((\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{u}_k)\mathbf{u}_k \tag{21}$$

下面进行降维, 特征向量 $\mathbf{u}_k$ 按照特征值 λ_k 的大小进行排列, 选取前 q 个维度, 剩余维度舍弃, 那么

$$\hat{\mathbf{x}}_i'' = \sum_{k=1}^{q}((\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{u}_k)\mathbf{u}_k \tag{22}$$

最小重构距离是指将降维后的 $\hat{\mathbf{x}}_i''$ 还原为 $\mathbf{x}_i''$ 所需代价最小, 因此可以构造目标函数

$$\mathcal{J} = \frac{1}{N}\sum_{i=1}^{N}||\mathbf{x}_i'' - \hat{\mathbf{x}}_i''||^2 \tag{23}$$

进行适当变换

$$\begin{aligned}\mathcal{J} &= \frac{1}{N}\sum_{i=1}^{N}||\sum_{k=q+1}^{p}((\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{u}_k)\mathbf{u}_k||^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}\sum_{k=q+1}^{p}((\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{u}_k)^2 \\ &= \sum_{k=q+1}^{p}\mathbf{u}_k^T \mathbf{S}\mathbf{u}_k\end{aligned} \tag{24}$$

可以得到如下最优化问题

$$\begin{aligned}&\hat{\mathbf{u}}_k = \arg\min \sum_{k=q+1}^{p}\mathbf{u}_k^T \mathbf{S}\mathbf{u}_k \\ &s.t. \quad \mathbf{u}_k^T \mathbf{u}_k = 1\end{aligned}$$

因为 $\mathbf{u}_k$ 之间相互独立, 因此可以分别求出, 即

$$\begin{aligned} \hat{\mathbf{u}}_k &= \arg\min \mathbf{u}_k^T \mathbf{S} \mathbf{u}_k \\ s.t. \quad & \mathbf{u}_k^T \mathbf{u}_k = 1 \end{aligned} \tag{25}$$

此优化问题与上一节一致, 使用拉格朗日乘子法, 很容易求得

$$\mathbf{S}\mathbf{u}_k = \lambda_k \mathbf{u}_k \tag{26}$$

3.3 SVD 角度看 PCA

在上几节课中介绍的方法都需要求解协方差矩阵的特征值与特征向量, 本节将从奇异值分解 (SVD) 的角度来看 PCA. 因为笔者之前没学过 SVD, 因此在正式讨论之前, 先对 SVD 进行学习.

3.3.1 特征分解和奇异值分解 (SVD)

分解是数学中常用的手段, 数学对象可以通过分解, 将其分解为多个部分或者找到一些属性方便理解和分析, 最典型的如质因数分解, 傅立叶变化也属于一种分解, 将信号从时域转换到频域, 方便分析. 同样, 也可以通过分解矩阵来得到一些矩阵中不明显的性质.

特征值分解是将方阵分解成一组特征向量和特征值的方法. 对于一个方阵 $\mathbf{A}$, 特征向量指的是与 $\mathbf{A}$ 相乘后相当于对该向量进行缩放的非零向量 $\mathbf{x}$

$$\mathbf{A}\mathbf{x} = \lambda \mathbf{x} \tag{27}$$

下面介绍一下相似的概念, 对于方阵 $\mathbf{A}$ 和 $\mathbf{B}$ 存在一个可逆矩阵 $\mathbf{P}$ 使得

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{B} \tag{28}$$

那么称两矩阵相似, 特别的[2], 如果矩阵 $\mathbf{B}$ 为对角矩阵, 那么 $\mathbf{P}$ 是由 $\mathbf{A}$ 的特征向量构成的矩阵, 且 $\mathbf{B} = \Lambda = diag(\lambda_1, \lambda_2, \cdots, \lambda_N)$, 称矩阵 $\mathbf{A}$可对角化, 由线性代数的知识可知, 方阵可对角化的充要条件为有 n 个线性无关的特征向量.

在 PCA 中, 进行对角化的协方差矩阵 S 具有更加特殊的性质, 它是一个 , 对于一个实对称矩阵, 其特征值一定为实数, 且不同特征值之间的特征向量相互正交(注意, 这保证了降为后构成正交基), 并且一定存在一个正交矩阵 Q 使得

$$\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \mathbf{Q}^T\mathbf{A}\mathbf{Q} = \Lambda \tag{29}$$

因此 $\mathbf{A}$ 可以被分解为实特征向量和实特征值

$$\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^T \tag{30}$$

且

$$\mathbf{Q} = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_N \end{pmatrix} \quad \Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_N \end{pmatrix}$$

[2]百度百科

下面介绍奇异值分解 (SVD), 只有方阵才能进行特征分解, 对于一般的实矩阵, 更加常用的是奇异值分解, 它将矩阵分解为奇异向量和奇异值. 与式 (30) 类似, 将矩阵 A 分解成三个矩阵的乘积.

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T \tag{31}$$

其中, $\mathbf{A}$ 是 $m \times n$ 的矩阵, $\mathbf{U}$ 是一个 $m \times m$ 的方阵, Σ 是一个 $m \times n$ 的矩阵, $\mathbf{V}$ 是一个 $n \times n$ 的方阵. $\mathbf{U}, \mathbf{V}$ 都是正交矩阵, Σ 为对角矩阵.

Σ 矩阵对角线上的元素称为奇异值, 矩阵 $\mathbf{U}$ 的列向量为左奇异向量 (即 $\mathbf{A}\mathbf{A}^T$ 的特征向量), 矩阵 $\mathbf{V}$ 的列向量为右奇异向量 (即 $\mathbf{A}^T\mathbf{A}$ 的特征向量). 非 0 奇异值是 $\mathbf{A}^T\mathbf{A}, \mathbf{A}\mathbf{A}^T$ 的特征值的平方根.

3.3.2 样本数据的 SVD 分解

首先, 对数据进行中心化 (均值变为 0)

$$\begin{aligned}
\mathbf{HX} &= (\mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T)\mathbf{X} \\
&= \mathbf{X} - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T\mathbf{X} \\
&= (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)^T - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)^T \\
&= (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)^T - \frac{1}{N}\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)^T \\
&= (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)^T - (\frac{1}{N}\sum_{i=1}^N \mathbf{x}_i, \frac{1}{N}\sum_{i=1}^N \mathbf{x}_i, \cdots, \frac{1}{N}\sum_{i=1}^N \mathbf{x}_i)^T \\
&= (\mathbf{x}_1 - \hat{\mathbf{x}}, \mathbf{x}_2 - \hat{\mathbf{x}}, \cdots, \mathbf{x}_N - \hat{\mathbf{x}})^T
\end{aligned}$$

随后将中心化后的 $\mathbf{HX}$ 进行奇异值分解

$$\mathbf{HX} = \mathbf{U}\Sigma\mathbf{V}^T \tag{32}$$

又由式 (11) 得

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

很显然, $\mathbf{V}$ 是 $\mathbf{S}$ 特征向量组成的矩阵即主成分, $\Sigma^T\Sigma$ 是特征值矩阵, 与上两节讨论等价.

下面再简单说一下主坐标分析 (PCoA), 令 $\mathbf{T} = \mathbf{HX}\mathbf{X}^T\mathbf{H}^T$ 得

$$\mathbf{T} = \mathbf{U}\Sigma\Sigma^T\mathbf{U}^T \tag{33}$$

使用 $\mathbf{S}$ 进行特征分解后得到主成份 (即主坐标轴, 点的坐标还得继续表示), 将 $\mathbf{X}$ 投影到主成

份的方向后, 得到坐标矩阵是

$$\mathbf{HX} \cdot \mathbf{V} = \mathbf{U}\Sigma\mathbf{V}^T \cdot \mathbf{V} = \mathbf{U}\Sigma \tag{34}$$

使用 $\mathbf{T}$ 进行特征分解, 便可以直接得到坐标

$$\begin{aligned}\mathbf{TU}\Sigma &= \mathbf{U}\Sigma\Sigma^T\mathbf{U}^T\mathbf{U}\Sigma \\ &= \mathbf{U}\Sigma\Sigma^T\Sigma\end{aligned}$$

$\mathbf{U}\Sigma$(即坐标矩阵) 是 $\mathbf{T}$ 的特征向量组成的矩阵, $\Sigma\Sigma^T$ 是 $\mathbf{T}$ 特征值组成的矩阵.

PCA 和 PCoA 的区别是 S 为 $p \times p$ 的矩阵, T 为 $N \times N$ 的矩阵, 根据数据大小和特征多少选择不同的方法.

3.4 P-PCA

本节内容从概率的角度看 PCA, 因此该方法也被称为 P-PCA(Probabilistic PCA), 假设原始数据 $\mathbf{x} \in \mathbb{R}^p$, 降维后的数据 $\mathbf{z} \in \mathbb{R}^p$, 通常把 $\mathbf{x}$ 称为 observed data, $\mathbf{z}$ 称为 latent variable.

下面做出如下假设

$$\begin{aligned}z &\sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p) \\ \mathbf{x} &= \mathbf{wz} + \mu + \epsilon \\ \epsilon &\sim \mathcal{N}(\mathbf{0}_p, \sigma \star \mathbf{I}_p) \\ \epsilon &\perp \mathbf{z}\end{aligned}$$

根据以上假设, 最终目的是求解出后验 $\mathbf{z}|\mathbf{x}$, 需要先求出 $\mathbf{x}|\mathbf{z}, \mathbf{x}$, 通常来说, 求解 P-PCA 有两个步骤

- Inference: $p(\mathbf{z}|\mathbf{x})$
- Learning: $\mathbf{w}, \mu, \sigma^2 \rightarrow$ EM 算法

本节讲述 Inference 部分, Learning 可以使用最大似然估计或 EM 算法 (之后会讲) 解决. 对于 $\mathbf{x}|\mathbf{z}$($\mathbf{z}$ 相当于常数)

$$\begin{aligned}E[\mathbf{x}|\mathbf{z}] &= E[\mathbf{wz} + \mu + \epsilon|\mathbf{z}] = \mathbf{wz} + \mu \\ Var[\mathbf{x}|\mathbf{z}] &= Var[\mathbf{wz} + \mu + \epsilon|\mathbf{z}] = \sigma^2\mathbf{I}_p\end{aligned}$$

对于 $\mathbf{x}$

$$\begin{aligned}E[\mathbf{x}] &= E[\mathbf{wz} + \mu + \epsilon] = \mu \\ Var[\mathbf{x}] &= Var[\mathbf{wz} + \mu + \epsilon] = \mathbf{w}Var[\mathbf{z}]\mathbf{w}^T + \sigma^2\mathbf{I}_p = \mathbf{ww}^T + \sigma^2\mathbf{I}_p\end{aligned} \tag{35}$$

那么

$$\begin{aligned}\mathbf{x}|\mathbf{z} &\sim \mathcal{N}(\mathbf{wz} + \mu, \sigma^2\mathbf{I}_p) \\ \mathbf{x} &\sim \mathcal{N}(\mu, \mathbf{ww}^T + \sigma^2\mathbf{I}_p)\end{aligned} \tag{36}$$

接下来求解 $\mathbf{z}|\mathbf{x}$, 需要用到第二部分数学基础中的内容, 构造一联合分布

$$\begin{pmatrix}\mathbf{x} \\ \mathbf{z}\end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mu \\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{ww}^T + \sigma^2\mathbf{I}_p & \Sigma_{\mathbf{xz}} \\ \Sigma_{\mathbf{zx}} & \mathbf{I}_p\end{bmatrix}\right) \tag{37}$$

又因为 $\Sigma_{\mathbf{xz}} = \Sigma_{\mathbf{zx}}^T$, 因此只需要求出一个即可

$$\begin{aligned}
\Sigma_{\mathbf{xz}} &= E[(\mathbf{x} - \mu)(\mathbf{z} - \mathbf{0}_p)^T] \\
&= E[(\mathbf{x} - \mu)\mathbf{z}^T] \\
&= E[(\mathbf{wz} + \epsilon)\mathbf{z}^T] \\
&= E[\mathbf{wzz}^T] + E[\epsilon\mathbf{z}^T] \\
&= \mathbf{w}E[\mathbf{zz}^T] + E[\epsilon]E[\mathbf{z}^T] \\
&= \mathbf{w}E[\mathbf{zz}^T] \\
&= \mathbf{w}(Var[\mathbf{z}] + E[\mathbf{z}]E[\mathbf{z}]^T) = \mathbf{w}\mathbf{I}_p \\
&= \mathbf{w}
\end{aligned} \tag{38}$$

因此

$$\begin{pmatrix} \mathbf{x} \\ \mathbf{z} \end{pmatrix} \sim \mathcal{N}(\begin{bmatrix} \mu \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{w}\mathbf{w}^T + \sigma^2\mathbf{I}_p & \mathbf{w} \\ \mathbf{w}^T & \mathbf{I}_p \end{bmatrix}) \tag{39}$$

接下来套用公式就能解出 $\mathbf{z}|\mathbf{x}$, 主要是领悟思想即可.