降维

我们知道，解决过拟合的问题除了正则化和添加数据之外，降维就是最好的方法。降维的思路来源于维度灾难的问题，我们知道 n 维球的体积为：

$$CR^n \tag{1}$$

那么在球体积与边长为 $2R$ 的超立方体比值为：

$$\lim_{n\to 0}\frac{CR^n}{2^nR^n}=0 \tag{2}$$

这就是所谓的维度灾难，在高维数据中，主要样本都分布在立方体的边缘，所以数据集更加稀疏。

降维的算法分为：

1. 直接降维，特征选择
2. 线性降维，PCA，MDS等
3. 分线性，流形包括 Isomap，LLE 等

为了方便，我们首先将协方差矩阵（数据集）写成中心化的形式：

$$\begin{aligned}
S&=\frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})(x_i-\overline{x})^T\\
&=\frac{1}{N}(x_1-\overline{x},x_2-\overline{x},\cdots,x_N-\overline{x})(x_1-\overline{x},x_2-\overline{x},\cdots,x_N-\overline{x})^T\\
&=\frac{1}{N}(X^T-\frac{1}{N}X^T\mathbb{I}_{N1}\mathbb{I}_{N1}^T)(X^T-\frac{1}{N}X^T\mathbb{I}_{N1}\mathbb{I}_{N1}^T)^T\\
&=\frac{1}{N}X^T(E_N-\frac{1}{N}\mathbb{I}_{N1}\mathbb{I}_{1N})(E_N-\frac{1}{N}\mathbb{I}_{N1}\mathbb{I}_{1N})^TX\\
&=\frac{1}{N}X^TH_NH_N^TX\\
&=\frac{1}{N}X^TH_NH_NX=\frac{1}{N}X^THX
\end{aligned} \tag{3}$$

这个式子利用了中心矩阵 H的对称性，这也是一个投影矩阵。

线性降维-主成分分析 PCA

损失函数

主成分分析中，我们的基本想法是将所有数据投影到一个字空间中，从而达到降维的目标，为了寻找这个子空间，我们基本想法是：

1. 所有数据在子空间中更为分散
2. 损失的信息最小，即：在补空间的分量少

原来的数据很有可能各个维度之间是相关的，于是我们希望找到一组 p 个新的线性无关的单位基 u_i，降维就是取其中的 q 个基。于是对于一个样本 x_i，经过这个坐标变换后：

$$\hat{x_i} = \sum_{i=1}^{p}(u_i^T x_i)u_i = \sum_{i=1}^{q}(u_i^T x_i)u_i + \sum_{i=q+1}^{p}(u_i^T x_i)u_i \tag{4}$$

对于数据集来说，我们首先将其中心化然后再去上面的式子的第一项，并使用其系数的平方平均作为损失函数并最大化：

$$\begin{aligned} J &= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{q}((x_i - \overline{x})^T u_j)^2 \\ &= \sum_{j=1}^{q} u_j^T S u_j \ , \ s.t. \ u_j^T u_j = 1 \end{aligned} \tag{5}$$

由于每个基都是线性无关的，于是每一个 u_j 的求解可以分别进行，使用拉格朗日乘子法：

$$\underset{u_j}{argmax}\, L(u_j, \lambda) = \underset{u_j}{argmax}\, u_j^T S u_j + \lambda(1 - u_j^T u_j) \tag{6}$$

于是：

$$S u_j = \lambda u_j \tag{7}$$

可见，我们需要的基就是协方差矩阵的本征矢。损失函数最大取在本征值前 q 个最大值。

下面看其损失的信息最少这个条件，同样适用系数的平方平均作为损失函数，并最小化：

$$\begin{aligned} J &= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=q+1}^{p}((x_i - \overline{x})^T u_j)^2 \\ &= \sum_{j=q+1}^{p} u_j^T S u_j \ , \ s.t. \ u_j^T u_j = 1 \end{aligned} \tag{8}$$

同样的：

$$\underset{u_j}{argmin}\, L(u_j, \lambda) = \underset{u_j}{argmin}\, u_j^T S u_j + \lambda(1 - u_j^T u_j) \tag{9}$$

损失函数最小取在本征值剩下的个最小的几个值。数据集的协方差矩阵可以写成 $S = U\Lambda U^T$，直接对这个表达式当然可以得到本征矢。

SVD 与 PCoA

下面使用实际训练时常常使用的 SVD 直接求得这个 q 个本征矢。

对中心化后的数据集进行奇异值分解：

$$HX = U\Sigma V^T, U^T U = E_N, V^T V = E_p, \Sigma : N \times p \tag{10}$$

于是：

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

因此，我们直接对中心化后的数据集进行 SVD，就可以得到特征值和特征向量 V，在新坐标系中的坐标就是：

$$HX \cdot V \tag{12}$$

由上面的推导，我们也可以得到另一种方法 PCoA 主坐标分析，定义并进行特征值分解：

$$T = HXX^T H = U\Sigma\Sigma^T U^T \tag{13}$$

由于：

$$TU\Sigma = U\Sigma(\Sigma^T\Sigma) \tag{14}$$

于是可以直接得到坐标。这两种方法都可以得到主成分，但是由于方差矩阵是 $p \times p$ 的，而 T 是 $N \times N$ 的，所以对样本量较少的时候可以采用 PCoA的方法。

p-PCA

下面从概率的角度对 PCA 进行分析，概率方法也叫 p-PCA。我们使用线性模型，类似之前 LDA，我们选定一个方向，对原数据 $x \in \mathbb{R}^p$ ，降维后的数据为 $z \in \mathbb{R}^q, q < p$。降维通过一个矩阵变换（投影）进行：

$$z \sim \mathcal{N}(\mathbb{O}_{q1}, \mathbb{I}_{qq}) \tag{15}$$
$$x = Wz + \mu + \varepsilon \tag{16}$$
$$\varepsilon \sim \mathcal{N}(0, \sigma^2\mathbb{I}_{pp}) \tag{17}$$

对于这个模型，我么可以使用期望-最大（EM）的算法进行学习，在进行推断的时候需要求得 $p(z|x)$，推断的求解过程和线性高斯模型类似。

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)} \tag{18}$$
$$\mathbb{E}[x] = \mathbb{E}[Wz + \mu + \varepsilon] = \mu \tag{19}$$
$$Var[x] = WW^T + \sigma^2\mathbb{I}_{pp} \tag{20}$$
$$\Longrightarrow p(z|x) = \mathcal{N}(W^T(WW^T + \sigma^2\mathbb{I})^{-1}(x - \mu), \mathbb{I} - W^T(WW^T + \sigma^2\mathbb{I})^{-1}W) \tag{21}$$

小结

降维是解决维度灾难和过拟合的重要方法，除了直接的特征选择外，我们还可以采用算法的途径对特征进行筛选，线性的降维方法以 PCA 为代表，在 PCA 中，我们只要直接对数据矩阵进行中心化然后求奇异值分解或者对数据的协方差矩阵进行分解就可以得到其主要维度。非线性学习的方法如流形学习将投影面从平面改为超曲面。