目录

第十章　降维与度量学习

0.1　k 近邻学习

k 近邻 (k-Nearest Neighbor，简称 kNN) 学习是无监督学习方法，也是“懒惰学习”的著名代表。假设样本独立同分布，可以得到结论：**最近邻分类器的泛化错误率不超过贝叶斯最优分类器的错误率的两倍**。

令训练样本为 $\boldsymbol{z}$，$c^* = \arg\max_{c\in\mathcal{Y}} P(c|\boldsymbol{x})$ 表示贝叶斯最优分类器的结果，则：

$$\begin{aligned} P(err) &= 1 - \sum_{c\in\mathcal{Y}} P(c|\boldsymbol{x})P(c|\boldsymbol{z}) \\ &\simeq 1 - \sum_{c\in\mathcal{Y}} P^2(c|\boldsymbol{x}) \\ &\leq 1 - P^2(c^*|\boldsymbol{x}) \\ &= (1 + P(c^*|\boldsymbol{x}))(1 - P(c^*|\boldsymbol{x})) \\ &\leq 2 \times (1 - P(c^*|\boldsymbol{x})) \end{aligned}$$

0.2　低维嵌入

使用 kNN 时需要满足样本密度足够大，然而高维数据很容易出现样本稀疏，距离计算困难的问题，此时称为“维数灾难”(curse of dimensionality)。缓解维数灾难的一个方法是降维，降维时需要保证原始空间样本中的距离信息在低维空间中同样存在。

0.2.1　多维缩放降维法

多维缩放 (Multiple Dimensional Scaling，简称 MDS) 降维方法是基于线性变换进行降维的线性降维方法。

令 m 个样本在原始空间的距离矩阵为 $\boldsymbol{D} \in \mathbb{R}^{m\times m}$，其第 i 行 j 列的元素为 $dist_{ij}$ 为样本 $\boldsymbol{x}_i$ 到样本 $\boldsymbol{x}_j$ 的距离，降维后在 d' 维空间中的表示是 $\boldsymbol{z} \in \mathbb{R}^{d'\times m}, d' \leq d$，且任意两个样本在 d' 维空间中的欧氏距离等于原始空间中的距离，即 $\|\boldsymbol{z}_i - \boldsymbol{z}_j\| = dist_{ij}$。

令 $\boldsymbol{B} = \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z} \in \mathbb{R}^{m\times m}$ 为降维后样本的内积矩阵，$b_{ij} = \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_j$，有：

$$dist_{ij}^2 = \|\boldsymbol{z}_i\|^2 + \|\boldsymbol{z}_j\|^2 - 2\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_j = b_{ii} + b_{jj} - 2b_{ij}$$

设降维后的样本 $\boldsymbol{Z}$ 被中心化，即 $\sum_{i=1}^m \boldsymbol{z}_i = \boldsymbol{0}$，显然矩阵 $\boldsymbol{B}$ 的行和列之和均为零，即 $\sum_{i=1}^m b_{ij} = \sum_{j=1}^m b_{ij} = 0$。可以得到：

$$\sum_{i=1}^m dist_{ij}^2 = tr(\boldsymbol{B}) + mb_{jj}$$

$$\sum_{j=1}^m dist_{ij}^2 = tr(\boldsymbol{B}) + mb_{ii}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{m} dist_{ij}^2 = 2m\ tr(\boldsymbol{B})$$

其中 $tr(\boldsymbol{B}) = \sum_{i=1}^{m} \|\boldsymbol{z}_i\|^2$。令

$$dist_{i\cdot}^2 = \frac{1}{m}\sum_{j=1}^{m} dist_{ij}^2$$

$$dist_{\cdot j}^2 = \frac{1}{m}\sum_{i=1}^{m} dist_{ij}^2$$

$$dist_{\cdot\cdot}^2 = \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} dist_{ij}^2$$

带入前面可得：

$$b_{ij} = -\frac{1}{2}(dist_{ij}^2 - dist_{i\cdot}^2 - dist_{\cdot j}^2 + dist_{\cdot\cdot}^2)$$

所以可以通过降维前后保持距离不变的距离矩阵 $\boldsymbol{D}$ 来求取内积矩阵 $\boldsymbol{B}$。

对矩阵 $\boldsymbol{B}$ 做特征值分解：$\boldsymbol{B} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}}$，其中 $\boldsymbol{\Lambda}$ 是特征值构成的对角矩阵。选取其中的 d^* 个非零特征值，构成对角矩阵 $\boldsymbol{\Lambda}_*$，并令 $\boldsymbol{V}_*$ 为其对应的特征向量矩阵，则有：

$$\boldsymbol{Z} = \boldsymbol{\Lambda}_*^{\boldsymbol{1/2}}\boldsymbol{V}_*^{\boldsymbol{T}} \in \mathbb{R}^{d^*\times m}$$

0.3 主成分分析

主成分分析 (Principal Component Analysis，简称 PCA) 也是一种常用的降维方法。主成分分析的目标是：

$$\max_{\boldsymbol{W}} \mathrm{tr}(\boldsymbol{W}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{W}) \qquad \text{s.t.} \quad \boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}$$

可以从**最近重构性** (样本点到划分超平面的距离足够近) 和**最大可分性** (样本点在超平面的投影尽可能分开) 两个角度得到上面的分析目标。从最大可分性的角度来看，样本点 $\boldsymbol{x}_i$ 在新空间中超平面上的投影为 $\boldsymbol{W}^T\boldsymbol{x}_i$，若所有样本点的投影尽可能分开，则应该使投影后样本点的方差最大化，而投影后的方差就是：

$$\sum_i \boldsymbol{W}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{W}$$

PCA 的求解方法：对协方差矩阵 $\boldsymbol{X}\boldsymbol{X}^T$ 进行特征值分解，并将取前 d' 大的 d' 个特征值对应的特征向量构成坐标转化矩阵 $\boldsymbol{W}$。

0.4 核化线性降维

非线性降维的一种常用方法是**基于核技巧对线性降维方法进行“核化”**。此时需要一个能够**显示**表示的核函数。

0.5 流形学习 (manifold learning)

流行是在局部与欧氏空间同胚的空间，即：在局部空间具有欧氏空间的性质，能用欧氏距离来进行距离计算。流行学习也可用于可视化。

0.5.1 等度量映射 (Isomap)

等度量映射 (Isometric Mapping，简称 Isomap) 的基本出发点，是认为低维流形嵌入到高维空间后，直接在高维空间中计算直线距离具有误导性，因为**高维空间中的直线距离在低维嵌入流行上是不可达的**。

低维嵌入流行上两点间的距离是"测地线"距离：样本空间曲面上的最短曲线距离。计算方法：利用流行在局部上与欧式空间同胚这个性质，对每个点基于欧氏距离找出其近邻点，然后建立近邻点连接图，途中近邻点之间存在连接，而非近邻点之间不存在连接。于是计算测地线距离就变成了**计算近邻连接图上两点之间最短路径问题**。可以使用 Dijkstra 算法、Floyd 算法等。

对近邻图的构造可以指定近邻的个数，也可以设置近邻距离阈值。**Isomap 保持了近邻样本之间的距离。**

0.5.2 局部线性嵌入 (LLE)

局部线性嵌入 (Locally Linear Embedding，简称 LLE) 保持了近邻样本之间的线性关系。

LLE 算法

输入: 样本集 $D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_m,y_m)\}$;
近邻参数 k;
低维空间维数 d'。

1: 初始化一组原型向量 $\{\boldsymbol{p}_1,\boldsymbol{p}_2,\cdots,\boldsymbol{p}_q\}$，可以随机选择 q 原型样本
2: **for** $i=1,2,\cdots,m$ **do**
3: 确定 $\boldsymbol{x}_i$ 的 k 近邻
4: 通过式子: $\min\limits_{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_m}\sum_{i=1}^{m}\|\boldsymbol{x}_i-\sum_{j\in Q_i}w_{ij}\boldsymbol{x}_j\|_2^2 \quad s.t. \quad \sum_{j\in Q_i}w_{ij}=1$ 求得 $w_{ij},j\in Q_i$
5: 对于 $j\notin Q_i$，令 $w_{ij}=0$
6: **end for**
7: 计算 $\boldsymbol{M}=(\boldsymbol{I}-\boldsymbol{W})^T(\boldsymbol{I}-\boldsymbol{W})$
8: 对 $\boldsymbol{M}$ 进行特征值分解
9: 得到 $\boldsymbol{M}$ 的最小 d' 个特征值对应的特征向量

输出: 样本空间集 D 在低维空间的投影 $Z=\{\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_m\}$

0.6 度量学习

度量学习 (metric learning)：每个空间对应了在样本属性上定义的一个距离度量，寻找合适的空间，也就是在寻找一个合适的距离度量。

要对距离度量进行学习，首先需要有一个便与学习的距离度量表达式，第九章总结了常用的距离计算方式。近邻成分分析 (Neighbourhood Component Analysis，简称 NCA) 是基于马氏距离的分析。

0.7 总结

懒惰学习方法主要有 k 近邻学习器，懒惰决策树；朴素贝叶斯分类器能以懒惰学习方式使用，也可以急切学习方式使用。

主成分分析是无监督的线性降维方法，监督线性降维方法最著名的是线性判别分析 (LDA)，其核化版本是 KLDA。

PCA 的优点：算法简单，具有线性误差；在计算过程中不需要人为的设定参数或是根据任何经验模型对计算进行干预，最后的结果只与数据相关，与用户是独立的；主成分可以用作多元回归和聚类分析的输入。PCA 的问题：存储空间大，计算复杂度高；有可能投影以后对数据的区分作用并不大，反而可能使得数据点揉杂在一起无法区分。

LDA 可以用于分类工作，但是对于样本维数大于样本数时不好处理。

常用的流行学习方法有：Isomap,LLE, 拉普拉斯特征映射，局部空间对齐等。

Isomap 虽然具有拓扑不稳定性，计算复杂，对噪声敏感，但仍然被广泛采用，效果良好。拉普拉斯特征映射思想简单，计算也简单，但是要求观测数据采样稠密，对噪声敏感性很大。

流形学习中非线性维数约简方法与线性维数约简方法相比的一个显著特点是分析中的**局部性**。现有非线性维数约简方法大多基于小的邻域学习, 期望通过在小邻域上的学习得到一个全局的坐标, 这往往是不现实的。因此非线性降维的一个研究方向是：如何将全局与局部数据学习结合起来。