课程8：
特征提取（传统方法）

k近邻学习

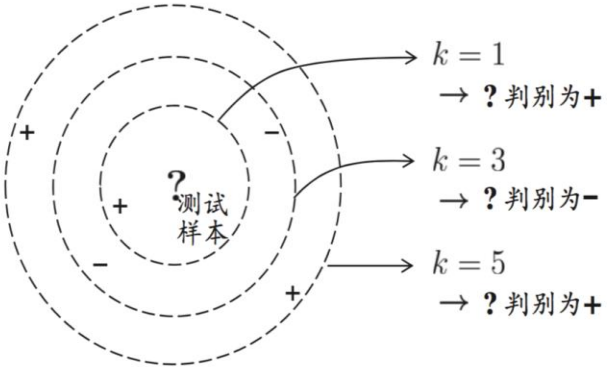


图 10.1 k 近邻分类器示意图. 虚线显示出等距线; 测试样本在 $k=1$ 或 $k=5$ 时被判别为正例, $k=3$ 时被判别为反例.

- 投票法：选择这k个样本中出现最多的类别标记作为预测结果。
- 平均法：将这k个样本的实值输出标记的平均值作为预测结果。

k近邻分类器中的k是一个重要参数，当k取不同值时，分类结果会有显著不同。另一方面，若采用不同的距离计算方式，则找出的“近邻”可能有显著差别，从而也会导致分类结果有显著不同。

低维嵌入

维数灾难（curse of dimensionality）

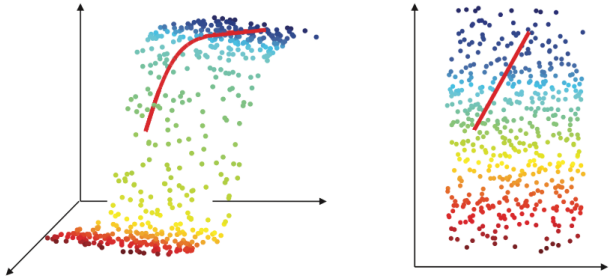
上述讨论基于一个重要的假设：任意测试样本附近的任意小的距离范围内总能找到一个训练样本，即训练样本的采样密度足够大，或称为“密采样”。然而，这个假设在现实任务中通常很难满足：

若属性维数为1，当δ=0.001，仅考虑单个属性，则仅需1000个样本点平均分布在归一化后的属性取值范围内，即可使得任意测试样本在其附近0.001距离范围内总能找到一个训练样本，此时最近邻分类器的错误率不超过贝叶斯最优分类器的错误率的两倍。若属性维数为20，若样本满足密采样条件，则至少需要$(10^3)^{20}=10^{60}$个样本。

低维嵌入

- 缓解维数灾难的一个重要途径是降维(dimension reduction)
  - 即通过某种数学变换，将原始高维属性空间转变为一个低维“子空间”(subspace)，在这个子空间中样本密度大幅度提高，距离计算也变得更为容易。

- 为什么能进行降维？
  - 数据样本虽然是高维的，但与学习任务 密切相关的也许仅是某个低维分布，即高维空间中的一个低维“嵌入” (embedding)，因而可以对数据进行有效的降维。



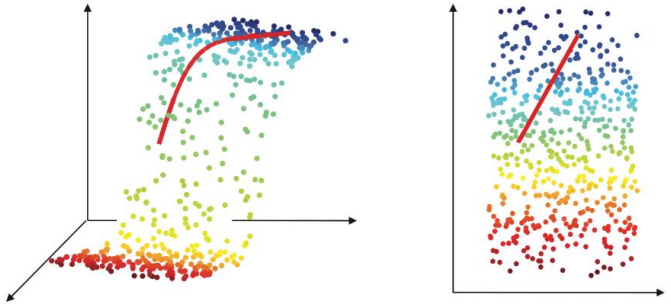
(a) 三维空间中观察到的样本点　　(b) 二维空间中的曲面

多维缩放

- 要求原始空间中样本之间的距离在低维空间中得以保持，即得到“多维缩放”
- 假定有m个样本，在原始空间中的距离矩阵为 $\mathbf{D} \in \mathbb{R}^{m \times m}$ ，其第i行j列的元素 $dist_{ij}$ 为样本 $\boldsymbol{x}_i$ 到 $\boldsymbol{x}_j$ 的距离。
- 目标是获得样本在 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 $\mathbf{B} = \mathbf{Z}^{\mathrm{T}}\mathbf{Z} \in \mathbb{R}^{m \times m}$ ，其中 $\mathbf{B}$ 为降维后的内积矩阵，$b_{ij} = \boldsymbol{z}_i^{\mathrm{T}} \boldsymbol{z}_j$ 有

$$
\begin{aligned}
dist_{ij}^2 &= ||\boldsymbol{z}_i||^2 + ||\boldsymbol{z}_j||^2 - 2\boldsymbol{z}_i^T \boldsymbol{z}_j \\
&= b_{ii} + b_{jj} - 2b_{ij}.
\end{aligned}
$$



(a) 三维空间中观察到的样本点　　(b) 二维空间中的曲面

多维缩放

为便于讨论，令降维后的样本 $\mathbf{Z}$ 被中心化，即 $\sum_{i=1}^{m} z_i = 0$。显然，矩阵 $\mathbf{B}$ 的行与列之和均为零，即

$$\sum_{i=1}^{m} b_{ij} = \sum_{j=1}^{m} b_{ij} = 0.$$

易知 $\sum_{i=1}^{m} dist_{ij}^2 = \mathrm{tr}(\mathbf{B}) + m b_{jj}$, $\sum_{j=1}^{m} dist_{ij}^2 = \mathrm{tr}(\mathbf{B}) + m b_{ii}$, $\sum_{i=1}^{m}\sum_{j=1}^{m} dist_{ij}^2 = 2m\,\mathrm{tr}(\mathbf{B})$,

其中 $\mathrm{tr}(\cdot)$ 表示矩阵的迹(trace)，$\mathrm{tr}(\mathbf{B}) = \sum_{i=1}^{m} \|\boldsymbol{z}_i\|^2$ 。令

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

由此即可通过降维前后保持不变的距离矩阵 $\mathbf{D}$ 求取内积矩阵 $\mathbf{B}$ ：

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

多维缩放

对矩阵$\mathbf{B}$ 做特征值分解(eigenvalue decomposition)$\mathbf{B}=\mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}}$，其中$\mathbf{\Lambda}=\mathrm{diag}(\lambda_1,\lambda_2,\ldots,\lambda_d)$ 为特征值构成的对角矩阵，

$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_d$为特征向量矩阵，假定其中有$d^*$个非零正特征值，它们构成对角矩阵 $\mathbf{\Lambda}_*=\mathrm{diag}(\lambda_1,\lambda_2,\ldots,\lambda_{d*})$，$\mathbf{V}$ 为特征向量矩阵。令 $\mathbf{V}_*$ 表示相应的特征矩阵，则$\mathbf{Z}$ 可表达为 $\mathbf{Z}=\lambda_*^{1/2}\mathbf{V}_*^{\mathrm{T}}\in\mathbb{R}^{d^*\times m}$。

$$\mathrm{B}=V\begin{pmatrix}\lambda_1 & & 0\\ & \ddots & \\ 0 & & \lambda_n\end{pmatrix}V'=V\begin{pmatrix}\sqrt{\lambda_1} & & 0\\ & \ddots & \\ 0 & & \sqrt{\lambda_n}\end{pmatrix}\underbrace{\begin{pmatrix}\sqrt{\lambda_1} & & 0\\ & \ddots & \\ 0 & & \sqrt{\lambda_n}\end{pmatrix}V'}_{Z}$$

多维缩放

- 对矩阵 $\mathbf{B}$ 做特征值分解(eigenvaluedecomposition) $\mathbf{B}=\mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}}$，其中 $\mathbf{\Lambda}=\mathrm{diag}(\lambda_1,\lambda_2,\ldots,\lambda_d)$ 为特征值构成的对角矩阵，

  $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_d$ 为特征向量矩阵，假定其中有 d^* 个非零正特征值它们构成对角阵 $\mathbf{\Lambda}_*=\mathrm{diag}(\lambda_1,\lambda_2,\ldots,\lambda_{d*})$，$\mathbf{V}$ 为特征向量矩阵。令 $\mathbf{V}_*$ 表示相应的特征矩阵，则 $\mathbf{Z}$ 可表达为 $\mathbf{Z}=\lambda_*^{1/2}\mathbf{V}_*^{\mathrm{T}}\in\mathbb{R}^{d^*\times m}$。

- 在现实应用中为了有效降维，往往仅需降维后的距离与原始空间中的距离尽可能接近，而不必严格相等。此时可取 $d'\ll d$ 个最大特征值构成对角矩阵 $\tilde{\mathbf{\Lambda}}=\mathrm{diag}(\lambda_1,\lambda_2,\ldots,\lambda_{d'})$，令 $\tilde{\mathbf{V}}$ 表示相应的特征向量矩阵，则 $\mathbf{Z}$ 可表达为

$$\mathbf{Z}=\tilde{\mathbf{\Lambda}}^{1/2}\tilde{\mathbf{V}}^{\mathrm{T}}\in\mathbb{R}^{d'\times m}.$$

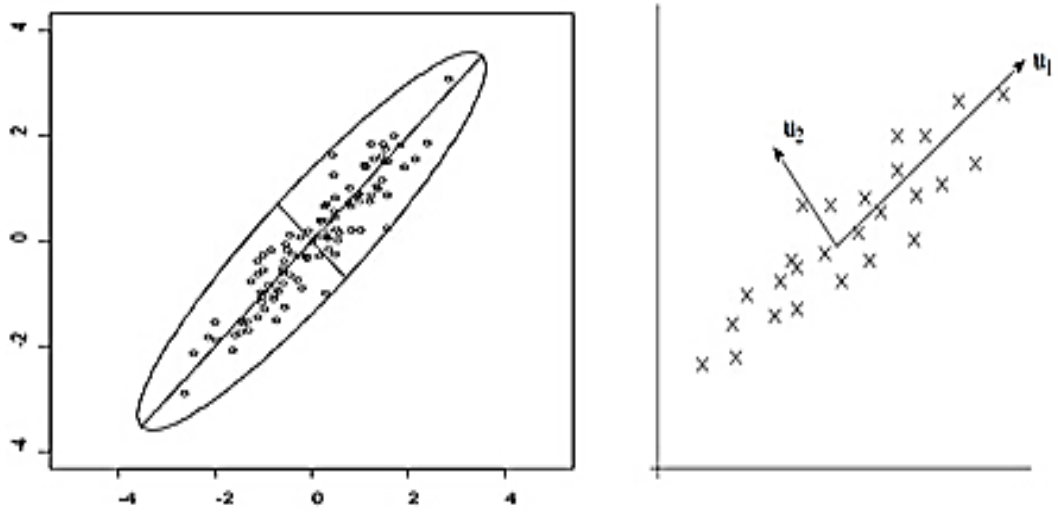
主成分分析（PCA）

- 对于正交属性空间中的样本点，如何用一个超平面对所有样本进行恰当的表达？

- 容易想到，若存在这样的超平面，那么它大概应具有这样的性质：
  - 最近重构性：样本点到这个超平面的距离都足够近；
  - 最大可分性：样本点在这个超平面上的投影能尽可能分开。

- 基于最近重构性和最大可分性，能分别得到主成分分析的两种等价推导。

主成分分析（PCA）



上图中，u1就是主成分方向，然后在二维空间中取和u1方向正交的方向，就是u2的方向。则n个数据在u1轴的离散程度最大（方差最大），数据在u1上的投影代表了原始数据的绝大部分信息，即使不考虑u2，信息损失也不多。而且，u1、u2不相关。只考虑u1时，二维降为一维。

主成分分析

最近重构性

- 对样本进行中心化，$\sum_i x_i = 0$，再假定投影变换后得到的新坐标为 $\{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_d\}$，其中 $\boldsymbol{w}_i$ 是标准正交基向量，

$$||\boldsymbol{w}_i||_2 = 1, \boldsymbol{w}_i^{\mathrm{T}} \boldsymbol{w}_j = 0 (i \neq j).$$

- 若丢弃新坐标系中的部分坐标，即将维度降低到 $d' < d$，则样本点在低维坐标系中的投影是 $\boldsymbol{z}_i = (z_{i1}; z_{i2}; \ldots; z_{id'})$，$z_{ij} = \boldsymbol{w}_j^{\mathrm{T}} \boldsymbol{x}_i$ 是 $\boldsymbol{x}_i$ 在低维坐标下第 j 维的坐标，若基于 $\boldsymbol{z}_i$ 来重构 $\boldsymbol{x}_i$，则会得到

$$\hat{\boldsymbol{x}}_i = \sum_{j=1}^{d'} z_{ij} \boldsymbol{w}_j.$$

主成分分析

最近重构性

- 考虑整个训练集，原样本点 $\boldsymbol{x}_i$ 与基于投影重构的样本点 $\hat{\boldsymbol{x}}_i$ 之间的距离为

$$\sum_{i=1}^{m}\left\|\sum_{j=1}^{d'} z_{ij}\boldsymbol{w}_j-\boldsymbol{x}_i\right\|_2^2=\sum_{i=1}^{m}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_i-2\sum_{i=1}^{m}\boldsymbol{z}_i^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\boldsymbol{x}_i+\text{const}$$

$$\propto -\mathrm{tr}\left(\mathbf{W}^{\mathrm{T}}\left(\sum_{i=1}^{m}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right)\mathbf{W}\right).$$

- 根据最近重构性应最小化上式。考虑到 $\boldsymbol{w}_j$ 是标准正交基，$\sum_i \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}$ 是协方差矩阵，有

$$\begin{aligned}\min_{\mathbf{W}} \quad & -\mathrm{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W})\\ \text{s.t.} \quad & \mathbf{W}^{\mathrm{T}}\mathbf{W}=\mathbf{I}.\end{aligned}$$

这就是主成分分析的优化目标。

主成分分析

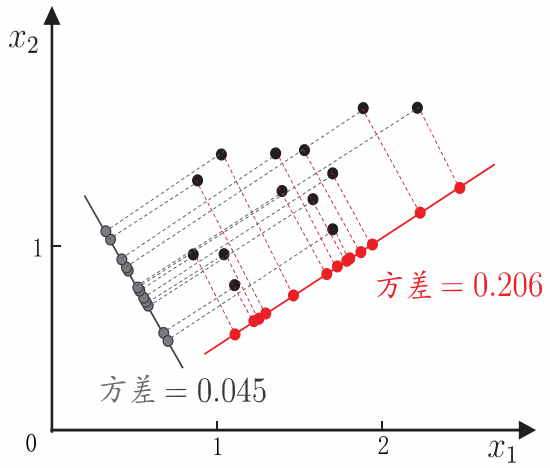
最大可分性

- 样本点 $\boldsymbol{x}_i$ 在新空间中超平面上的投影是 $\mathbf{W}^{\mathrm{T}}\boldsymbol{x}_i$ 若所有样本点的投影能尽可能分开，则应该使得投影后样本点的方差最大化。若投影后样本点的方差是 $\sum_i \mathbf{W}^{\mathrm{T}}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\mathbf{W}$ ，于是优化目标可写为

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

显然与 $\min_{\mathbf{W}} \quad -\mathrm{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W})$ 等价。

$$\text{s.t.} \quad \mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}.$$



主成分分析

PCA的求解

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

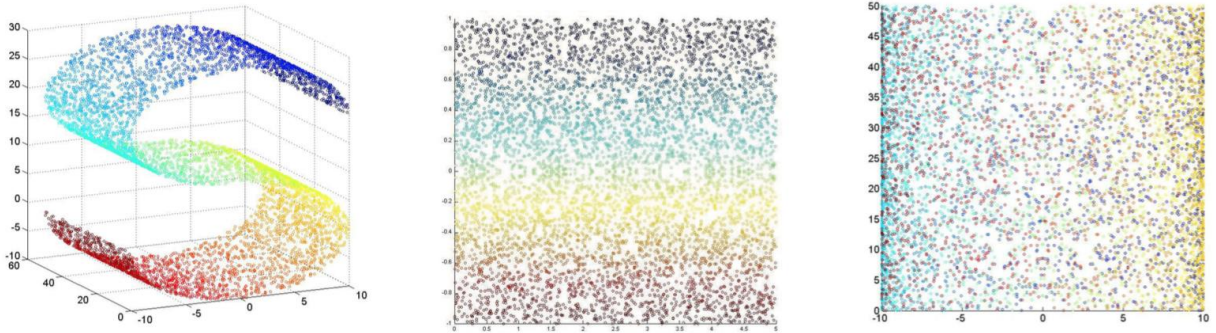
- 对优化式使用拉格朗日乘子法可得

$$\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W} = \lambda\mathbf{W}.$$

只需对协方差矩阵 $\mathbf{X}\mathbf{X}^{\mathrm{T}}$进行特征值分解，并将求得的特征值排序：$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$，再取前 d' 个特征值对应的特征向量构成 $\mathbf{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{d'})$，这就是主成分分析的解。

核化线性降维

- 线性降维方法假设从高维空间到低维空间的函数映射是线性的，然而，在不少现实任务中，可能需要非线性映射才能找到恰当的低维嵌入：



(a) 三维空间中的观察　(b) 本真二维结构　(c) PCA 降维结果

三维空间中观察到的3000个样本点，是从本真二维空间中矩形区域采样后以S形曲面嵌入，此情况下线性降维会丢失低维结构。图中数据点的染色显示出低维空间的结构。

核化主成分分析(Kernelized PCA，简称KPCA)

- 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”(kernelized)。

- 假定我们将在高维特征空间中把数据投影到由 $\mathbf{W}$ 确定的超平面上，即PCA欲求解

$$\left(\sum_{i=1}^{m} \boldsymbol{z}_i \boldsymbol{z}_i^{\mathrm{T}}\right) \mathbf{W} = \lambda \mathbf{W}.$$

- 其中 $\boldsymbol{z}_i$ 是样本点 $\boldsymbol{x}_i$ 在高维特征空间中的像。令 $\boldsymbol{\alpha}_i = \frac{1}{\lambda} \boldsymbol{z}_i^{\mathrm{T}} \mathbf{W}$ ，

$$\mathbf{W} = \frac{1}{\lambda}\left(\sum_{i=1}^{m} \boldsymbol{z}_i \boldsymbol{z}_i^{\mathrm{T}}\right) \mathbf{W} = \sum_{i=1}^{m} \boldsymbol{z}_i \frac{\boldsymbol{z}_i^{\mathrm{T}} \mathbf{W}}{\lambda} = \sum_{i=1}^{m} \boldsymbol{z}_i \boldsymbol{\alpha}_i.$$

核化主成分分析

- 假定 $\boldsymbol{z}_i$ 是由原始属性空间中的样本点 $\boldsymbol{x}_i$ 通过映射 ϕ 产生，即

$$\boldsymbol{z}_i = \phi(\boldsymbol{x}_i),\ i = 1, 2, \ldots, m.$$

- 若ϕ 能被显式表达出来，则通过它将样本映射至高维空间，再在特征空间中实施PCA即可，即有

$$\left(\sum_{i=1}^{m} \phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}_i)^{\mathrm{T}}\right)\mathbf{W} = \lambda\mathbf{W}.$$

并且

$$\mathbf{W} = \sum_{i=1}^{m} \phi(\boldsymbol{x}_i)\boldsymbol{\alpha}_i.$$

核化主成分分析

- 一般情形下，我们不清楚 ϕ 的具体形式，于是引入核函数

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^{\mathrm{T}} \phi(\boldsymbol{x}_j).$$

- 又由 $\mathbf{W} = \sum_{i=1}^{m} \phi(\boldsymbol{x}_i)\boldsymbol{\alpha}_i$，代入优化式 $\left(\sum_{i=1}^{m} \phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}_i)^{\mathrm{T}}\right)\mathbf{W} = \lambda\mathbf{W}$，有

$$\mathbf{K}\mathbf{A} = \lambda\mathbf{A}.$$

其中 $\mathbf{K}$ 为 κ 对应的核矩阵，$(\mathbf{K})_{ij} = \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$，$\mathbf{A} = (\boldsymbol{\alpha}_1; \boldsymbol{\alpha}_2; \ldots; \boldsymbol{\alpha}_m)$.

- 上式为特征值分解问题，取 $\mathbf{K}$ 最大的 d' 个特征值对应的特征向量得到解。

核化主成分分析

对新样本 $\boldsymbol{x}$，其投影后的第 j $(j=1,2,\ldots,d')$ 维坐标为

$$z_j = \boldsymbol{w}_j^{\mathrm{T}}\phi(\boldsymbol{x}) = \sum_{i=1}^{m}\alpha_i^j\phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x})$$

$$= \sum_{i=1}^{m}\alpha_i^j\kappa(\boldsymbol{x}_i,\boldsymbol{x}).$$

其中 $\boldsymbol{\alpha}_i$ 已经过规范化，α_i^j 是 $\boldsymbol{\alpha}_i$ 的第 j 个分量。由该式可知，为获得投影后的坐标，KPCA需对所有样本求和，因此它的计算开销较大。

流形学习

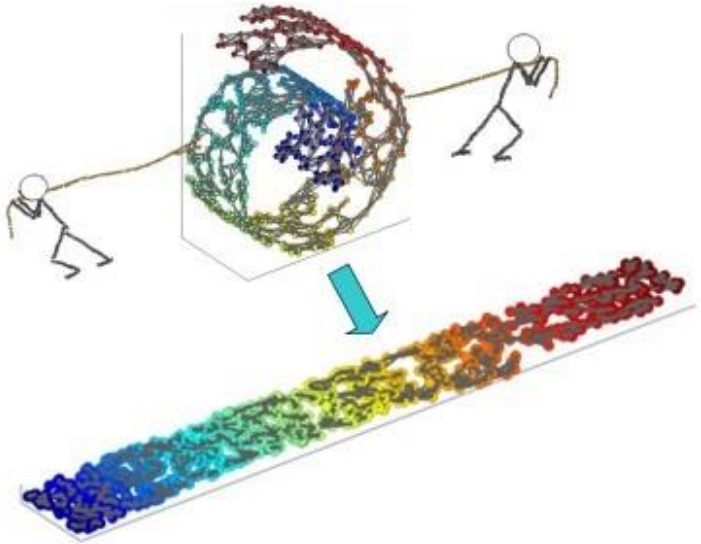
- 流形学习(manifold learning)是一类借鉴了拓扑流形概念的降维方法。“流形”是在局部与欧氏空间同胚的空间，换言之，它在局部具有欧氏空间的性质，能用欧氏距离来进行距离计算。

- 若低维流形嵌入到高维空间中，则数据样本在高维空间的分布虽然看上去非常复杂，但在局部上仍具有欧氏空间的性质，因此，可以容易地在局部建立降维映射关系，然后再设法将局部映射关系推广到全局。

- 当维数被降至二维或三维时，能对数据进行可视化展示，因此流形学习也可被用于可视化。

流形学习

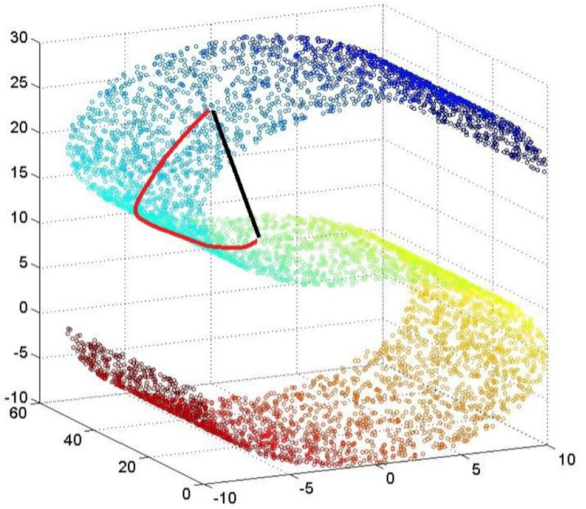
一个形象的流形降维过程如下图。我们有一块卷起来的布，我们希望将其展开到一个二维平面，我们希望展开后的布能够在局部保持布结构的特征，其实也就是将其展开的过程，就想两个人将其拉开一样。



流形学习

等度量映射(Isometric Mapping)

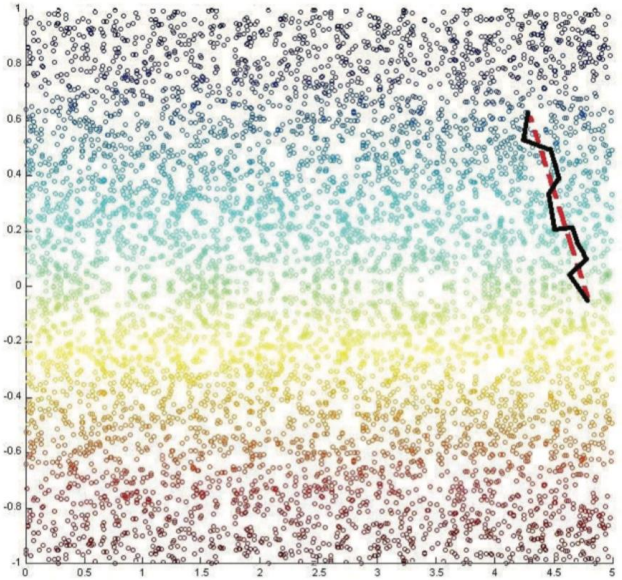
- 低维流形嵌入到高维空间之后，直接在高维空间中计算直线距离具有误导性，因为高维空间中的直线距离在低维嵌入流形上不可达。而低维嵌入流形上两点间的本真距离是“测地线”(geodesic)距离。



(a) 测地线距离与高维直线距离

流形学习

- 测地线距离的计算：利用流形在局部上与欧氏空间同胚这个性质，对每个点基于欧氏距离找出其近邻点，然后就能建立一个近邻连接图，图种近邻点之间存在连接，而非近邻点之间不存在连接，于是，计算两点之间测地线距离的问题，就转变为计算近邻连接图上两点之间的最短路径问题。
- 最短路径的计算可通过Dijkstra算法或Floyd算法实现。得到距离后可通过多维缩放方法获得样本点在低维空间中的坐标。

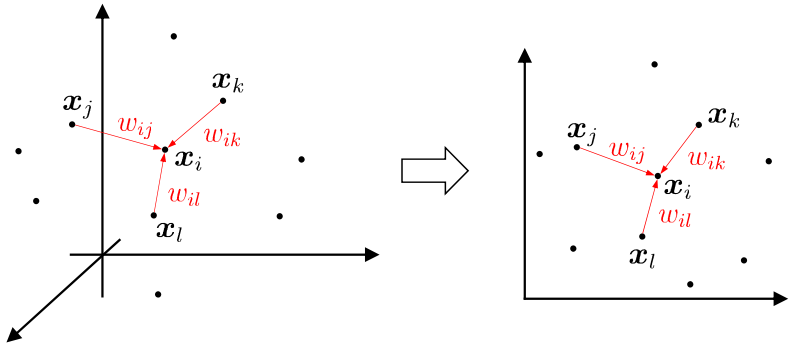


测地线距离与近邻距离

流形学习

局部线性嵌入（Locally Linear Embedding，LLE）

- 局部线性嵌入试图保持邻域内的线性关系，并使得该线性关系在降维后的空间中继续保持。



$$\boldsymbol{x}_i = w_{ij}\boldsymbol{x}_j + w_{ik}\boldsymbol{x}_k + w_{il}\boldsymbol{x}_l$$

流形学习

- LLE先为每个样本 $\boldsymbol{x}_i$ 找到其近邻下标集合 Q_i ，然后计算出基于 Q_i 的中的样本点对 $\boldsymbol{x}_i$ 进行线性重构的系数 $\boldsymbol{w}_i$ ：

$$\min_{\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_m} \sum_{i=1}^{m} \left\| \boldsymbol{x}_i - \sum_{j\in Q_i} w_{ij}\boldsymbol{x}_j \right\|_2^2$$
$$\text{s.t.} \sum_{j\in Q_i} w_{ij} = 1,$$

其中 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 均为已知，令 $C_{jk} = (\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}}(\boldsymbol{x}_i - \boldsymbol{x}_k)$ ，w_{ij} 有闭式解

$$w_{ij} = \frac{\sum_{k\in Q_i} C_{jk}^{-1}}{\sum_{l,s\in Q_i} C_{ls}^{-1}}.$$

流形学习

- LLE在低维空间中保持 $\boldsymbol{w}_i$ 不变，于是 $\boldsymbol{x}_i$ 对应的低维空间坐标 $\boldsymbol{z}_i$ 可通过下式求解：

$$\min_{\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_m} \sum_{i=1}^{m} \left\| \boldsymbol{z}_i - \sum_{j\in Q_i} w_{ij}\boldsymbol{z}_j \right\|_2^2$$

- 令 $\mathbf{Z} = (\mathrm{z}_1, \mathrm{z}_2, \ldots, \mathrm{z}_m) \in \mathbb{R}^{d'\times m}, (\mathbf{W})_{ij} = w_{ij},$

$$\mathbf{M} = (\mathbf{I} - \mathbf{W})^{\mathrm{T}}(\mathbf{I} - \mathbf{W}),$$

则优化式可重写为右式，并通过特征值分解求解。

$$\min_{\mathbf{Z}} \operatorname{tr}(\mathbf{Z}\mathbf{M}\mathbf{Z}^{\mathrm{T}})$$
$$\text{s.t. } \mathbf{Z}\mathbf{Z}^{\mathrm{T}} = \mathbf{I}.$$

K-L变换

从n维特征中选取m维特征，如何在信息损失最小的情况下选取特征(因为必然会删去n-m维特征)，使得剩下的特征更加有利于分类，离散K-L变换（Karhunen-Loeve变换）就是常用的方法。

K-L变换推导过程：

- 为了找到满足条件的变换矩阵U，令：$y = U^T x$ 因为新向量y各分量之间是相互独立的，因此有：

$$E(y_i y_j) = 0, i \neq j$$

- 又从自相关矩阵的定义有：

$$\mathrm{R}_y = E(yy^T) = E(U^T xx^T U) = U^T R_x U$$

而Rx是对称矩阵，因此它的特征向量是相互正交的。如果将U的列向量取为Rx的特征向量，这时 Ry可以转化为对角矩阵：

$$R_y = U^T R_x U = \Lambda$$

其中 ∧是对角矩阵，对角线元素是 Rx的特征值 λi，i=1，2，...，n，由此可以确定变换矩阵A，它的列向量就是特征向量，这些特征向量之间是相互正交的。利用变换矩阵 A对原输入向量x进行变换，获得新向量 y的过程就是K-L变换。

度量学习

- 欲对距离度量进行学习，必须有一个便于学习的距离度量表达形式。对两个 d 维样本 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$，它们之间的平方欧氏距离可写为

$$\mathrm{dist}_{\mathrm{ed}}^2(\boldsymbol{x}_i, \boldsymbol{x}_j) = ||\boldsymbol{x}_i - \boldsymbol{x}_j||_2^2 = dist_{ij,1}^2 + dist_{ij,2}^2 + \cdots + dist_{ij,d}^2,$$

- 其中 $dist_{ij,k}$ 表示 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 在第 k 维上的距离。若假定不同属性的重要性不同，则可引入属性权重 $\boldsymbol{w}$ 得到

$$\begin{aligned}\mathrm{dist}_{\mathrm{wed}}^2(\boldsymbol{x}_i, \boldsymbol{x}_j) &= ||\boldsymbol{x}_i - \boldsymbol{x}_j||_2^2 = w_1 \cdot dist_{ij,1}^2 + w_2 \cdot dist_{ij,2}^2 + \cdots + w_d \cdot dist_{ij,d}^2 \\ &= (\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} \mathbf{W} (\boldsymbol{x}_i - \boldsymbol{x}_j),\end{aligned}$$

- 其中 $w_i \geq 0$, $\mathbf{W} = \mathrm{diag}(\boldsymbol{w})$ 是一个对角矩阵，$(\mathbf{W})_{ii} = w_i$，可通过学习确定。

度量学习

- $\mathbf{W}$ 的非对角元素均为零，这意味着坐标轴是正交的，即属性之间无关；但现实问题中往往不是这样，例如考虑西瓜的“重量”和“体积”这两个属性，它们显然是正相关的，其对应的坐标轴不再正交。为此将 $\mathbf{W}$ 替换为一个普通的半正定对称矩阵 $\mathbf{M}$ ，于是就得到了马氏距离(Mahalanobis distance)。

$$\mathrm{dist}_{\mathrm{mah}}^2(\boldsymbol{x}_i,\boldsymbol{x}_j)=(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}\mathbf{W}(\boldsymbol{x}_i-\boldsymbol{x}_j)=||\boldsymbol{x}_i-\boldsymbol{x}_j||_{\mathbf{M}}^2,$$

其中 $\mathbf{M}$ 亦称“度量矩阵”，而度量学习则是对 $\mathbf{M}$ 进行学习。注意到为了保持距离非负且对称，$\mathbf{M}$必须是（半）正定对称矩阵，即必有正交基 $\mathbf{P}$ 使得 $\mathbf{M}$ 能写为 $\mathbf{M}=\mathbf{P}\mathbf{P}^{\mathrm{T}}$ 。

- 对 $\mathbf{M}$ 进行学习当然要设置一个目标。假定我们是希望提高近邻分类器的性能，则可将 $\mathbf{M}$ 直接嵌入到近邻分类器的评价指标中去，通过优化该性能指标相应地求得 $\mathbf{M}$ 。

度量学习

近邻成分分析(Neighbourhood Component Analysis, NCA)

- 近邻成分分析在进行判别时通常使用多数投票法，邻域中的每个样本投1票，邻域外的样本投0票。不妨将其替换为概率投票法。对于任意样本 $\boldsymbol{x}_i$ ，它对分类结果 $\boldsymbol{x}_j$ 影响的概率为

$$p_{ij} = \frac{\exp(-||\boldsymbol{x}_i - \boldsymbol{x}_j||_{\mathrm{M}}^2)}{\sum_l \exp(-||\boldsymbol{x}_i - \boldsymbol{x}_l||_{\mathrm{M}}^2)},$$

- 当 $i = j$ 时，p_{ij} 最大。显然，$\boldsymbol{x}_j$ 对 $\boldsymbol{x}_i$ 的影响随着它们之间距离的增大而减小。若以留一法(LOO)正确率的最大化为目标，则可计算 $\boldsymbol{x}_i$ 的留一法正确率，即它被自身之外的所有样本正确分类的概率为

$$p_i = \sum_{j \in \Omega_i} p_{ij},$$

其中 Ω_i 表示与 $\boldsymbol{x}_i$ 属于相同类别的样本的下标集合。

度量学习

- 整个样本集上的留一法正确率为

$$\sum_{i=1}^{m} p_i = \sum_{i=1}^{m} \sum_{j \in \Omega_i} p_{ij}.$$

- 由 $p_{ij} = \frac{\exp(-||\boldsymbol{x}_i - \boldsymbol{x}_j||_{\mathbf{M}}^2)}{\sum_l \exp(-||\boldsymbol{x}_i - \boldsymbol{x}_l||_{\mathbf{M}}^2)}$ 和 $\mathbf{M} = \mathbf{P}\mathbf{P}^{\mathrm{T}}$，则NCA的优化目标为

$$\min_{\mathbf{P}} \quad 1 - \sum_{i=1}^{m} \sum_{j \in \Omega_i} \frac{\exp\left(-||\mathbf{P}^{\mathrm{T}}\boldsymbol{x}_i - \mathbf{P}^{\mathrm{T}}\boldsymbol{x}_j||_2^2\right)}{\sum_l \exp\left(-||\mathbf{P}^{\mathrm{T}}\boldsymbol{x}_i - \mathbf{P}^{\mathrm{T}}\boldsymbol{x}_l||_2^2\right)}.$$

求解即可得到最大化近邻分类器LOO正确率的距离度量矩阵 $\mathbf{M}$。

度量学习

- 实际上，我们不仅能把错误率这样的监督学习目标作为度量学习的优化目标，还能在度量学习中引入领域知识。

- 若已知某些样本相似、某些样本不相似，则可定义“必连”(must-link)约束集合 C与“勿连”(cannot-link)约束集合 M：

$(x_i, x_j) \in \mathcal{C}$ 表示 x_i 与 x_j 相似，$(x_i, x_j) \in \mathrm{M}$ 表示 x_i 与 x_j 不相似。显然，我们希望相似的样本之间距离较小，不相似的样本之间距离较大，于是可通过求解下面这个凸优化问题获得适当的度量矩阵 M：

$$\min_{M} \quad \sum_{(x_i, x_j) \subseteq M} \| x_i - x_j \|_M^2$$

$$s.t \quad \sum_{(x_i, x_j) \subseteq M} \| x_i - x_j \|_M^2 \geq 1$$

- 其中约束 $\mathrm{M} \geq 0$ 表明 M 必须是半正定的。上式要求在不相似样本间的距离不小于1的前提下，使相似样本间的距离尽可能小。

度量学习

- 不同的度量学习方法针对不同目标获得“好”的半正定对称距离度量矩阵 $\mathbf{M}$，若是一个低秩矩阵 $\mathbf{M}$，则通过对$\mathbf{M}$进行特征值分解，总能找到一组正交基，其正交基数目为矩阵 $\mathbf{M}$ 的秩 $\mathrm{rank}(\mathbf{M})$，小于原属性数$d$。于是，度量学习学得的结果可衍生出一个降维矩阵 $\mathbf{P} \in \mathbb{R}^{d \times \mathrm{rank}(\mathbf{M})}$，能用于降维之目的。

参考文献

1. G. Baudat, F. Anouar. Generlized Discriminant Analysis Using a Kernel Approach[J]. Neural Computation, 2000:2385-2404.
2. M. Belkin, P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation[J]. Neural Computation, 2003:1373-1396.
3. M. Belkin, P. Niyogi, V.Sindhwani. Manifold regulation: A geometric framework for learning from labeled and unlabeled example[J]. Journal of Machine Learning Research, 2006(7):2399-2434
4. T. M. Cover, P. E. Hart. Nearest neighbor pattern classification[J]. IEEE Transactions on Information Theory, 1967:21-27

相关论文会放到课程网页中，如有需要请自行下载。