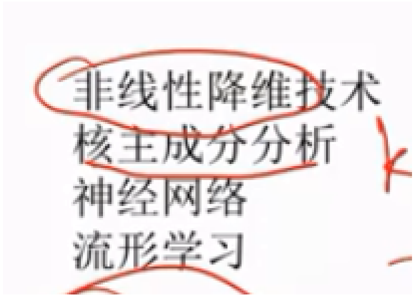
9 数据降维2--非线性降维

9.1 本集内容简介

非线性降维算法
流形的概念
流形学习的概念
局部线性嵌入
拉普拉斯特征映射
局部保持投影
等距映射
实验环节

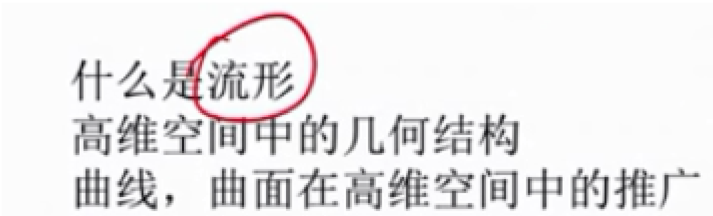
9.2 非线性降维技术简介



这节课的重点是流行学习算法

9.3 流行简介

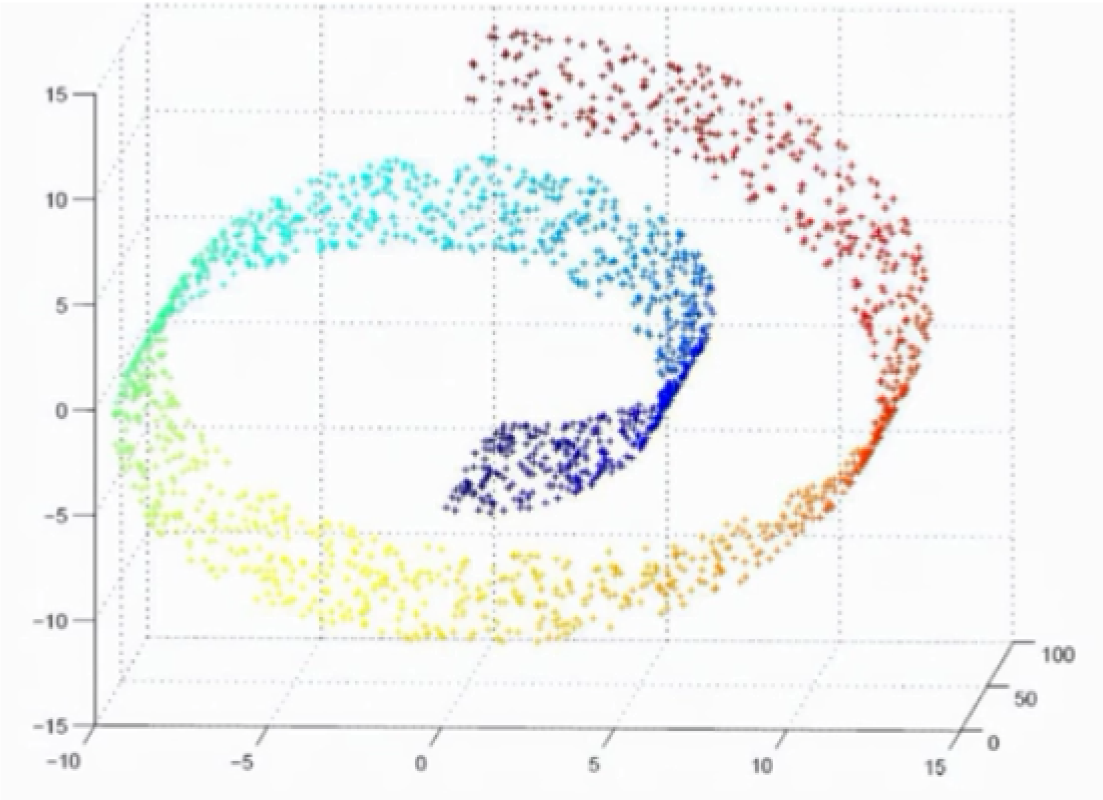
什么是流行？



在微分几何和拓扑里面

就是空间中，有一个类似集合形状的东西。

三维空间里的一个例子



9.4 流行学习简介

> 什么是流形学习？
> 假设数据在高维空间中的分布服从某种几何形状
> 利用这种几何约束来完成对数据的处理，如降维，分类，聚类
>
> 局部线性嵌入
> [1] Roweis, Sam T and Saul, Lawrence K. Nonlinear dimensionality reduction by locally linear embedding. Science, 290(5500). 2000: 2323-2326.
>
> 拉普拉特征映射
> [2] Belkin, Mikhail and Niyogi, Partha. Laplacian eigenmaps for dimensionality reduction and data representation. Neural computation. 15(6). 2003:1373-1396.
>
> 局部保持投影
> [3] He Xiaofei and Niyogi, Partha. Locality preserving projections. NIPS. 2003:234-241.
>
> 等距映射
> [4] Tenenbaum, Joshua B and De Silva, Vin and Langford, John C. A global geometric framework for nonlinear dimensionality reduction. Science, 290(5500). 2000: 2319-2323.

有兴趣可以看论文

9.5 局部线性嵌入

局部线性嵌入（LLE）

每个样本可以用它邻居的线性组合近似重构： $\mathrm{x}_i \approx \sum_j w_{ij}\mathrm{x}_j$

求解下面的最优化问题可以得到重构系数：

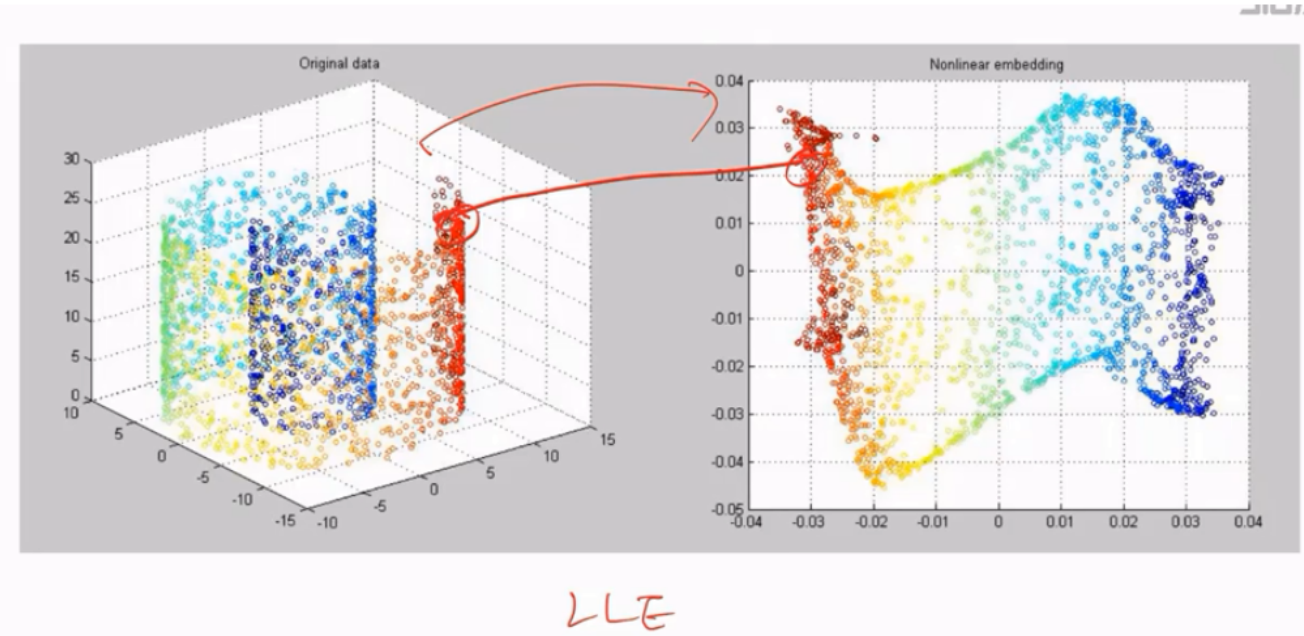
$$\varepsilon(\mathrm{W}) = \sum_i \left\| \mathrm{x}_i - \sum_j w_{ij}\mathrm{x}_j \right\|^2$$

$$\sum_j w_{ij} = 1$$

将向量映射到低维空间，保持这种线性重构关系：

$$\Phi(\mathrm{Y}) = \sum_i \left\| \mathrm{y}_i - \sum_j w_{ij}\mathrm{y}_j \right\|^2$$

1. 简写LLE
    1. 核心思想
2. 公式很难记住



9.6 拉普拉斯的特征映射

> 拉普拉斯特征映射
> 基于图论的方法
> 用样本构造图，然后计算拉普拉斯矩阵，最后对矩阵进行特征值分解

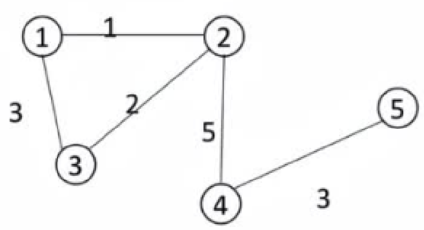
1. 计算拉普拉斯矩阵
2. 对矩阵进行特征值分解

图

<V,E,W>

- V 定点
- E 边
- W 权重

图
顶点，边，权重
有向图，无向图
邻接矩阵
度，加权度



邻接矩阵

$$\begin{bmatrix} 0 & 1 & 3 & 0 & 0 \\ 1 & 0 & 2 & 5 & 0 \\ 3 & 2 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 & 3 \\ 0 & 0 & 0 & 3 & 0 \end{bmatrix}$$

加权度矩阵 $d_i = \sum_j w_{ij}$

$$\begin{bmatrix} 4 & 0 & 0 & 0 & 0 \\ 0 & 8 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{bmatrix}$$

拉普拉斯矩阵

$$\mathrm{L} = \mathrm{D} - \mathrm{W} \quad \begin{bmatrix} 4 & -1 & -3 & 0 & 0 \\ -1 & 8 & -2 & -5 & 0 \\ -3 & -2 & 5 & 0 & 0 \\ 0 & -5 & 0 & 8 & -3 \\ 0 & 0 & 0 & -3 & 3 \end{bmatrix}$$

- D；加权度矩阵，把相邻边的权重加起来
- W：邻接矩阵，主对角线上都是0
- L：拉普拉斯矩阵

拉普拉斯矩阵半正定

注意公式推导做了拆分

拉普拉斯矩阵半正定

$L = D - W$

$$
\begin{aligned}
f^T L f &= f^T D f - f^T W f \\
&= \sum_{i=1}^{n} d_i f_i^2 - \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} f_i f_j \\
&= \frac{1}{2}\left(\sum_{i=1}^{n} d_i f_i^2 - 2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} f_i f_j + \sum_{j=1}^{n} d_j f_j^2 \right) \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \left(f_i - f_j \right)^2
\end{aligned}
$$

算法流程

有了以上基础之后

算法流程
构造带权重的图
计算图的拉普拉斯矩阵
求解广义特征值问题

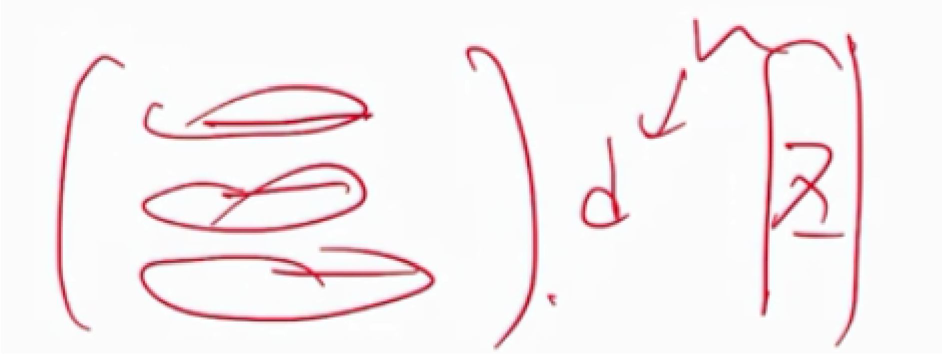
构造带权重的图

$$\left\| x_i - x_j \right\|^2 < \varepsilon$$

近邻规则

计算权重

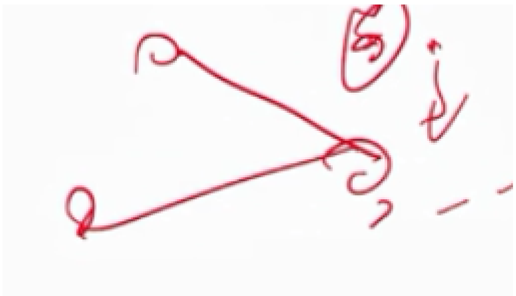
$$w_{ij} = \exp\left(-\frac{\left\| x_i - x_j \right\|^2}{t} \right)$$



怎么判断图的连通性

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 < \varepsilon$$

近邻规则



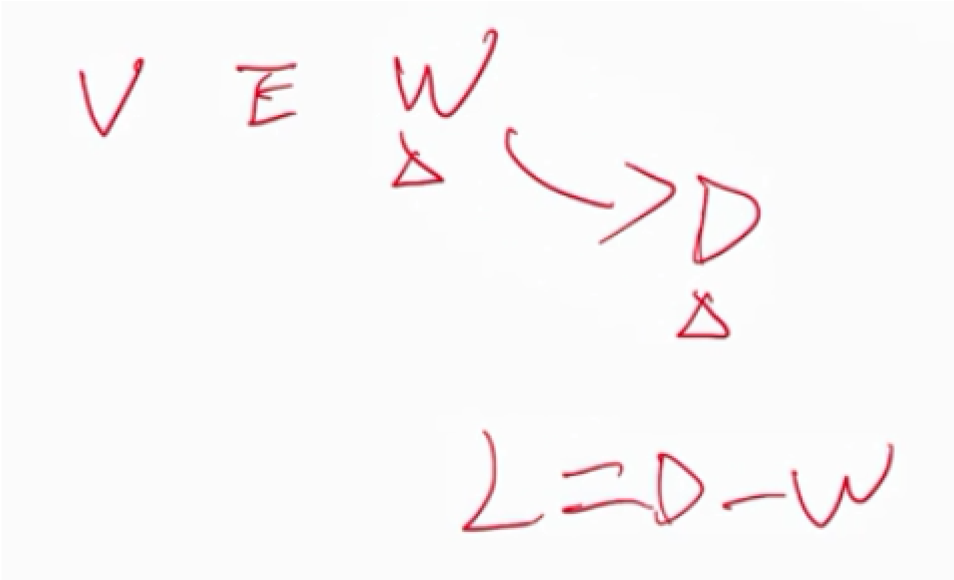
特征映射这部分注意推导

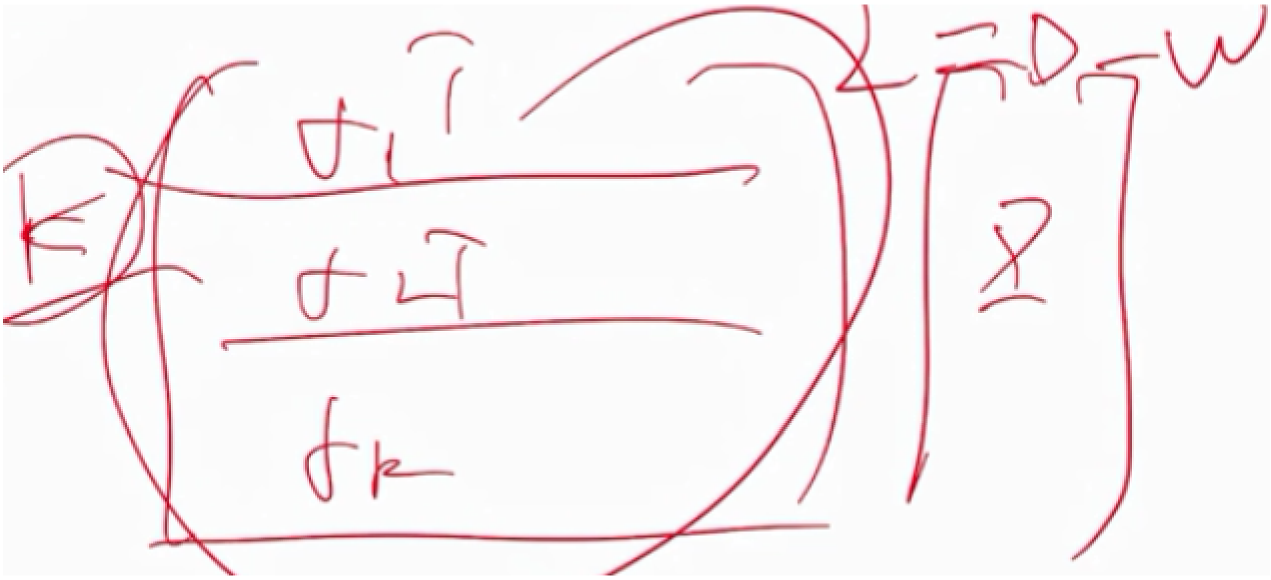
特征映射

$$\mathrm{Lf} = \lambda \mathrm{Df}$$

$$0 = \lambda_0 \leq \ldots \leq \lambda_{k-1}$$

$$\mathrm{f}_0, \ldots, \mathrm{f}_{k-1}$$





处理特征值问题

9.7 局部保持投影

局部保持投影也是基于图论的算法。

局部保持投影

$$XLX^{T}a = \lambda XDX^{T}a$$

$$\lambda_0 < \ldots . < \lambda_{l-1}$$

$$a_0, \ldots, a_{l-1}$$

$$x_i \rightarrow y_i = A^{T}x_i, A = (a_0, a_1, \ldots, a_{l-1})$$

拉普拉斯的映射和PCA是一样的

9.8 等距映射

等距映射
利用了测地线的概念
将数据投影到低维空间之后，尽量保持原有的距离关系

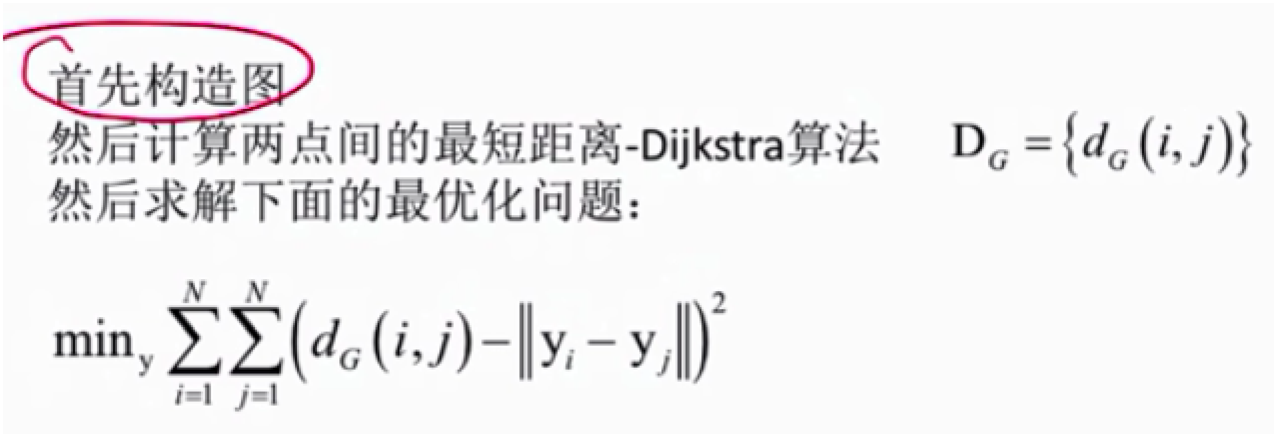
什么是测地线？

什么是测地线？

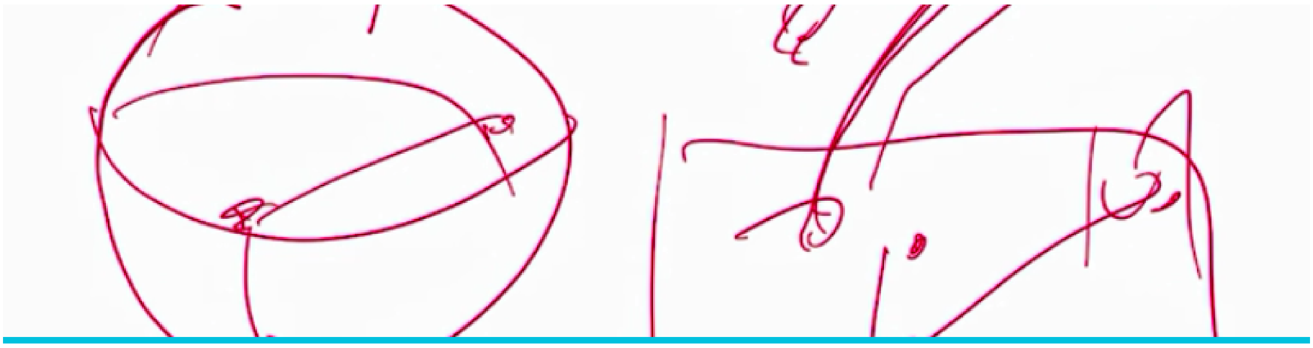
例如美国和中国远，

拿地图做例子

- 以前高中几何，球面上两点最短的距离是弧长。



迪杰特斯拉算法



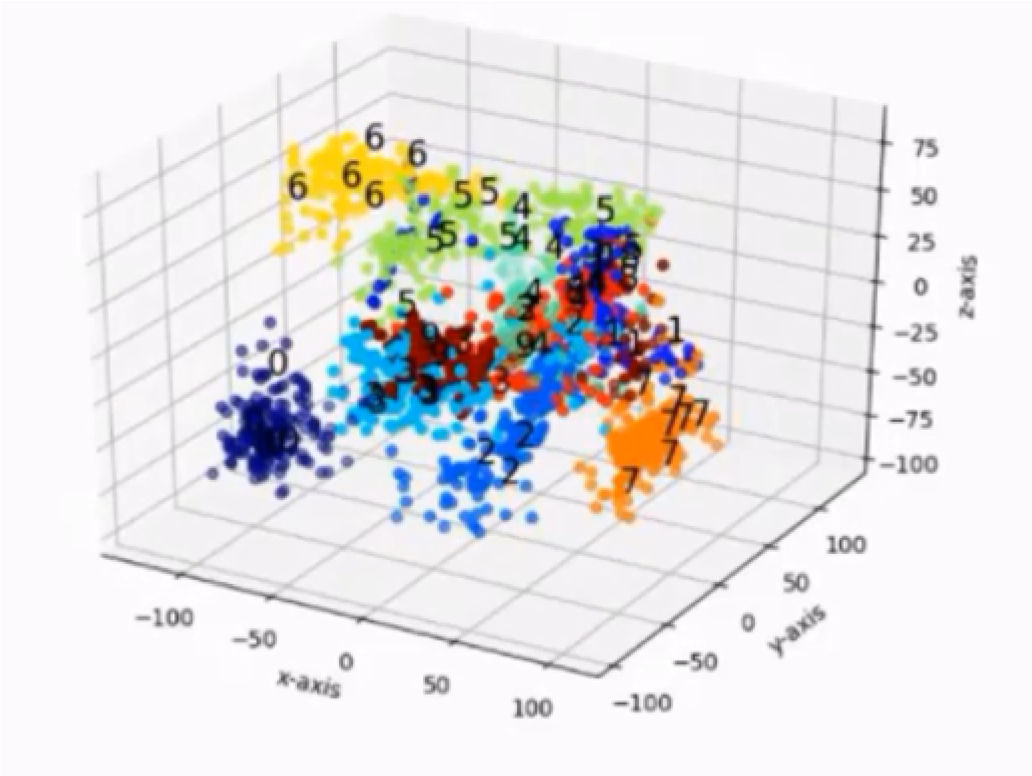
总结：

- 计算图
- 计算测地线
- 求解最优化问题，保持相对距离

9.9 实验环节

在云端实验室

这节课所学的算法都有



流行学算法，2000年诞生之后，挺火热的。但是很少有实际的应用。

9.10 本集总结