



第十讲 降维与度量学习

高级机器学习



提纲

- k-近邻学习
- 低维嵌入
- 流形学习
- 度量学习

k 近邻学习器

课外知识：
懒惰学习 = 事先没有分类器，见到测试样本再开始准备分类器

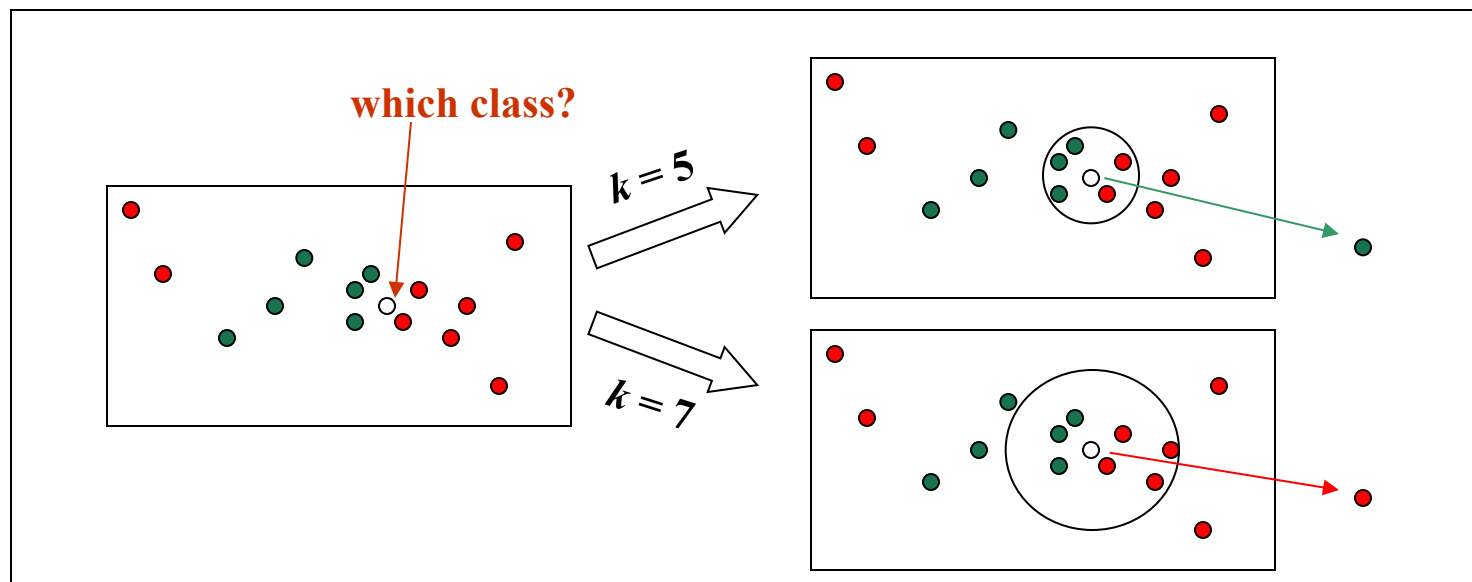
k 近邻 (k -Nearest Neighbor, k NN)

懒惰学习 (lazy learning) 的代表

基本思路：

近朱者赤，近墨者黑

(投票法；平均法)



关键： k 值选取；距离计算

最近邻学习器重要性质

给定测试样本 \mathbf{x} , 若其最近邻样本为 \mathbf{z} , 则最近邻分类器出错的概率就是 \mathbf{x} 和 \mathbf{z} 类别标记不同的概率,

$$P(err) = 1 - \sum_{c \in \mathcal{Y}} P(c | \mathbf{x})P(c | \mathbf{z}).$$

$$P(err) = 1 - \sum_{c \in \mathcal{Y}} P(c | \mathbf{x})P(c | \mathbf{z})$$

$$\simeq 1 - \sum_{c \in \mathcal{Y}} P^2(c | \mathbf{x})$$

$$\leq 1 - P^2(c^* | \mathbf{x})$$

$$= (1 + P(c^* | \mathbf{x})) (1 - P(c^* | \mathbf{x}))$$

$$\leq 2 \times (1 - P(c^* | \mathbf{x})).$$

最近邻分离器的泛化错误率

不会超过

贝叶斯最优分类器错误率的两倍!

kNN如此完美; 但是在真实的应用中, 我们是否能够准确的找到 k 近邻呢?

维数灾难

但是在真实的应用中，我们是否能够准确的找到 k 近邻呢？

密采样(dense sampling)假设：样本的每个 ϵ -邻域都有近邻

如果近邻的距离阈值设为 10^{-3}

假定维度为20，如果样本需要满足密采样条件
需要的样本数量近 10^{60}

想象一下：一张并不是很清晰的图像：70余万维
我们为了找到恰当的近邻，需要多少样本？

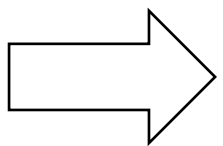


维数灾难

高维空间给距离计算带来很大的麻烦

当维数很高时，甚至连计算内积都不再容易

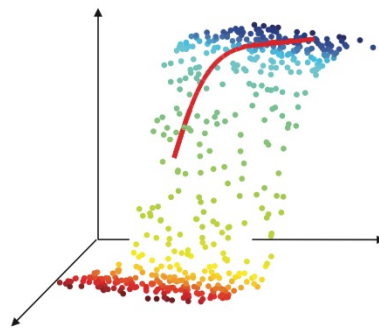
更严重的是：样本变得稀疏，近邻容易不准



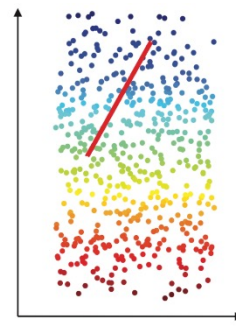
降维（将高维样本降到低维）

为什么能进行降维？

数据样本虽是高维的，但与学习任务密切相关的**也许**仅是某个低维空间，即高维空间中的一个**低维“嵌入” (embedding)**



(a) 三维空间中观察到的样本点



(b) 二维空间中的曲面

图 10.2 低维嵌入示意图

提纲

- k-近邻学习
- 低维嵌入
- 流形学习
- 度量学习

多维缩放方法 (MDS, Multiple Dimensional Scaling)

MDS旨在寻找一个低维子空间，使得距离和样本原有距离近似不变

从距离矩阵到内积矩阵

假定 \mathbf{z} 已被中心化，即 $\sum_{i} \mathbf{z}_i = 0$

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

$$\sum_{i=1}^m dist_{ij}^2 = \text{tr}(\mathbf{B}) + mb_{jj}, \quad dist_{i.}^2 = \frac{1}{m} \sum_{j=1}^m dist_{ij}^2,$$

$$\sum_{j=1}^m dist_{ij}^2 = \text{tr}(\mathbf{B}) + mb_{ii}, \quad dist_{.j}^2 = \frac{1}{m} \sum_{i=1}^m dist_{ij}^2,$$

$$\sum_{i=1}^m \sum_{j=1}^m dist_{ij}^2 = 2m \text{tr}(\mathbf{B}), \quad dist_{..}^2 = \frac{1}{m^2} \sum_{i=1}^m \sum_{j=1}^m dist_{ij}^2,$$

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

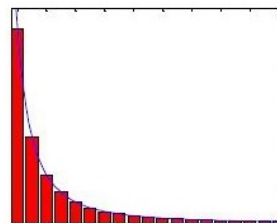
寻找低维子空间尽量保持样本内积不变

思路：特征值分解 (常用技巧)

设样本之间的内积矩阵均为 \mathbf{B}
对 \mathbf{B} 进行特征值分解：

$$\mathbf{B} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^T$$

由谱分解的数学性质，我们知道：



特征谱

谱分布长尾：存在相当数量的小特征值

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

$$\mathbf{B} = \mathbf{Z}^T \mathbf{Z} \in \mathbb{R}^{m \times m}$$

主要思路：内积保距

提纲

- k-近邻学习
- 低维嵌入
- 流形学习
- 度量学习

流形学习 - ISOMAP

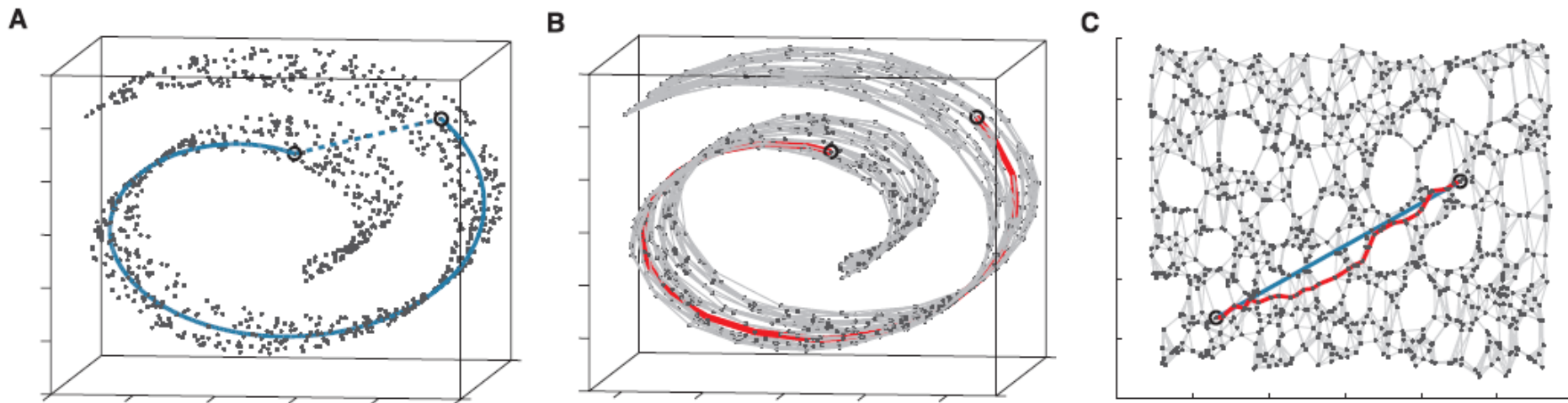


Fig. 3. The "Swiss roll" data set, illustrating how Isomap exploits geodesic paths for nonlinear dimensionality reduction. (A) For two arbitrary points (circled) on a nonlinear manifold, their Euclidean distance in the high-dimensional input space (length of dashed line) may not accurately reflect their intrinsic similarity, as measured by geodesic distance along the low-dimensional manifold (length of solid curve). (B) The neighborhood graph G constructed in step one of Isomap (with $K = 7$ and $N =$

1000 data points) allows an approximation (red segments) to the true geodesic path to be computed efficiently in step two, as the shortest path in G . (C) The two-dimensional embedding recovered by Isomap in step three, which best preserves the shortest path distances in the neighborhood graph (overlaid). Straight lines in the embedding (blue) now represent simpler and cleaner approximations to the true geodesic paths than do the corresponding graph paths (red).

www.sciencemag.org SCIENCE VOL 290 22 DECEMBER 2000

基本步骤:

- 构造近邻图

- 基于最短路径算法近似任意两点之间的测地线(geodesic)距离

- 基于距离矩阵通过MDS获得低维嵌入

关键思路：测地线距离（近似）、保距

流形学习 - LLE

基本步骤：

- 为每个样本构造近邻集合 Q_i
- 为每个样本计算基于 Q_i 的线性重构系数

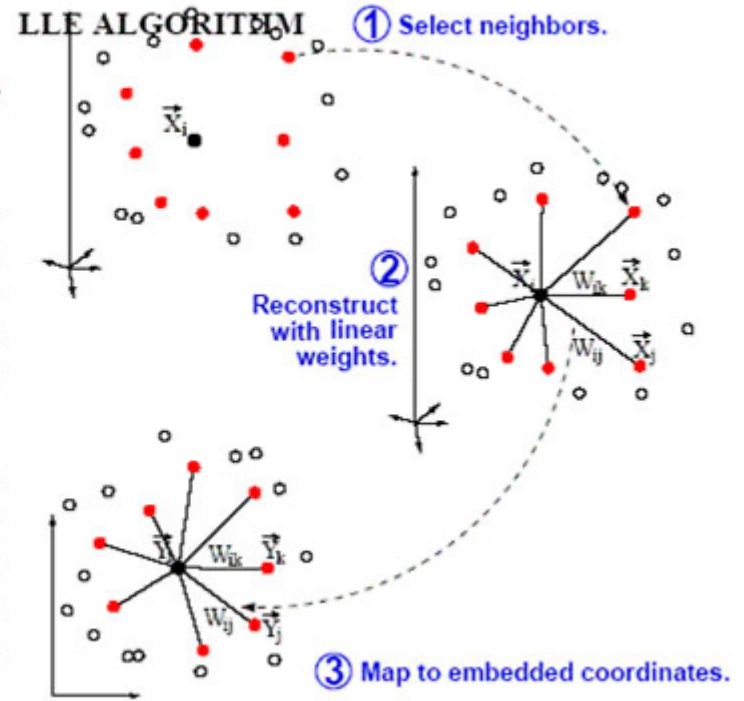
$$\min_{w_1, w_2, \dots, w_m} \sum_{i=1}^m \left\| \mathbf{x}_i - \sum_{j \in Q_i} w_{ij} \mathbf{x}_j \right\|_2^2$$

s.t. $\sum_{j \in Q_i} w_{ij} = 1,$

- 在低维空间中保持 w_{ij} 不变，求解下式

$$\min_{z_1, z_2, \dots, z_m} \sum_{i=1}^m \left\| z_i - \sum_{j \in Q_i} w_{ij} z_j \right\|_2^2$$

- LLE ALGORITHM**
1. Compute the neighbors of each data point, \bar{X}_i .
 2. Compute the weights W_{ij} that best reconstruct each data point \bar{X}_i from its neighbors, minimizing the cost in Equation (1) by constrained linear fits.
 3. Compute the vectors \bar{Y}_i best reconstructed by the weights W_{ij} , minimizing the quadratic form in Equation (2) by its bottom nonzero eigenvectors.



www.sciencemag.org SCIENCE VOL 290 22 DECEMBER 2000

关键思路：重构权值

提纲

- k-近邻学习
- 低维嵌入
- 流形学习
- 度量学习

距离度量的种类

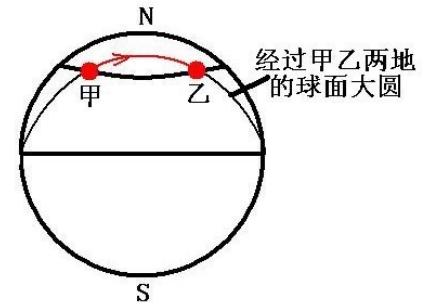
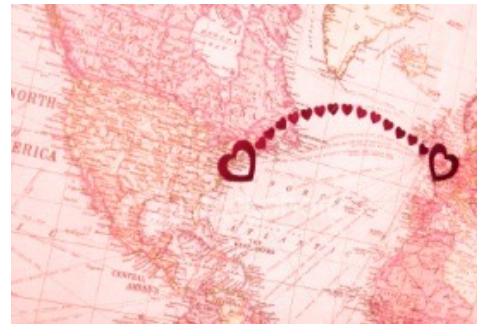
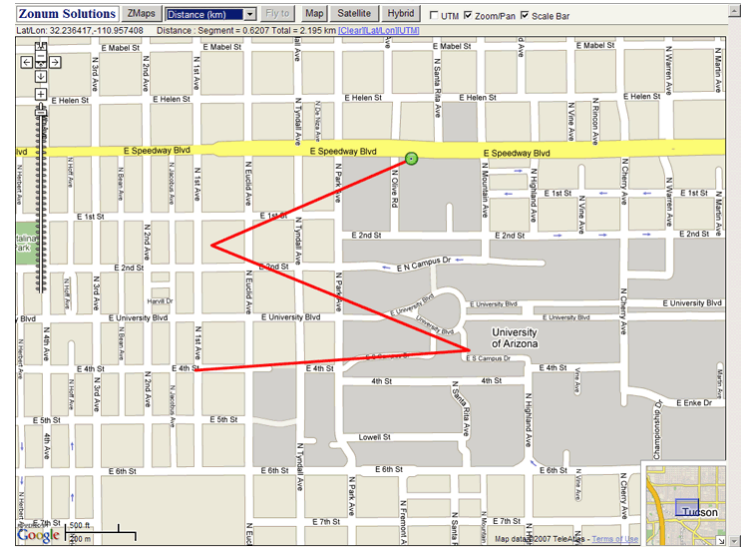
选哪个距离度量呢？能否学出来一个最好的距离度量？——距离度量学习

□ 欧式距离

$$d(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

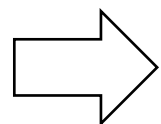
□ 曼哈顿距离

□ 测地距离 (Geodesic Distance)



距离度量学习 (distance metric learning)

降维希望找到一个“合适”低维空间下的距离度量



能否直接“学出”合适的距离？

首先，要能够通过“参数化”来学习距离度量

马氏距离 (Mahalanobis distance) 是一个很好的选择：

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

其中 \mathbf{M} 称为“度量矩阵”，数学上它是一个半正定对称矩阵

距离度量学习就是要对 \mathbf{M} 进行学习

距离度量学习 (distance metric learning)

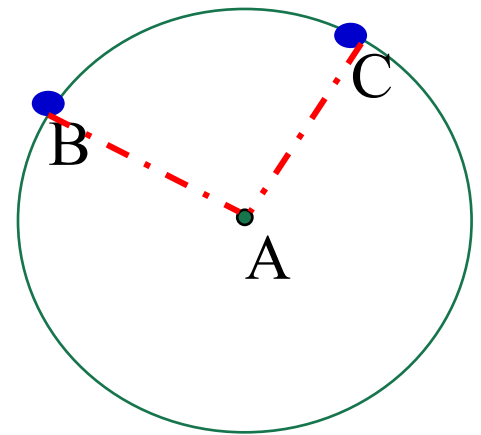
为什么是马氏距离？

“什么是距离？”

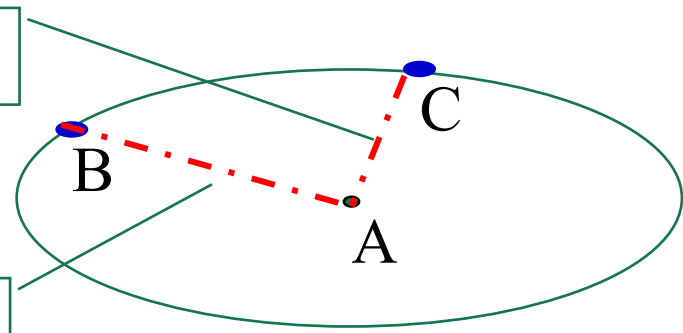
欧氏距离的缺陷——各个方向同等重要

但是：

- 有缘千里来相会
(表面欧氏距离大但实际距离小)
- 无缘对面手难牵
(表面欧式距离小但实际距离大)
- 马氏距离应运而生



咫尺天涯



天涯咫尺

距离度量学习 (distance metric learning)

M怎么学习？ 要有个合适的目标函数

□ **目标1：** 结合具体分类器的性能

例如，以近邻分类器的性能为目标，则得到经典的**NCA**算法

□ **目标2：** 结合领域知识

例如，若已知“必连” (must-link) 约束集合 \mathcal{M} 与“勿连” (cannot-link) 约束集合 \mathcal{C} ，则可通过求解下述凸优化问题得到 **M**：

$$\begin{aligned} \min_{\mathbf{M}} \quad & \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{M}} \|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{M}}^2 \\ \text{s.t.} \quad & \sum_{(\mathbf{x}_i, \mathbf{x}_k) \in \mathcal{C}} \|\mathbf{x}_i - \mathbf{x}_k\|_{\mathbf{M}}^2 \geq 1, \\ & \mathbf{M} \succeq 0, \end{aligned}$$

距离度量学习 – NCA: Neighborhood Component Analysis

近邻分类器在进行判别时通常使用多数投票法, 替换为概率投票法. 对于任意样本 x_j , 它对 x_i 分类结果影响的概率为

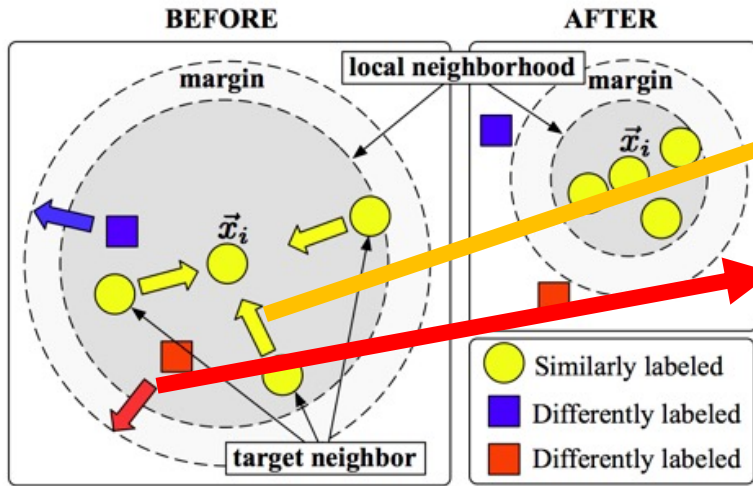
$$p_{ij} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{M}}^2)}{\sum_l \exp(-\|\mathbf{x}_i - \mathbf{x}_l\|_{\mathbf{M}}^2)},$$

x_i 样本的LOO正确率:
$$p_i = \sum_{j \in \Omega_i} p_{ij},$$

训练集上的LOO正确率:
$$\sum_{i=1}^m p_i = \sum_{i=1}^m \sum_{j \in \Omega_i} p_{ij}.$$

NCA的优化目标:
$$\mathbf{M} = \mathbf{P}\mathbf{P}^T \quad \min_{\mathbf{P}} 1 - \sum_{i=1}^m \sum_{j \in \Omega_i} \frac{\exp(-\|\mathbf{P}^T \mathbf{x}_i - \mathbf{P}^T \mathbf{x}_j\|_2^2)}{\sum_l \exp(-\|\mathbf{P}^T \mathbf{x}_i - \mathbf{P}^T \mathbf{x}_l\|_2^2)}.$$

距离度量学习 – LMNN: Large Margin Nearest Neighbors



$$\epsilon_{\text{pull}}(\mathbf{L}) = \sum_{j \rightsquigarrow i} \|\mathbf{L}(\vec{x}_i - \vec{x}_j)\|^2$$

+

$$\epsilon_{\text{push}}(\mathbf{L}) = \sum_{i, j \rightsquigarrow i} \sum_l (1 - y_{il}) [1 + \|\mathbf{L}(\vec{x}_i - \vec{x}_j)\|^2 - \|\mathbf{L}(\vec{x}_i - \vec{x}_l)\|^2]_+$$

=

$$\epsilon(\mathbf{L}) = (1 - \mu) \epsilon_{\text{pull}}(\mathbf{L}) + \mu \epsilon_{\text{push}}(\mathbf{L})$$

Minimize $(1 - \mu) \sum_{i, j \rightsquigarrow i} (\vec{x}_i - \vec{x}_j)^\top \mathbf{M}(\vec{x}_i - \vec{x}_j) + \mu \sum_{i, j \rightsquigarrow i, l} (1 - y_{il}) \xi_{ijl}$ **subject to:**

(1) $(\vec{x}_i - \vec{x}_l)^\top \mathbf{M}(\vec{x}_i - \vec{x}_l) - (\vec{x}_i - \vec{x}_j)^\top \mathbf{M}(\vec{x}_i - \vec{x}_j) \geq 1 - \xi_{ijl}$

(2) $\xi_{ijl} \geq 0$

(3) $\mathbf{M} \succeq 0$.

小结

□ k-近邻学习

- 知道怎么实现，以及知道它重要的理论性质和现实不可行之原因

□ 低维嵌入

- 知道MDS的思路和能够回答为什么

□ 流形学习

- 知道ISOMAP和LLE是怎么实现（核心思路）

□ 度量学习

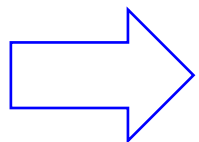
- 知道度量学习的主要思路——学习参数化马氏距离

主成分分析 (Principal Component Analysis, PCA)

正交属性空间中的样本点，如何使用一个超平面对所有样本进行恰当的表达？

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



主成分分析的两种等价推导

PCA - 最近重构性

对样本进行中心化: $\sum_i \mathbf{x}_i = \mathbf{0}$

假定投影变换后得到的新坐标系为 $\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_d\}$, 其中 \mathbf{w}_i 是标准正交基向量

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

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 $\mathbf{z}_i = (z_{i1}; z_{i2}; \dots; z_{id'})$ $z_{ij} = \mathbf{w}_j^T \mathbf{x}_i$

若基于 \mathbf{z}_i 来重构 \mathbf{x}_i , 则会得到 $\hat{\mathbf{x}}_i = \sum_{j=1}^{d'} z_{ij} \mathbf{w}_j$.

PCA - 最近重构性 (续)

原样本点 \mathbf{x}_i 与基于投影重构的样本点 $\hat{\mathbf{x}}_i$ 之间的距离为

$$\begin{aligned} \sum_{i=1}^m \left\| \sum_{j=1}^{d'} z_{ij} \mathbf{w}_j - \mathbf{x}_i \right\|_2^2 &= \sum_{i=1}^m \mathbf{z}_i^T \mathbf{z}_i - 2 \sum_{i=1}^m \mathbf{z}_i^T \mathbf{W}^T \mathbf{x}_i + \text{const} \\ &\propto -\text{tr} \left(\mathbf{W}^T \left(\sum_{i=1}^m \mathbf{x}_i \mathbf{x}_i^T \right) \mathbf{W} \right). \end{aligned}$$

\mathbf{w}_j 是正交基, $\sum_i \mathbf{x}_i \mathbf{x}_i^T$ 是协方差矩阵, 于是由最近重构性, 有:

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

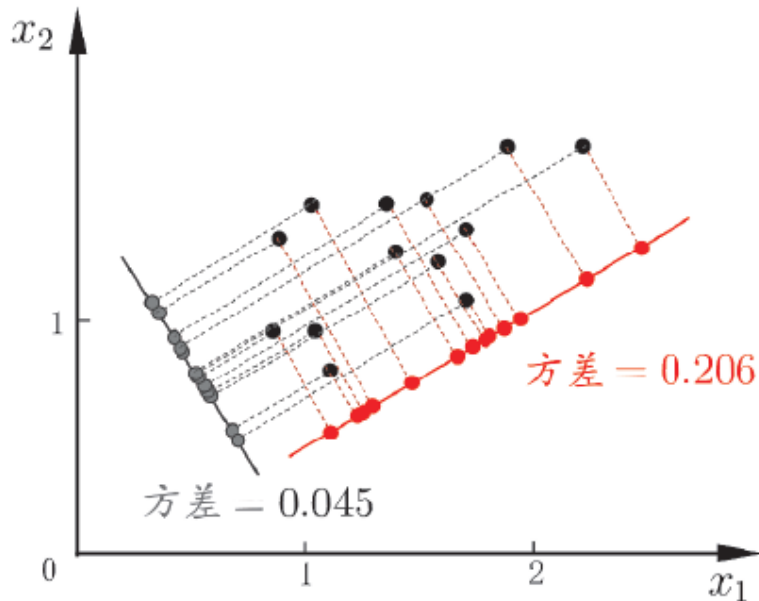
主要思路: 重构误差

这就是主成分分析的优化目标

PCA - 最大可分性

样本点 \mathbf{x}_i 在新空间中超平面上的投影是 $\mathbf{W}^T \mathbf{x}_i$ ，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

投影后样本点的方差是 $\sum_i \mathbf{W}^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{W}$



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

等价于：
$$\min_{\mathbf{W}} -\text{tr}(\mathbf{W}^T \mathbf{X} \mathbf{X}^T \mathbf{W})$$
$$\text{s.t. } \mathbf{W}^T \mathbf{W} = \mathbf{I}.$$

PCA 求解

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

使用拉格朗日乘子法可得

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

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

关键变量：子空间方差

PCA 应用

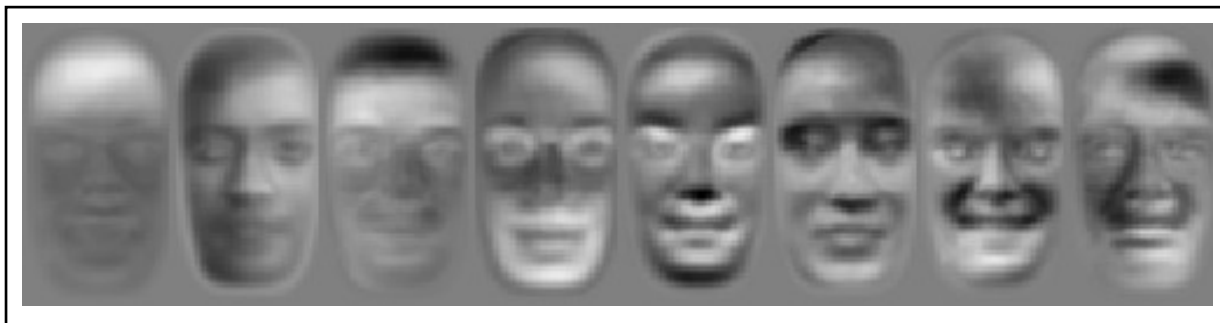
d' 的设置:

- 用户指定
- 在低维空间中对k近邻或其他分类器进行交叉验证
- 设置重构阈值, 例如 $t=95\%$, 然后选取最小的 d' 使得
$$\frac{\sum_{i=1}^{d'} \lambda_i}{\sum_{i=1}^d \lambda_i} \geq t.$$

PCA 是最常用的降维方法, 在不同领域有不同的称谓

例如在人脸识别中该技术被称为“特征脸”(eigenface)

因为若将前 d' 个特征值对应的特征向量还原为图像, 则得到



小结

□ k-近邻学习

- 知道怎么实现，以及知道它重要的理论性质和现实不可行之原因

□ 低维嵌入

- 知道MDS的思路和能够回答为什么

□ 流形学习

- 知道ISOMAP和LLE是怎么实现（核心思路）

□ 度量学习

- 知道度量学习的主要思路——学习参数化马氏距离

□ PCA

- PCA原理和计算