十、降维与度量学习

主讲教师：周志华

k近邻学习器

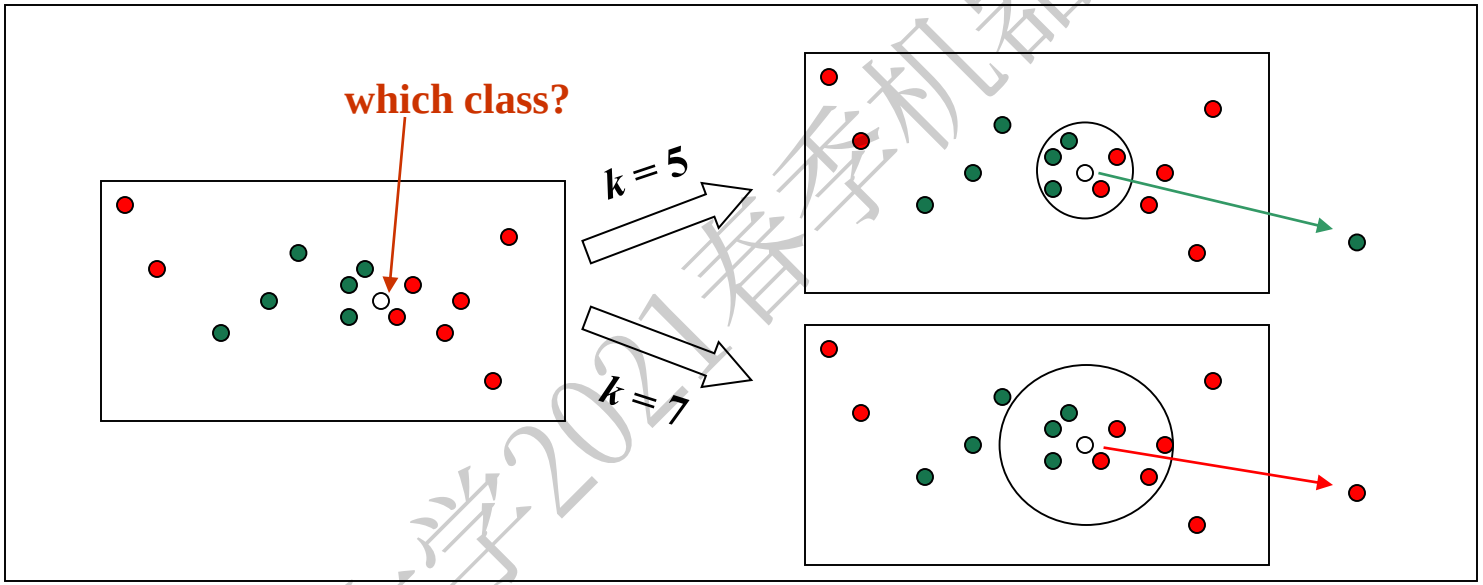
k 近邻 (k-Nearest Neighbor, kNN)

懒惰学习 (lazy learning) 的代表

基本思路：

近朱者赤，近墨者黑

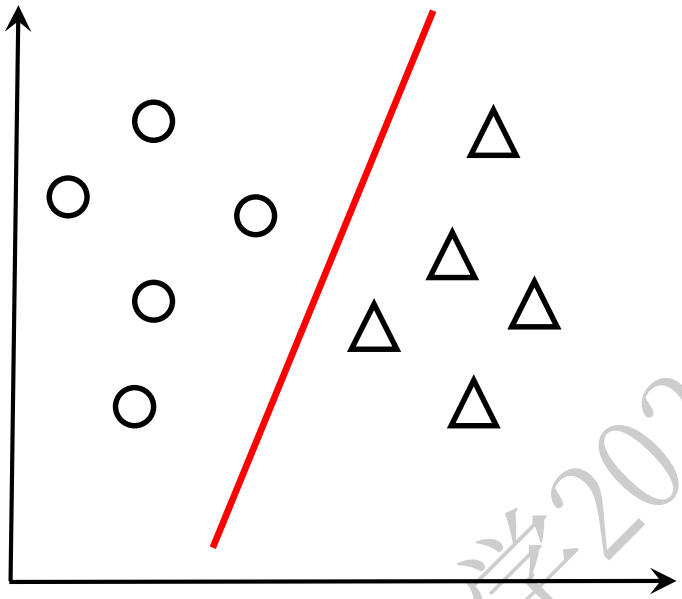
(投票法；平均法)



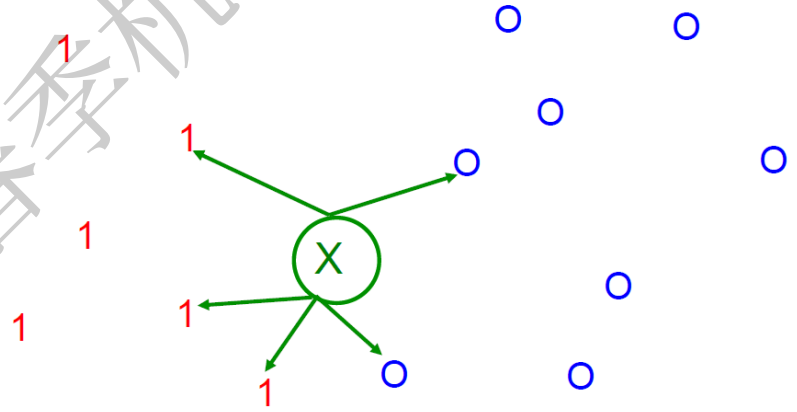
关键： k 值选取；距离计算

k近邻学习器

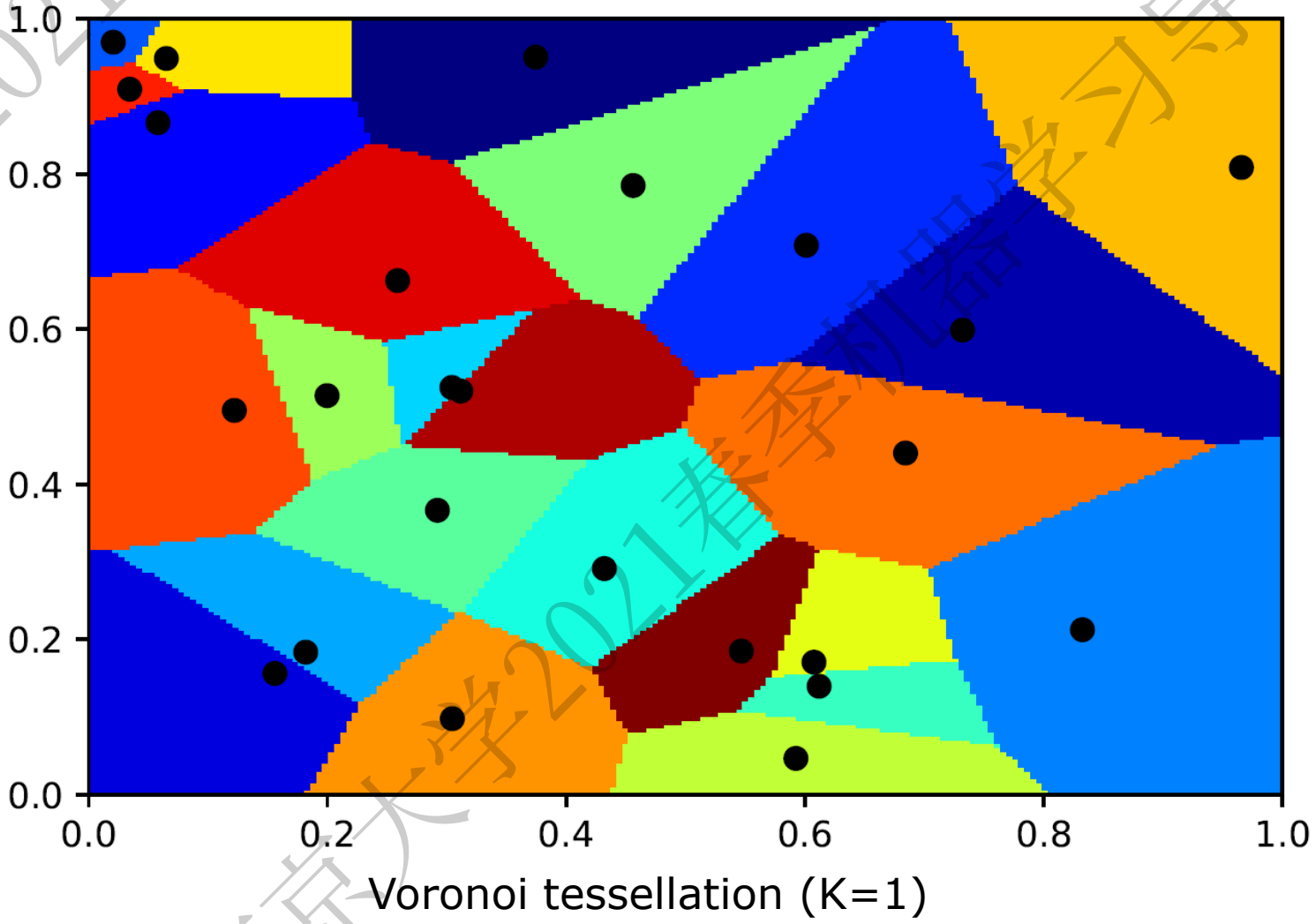
参数化模型



非参数化模型



*k*近邻学习器



最近邻学习器和贝叶斯最优分类器

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

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

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！

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

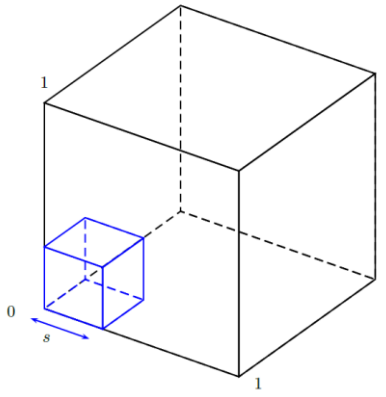
但是在真实的应用中，我们是否能够准确的找到k近邻呢？

密采样(dense sampling)

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

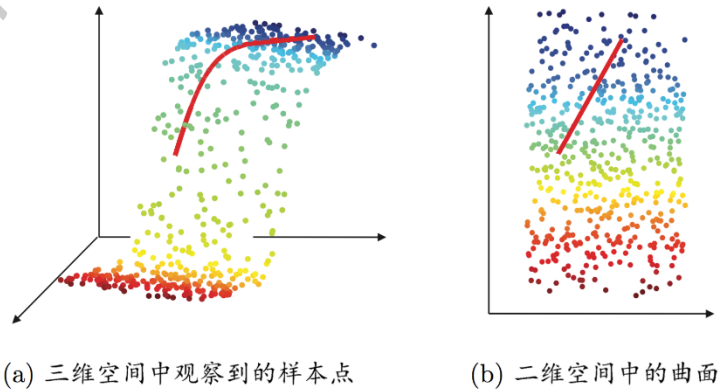
高维空间给距离计算带来很大的麻烦，当维数很高时甚至连计算**内积**都不再容易

考虑到：计算开销，可视化，特征提取等方面

⇒ 降维

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分布，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

线性降维

假设从高维空间到低维空间的函数映射是线性的，寻找**线性投影W**，将数据投影到低维

主成分分析 (Principal Component Analysis, PCA)

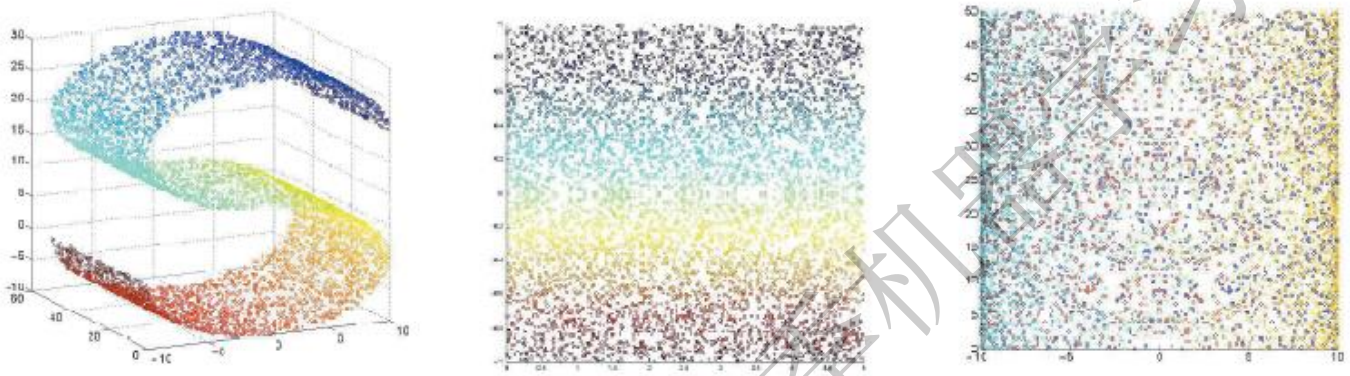
正交属性空间中的样本点，如何使用一个超平面对所有样本进行恰当的表达？

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

非线性降维

- 然而在许多现实任务中，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

图 10.6　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，此情形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．

非线性降维的常用方法：

- 核化线性降维：如KPCA, KLDA, …
- 流形学习 (manifold learning)

多维缩放方法（MDS）

MDS (Multiple Dimensional Scaling) 旨在寻找一个低维子空间，
样本在此子空间内的距离和样本原有距离尽量保持不变

考虑问题：如何在距离矩阵和内积矩阵之间建立联系？

考虑变形问题：如何在低维子空间和高维空间之间保持样本之间的内积不变？

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

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

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

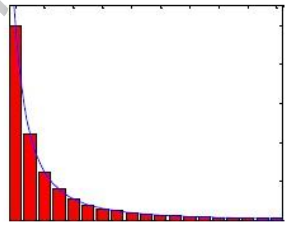
$$\sum_{i=1}^{m}\sum_{j=1}^{m} dist_{ij}^2 = 2m\,\mathrm{tr}(\mathbf{B})\,, \qquad dist_{\cdot\cdot}^2 = \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} dist_{ij}^2\,,$$

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

设样本之间的内积矩阵均为B
对B进行特征值分解：

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

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

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

关键变量：距离、内积，保距

流形学习 - ISOMAP

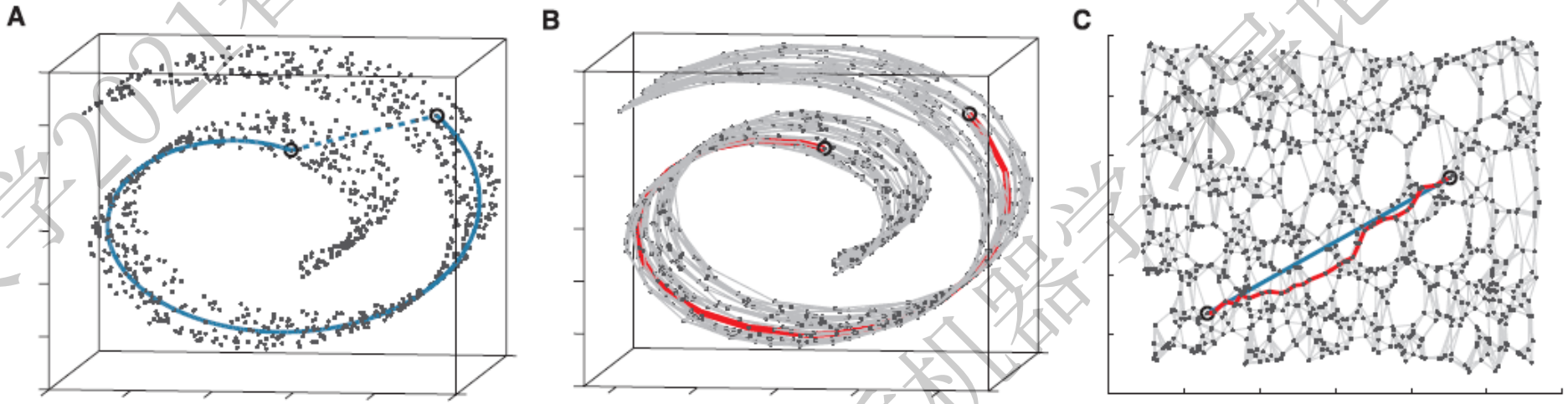


Fig. 3. The "Swiss roll" data set, illustrating how Isomap exploits geodesic paths for nonlinear dimensionality reduction. (**A**) For two arbitrary points (circled) on a nonlinear manifold, their Euclidean distance in the high-dimensional input space (length of dashed line) may not accurately reflect 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：

关键变量：测地线距离（近似）、保距

- 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

局部线性嵌入 - LLE

基本步骤：

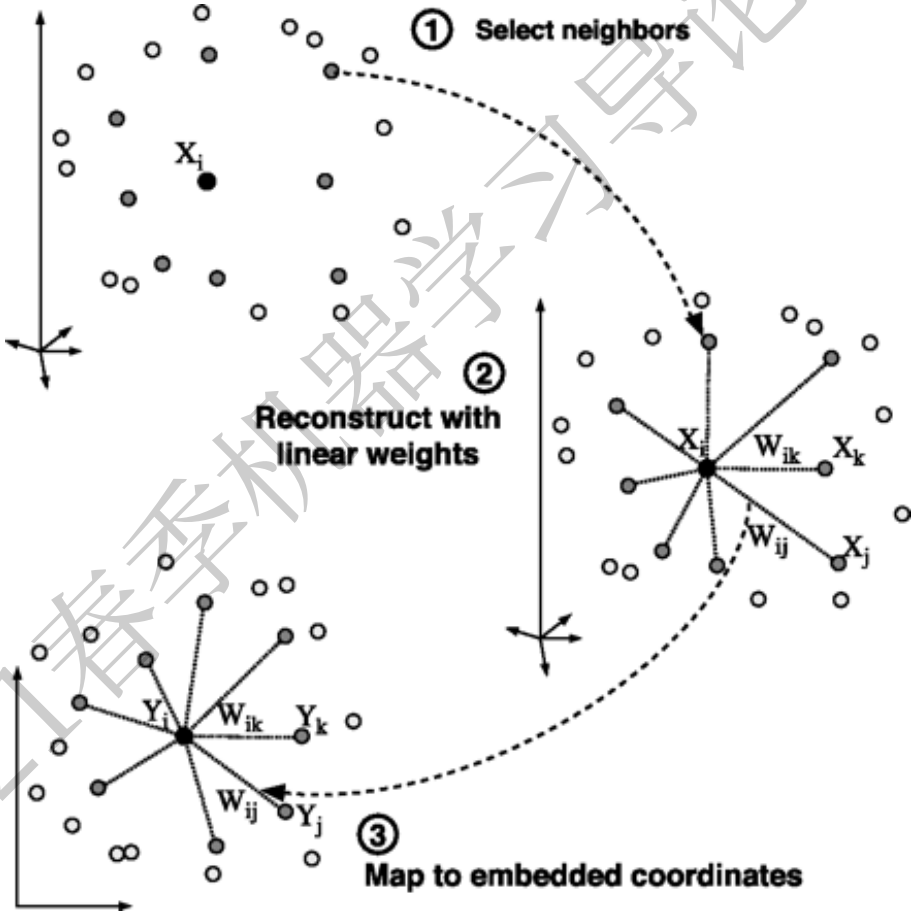
- 为每个样本构造近邻集合Q_i
- 为每个样本计算基于Q_i的线性重构系数

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

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

- 在低维空间中保持w_{ij}不变，求解下式

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



关键变量：重构权值

距离度量

距离度量 (distance metric) 需满足的基本性质：

非负性: $\mathrm{dist}(\boldsymbol{x}_i, \boldsymbol{x}_j) \geqslant 0$;

同一性: $\mathrm{dist}(\boldsymbol{x}_i, \boldsymbol{x}_j) = 0$ 当且仅当 $\boldsymbol{x}_i = \boldsymbol{x}_j$;

对称性: $\mathrm{dist}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \mathrm{dist}(\boldsymbol{x}_j, \boldsymbol{x}_i)$;

直递性: $\mathrm{dist}(\boldsymbol{x}_i, \boldsymbol{x}_j) \leqslant \mathrm{dist}(\boldsymbol{x}_i, \boldsymbol{x}_k) + \mathrm{dist}(\boldsymbol{x}_k, \boldsymbol{x}_j)$.

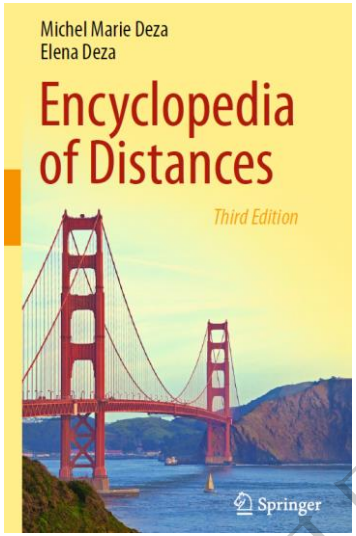
常用距离形式：

闵可夫斯基距离 (Minkowski distance)

$$\mathrm{dist}_{\mathrm{mk}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\sum_{u=1}^{n} |x_{iu} - x_{ju}|^p \right)^{\frac{1}{p}}$$

$p = 2$: 欧氏距离(Euclidean distance)

$p = 1$: 曼哈顿距离(Manhattan distance)



距离度量学习 (distance metric learning)

降维的主要目的是希望找到一个“合适的”低维空间

每个空间对应了在样本属性上定义的一个距离度量

⇨ 能否直接“学出”合适的距离？

首先，要有可以通过学习来“参数化”的距离度量形式

马氏距离 (Mahalanobis distance) 是一个很好的选择：

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

其中 $\mathbf{M}$ 是一个半正定对称矩阵，亦称“度量矩阵”

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)

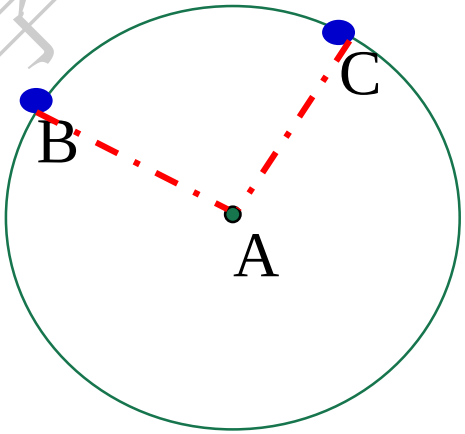
为什么要马氏距离 ？

我们回顾一下“什么是距离？”再思考一下“距离度量 ”

度量的是什么？

It’s a long distance to **walk**….

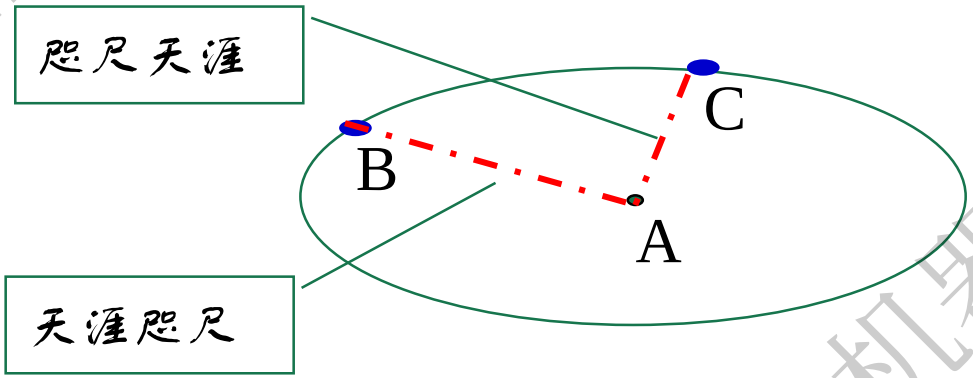
旅行的开销！



欧氏距离的缺陷 —— 各向同性

但是：

- 有缘千里来相会
  (欧氏距离大但开销少)
- 无缘对面手难牵
  (优势距离小但开销大)
- 马氏距离应运而生



距离度量学习 (distance metric learning)

其次，对 $\mathbf{M}$ 进行学习的目标是什么？

- 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$\mathcal{C}$，则可通过求解这个凸优化问题得到 $\mathbf{M}$:

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

如何进行相似性的指导

使用二元组、三元组构成弱监督信息

更相似？

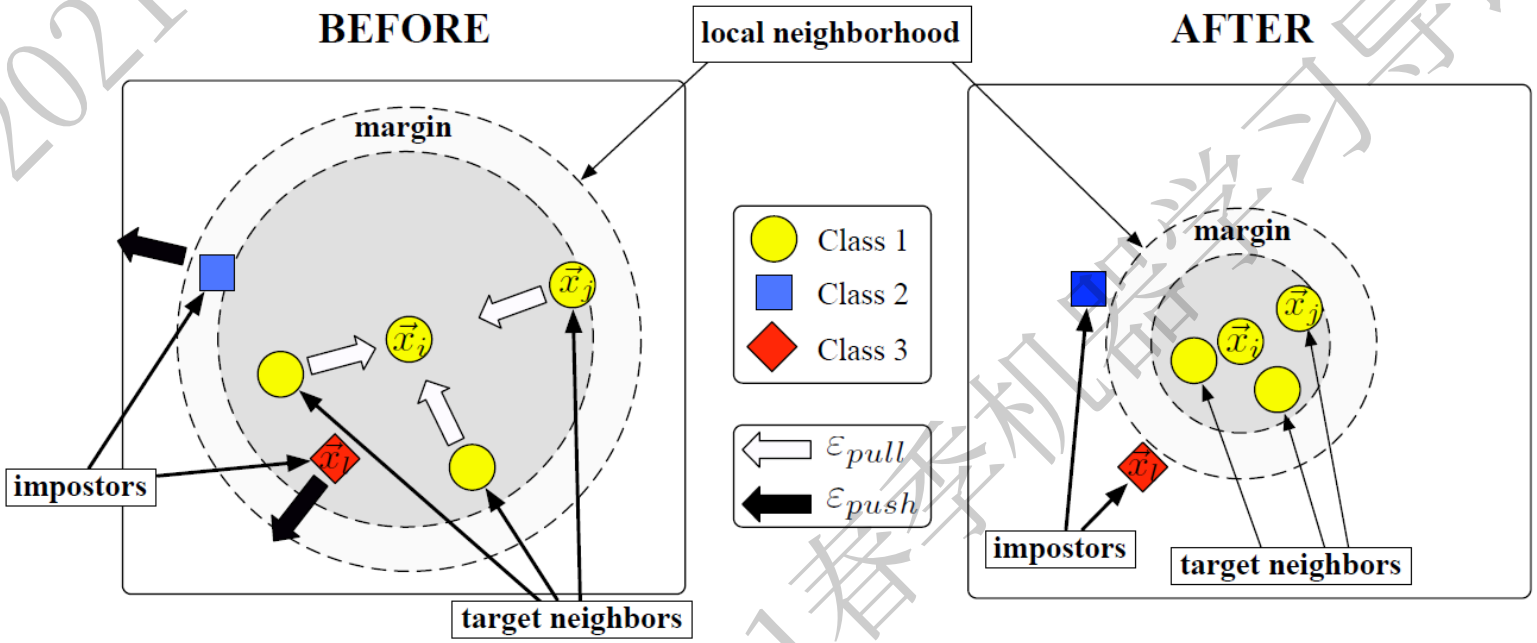
更不相似？

相似性关系能够被广泛获取

识图系统中
用户的点击

社交网络中用
户的好友关系

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



不相似样本远离

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

相似样本接近

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

$$\varepsilon(\mathbf{L}) = (1-\mu)\,\varepsilon_{pull}(\mathbf{L}) + \mu \varepsilon_{push}(\mathbf{L}).$$

习题2-2 广义瑞利商

在面对多类样本时, FDA 需要求解广义瑞利商:

$$\max_{\boldsymbol{w}} \quad \frac{\boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w}}. \tag{5}$$

(1) **[15pts]** 请证明: 瑞利商满足

$$\lambda_{\min}(\boldsymbol{A}) \leq \frac{\boldsymbol{w}^T \boldsymbol{A} \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{w}} \leq \lambda_{\max}(\boldsymbol{A}), \tag{6}$$

其中 $\boldsymbol{A}$ 为实对称矩阵, $\lambda(\boldsymbol{A})$ 为 $\boldsymbol{A}$ 的特征值.

Solution. (1) 设 $\boldsymbol{A}\boldsymbol{\xi}_i = \lambda_i \boldsymbol{\xi}_i$, $i = 1, 2, \cdots, n$, $\{\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_n\}$ 构成一组单位正交基, 从而存在一组 $\{\alpha_i\}_{i=1}^n$ 使得

$$\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i \boldsymbol{\xi}_i,$$

代入瑞利商的定义, 可得

$$\frac{\boldsymbol{w}^T \boldsymbol{A} \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{w}} = \frac{\left(\sum_i \alpha_i \boldsymbol{\xi}_i^T\right) \boldsymbol{A} \left(\sum_i \alpha_i \boldsymbol{\xi}_i\right)}{\left(\sum_i \alpha_i \boldsymbol{\xi}_i^T\right) \left(\sum_i \alpha_i \boldsymbol{\xi}_i\right)} = \frac{\sum_i \alpha_i^2 \lambda_i}{\sum_i \alpha_i^2}.$$

可见瑞利商是 $\boldsymbol{A}$ 的特征值的加权平均, 所以

$$\lambda_{\min}(\boldsymbol{A}) \leq \frac{\boldsymbol{w}^T \boldsymbol{A} \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{w}} \leq \lambda_{\max}(\boldsymbol{A}).$$

习题2-2 广义瑞利商

在面对多类样本时, FDA 需要求解广义瑞利商:

$$\max_{\boldsymbol{w}} \quad \frac{\boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w}}. \tag{5}$$

(2) **[15pts]** 请证明: 如果 $\boldsymbol{A}$ 为实对称矩阵, $\boldsymbol{B}$ 为正定矩阵, 那么广义瑞利商满足

$$\lambda_{\min}(\boldsymbol{B}^{-1}\boldsymbol{A}) \leq \frac{\boldsymbol{w}^T \boldsymbol{A} \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{B} \boldsymbol{w}} \leq \lambda_{\max}(\boldsymbol{B}^{-1}\boldsymbol{A}). \tag{7}$$

(2) 对 $\boldsymbol{B}$ 进行正交对角化, 可得 $\boldsymbol{B} = \boldsymbol{P}^T \boldsymbol{\Lambda} \boldsymbol{P}$, 其中 $\boldsymbol{P}$ 为特征向量构成的正交矩阵, $\boldsymbol{\Lambda}$ 为对角线元素为相应特征值 (正数) 的对角矩阵. 于是,

$$\frac{\boldsymbol{w}^T \boldsymbol{A} \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{B} \boldsymbol{w}} = \frac{\tilde{\boldsymbol{w}}^T \boldsymbol{C} \tilde{\boldsymbol{w}}}{\tilde{\boldsymbol{w}}^T \tilde{\boldsymbol{w}}},$$

其中 $\tilde{\boldsymbol{w}} = \boldsymbol{\Lambda}^{\frac{1}{2}} \boldsymbol{P} \boldsymbol{w}$, $\boldsymbol{C} = \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{P} \boldsymbol{A} \boldsymbol{P}^T \boldsymbol{\Lambda}^{-\frac{1}{2}}$. 由前一问知

$$\lambda_{\min}(\boldsymbol{C}) \leq \frac{\tilde{\boldsymbol{w}}^T \boldsymbol{C} \tilde{\boldsymbol{w}}}{\tilde{\boldsymbol{w}}^T \tilde{\boldsymbol{w}}} \leq \lambda_{\max}(\boldsymbol{C}).$$

因此 $\boldsymbol{C}$ 与 $\boldsymbol{B}^{-1}\boldsymbol{A}$ 有相同的特征值, 所以

$$\lambda_{\min}(\boldsymbol{B}^{-1}\boldsymbol{A}) \leq \frac{\boldsymbol{w}^T \boldsymbol{A} \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{B} \boldsymbol{w}} \leq \lambda_{\max}(\boldsymbol{B}^{-1}\boldsymbol{A}).$$

习题4-1 神经网络

在训练神经网络之前，我们需要确定的是整个网络的结构，在确定结构后便可以输入数据进行端到端的学习过程。考虑西瓜书第101-102页以及书中图5.7中描述的神经网络，即：输入是d维向量$\mathbf{x} \in \mathbb{R}^d$，隐藏层由$q$个隐层单元组成，输出层为$l$个输出单元，其中隐层第$h$个神经元的阈值用$\gamma_h$表示，输出层第$j$个神经元的阈值用$\theta_j$表示，输入层第$i$个神经元与隐层第$h$个神经元之间的连接权重为$v_{ih}$，隐层第$h$个神经元与输出层第$j$个神经元之间的连接权重为$w_{hj}$，记隐层第$h$个神经元接收到的输入为$\alpha_h = \sum_{i=1}^{d} v_{ih} x_i$，输出为$b_h = f(\alpha_h - \gamma_h)$，输出层第$j$个神经元接收到的输入为$\beta_j = \sum_{h=1}^{q} w_{hj} b_h$，输出为$\hat{y}_j = f(\beta_j - \theta_j)$，$f$为对应的激活函数。

(2) **[20 pts]** 神经网络学习分类问题时，模型输出层更加常用的设置是$Softmax$加交叉熵损失，即：假定隐层单元的激活函数是$Sigmoid$函数不变，对输出层，令$z_j = \beta_j - \theta_j$，则输出为$\hat{y}_j = \frac{e^{z_j}}{\sum_{i=1}^{l} e^{z_i}}$，损失函数$E = -\sum_{j=1}^{l} y_j \log \hat{y}_j$，$y_j$为该输出对应真实标记。请给出此时的梯度$\frac{\partial E}{\partial w_{hj}}$，$\frac{\partial E}{\partial \theta_j}$，$\frac{\partial E}{\partial v_{ih}}$和$\frac{\partial E}{\partial \gamma_h}$。（需给出计算步骤，可以像西瓜书一样定义新的符号$g_j$、$e_h$，但已有符号需使用题目中给定符号表示）

习题4-3 朴素贝叶斯分类器

通过对课本的学习，我们了解了采用“属性条件独立性假设”的朴素贝叶斯分类器。现在我们有如下表所示的一个数据集，其中x_1与x_2为特征，其取值集合分别为$x_1 = \{-1, 0, 1\}$，$x_2 = \{B, M, S\}$，y为类别标记，其取值集合为$y = \{0, 1\}$：

表 1: 数据集

编号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_1	-1	-1	-1	-1	-1	0	0	0	0	0	0	1	1	1	1
x_2	B	M	M	B	B	B	M	M	S	S	S	M	M	S	S
y	0	0	1	1	0	0	0	1	1	1	1	1	1	1	0

(1) **[5pts]**通过查表直接给出的$x = \{0, B\}$的类别；

(1) **[5pts]** 找到表中编号为6的样本满足$x = \{0, B\}$，故$x = \{0, B\}$的类别是0。

习题4-3 朴素贝叶斯分类器

通过对课本的学习，我们了解了采用“属性条件独立性假设”的朴素贝叶斯分类器。现在我们有如下表所示的一个数据集，其中x_1与x_2为特征，其取值集合分别为$x_1 = \{-1, 0, 1\}$，$x_2 = \{B, M, S\}$，y为类别标记，其取值集合为$y = \{0, 1\}$：

表 1: 数据集

编号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_1	-1	-1	-1	-1	-1	0	0	0	0	0	0	1	1	1	1
x_2	B	M	M	B	B	B	M	M	S	S	S	M	M	S	S
y	0	0	1	1	0	0	0	1	1	1	1	1	1	1	0

(2) **[15pts]** 使用所给训练数据，学习一个朴素贝叶斯分类器，并用学习得到的分类器确定$x = \{0, B\}$的标记，要求写出详细计算过程；

$$P(y=0)P(x_1=0|y=0)P(x_2=B|y=0) = \frac{1}{15}$$
$$P(y=1)P(x_1=0|y=1)P(x_2=B|y=1) = \frac{4}{135}$$

由于$\frac{1}{15} > \frac{4}{135}$，因此朴素贝叶斯分类器将样本类别标记判定是0。

习题4-3 朴素贝叶斯分类器

通过对课本的学习，我们了解了采用“属性条件独立性假设”的朴素贝叶斯分类器。现在我们有如下表所示的一个数据集，其中x_1与x_2为特征，其取值集合分别为$x_1 = \{-1, 0, 1\}$，$x_2 = \{B, M, S\}$，y为类别标记，其取值集合为$y = \{0, 1\}$：

表 1: 数据集

编号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_1	-1	-1	-1	-1	-1	0	0	0	0	0	0	1	1	1	1
x_2	B	M	M	B	B	B	M	M	S	S	S	M	M	S	S
y	0	0	1	1	0	0	0	1	1	1	1	1	1	1	0

(3) **[15pts]** 使用“拉普拉斯修正”，再学习一个朴素贝叶斯分类器，以及重新计算$x = \{0, B\}$的标记，要求写出详细计算过程。

$$P(y=0) = \frac{6}{15}, P(y=1) = \frac{9}{15} \Rightarrow \begin{aligned} P(y=0) &= \frac{7}{17} \\ P(y=1) &= \frac{10}{17} \end{aligned}$$

$$\begin{aligned} P(y=0)P(x_1=0|y=0)P(x_2=B|y=0) &\approx 0.041 \\ P(y=1)P(x_1=0|y=1)P(x_2=B|y=1) &\approx 0.033 \end{aligned}$$

习题5-1 PCA

$\boldsymbol{x} \in \mathbb{R}^D$ 是一个随机向量，其均值和协方差分别是 $\boldsymbol{\mu_x} = \mathbb{E}(\boldsymbol{x}) \in \mathbb{R}^D$，$\Sigma_x = \mathbb{E}(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top \in \mathbb{R}^{D \times D}$。定义随机变量 $y_i = \boldsymbol{u}_i^\top \boldsymbol{x} + a_i \in \mathbb{R}, i = 1, \ldots, d \leq D$ 为 $\boldsymbol{x}$ 的主成分，其中 $\boldsymbol{u}_i \in \mathbb{R}^D$ 是单位向量($\boldsymbol{u}_i \top \boldsymbol{u}_i = 1$)，$a_i \in \mathbb{R}$，$\{y_i\}_{i=1}^d$ 是互不相关的零均值随机变量，它们的方差满足 $\mathrm{Var}(y_1) \geq \mathrm{Var}(y_2) \geq \cdots \geq \mathrm{Var}(y_d)$。假设 Σ_x 没有重复的特征值，请证明：

1. **[5pts]** $a_i = -\boldsymbol{u}_i^\top \boldsymbol{\mu_x}, i = 1, \ldots, d$。

1. 由 y_i 是零均值的随机变量可得

$$\mathbb{E}[y_i] = \boldsymbol{\mu}_i^\top \mathbb{E}[\boldsymbol{x}_i] + \mathbb{E}[a_i] = \boldsymbol{u}_i^\top \boldsymbol{\mu_x} + a_i = 0, \quad i = 1, \ldots, d$$

则 $a_i = -\boldsymbol{u}_i^\top \boldsymbol{\mu_x}, i = 1, \ldots, d$

习题5-1 PCA

$\boldsymbol{x} \in \mathbb{R}^D$ 是一个随机向量，其均值和协方差分别是 $\boldsymbol{\mu_x} = \mathbb{E}(\boldsymbol{x}) \in \mathbb{R}^D$，$\Sigma_x = \mathbb{E}(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top \in \mathbb{R}^{D \times D}$。定义随机变量 $y_i = \boldsymbol{u}_i^\top \boldsymbol{x} + a_i \in \mathbb{R}, i = 1, \ldots, d \leq D$ 为 $\boldsymbol{x}$ 的主成分，其中 $\boldsymbol{u}_i \in \mathbb{R}^D$ 是单位向量($\boldsymbol{u}_i \top \boldsymbol{u}_i = 1$)，$a_i \in \mathbb{R}$，$\{y_i\}_{i=1}^d$ 是互不相关的零均值随机变量，它们的方差满足 $\mathrm{Var}(y_1) \geq \mathrm{Var}(y_2) \geq \cdots \geq \mathrm{Var}(y_d)$。假设 Σ_x 没有重复的特征值，请证明：

2. **[10pts]** $\boldsymbol{u}_1$ 是 Σ_x 最大的特征值对应的特征向量。

$$\mathrm{Var}(y_1) = \mathbb{E}\left[\left(\boldsymbol{u}_1^\top(\boldsymbol{x} - \boldsymbol{u}_x)\right)^2\right] = \mathbb{E}\left[\boldsymbol{u}_1^\top(\boldsymbol{x} - \boldsymbol{u}_x)(\boldsymbol{x} - \boldsymbol{u}_x)^\top \boldsymbol{u}_1\right] = \boldsymbol{u}_1^\top \Sigma_x \boldsymbol{u}_1$$

则优化问题为

$$\max_{u_1 \in \mathbb{R}^D} \boldsymbol{u}_1^\top \Sigma_x \boldsymbol{u}_1 \quad \text{s.t.} \quad \boldsymbol{u}_1^\top \boldsymbol{u}_1 = 1$$

由拉格朗日函数

$$\mathcal{L} = \boldsymbol{u}_1^\top \Sigma_x \boldsymbol{u}_1 + \lambda_1 \left(1 - \boldsymbol{u}_1^\top \boldsymbol{u}_1\right)$$

其中 λ_1 为拉格朗日乘子。

令上式对 $\boldsymbol{u}_1$ 和 λ_1 求导分别等于0，可得

$$\Sigma_x \boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1 \quad \text{and} \quad \boldsymbol{u}_1^\top \boldsymbol{u}_1 = 1$$

因此 $\boldsymbol{u}_1$ 是 λ_1 所对应的 Σ_x 的特征向量。

习题5-1 PCA

$\boldsymbol{x} \in \mathbb{R}^D$ 是一个随机向量，其均值和协方差分别是 $\boldsymbol{\mu}_{\boldsymbol{x}} = \mathbb{E}(\boldsymbol{x}) \in \mathbb{R}^D$，$\Sigma_x = \mathbb{E}(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top \in \mathbb{R}^{D \times D}$。定义随机变量 $y_i = \boldsymbol{u}_i^\top \boldsymbol{x} + a_i \in \mathbb{R}, i = 1, \ldots, d \leq D$ 为 $\boldsymbol{x}$ 的主成分，其中 $\boldsymbol{u}_i \in \mathbb{R}^D$ 是单位向量($\boldsymbol{u}_i \top \boldsymbol{u}_i = 1$)，$a_i \in \mathbb{R}$，$\{y_i\}_{i=1}^d$ 是互不相关的零均值随机变量，它们的方差满足 $\operatorname{Var}(y_1) \geq \operatorname{Var}(y_2) \geq \cdots \geq \operatorname{Var}(y_d)$。假设 Σ_x 没有重复的特征值，请证明：

3. **[15pts]** $\boldsymbol{u}_2^\top \boldsymbol{u}_1 = 0$，且 $\boldsymbol{u}_2$ 是 Σ_x 第二大特征值对应的特征向量。

类似于 $\boldsymbol{u}_1$ 的求解过程，求解 $\boldsymbol{u}_2$ 对应的最大化目标是 $\boldsymbol{u}_2^\top \Sigma_x \boldsymbol{u}_2$，多一个约束条件 $\boldsymbol{u}_1^\top \boldsymbol{u}_2 = 0$。即优化问题为

$$\max_{u_2 \in \mathbb{R}^D} \boldsymbol{u}_2^\top \Sigma_x \boldsymbol{u}_2 \quad \text{s.t.} \quad \boldsymbol{u}_2^\top \boldsymbol{u}_2 = 1 \quad \text{and} \quad \boldsymbol{u}_1^\top \boldsymbol{u}_2 = 0$$

定义拉格朗日函数

$$\mathcal{L} = \boldsymbol{u}_2^\top \Sigma_x \boldsymbol{u}_2 + \lambda_2 \left(1 - \boldsymbol{u}_2^\top \boldsymbol{u}_2\right) + \gamma \boldsymbol{u}_1^\top \boldsymbol{u}_2$$

对 $\boldsymbol{u}_2, \lambda_2, \gamma$ 分别求导令导数等于0，得

$$\Sigma_x \boldsymbol{u}_2 + \frac{\gamma}{2} \boldsymbol{u}_1 = \lambda_2 \boldsymbol{u}_2, \quad \boldsymbol{u}_2^\top \boldsymbol{u}_2 = 1 \quad \text{and} \quad \boldsymbol{u}_1^\top \boldsymbol{u}_2 = 0$$

习题5-1 PCA

$\boldsymbol{x} \in \mathbb{R}^D$ 是一个随机向量，其均值和协方差分别是 $\boldsymbol{\mu_x} = \mathbb{E}(\boldsymbol{x}) \in \mathbb{R}^D$，$\Sigma_x = \mathbb{E}(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top \in \mathbb{R}^{D \times D}$。定义随机变量 $y_i = \boldsymbol{u}_i^\top \boldsymbol{x} + a_i \in \mathbb{R}, i = 1, \ldots, d \leq D$ 为 $\boldsymbol{x}$ 的主成分，其中 $\boldsymbol{u}_i \in \mathbb{R}^D$ 是单位向量($\boldsymbol{u}_i \top \boldsymbol{u}_i = 1$)，$a_i \in \mathbb{R}$，$\{y_i\}_{i=1}^d$ 是互不相关的零均值随机变量，它们的方差满足 $\mathrm{Var}(y_1) \geq \mathrm{Var}(y_2) \geq \cdots \geq \mathrm{Var}(y_d)$。假设 Σ_x 没有重复的特征值，请证明：

3. **[15pts]** $\boldsymbol{u}_2^\top \boldsymbol{u}_1 = 0$，且 $\boldsymbol{u}_2$ 是 Σ_x 第二大特征值对应的特征向量。

由 y_1, y_2 不相关，得 $\mathrm{Cov}(y_1, y_2) = \mathbb{E}(y_1 y_2) - \mathbb{E}(y_1)\mathbb{E}(y_2) = 0$。

又 y_1, y_2 均值为0，所以 $\mathbb{E}(y_1 y_2) = \mathbb{E}(y_1)\mathbb{E}(y_2) = 0$。即

$$\mathbb{E}\left[\left(\boldsymbol{u}_1^\top(\boldsymbol{x} - \boldsymbol{u}_x)\right)\left(\boldsymbol{u}_2^\top(\boldsymbol{x} - \boldsymbol{u}_x)\right)\right] = \mathbb{E}\left[\boldsymbol{u}_1^\top(\boldsymbol{x} - \boldsymbol{u}_x)(\boldsymbol{x} - \boldsymbol{u}_x)^\top \boldsymbol{u}_2\right] = \boldsymbol{u}_1^\top \Sigma_x \boldsymbol{u}_2 = \lambda_1 \boldsymbol{u}_1^\top \boldsymbol{u}_2 = 0$$

由于 $\lambda_1 \neq 0$，因此 $\boldsymbol{u}_1^\top \boldsymbol{u}_2 = 0$。

$$\Sigma_x \boldsymbol{u}_2 + \frac{\gamma}{2}\boldsymbol{u}_1 = \lambda_2 \boldsymbol{u}_2, \quad \boldsymbol{u}_2^\top \boldsymbol{u}_2 = 1 \quad \text{and} \quad \boldsymbol{u}_1^\top \boldsymbol{u}_2 = 0$$

第一个式子左乘以 $\boldsymbol{u}_1^\top$ 可得，$\boldsymbol{u}_1^\top \Sigma_x \boldsymbol{u}_2 + \frac{\gamma}{2}\boldsymbol{u}_1^\top \boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1^\top \boldsymbol{u}_2 + \frac{\gamma}{2} = \lambda_2 \boldsymbol{u}_1^\top \boldsymbol{u}_2$，因此 $\gamma = 2(\lambda_2 - \lambda_1)\boldsymbol{u}_1^\top \boldsymbol{u}_2 = 0$，故 $\Sigma_x \boldsymbol{u}_2 = \lambda_2 \boldsymbol{u}_2$。则最大化的目标为 $\boldsymbol{u}_2^\top \Sigma_x \boldsymbol{u}_2 = \lambda_2 = \mathrm{Var}(y_2)$。

因为 Σ_x 没有重复的特征值，所以当 λ_2 为 Σ_x 第二大的特征值，$\boldsymbol{u}_2$ 为 Σ_x 第二大的特征值对应的特征向量时，$\boldsymbol{u}_2^\top \Sigma_x \boldsymbol{u}_2$ 取得最大值。

习题5-2 聚类

2 [30pts] Clustering

考虑p维特征空间里的混合模型

$$g(x) = \sum_{k=1}^{K} \pi_k g_k(x)$$

其中$g_k = N(\mu_k, \mathbf{I} \cdot \sigma^2)$，$\mathbf{I}$是单位矩阵，$\pi_k > 0$，$\sum_k \pi_k = 1$。$\{\mu_k, \pi_k\}, k = 1, \ldots, K$和$\sigma^2$是未知参数。

设有数据$x_1, x_2, \ldots, x_N \sim g(x)$,

1. **[10pts]** 请写出数据的对数似然。
2. **[15pts]** 请写出求解极大似然估计的EM算法。
3. **[5pts]** 请简要说明如果σ的值已知，并且$\sigma \to 0$，那么该EM算法就相当于K-means聚类。

习题5-3 集成学习

[20pts] Bagging 产生的每棵树是同分布的，那么 B 棵树均值的期望和其中任一棵树的期望是相同的。因此，Bagging 产生的偏差和其中任一棵树的偏差相同，Bagging 带来的性能提升来自于方差的降低。

我们知道，方差为 σ^2 的 B 个独立同分布的随机变量，其均值的方差为 $\frac{1}{B}\sigma^2$。如果这些随机变量是同分布的，但不是独立的，设两两之间的相关系数 $\rho > 0$，请推导均值的方差为 $\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$。

$$
\begin{aligned}
\operatorname{Var}\left(\frac{\sum_{i=1}^{B} X_i}{B}\right) &= \frac{1}{B^2}\operatorname{Var}\left(\sum x_i\right) \\
&= \frac{1}{B^2}\sum_{i=1}^{B}\operatorname{Var}\left(X_i\right) + \frac{1}{B^2}\sum_{i\neq j}^{B}\operatorname{Cov}\left(X_i, X_j\right) \\
&= \frac{\sigma^2}{B} + \frac{B-1}{B}\sigma^2\rho \\
&= \sigma^2\rho + \frac{1-\rho}{B}\sigma^2.
\end{aligned}
$$

随着B的增加，第二项逐渐消失，所以树与树之间的相关性限制了平均带来的好处

前往……