Case VI:
How to Measure the Difference

主讲教师：詹德川

k 近邻学习器

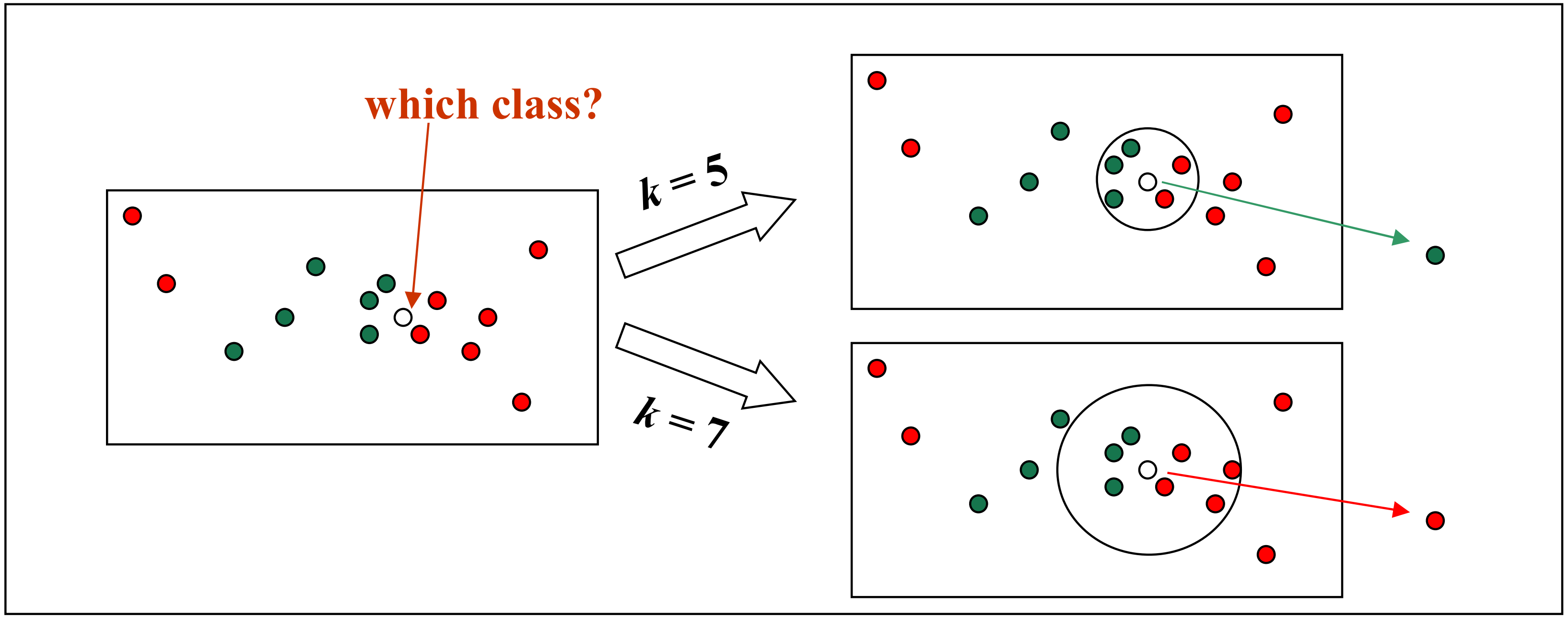
k 近邻 (k-Nearest Neighbor, kNN)

懒惰学习 (lazy learning) 的代表

基本思路：

近朱者赤，近墨者黑

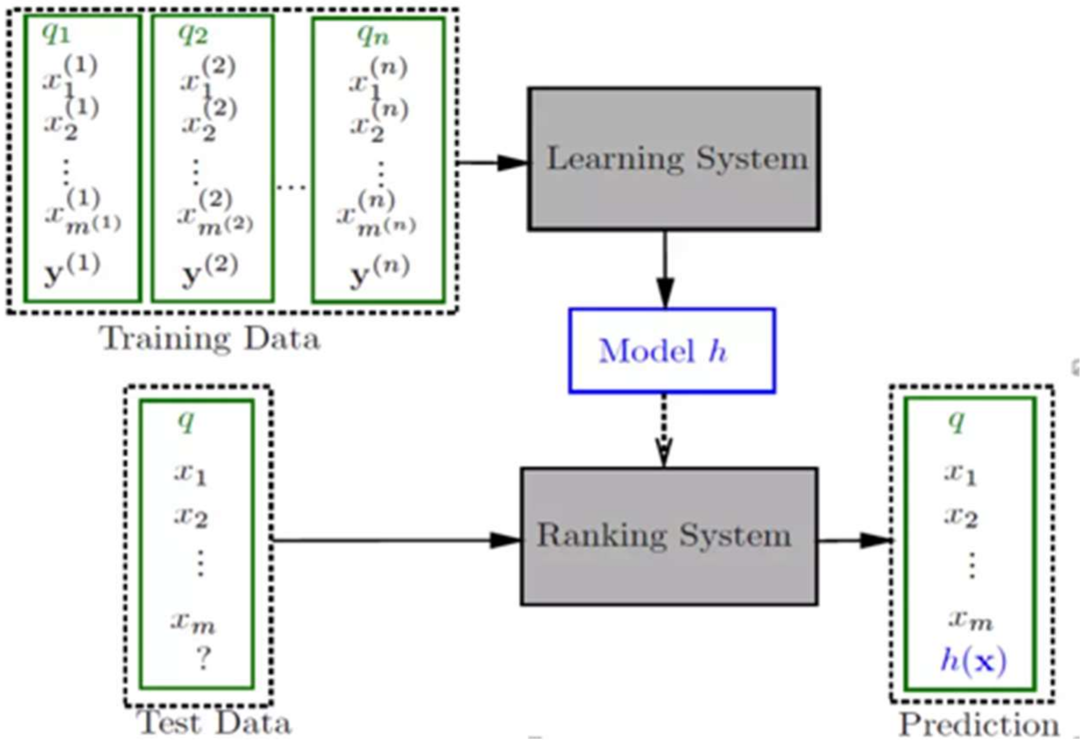
(投票法；平均法)



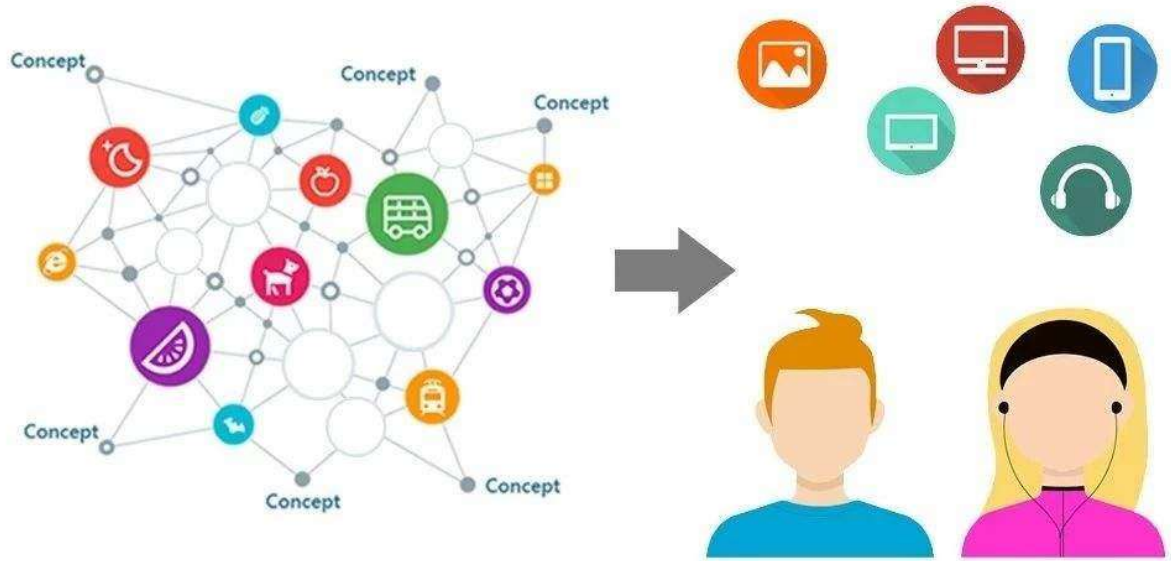
关键： k 值选取；距离计算

Information Retrieval

Ranking based
Similarity Measure



Recommendation Systems



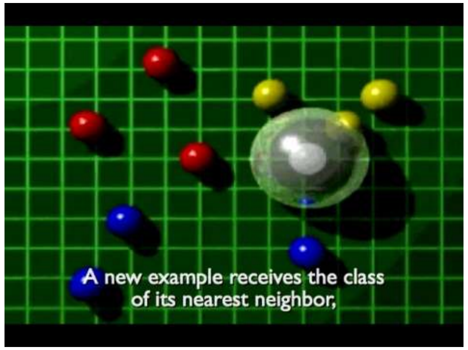
Similarity between
Concepts
Users
Items
Hybrids

距离度量的重要性

- 人类很早就意识到距离的重要性，并且在文明发展之初就以统一度量衡作为进步的标识

- 在信息检索中如何比较提交的查询和检索结果之间是否相似？
  - 使用距离来表示样本之间的不相似度

- 使用最近邻分类器如何判别一个样本属于哪个类别？
  - 近朱者，赤；近墨者，黑



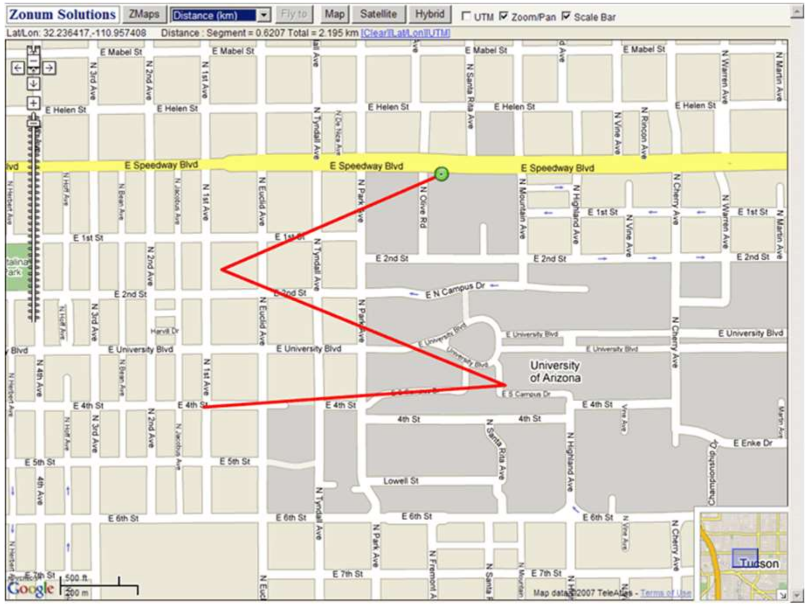
- Kernel Machine核的生成也和距离度量有着密切的关系

距离度量的种类

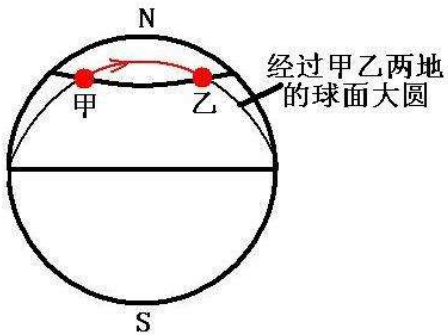
- Euclidean Distance

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

- Block Distance



- Geodesic Distance

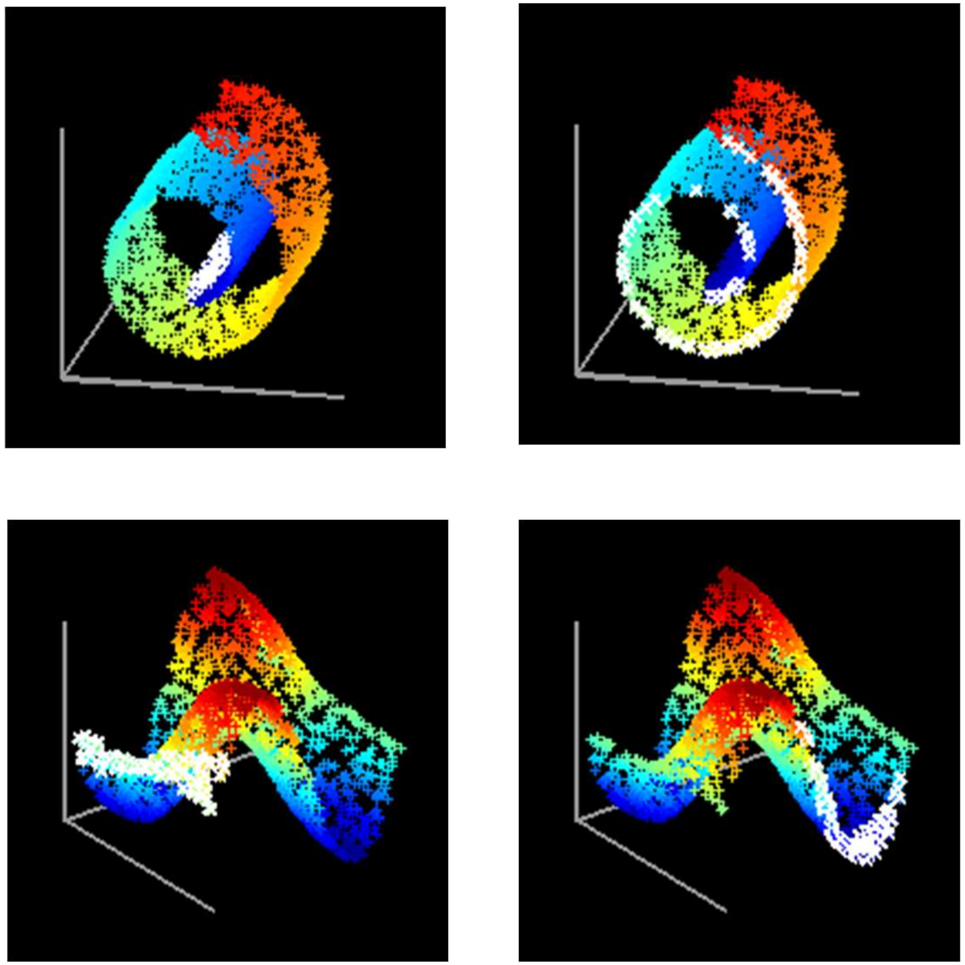


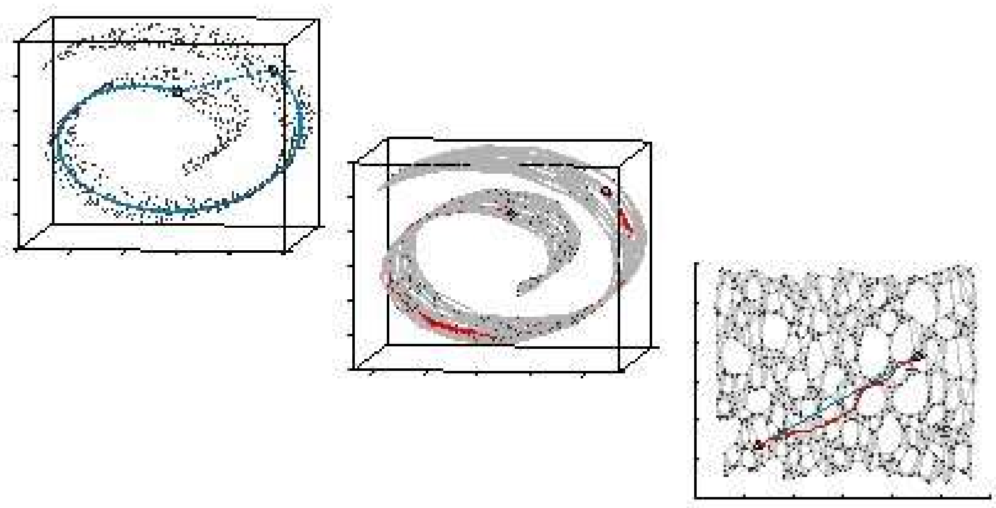
研究者对距离度量的研究

- 为特定应用设计距离度量
- 距离度量学习

Isomap

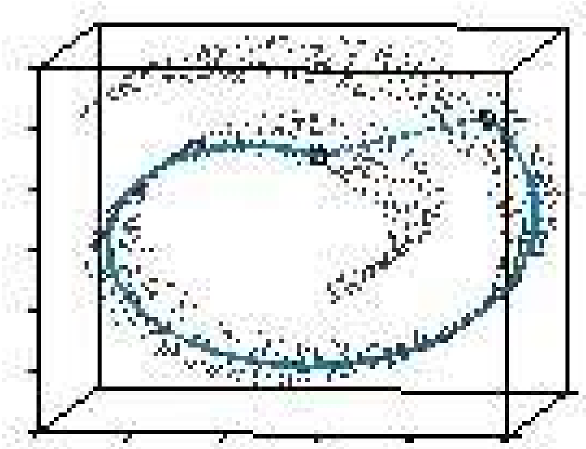
- 背景
  - 高维数据具有更加自然的低维结构
  - 使用低维结构上面的直线距离更加能够反映样本之间不相似的程度



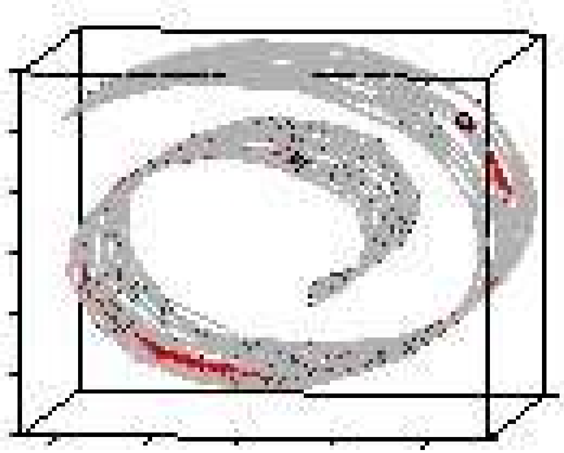


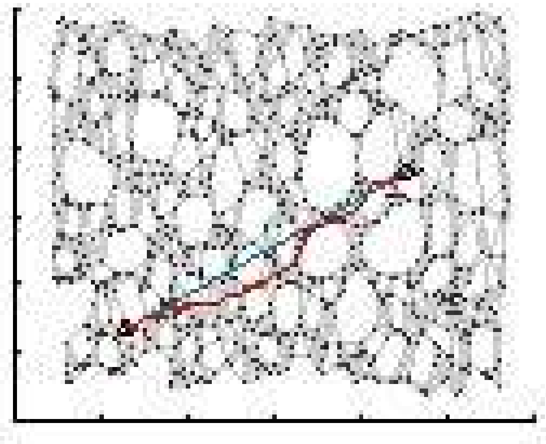
Isomap

要求的是两点之间的测地线距离



无法获得测地线！怎么办！





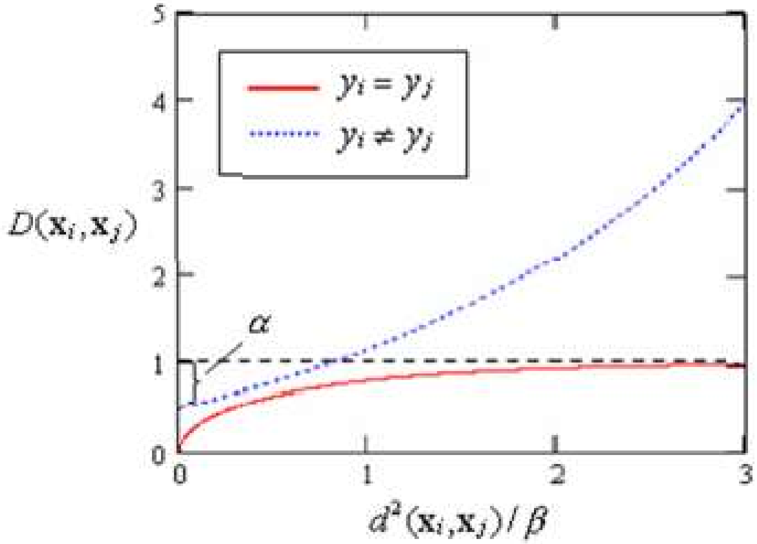
近似获得测地线距离：

Isomap的缺陷及对其的改进

- Isomap能够度量样本之间在低维本真空间上的距离
- 却容易受到噪音的影响，并且不利于分类

- 改进方法：
  - 引入类别信息，对距离度量进行改进 Supervised-Isomap [Geng, Zhan and Zhou 05].

$$D(\mathbf{x}_i,\mathbf{x}_j)=\begin{cases}\sqrt{1-e^{\frac{-d^2(\mathbf{x}_i,\mathbf{x}_j)}{\beta}}} & y_i=y_j\\ \sqrt{e^{\frac{d^2(\mathbf{x}_i,\mathbf{x}_j)}{\beta}}}-\alpha & y_i\neq y_j\end{cases},$$



研究者对距离度量的研究

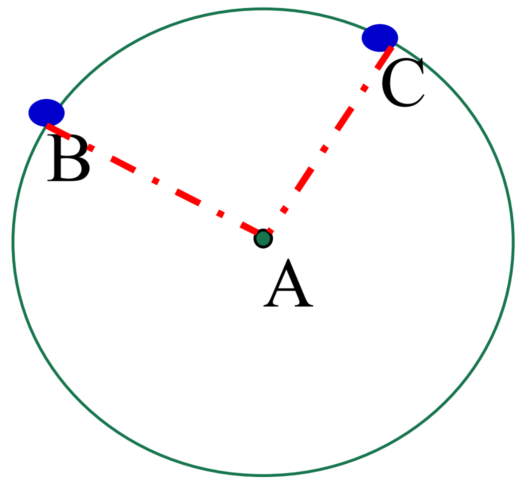
- 为特定应用设计距离度量

- 距离度量学习
  - 顾名思义，即利用学习的方法获得更好的度量距离的方式
  - 大多数研究者针对马氏距离的度量矩阵A进行学习

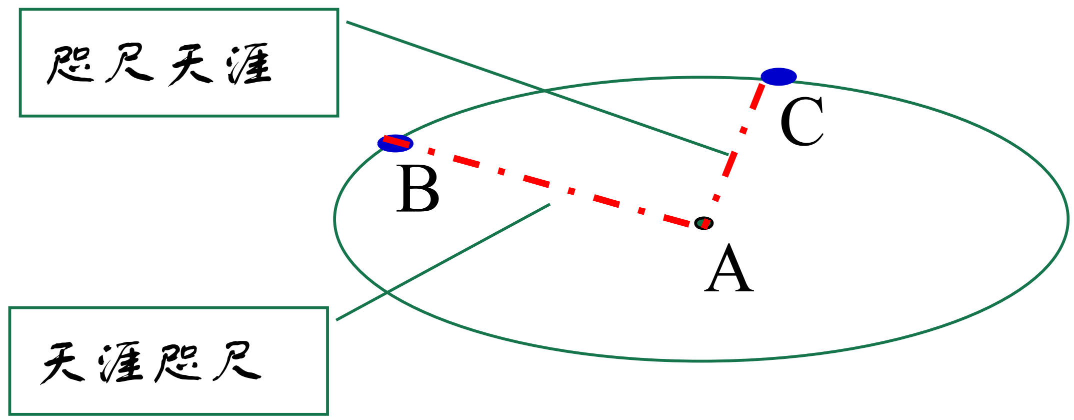
$$d_{ij} = \sqrt{\|\mathbf{x}_i - \mathbf{x}_j\|_{\boxed{A}}^2} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top \boxed{A} (\mathbf{x}_i - \mathbf{x}_j)}$$

对马氏距离的进一步解释

- 为什么要马氏距离 $d_{ij} = \sqrt{\|\mathbf{x}_i - \mathbf{x}_j\|_A^2} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top A(\mathbf{x}_i - \mathbf{x}_j)}$ ？
- 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



利用边信息的距离度量学习方法

基本思想:

- 一个好的距离度量方法，应该能够使得同类之间的距离小于1，异类之间的距离大于1

DML实现方法

$$\min \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{S}} \|\mathbf{x}_i - x_j\|_A^2$$

$$s.t. \quad A \in S_+, \quad \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{D}} \|\mathbf{x}_i - \mathbf{x}_j\|_A^2 \geq 1.$$

最小化同类之间的距离

保持异类样本之间的距离大于某个阈值

限制学习到的距离度量矩阵是半正定的；保证样本之间的距离为正(或者0)，也即：d_{ij}>=0

是否马氏距离就一定符合人们的认知？

- 在某些应用中，距离的定义偏重于一些特定的属性，并且对于不同的样本，这些属性是各不相同的
- 距离的定义应该是样本自适应的
- 例如：
  - 当我们将描述天空的图片和其他图片进行比较的时候
    - 关注的是图片的颜色（蓝色）、纹理（有着特殊的光线）等

  - 当我们将菲尔普斯II和其他游泳运动员比较的时候
    - 关注的是他脚的形状，游泳的速度

是否马氏距离就一定符合人们的认知？


图像1


图像2


图像3


查询1：森林


查询2：豹子

距离计算方法和特定的样本个体相关

- 当用户提交“查询1”时，图像2与图像1应该比与图像3更接近，因为前两者都包含了描述“森林”这个查询概念的特征；而在用户提交“查询2”时，图像2与图像3应该比与图像1更接近，因为图像2和图像3都包含了描述“豹子”这个概念的图像特征。

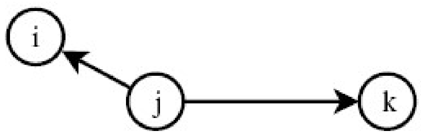
样本自适应方面已有的工作

- 问题：不同样本有着各自不同的视角、语义着重
- 或者说，将样本表示在高维空间中，样本的距离度量和其本身的局部特性相关
- 解决方案：对每个样本赋以不同的距离度量

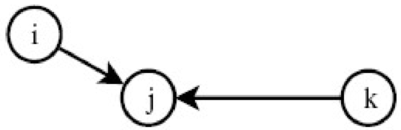
- QSim [Zhou and Dai, ICDM'06] [Athitsos et al., TDS'07]
- Local distance functions [Frome et al., NIPS'06, ICCV'07]

以往样本自适应方法的缺陷

- Qsim：
  - 用于基于内容的图像检索，对象之间的距离会受到提交的检索的影响
  - 主要问题：该方法完全基于启发式想法
    - 启发式想法 —— 拍脑袋想办法：
    - 人在解决问题时所采取的一种根据经验规则进行发现的方法
    - 利用过去的经验,选择已经行之有效的方法，而不是系统地、以确定的步骤去寻求答案
- Local Distance Functions：
  - [Frome et al. NIPS'06]



  - [Frome et al. ICCV'07]

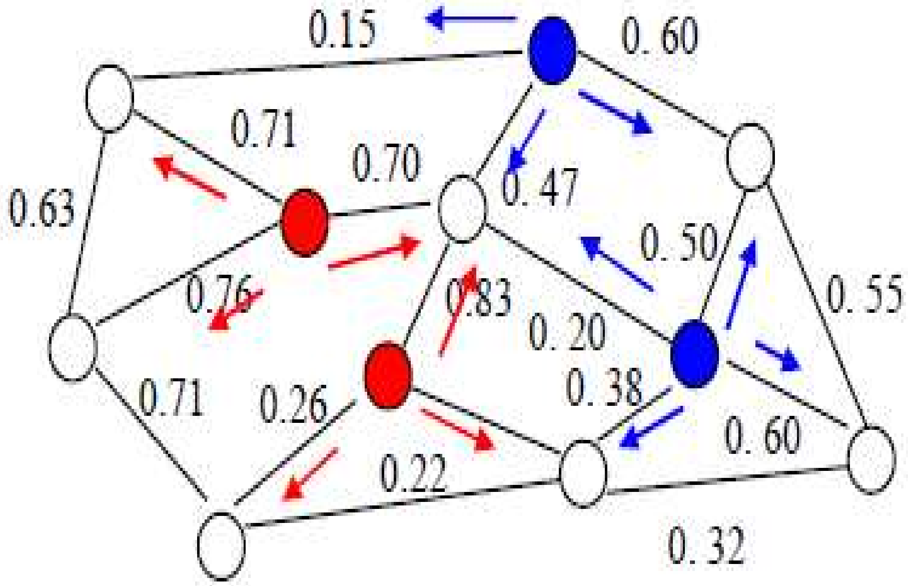


是否能够同时为标记样本和未标记样本学习得到样本自适应距离度量呢？

答案是肯定的，我们可以尝试使用类似于标记传播的思想来进行距离度量传播！

什么是标记传播？

- 一种基于图的半监督学习算法
  - 边的权重往往和点之间的欧氏距离相关

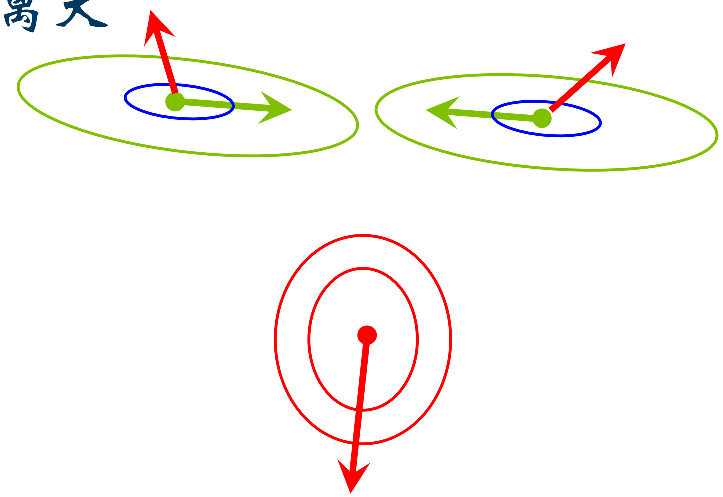


如何生成和传递样本自适应距离？

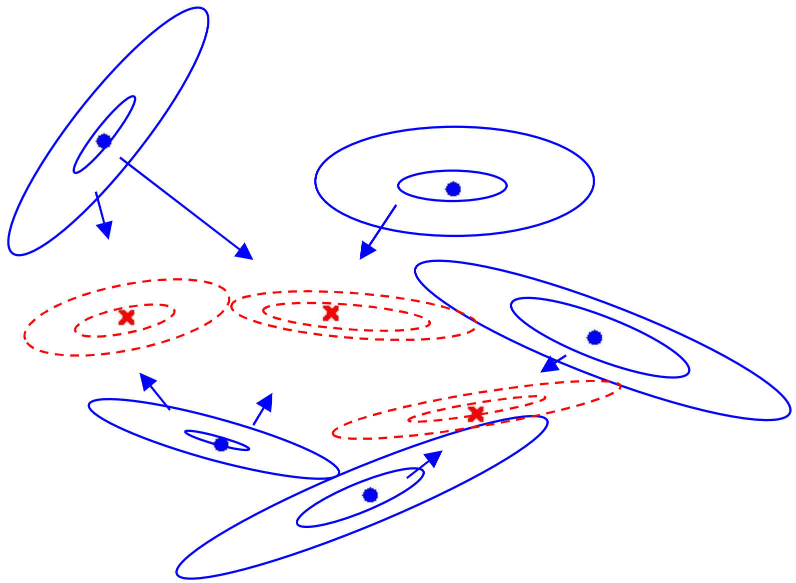
- 从标记样本中生成样本自适应距离度量，并且通过邻域关系（图）将这种样本自适应度量传播出去

- 标记样本上的样本自适应距离度量生成方式：

  生成的度量应该使得同类样本之间的距离小，异类样本之间的距离大



- 未标记样本的自适应距离度量应该遵从相近相似的原则（Metric Propagation）从近邻样本处获得



形式化模型

- 进行度量（标记）传播有两种选择：1. 采用随机游走的策略进行迭代求解；2. 将度量（标记）传播形式化成一个优化问题，这里我们选择的是后者

$$\min_{\mathbf{W}} \quad \lambda \sum_{i=1}^{n} \sum_{\boxed{j \in \mathcal{C}_i}} \boxed{\ell\left(\hat{y}_{ij}, D_i(\mathbf{x}_j)\right)} + \boxed{\Omega(\mathbf{W}, \mathbf{G})}$$

$$s.t. \quad \mathbf{w}_i \geq 0, i = 1, \cdots, n+u,$$

学习的度量在标记样本上犯下的错误

该正则化项负责隐式地进行度量传播

并不需要考虑所有的样本，仅仅只要考虑特定样本邻域内的关系即可

受[Zhu 2003]的启发，该正则化项可以定义为：

对于同类样本来 … 的

但是对于和特定 … hinge … 需要考虑

$$\Omega(\mathbf{W}, \mathbf{G}) = \sum_{i,j=1}^{n+u} E_{ij} ||\mathbf{w}_i - \mathbf{w}_j||^2 = 2tr(\mathbf{W}^\top \mathbf{L} \mathbf{W}).$$

形式化模型的特例和泛化

虽然在我们的工作中仅仅考虑到了样本对之间的信息，
但是整个ISD框架可以用于更普遍的情况

$$\min_{\mathbf{W}} \quad \lambda \sum_{i=1}^{n} \sum_{\boxed{j \in \mathcal{C}_i}} \boxed{\ell\left(\hat{y}_{ij}, D_i(\mathbf{x}_j)\right)} + \boxed{\Omega(\mathbf{W}, \mathbf{G})}$$

$$s.t. \quad \mathbf{w}_i \geq 0, i = 1, \cdots, n+u,$$

可以在此使用其他不同的监督信息，
例如triplets information

FSM [Frome et al. NIPS'06] is a special case of ISD

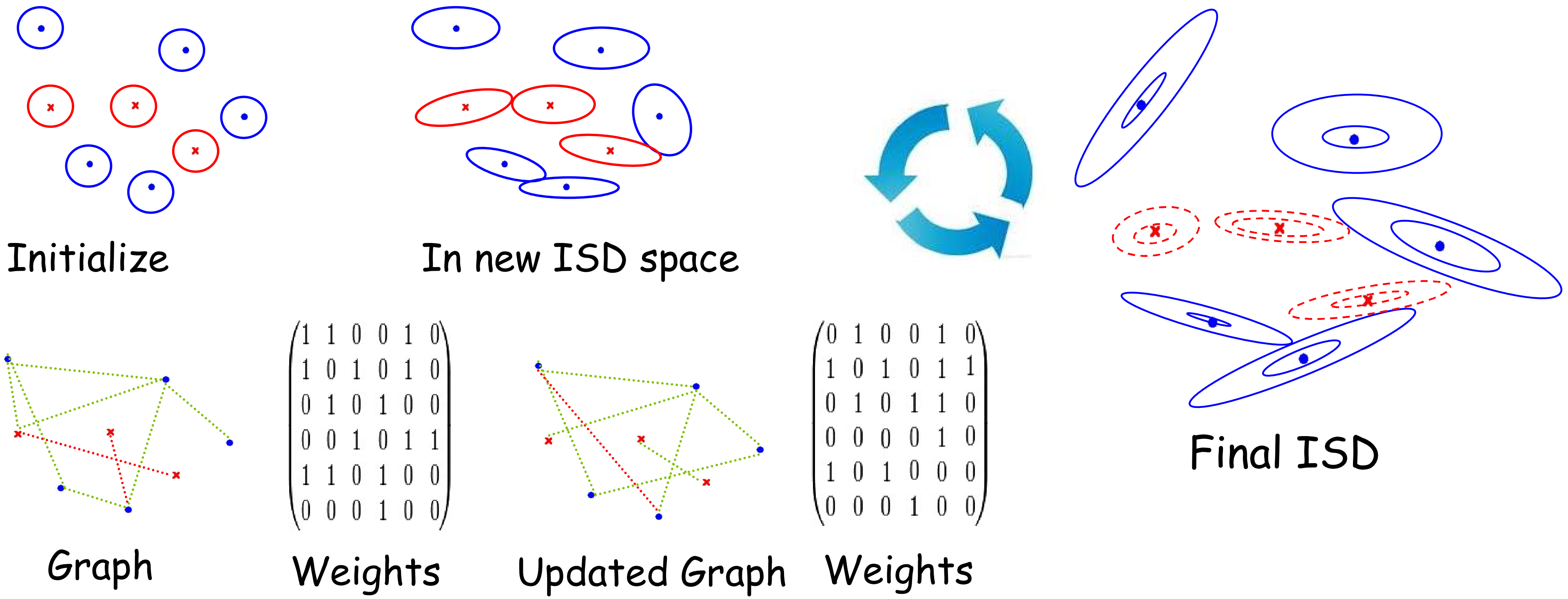
$$\Omega(\mathbf{W}, \mathbf{G}) = \sum_{i,j=1}^{n+u} E_{ij} ||\mathbf{w}_i - \mathbf{w}_j||^2 = 2tr(\mathbf{W}^\top \mathbf{L}\mathbf{W}).$$

L is set to identity matrix

图的构建和精化

$$
\begin{aligned}
\min_{\mathbf{W}} \quad & \lambda \sum_{i=1}^{n} \sum_{j \in \mathcal{C}_i} \ell\left(\hat{y}_{ij}, D_i(\mathbf{x}_j)\right) + \Omega(\mathbf{W}, \boxed{\mathbf{G}}) \\
s.t. \quad & \mathbf{w}_i \geq 0, i = 1, \cdots, n+u,
\end{aligned}
$$

我们假设G是预先给定的，但是如果没有给定……



损失函数的选择 ISD-L1

$$\ell\left(\hat{y}_{ij}, D_i(\mathbf{x}_j)\right) = \max\left(0, \boxed{\hat{y}_{ij}\left(D_i(\mathbf{x}_j) - \eta\right)}\right)$$

$$\boxed{\begin{aligned} D_i(\mathbf{x}) &= \mathbf{w}_i^\top \delta_{\mathbf{x}_i,\mathbf{x}} \\ \delta_{\mathbf{x}_i,\mathbf{x}_j} &= (\mathbf{x}_i - \mathbf{x}_j) \odot (\mathbf{x}_i - \mathbf{x}_j) \end{aligned}}$$

通过引入裕量

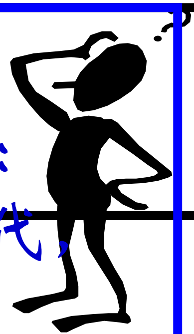
$$\min_{\mathbf{W},\ \xi_{i,j}} \quad \lambda \sum\nolimits_{i=1}^{n} \sum\nolimits_{j \in \mathcal{C}_i} \xi_{i,j} + 2tr(\mathbf{W}^\top \mathbf{L}\mathbf{W})$$

$$\text{s.t} \quad \hat{y}_{ij}(\mathbf{w}_i \delta_{\mathbf{x}_i,\mathbf{x}_j} - 1) \leq \xi_{i,j}, i = 1, \cdots, n$$

$$\boldsymbol{\xi} \geq \mathbf{0},\ \ \mathbf{w}_i \geq 0, i = 1, \cdots, n + u$$

考虑到W的数量很多，同时求解有困难；
计算量过大！

凸问题→意味着全局最小=局部最小
我们可以通过alternating descent的方法进行优化，也即：
每次固定其他的ws，优化其中的一个w，然后再反复迭代，
直到收敛



求解空间的选择

Primal:

$$\min_{\mathbf{w}_i,\ \xi_j} \quad \sum_{i,j=1}^{n+u} E_{ij}(\mathbf{w}_i^\top \mathbf{w}_i - 2\mathbf{w}_i^\top \mathbf{w}_j) + \lambda \sum_{j\in\mathcal{C}_i} \xi_j$$

$$s.t. \quad \hat{y}_{ij}(\mathbf{w}_i^\top \delta_{\mathbf{x}_i,\mathbf{x}_j} - 1) \le \xi_j, \quad j \in \mathcal{C}_i$$

$$\boldsymbol{\xi} \ge 0, \ \mathbf{w}_i \ge 0$$

Dual:

$$\min_{\boldsymbol{\alpha}} \quad (\hat{\mathbf{D}}_i\boldsymbol{\alpha} - \boldsymbol{\gamma})^\top(\hat{\mathbf{D}}_i\boldsymbol{\alpha} - \boldsymbol{\gamma}) - 4(\hat{\mathbf{D}}_i\boldsymbol{\alpha} - \boldsymbol{\gamma})^\top \boldsymbol{C}_i + 4\theta_i\boldsymbol{\alpha}^\top \boldsymbol{y}_{i.}$$

$$s.t. \quad 0 \le \boldsymbol{\alpha} \le \lambda, \boldsymbol{\gamma} \ge 0$$

$$\hat{\mathbf{D}}_i = \mathbf{D}_{i.}\mathbf{Y}_i \qquad \mathbf{D}_{i.} = \left[\delta_{\mathbf{x}_i,\mathbf{x}_1}, \delta_{\mathbf{x}_i,\mathbf{x}_2}, \cdots, \delta_{\mathbf{x}_i,\mathbf{x}_p}\right]$$

$$\boldsymbol{C}_i = \sum_j E_{ij}\mathbf{w}_j \qquad \theta_i = \sum_j E_{ij}$$

加速方法

在加速算法方面，我们已经做出的努力是：

- 使用alternating descend方法进行优化
- 对同类之间产生的约束的数量进行了消减

但是异类之间产生的约束的数量仍然可能十分巨大，从而导致算法异常耗时

从nu-SVM中获得灵感，我们是否可以利用类似的方法得到一种更有效率的方法呢？ 答案是肯定的

$$\ell\left(\hat{y}_{i,j}, D_i(\mathbf{x}_j)\right) = \max\left(0, \hat{y}_{i,j}(D_i(\mathbf{x}_j) - \eta)\right)^2$$

加速方法 ISD-L2

$$\min_{\mathbf{W},\xi_{i,j},\rho} \quad \lambda \sum_{i=1}^{n} \sum_{j\in\mathcal{C}_i} \xi_{i,j}^2 + 2tr(\mathbf{W}^\top \mathbf{L}\mathbf{W}) - \rho$$

$$\text{s.t} \quad \hat{y}_{ij}(\mathbf{w}_i \delta_{\mathbf{x}_i,\mathbf{x}_j} - 1) \le \xi_{i,j} - \rho$$

Dual:

$$\min_{\boldsymbol{\alpha}} \quad \boldsymbol{\alpha}^\top (\hat{\mathbf{D}}_i^\top \hat{\mathbf{D}}_i + \frac{\theta_i}{\lambda}\mathbf{I})\boldsymbol{\alpha} + 4(\theta_i \boldsymbol{y}_i^\top - \boldsymbol{C}_i^\top \hat{\mathbf{D}}_i)\boldsymbol{\alpha}$$

$$s.t. \quad \boxed{\boldsymbol{\alpha}^\top \mathbf{1} = 1,} \ \boldsymbol{\alpha} \ge 0$$

注意：这里存在一个等式约束

为了简化问题，我们可以首先把这个约束条件给去掉，然后先求解出w之后再用下式让w满足最终

因为等式约束的存在，这个Dual问题可以快速地使用SMO (Sequential Minimal Optimization) 求解

$$\hat{\mathbf{w}}_i = (\mathbf{C}_i - \hat{\mathbf{D}}_i \boldsymbol{\alpha}/2)/\theta_i$$

$$\mathbf{w}_i = \max(0, \hat{\mathbf{w}}_i)$$

如果标记样本特别少，怎么办？

$$\min \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{S}} \|\mathbf{x}_i - x_j\|_A^2$$

最小化同类之间的距离

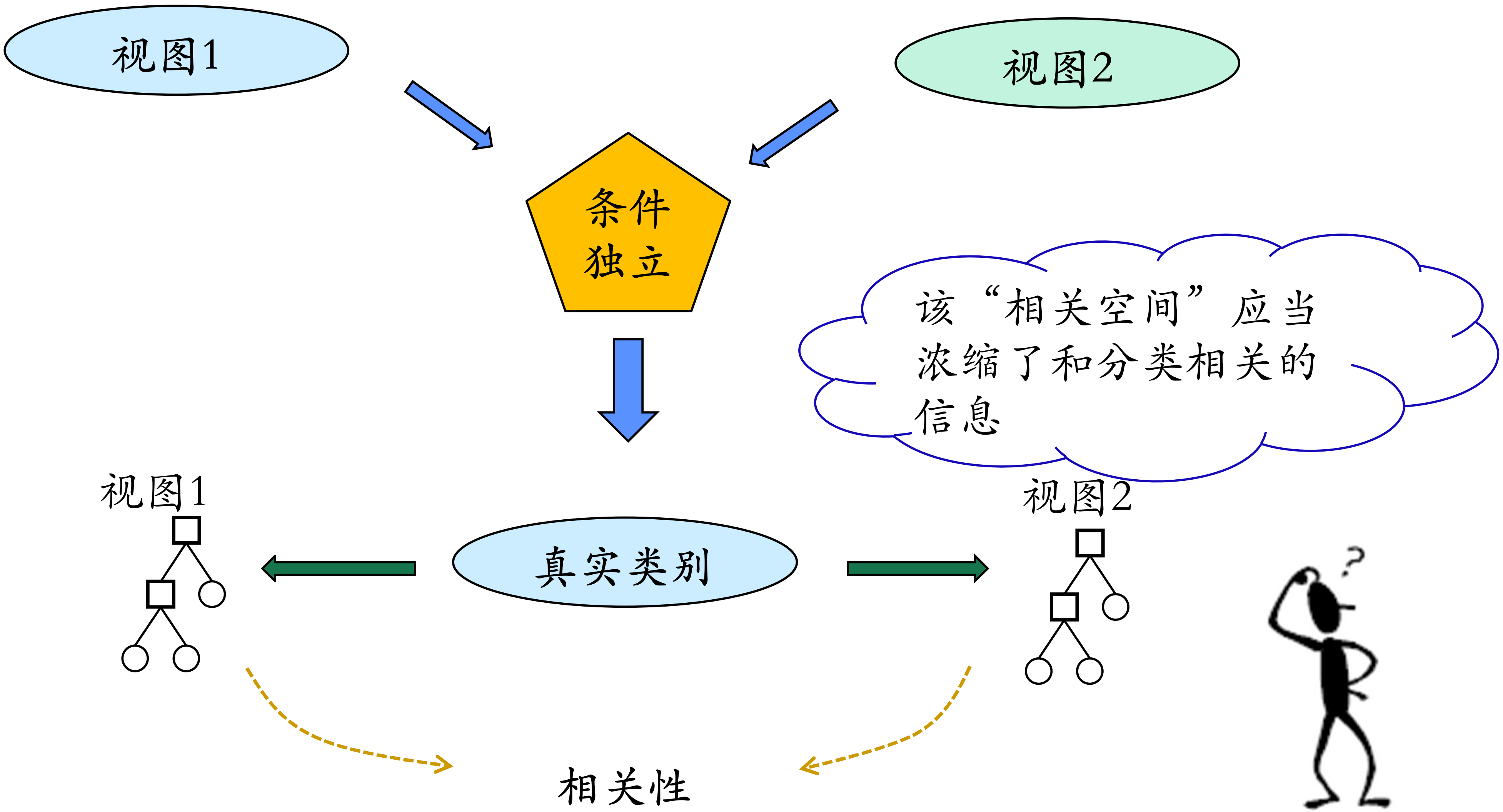
$$s.t.\quad A \in S_+, \quad \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{D}} \|\mathbf{x}_i - \mathbf{x}_j\|_A^2 \geq 1.$$

保持异类样本之间的距离大于某个阈值

当标记样本极少的时候，无法得到这些约束

但是，如果你在web上面搜索的时候会提交多个查询样本么？（将查询样本看成标记正例）

基于双视图关联距离度量的半监督学习



基于双视图关联距离度量的半监督学习

1. KCCA寻找“相关性”大的空间
2. 在此空间中定义距离度量
3. 使用该距离度量扩充标记样本集
4. 进行半监督学习Co-training

THANKS