第2章 归类理论

伯牙鼓琴，钟子期听之，方鼓琴而志在泰山，钟子期曰：‘善哉乎鼓琴！巍巍乎若泰山’。少时而志在流水。钟子期曰：‘善哉乎鼓琴，汤汤乎若流水’。钟子期死，伯牙摔琴绝弦，终身不复鼓琴，以为世无足复为鼓琴者。

——《吕氏春秋·本味》

北京交通大学《机器学习》课程组

为何要公理化?

1、阿尔伯特 爱因斯坦说过： 所有科学中最重要的目标就是，从最少数量的假设和公理出发，用逻辑演绎推理的方法解释最大量的经验事实。

2、从1956年达特茅斯会议以来，AI已经取得了巨大的进展，并改善了人们的现实生活。

3、2017年10月26日丘成桐在中国计算机大会上作大会报告，明确指出："我是很鼓励我们做人工智能的也能重复这个做法--从现在复杂多样的网络中找到它简单的公理"

Scientists、Mathematicians

2015年，Pedro Domingos 出版了The master algorithm （终极算法）将机器学习分为5大流派：

- 符号学派：决策树
- 连接学派： 深度学习
- 进化学派：进化算法
- 贝叶斯学派：概率图，朴素贝叶斯
- 类推学派：K近邻，K-means等

机器学习的假设

早在2004年，周志华就在基金委做的报告中明确指出，机器学习“以Tom Mitchell的经典教科书为例，很难看到基础学科（例如数学、物理学）教科书中那种贯穿始终的体系，也许会让人感到这不过是不同方法和技术的堆砌”。

- 问题： 机器学习是否存在共同的假设，更加明确的说法是，机器学习算法有共同的假设吗？
- 一个学科的特定假设，就是其公理。现在的问题是：机器学习是否可以公理化？

Data Scientists vs Data Analyst

机器学习能否公理化?

- 已有学习理论
  - PAC学习理论，统计学习理论，概率图理论
- 人类的学习机理
  - 六七岁的儿童学习能力已经很强了，但并不懂机器学习理论和算法
- 机器学习是否可以公理化
  - 能够统管所有机器、生物和人的学习理论［M.I. Jordan & T.M. Mitchell, Science 2015］

Psychologist、Data Scientists

提要

1、引言

2、类表示与类表示公理

3、归类公理

4、归类结果分类

5、归类方法设计准则

机器学习的共同目标

- 无论机器学习还是人的学习，其共同目标是学习知识。
- 知识的基本单位是概念，知识自身也是一个概念。
  因此，学习知识就是学习概念
- 学习概念就是是会使用概念，概念的使用称为归类。
  因此，学习与归类问题等价

什么是概念?

1、经典理论（亚里士多德）

- 概念有一个命题表示

1、经典理论（亚里士多德）

概念的两种表示：

- 内涵表示（intension）
  - ✓ 反映和揭示概念的本质属性，是人类主观世界对概念的认知，可存在于人的心智之中，用命题来表示

e.g. 素数：只能被1和其自身整除的自然数

- 外延表示
  - ✓ 包含了与概念对应的各种具体实例，是一个由具有概念本质属性的对象构成的集合
  - ✓ 外部可测的，可度量的

e.g. 素数集合{2,3,5,7,11,13,17,19,23,29,31，……}

什么是概念?

1、经典理论（亚里士多德）

- 概念有一个命题表示

2、原型理论（Rosch, 1978）

- 概念有一个原型表示

3、样例理论 (Medin & Schaffer, 1978)

- 概念有多个样例表示

4、知识理论 (Murphy & Medin, 1985)

- 概念是一个知识框架的组成部分

本书假设概念在人心智中是存在的。

概念在人心智中的表示称为认知表示。

归类

- 概念的使用问题称为归类
- 人接触世界，会将各种对象自动表示为心智中的各种概念，即人具有归类能力。
- 归类是人类一项最重要&最基本的认知能力，归类正确与否明确显示了人是否掌握了该类对应的概念
- 自然希望计算机也具有这种能力。为此，机器学习作为一门学科应运而生。
- 语义上，概念<=>类
- 综上所述，类的表示：内蕴表示+外部表示

典型的归类问题——例1

认知假设

1、为什么黑板前的小孩将黑板下的小孩的手势认作为OK呢?

- 一个最简单的解释是相似性 （像OK， 归为OK）

2、黑板下的小孩又为什么将自己的手势认作为3呢?

- 同样地，最简单的解释也是相似性。 （归为3，像3）

3、观察到的事实：输入：对象像归于的类，输出：对象归为最像的类。（人接触世界，会将各种对象自动根据相似性分门别类，即人具有归类能力。）

归类的认知假设

- 最直观的假设
  - 每个对象应该最相似于其归于的类。
  - 每个对象应该归为与其最相似的类。
- 更简单的归纳为：
  - 归哪类，像哪类。像哪类，归哪类。

"归哪类，像哪类；像哪类，归哪类。"之语义分析

- 复句分析：
  - "归哪类，像哪类"表示的是**信息输入者**的归类准则。
  - "像哪类，归哪类"表示的是**信息接受者**的归类准则。
- 单句分析：
  - 归哪类：对外可见，是类的外显表示；
  - 像哪类：主观内在，不一定外显，属于类的内部表示。

“归哪类，像哪类；像哪类，归哪类。” 之语义分析

- 词分析：
  - 归：对对象归类的外显指称，是人使用类外延表示的方式。
  - 像：对对象归类的内在指称，是人使用类认知表示的方式。

提要

1、引言

2、类表示与类表示公理

3、归类公理

4、归类结果分类

5、归类方法设计准则

机器学习的简化定义

只给予某些概念的一个有限外延子集，希望得到这些概念的内涵表示。

类表示的重要符号

- 类表示八元组

$$(\boldsymbol{X}, \boldsymbol{U}, \underline{\boldsymbol{X}}, \boldsymbol{Sim}_{\boldsymbol{X}})$$

$$(\boldsymbol{Y}, \boldsymbol{V}, \underline{\boldsymbol{Y}}, \boldsymbol{Sim}_{\boldsymbol{Y}})$$

归类算法的外部表示

1、归类输入的外部表示

- 由一个有限抽样对象集合O={$o\_1,o\_2,\cdots,o\_N$}的归类输入外部信息组成，包括：
  - ✓ 对象的特性输入表示：X
  - ✓ 对应的类外延表示：U

2、归类输出的外部表示

- ✓ 对象特性输出表示：Y
- ✓ 对应的类外延表示：V

归类输入的外部表示—对象特性输入表示 *X*

- 特征矩阵： $[x_{k\tau}]_{n\times p}$
- 相异性矩阵：$[d_{kl}]_{n\times n}$
- 相似性矩阵：$[s_{kl}]_{n\times n}$
- 图像、声音、文字、手势等

归类输入的外部表示—类外延表示 *U*

每个对象的归类情况由划分矩阵表示

划分矩阵: $U=\left[u_{ik}\right]_{c\times n}, u_{ik}\geq 0$

- 硬划分: $\sum_{i=1}^{c} u_{ik}=1, u_{ik}\in\{0,1\}, \sum_{k=1}^{n} u_{ik}>1,$

- 软划分: $\sum_{i=1}^{c} u_{ik}=1, u_{ik}\geq 0, \sum_{k=1}^{n} u_{ik}>0,$

- 可能性划分: $\sum_{i=1}^{c} u_{ik}>0, u_{ik}\geq 0, \sum_{k=1}^{n} u_{ik}>0$

(模糊划分/概率划分)

归类输出的外部表示—对象特性输出表示 *Y*

- 特征矩阵： $Y = [y_{k\tau}]_{n\times d}$
- 相异性矩阵：$[d_{kl}]_{n\times n}$
- 相似性矩阵：$[s_{kl}]_{n\times n}$
- 图像、语音、文本、手语等

归类输出的外部表示—类外延表示 *V*

每个对象的归类情况由划分矩阵表示

划分矩阵: $V=\left[v_{ik}\right]_{c\times n}, v_{ik}\geq 0$

- 硬划分: $\sum_{i=1}^{c} v_{ik}=1, v_{ik}\in\{0,1\}, \sum_{k=1}^{n} v_{ik}>1,$
- 软划分: $\sum_{i=1}^{c} v_{ik}=1, v_{ik}\geq 0, \sum_{k=1}^{n} v_{ik}>0,$
- 可能性划分: $\sum_{i=1}^{c} v_{ik}>0, v_{ik}\geq 0, \sum_{k=1}^{n} v_{ik}>0$

(模糊划分/概率划分)

归类算法的外部表示

- 归类输入的外部表示:

$$(X, U),\text{其中}\ X = \{x_1, x_2, \dots, x_n\}$$

- 归类输出的外部表示:

$$(Y, V),\text{其中}\ Y = \{y_1, y_2, \dots, y_n\}$$

指派算子（归）（assignment operator)

- 归类输入

$$\vec{X} = \{\overrightarrow{x_1}, \overrightarrow{x_2}, \dots, \overrightarrow{x_n}\},$$

其中，$\overrightarrow{x_k} = arg \max_i u_{ik}$

- 归类输出

$$\vec{Y} = \{\overrightarrow{y_1}, \overrightarrow{y_2}, \dots, \overrightarrow{y_n}\},$$

其中，$\overrightarrow{y_k} = arg \max_i v_{ik}$

内部表示：类的认知表示

经典理论（亚里士多德）：

类有一个命题表示

原型理论（Rosch, 1978)

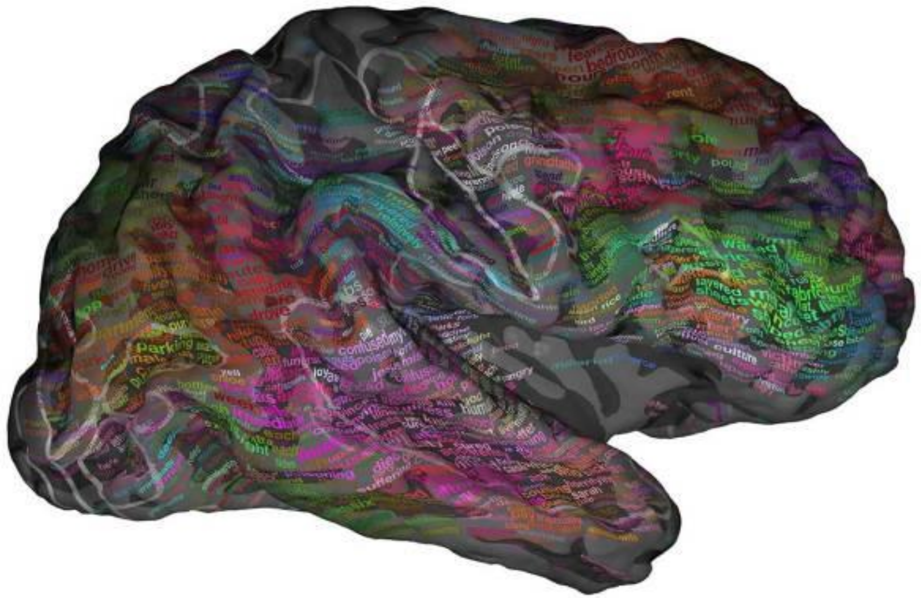
类有一个原型表示

样例理论 (Medin & Schaffer, 1978)

类有多个样本表示

知识理论 (Murphy & Medin, 1985)

类是一个知识框架的组成部分



Huth A G, et al, Natural speech reveals the semantic maps that tile human cerebral cortex. Nature, 532(7600):453-458, 2016

归类算法的内部表示

- 归类输入的内部表示

$$\left(\underline{X}, Sim_X\right)$$

- 归类输出的内部表示

$$\left(\underline{Y}, Sim_Y\right)$$

类的认知表示（主观内在）$\underline{X}$ 和 $\underline{Y}$

- 输入认知表示

$$\underline{X} = \{\underline{X}_1, \underline{X}_2, \cdots, \underline{X}_C\}$$

- 输出认知表示

$$\underline{Y} = \{\underline{Y}_1, \underline{Y}_2, \cdots, \underline{Y}_C\}$$

输入类相似性映射 Sim_X

- 类的认知表示也具有归类能力。
  - 可以根据与类的相似度将对象归类
- 输入类相似性映射

函数Sim_X：$X \times \{\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_c\} \mapsto R_+$ 是输入类相似性映射，满足条件：

$Sim_X(x_k, \underline{X}_i)$值增加表示 x_k 和 $\underline{X}_i$ 的相似性增大，函数$Sim_X(x_k, \underline{X}_i)$值减小表示 x_k 和 $\underline{X}_i$ 的相似性减小。

输出类相似性映射 Sim_Y

- 类的认知表示也具有归类能力.
  - 可以根据与类的相似度将对象归类
- 输出类相似性映射

函数 Sim_Y： $Y \times \{\underline{Y_1}, \underline{Y_2}, \dots, \underline{Y_c}\} \mapsto R_+$ 是 输出类相似性映射，满足条件：

$Sim_Y(y_k, \underline{Y_i})$值增加表示 y_k 和 $\underline{Y_i}$ 的相似性增大，函数$Sim_Y(y_k, \underline{Y_i})$值减小表示 y_k 和 $\underline{Y_i}$ 的相似性减小。

相似算子（像）(similarity operator)

- 输入相似指称： $\tilde{X} = \{\widetilde{x_1}, \widetilde{x_2}, \ldots, \widetilde{x_n}\}$,

  其中，$\widetilde{x_k} = argmax_i Sim_X\left(x_k, \underline{X_i}\right)$

- 输出相似指称： $\tilde{Y} = \{\widetilde{y_1}, \widetilde{y_2}, \ldots, \widetilde{y_n}\}$,

  其中，$\widetilde{y_k} = argmax_i Sim_Y\left(y_k, \underline{Y_i}\right)$

归类输入

- 一个归类算法，如果其外部输入表示是（X,U），对应的输入类内部表示为 $(\underline{X}, Sim_X)$，则归类输入四元组为 $\{X, U, \underline{X}, Sim_X\}$。

归类结果

- 一个归类算法，如果其外部输出表示是（Y,V），对应的输出类内部表示为$\left(\underline{Y}, Sim_Y\right)$，则归类结果四元组为$\left\{Y, V, \underline{Y}, Sim_Y\right\}$。

类表示存在公理(Existence Axiom of Category Representation)

- 如果一个归类算法的归类外显输入输出给定，则其对应的类内部表示存在。

- 一个归类算法，如果其归类外显输入是(X,U)，归类外显输出是(Y,V)，则存在其对应的归类内在输入为$(\underline{X},Sim_X)$和对应的归类内在输出$(\underline{Y},Sim_Y)$。

类内在表示

- 输入的类内在表示

$$\left(\underline{X}, Sim_X\right)$$

- 输出的类内在表示

$$\left(\underline{Y}, Sim_Y\right)$$

白箱算法与黑箱算法

- 如果 $\underline{Y}$ 被归类算法显式输出，则算法称为白箱算法。在白箱算法中，$\underline{Y}$ 对于使用者和设计者都是可见的。
- 如果 $\underline{Y}$ 不被归类算法显式输出，则算法称为黑箱算法。在黑箱算法中，$\underline{Y}$ 对于使用者是不可见的，但对于设计者可见。

类表示唯一公理(Uniqueness Axiom of Category Representation)

- 对一个归类算法，其输入输出对应的类表示（语义）应该相同。

- 如果归类算法的归类输入$(X, U, \underline{X}, Sim_X)$，其对应归类结果为$(Y, V, \underline{Y}, Sim_Y)$，则有$\vec{X} = \vec{Y}, \underline{X} = \underline{Y}, \tilde{X} = \tilde{Y}$.

归类的最高理想：高山流水遇知音

伯牙鼓琴，钟子期听之，方鼓琴而志在泰山，钟子期曰：‘善哉乎鼓琴！巍巍乎若泰山’。少时而志在流水。钟子期曰：‘善哉鼓琴，汤汤乎若流水’。钟子期死，伯牙摔琴绝弦，终身不复鼓琴，以为世无足复为鼓琴者。

类表示唯一公理的讨论

- 类表示唯一公理是学习算法能够成功学习的先验假设。
- 类表示唯一公理要求太高， 即使人也很难达到。

对一个归类算法，其输入输出对应的类表示（语义）应该相同

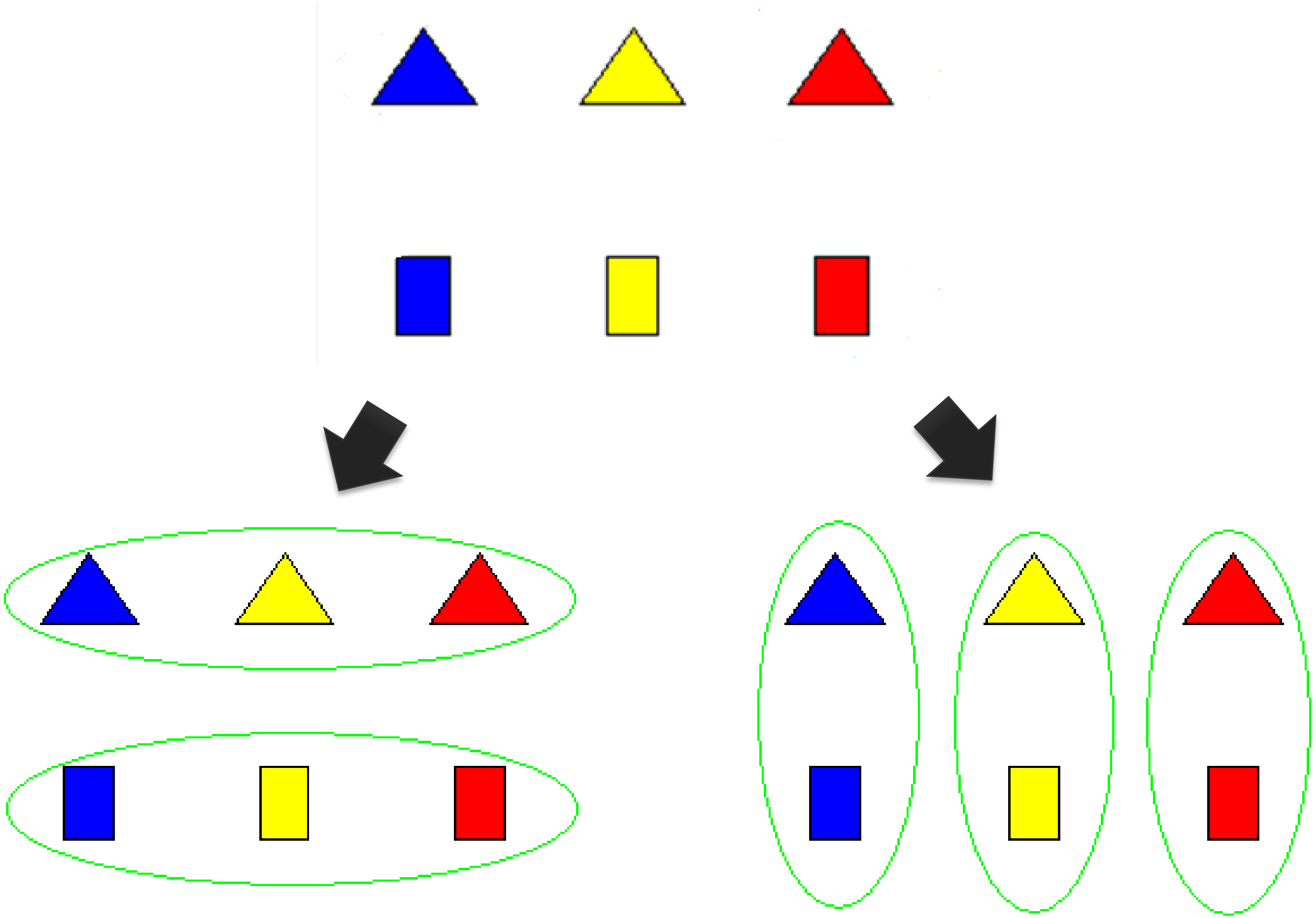
如果归类算法的归类输入 $(X, U, \underline{X}, Sim_X)$,

其对应归类结果为 $(Y, V, \underline{Y}, Sim_Y)$，则有

$$\vec{X} = \vec{Y}, \underline{X} = \underline{Y}, \tilde{X} = \tilde{Y}$$

知音难得

类相似性映射是任务依赖的、主观的

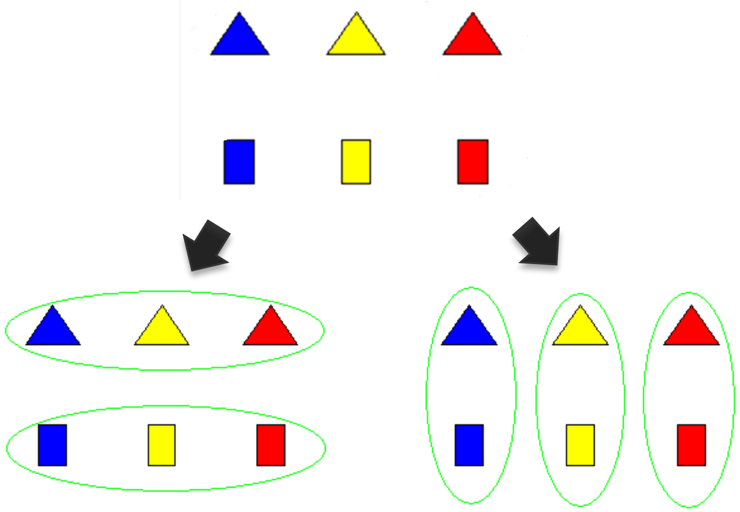


相似性悖论：

- 输入类相似性映射与输出类相似性映射在归类意义下不等价。即：

$$Sim_X \ <\neq>\ Sim_Y$$

- 在图像检索中也经常发生此类情形



美女与野兽

回顾 – 类表示

- 归类表示
  - 归类输入
    - 外部表示 (X,U)
      - X 对象特性输入表示 $X=\{x_1,x_2,\dots,x_N\}$归为C个子集$\{X_1,X_2,\dots,X_C\}$

        x_k：对象O_k的的特性输入； X_i：X中属于第i个输入类的对象子集
      - U 对应的类外延表示 $U=[uik]_{C\times N}$为划分矩阵或隶属度矩阵
        - 硬划分
        - 软划分
        - 可能性划分

        u_{ik}：对象O_k属于第i个输入类的隶属度

        指派（归）算子（外部指称）： $\vec{X}=\{\overrightarrow{x_1},\overrightarrow{x_2},\dots,\overrightarrow{x_n}\}$，其中$\overrightarrow{x_k}=arg\max_i u_{ik}$
    - 内部表示 $(\underline{X},Sim_X/Ds_X)$
      - $\underline{X}$ 输入认知表示 $\underline{X}=\{\underline{X_1},\underline{X_2},\dots,\underline{X_C}\}$ 其中$\underline{X_i}$为第i类的输入认知表示
      - Sim_X 输入相似性映射 $X\times\{\underline{X_1},\underline{X_2},\dots,\underline{X_C}\}\mapsto R_+\xrightarrow{\text{像哪类}}$相似（像）算子（内蕴指称）：

        $\tilde{X}=\{\widetilde{x_1},\widetilde{x_2},\dots,\widetilde{x_n}\}$，其中$\widetilde{x_k}=argmax_i Sim_X\left(x_k,\underline{X_i}\right)$
      - Ds_X 输入相异性映射 $X\times\{\underline{X_1},\underline{X_2},\dots,\underline{X_C}\}\mapsto R_+\xrightarrow{\text{像哪类}}$相似算子（内蕴指称）：

        $\tilde{X}=\{\widetilde{x_1},\widetilde{x_2},\dots,\widetilde{x_n}\}$，其中$\widetilde{x_k}=argmin_i Ds_X\left(x_k,\underline{X_i}\right)$
  - 归类输出
    - 内部表示 $(\underline{Y},Sim_Y/Ds_Y)$
      - $\underline{Y}$ 输出认知表示 $\underline{Y}=\{\underline{Y_1},\underline{Y_2},\dots,\underline{Y_C}\}$ 其中$\underline{Y_i}$为第i类的输入认知表示
      - Sim_Y 输出相似性映射 $Y\times\{\underline{Y_1},\underline{Y_2},\dots,\underline{Y_C}\}\mapsto R_+\xrightarrow{\text{像哪类}}$相似（像）算子（内蕴指称）：

        $\tilde{Y}=\{\widetilde{y_1},\widetilde{y_2},\dots,\widetilde{y_n}\}$，其中$\widetilde{y_k}=argmax_i Sim_Y\left(y_k,\underline{Y_i}\right)$
      - Ds_Y 输出相异性映射 $Y\times\{\underline{Y_1},\underline{Y_2},\dots,\underline{Y_C}\}\mapsto R_+\xrightarrow{\text{像哪类}}$相似算子（内蕴指称）：

        $\tilde{Y}=\{\widetilde{y_1},\widetilde{y_2},\dots,\widetilde{y_n}\}$，其中$\widetilde{y_k}=argmin_i Ds_Y\left(y_k,\underline{Y_i}\right)$
    - 外部表示 (Y,V)
      - Y 对象特性输出表示 $Y=\{y_1,y_2,\dots,y_N\}$归为C个子集$\{Y_1,Y_2,\dots,Y_C\}$

        y_k：对象O_k的的特性输出； Y_i：Y中属于第i个输出类的对象子集
      - V 对应的类外延表示 $V=[v_{ik}]_{C\times N}$为划分矩阵或隶属度矩阵
        - 硬划分——硬归类
        - 软划分, 可能性划分 } 软归类

        v_{ik}：对象O_k属于第i个输出类的隶属度

        指派（归）算子（外部指称）： $\vec{Y}=\{\overrightarrow{y_1},\overrightarrow{y_2},\dots,\overrightarrow{y_n}\}$，其中$\overrightarrow{y_k}=arg\max_i v_{ik}$

回顾——类表示存在公理

- 如果一个归类算法的归类外显输入输出给定，则其对应的类内部表示存在。
- 一个归类算法，如果其归类外显输入是(X,U)，归类外显输出是(Y,V)，则存在其对应的归类内在输入为$(\underline{X},Sim_X)$和对应的归类内在输出$(\underline{Y},Sim_Y)$。

回顾——类表示唯一公理

- 对一个归类算法，其输入输出对应的类表示（语义）应该相同。

- 如果归类算法的归类输入$(X, U, \underline{X}, Sim_X)$，其对应归类结果为$(Y, V, \underline{Y}, Sim_Y)$，则有$\vec{X} = \vec{Y}, \underline{X} = \underline{Y}, \tilde{X} = \tilde{Y}$.

提要

1、引言

2、类表示与类表示公理

3、归类公理

4、归类结果分类

5、归类方法设计准则

三大归类公理

1、样本可分性公理（Sample Separation Axiom, SS）

- 一个对象总有唯一一个类与其最相似。

2、类可分性公理（Categorization Separation Axiom, CS）

- 一个类至少有一个对象与其最相似。

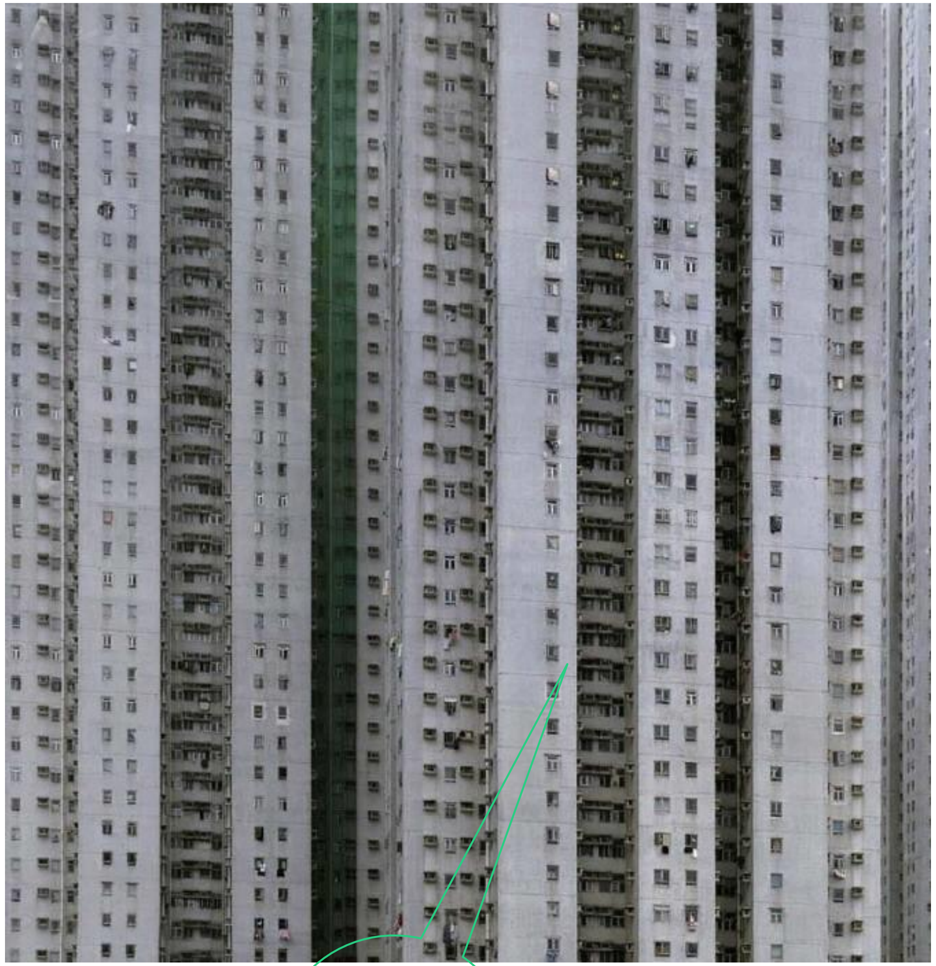
3、归类等价公理（Categorization Equivalency Axiom, CE）

- 对于任意一个类，其认知表示与外延表示的归类能力等价。

样本可分性公理（SS）

■ 归类不重

- 一个对象总有唯一一个类与其最相似。
- 归类结果 $(Y, V, \underline{Y}, Sim_Y)$ 应满足 $\forall k \exists i (\tilde{y}_k = i)$



各得其所

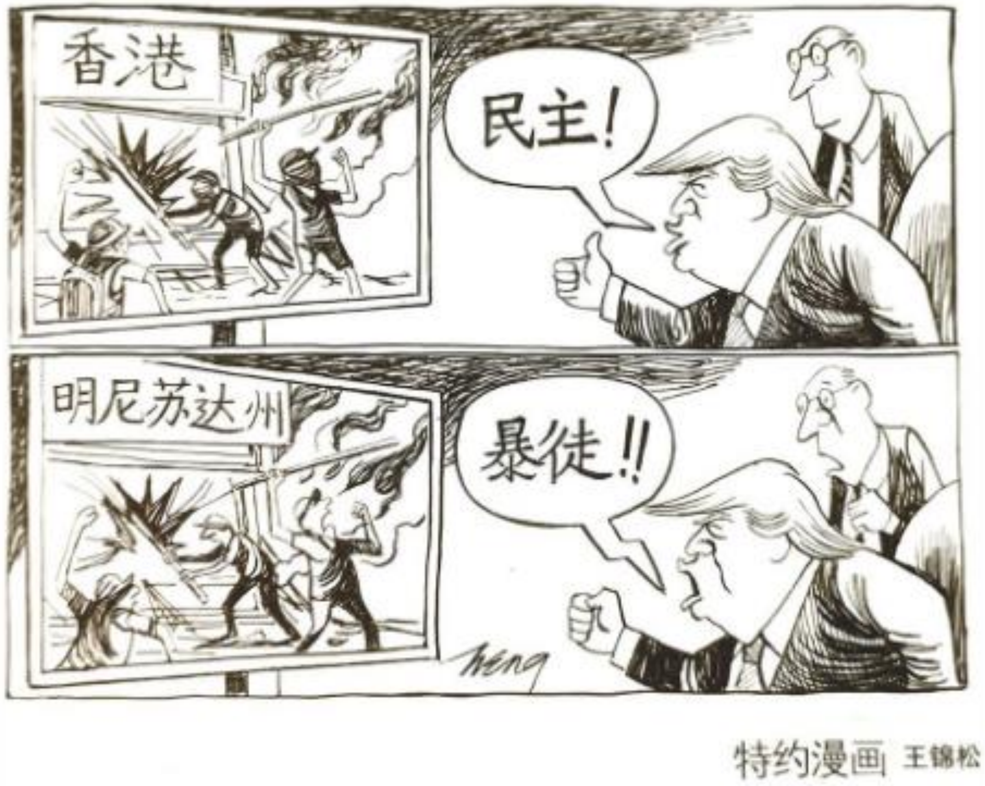
类内非空

- 一个类至少有一个对象与其最相似。
- 归类结果 $(Y, V, \underline{Y}, Sim_Y)$应满足$\forall i \exists k (y_k = i)$

空类

如果违反样本可分公理，则

HO Ching
1 hr ·



Yong Siu Quek
6 hrs ·

今天新加坡联合早报刊登的这张漫画，真有意思。

See Translation

定理2.1

如果一个归类结果有c个类，并满足类可分性公理，则有:

(1)每个类的内部表示都不同。

(2)至少存在c个不同的对象。

定理2.2

如果 $\forall k \forall i \forall j \left((j \neq i) \rightarrow \left(\mathrm{Sim}_Y(yk, \underline{Y_i}) \neq \mathrm{Sim}_Y(yk, \underline{Y_j}) \right) \right)$，则样本可分性公理成立。

类可分性公理的性质

定理2.3

如果一个归类结果$(Y, V, \underline{Y}, Sim_Y)$满足归类公理，则有：

(1) $\forall k \exists i(i = \overrightarrow{y_k})$

(2) $\forall i \exists k(i = \overrightarrow{y_k})$

证明思路: 样本可分公理/类可分公理 + 归类等价公理。

类可分性公理的性质

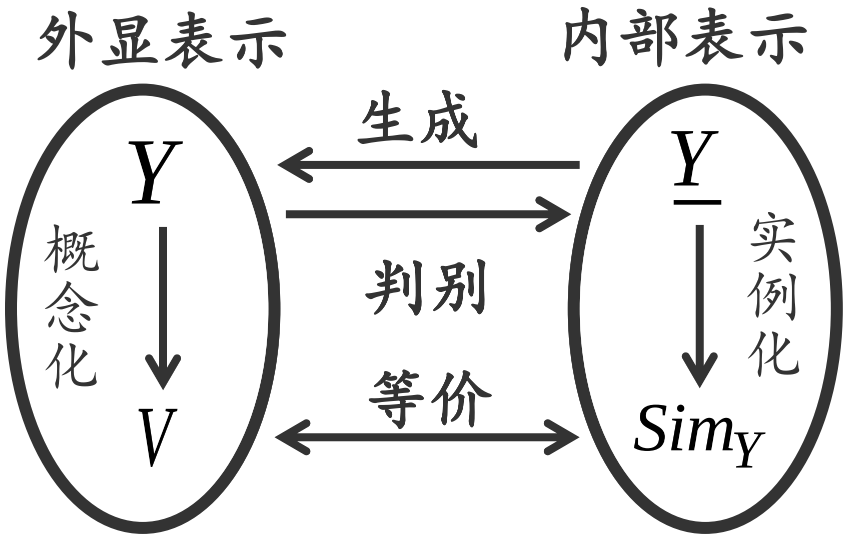
■ 定理2.4

如果一个归类输入$\left(Y, V, \underline{Y}, Sim_Y\right)$及其对应的归类输出$\left(Y, V, \underline{Y}, Sim_Y\right)$满足归类等价公理，则 $\vec{X}=\vec{Y}$ 等价于$\tilde{X}=\tilde{Y}$ 。

归类等价公理(CE)

■ 归类等价

- 对于任意一个类, 其认知表示与外延表示的归类能力等价。
- 归类结果$(Y,V,\underline{Y},Sim_Y)$ 应满足$\vec{Y}=\tilde{Y}$

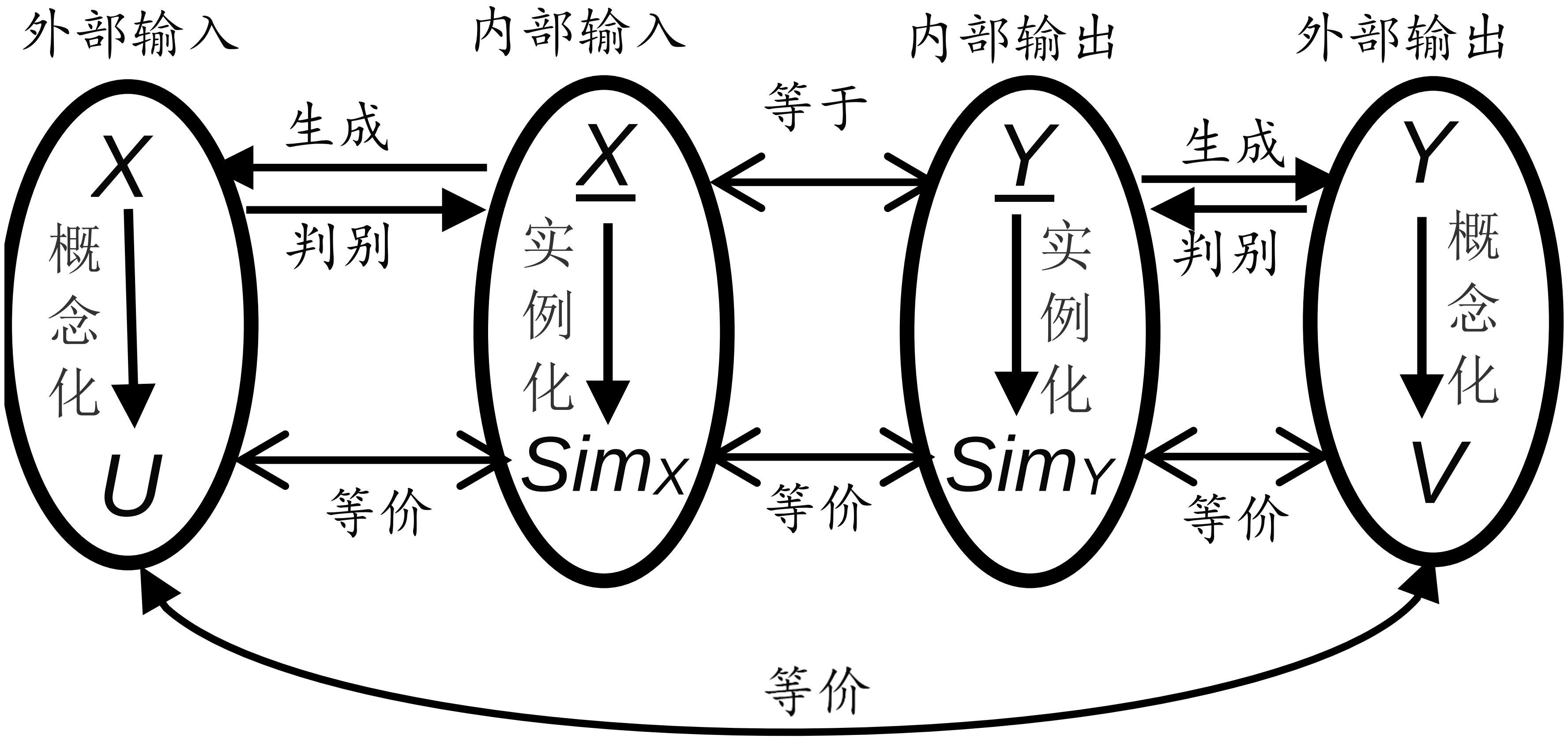


直心是道场；知行合一

类相异函数描述下的归类公理

- **一个归类结果$(Y, V, \underline{Y}, Ds_Y)$ 应满足：**
  - 类可分（CS）：$\forall i \exists k \left(i = \tilde{y}_k\right)$
  - 样本可分（SS）：$\forall k \exists i \left(i = \tilde{y}_k\right)$
  - 归类等价（CE）：$\vec{Y} = \tilde{Y}$

类表示公理和归类公理



外部输入
内部输入
内部输出
外部输出
生成
判别
等于
生成
判别
X
概念化
U
X
实例化
Simx
Y
实例化
SimY
Y
概念化
V
等价
等价
等价
等价

样本可分性公理与其他归类理论

1、原型理论

- 一个对象归为A类而不是其他类，仅仅因为该对象更像A类的原型表示而不是其他类的原型表示。

2、样例理论

- 一个对象归为A类而不是其他类，仅仅因为该对象更像A类的样例表示而不是其他类的样例表示。

上述归类理论与样本可分性公理完全一致。

提要

1、引言

2、类表示与类表示公理

3、归类公理

4、归类结果分类（自学）

5、归类方法设计准则

提要

1、引言

2、类表示与类表示公理

3、归类公理

4、归类结果分类

5、归类方法设计准则

类表示存在公理和归类等价公理要求

- 类表示存在公理仅仅要求输入输出有对应的类内部表示，是归类算法能够设计的基础。

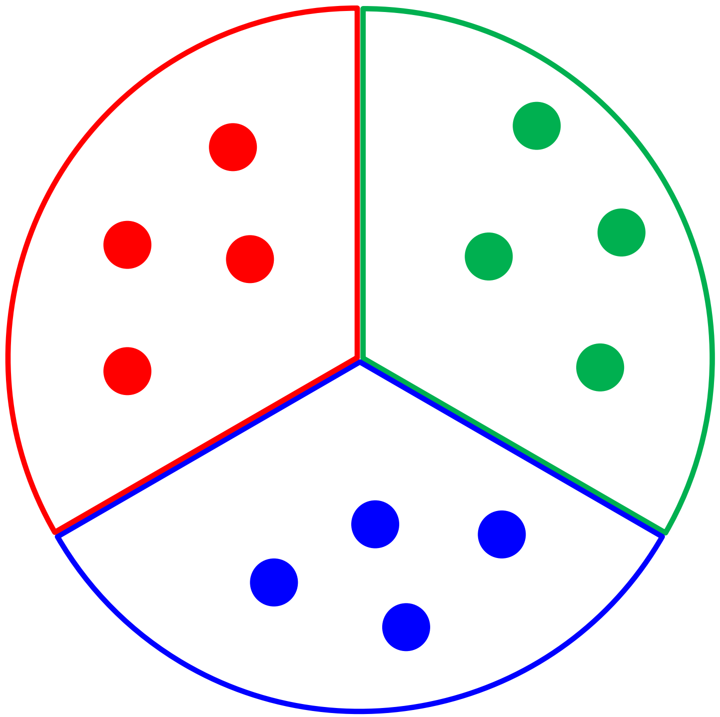
- 归类等价公理要求一个归类算法的外显指称与内蕴指称一致（表里如一），是实现学习目的必要条件。

- 因此，类表示存在公理和归类等价公理在归类算法设计中不需考虑。

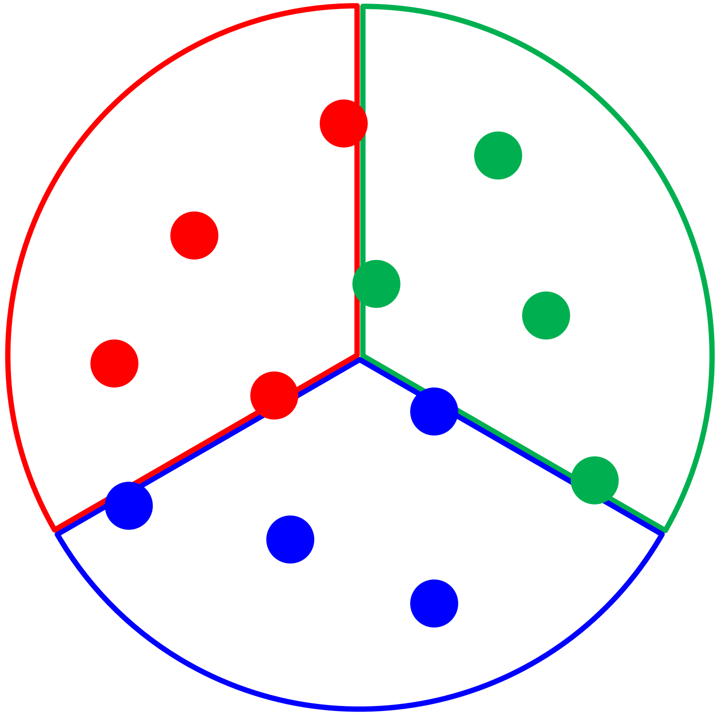
类表示唯一公理要求

- 类表示唯一公理要求信息提供者（言者）与信息接收者（听者）之间完全心意相通、完美匹配。
- 如果类表示唯一公理成立，则归类错误率为零。
- 因此，类表示唯一公理要求太高，在归类算法设计过程中需要适当放低要求。

样本可分性公理要求



理想情况

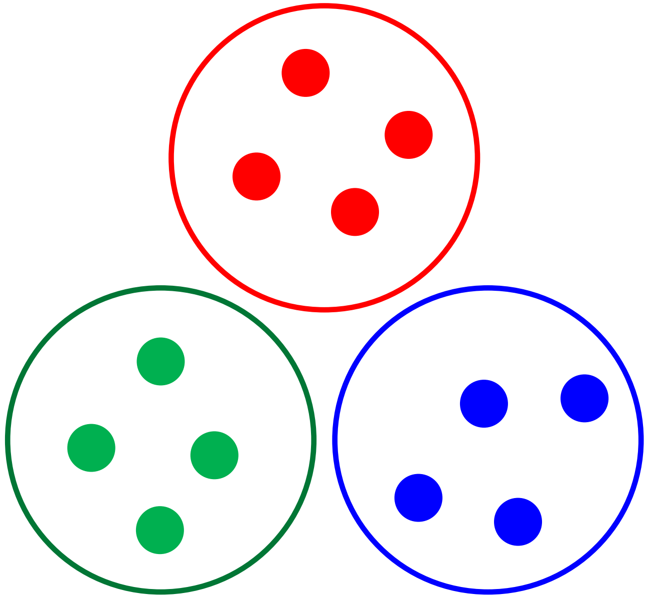


极端情况

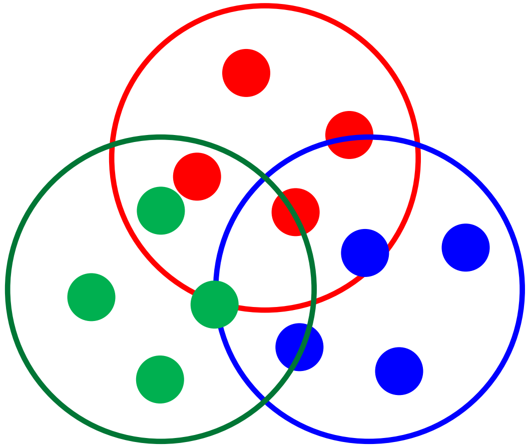
样本可分性公理要求太低！

期望类内样本越紧密越好。

类可分性公理要求



理想情况



极端情况

类可分性公理要求太低！

期望类间距离越远越好。

归类方法设计原则

- 类表示唯一公理要求太高，设计归类算法时应使得归类结果尽量接近类表示唯一公理。由此得出一个设计原则：
  - **类一致性原则**
- 可分性公理要求太低，设计归类算法时应使得归类结果尽量远离违反可分性公理的情形。由此得出两个设计原则：
  - **类紧致性原则**
  - **类分离性原则**

■ 类一致性准则

一个好的归类结果应该使得类表示唯一性公理尽可能成立，即使输入类表示与输出类表示尽量一致。

■ 类一致性判据

$J_E:\{X,\vec{X},\underline{X},\tilde{X}\}\times\{Y,\vec{Y},\underline{Y},\tilde{Y}\}\mapsto R_+$ 是类表示一致性判据，如果 J_E 的最优值对应着 $\{\vec{X},\underline{X},\tilde{X}\}$ 与 $\{\vec{Y},\underline{Y},\tilde{Y}\}$ 之间具有最小误差的归类结果，类一致性判据可以用来判定归类结果是否遵守类表示唯一公理，甚至可以设计归类算法。

类紧致性原则

每个对象与最相似类和其次相似类的相似程度差别要大。由此可知：

- 类内相似度最大
- 类内方差最小

类紧致判据

对于归类输入，$J_C:\{X,U,\underline{X},Ds_X\}\longmapsto R_+$称为类紧致判据，如果$J_C\{X,U,\underline{X},Ds_X\}$的最优值对应的归类输入有最大的紧致性，类紧致性判据可以用来判定归类结果是否遵守类表示唯一公理，甚至可以设计归类算法。

类分离性准则

类分离性原则

一个好的分类结果应该使得类间的距离最大。

类分离性判据

$J_s:\{Y,V\}\times\left\{\left\{\underline{Y}_1,\underline{Y}_2,\cdots,\underline{Y}_c\right\}\middle|\forall i,\underline{Y}_i\ represents\ Y_i\right\}\mapsto R_+$ 是类分离判据，如果 $J_S\{Y,V,\underline{Y}\}$的最优值对应着具有最大类间距离的归类结果，类分离性判据可以用来判定归类结果是否遵守类可分性公理，甚至可以设计归类算法。

奥卡姆剃刀原则

- “如无必要，勿增实体”。
- 简单就是美。
- 对于归类算法来说，如果很多类表示都可行，则选择简单的类表示。
- 何为简单的类表示？如何度量？
  - 一般情况下， 归类算法的输入输出有8元组
  - 单类问题比多类问题简单
  - 特性输入表示与特性输出表示相同时比不同简单
  - 单源数据比多源数据简单
- 对于性能相同或者相近的模型或理论，选择简单的。

BJTU “Machine Learning” Group

于　剑：jianyu@bjtu.edu.cn；

景丽萍：lpjing@bjtu.edu.cn；

田丽霞：lxtian@bjtu.edu.cn;

黄惠芳：hfhuang@bjtu.edu.cn；

李晓龙：hlli@bjtu.edu.cn;

吴　丹：wudan@bjtu.edu.cn；

万怀宇：hywan@bjtu.edu.cn;

王　晶：wj@bjtu.edu.cn.