数学的数学: 范畴论及其应用

田泽禹
tianzy1019@mails.jlu.edu.cn

吉林大学

第六届吉大-哈工大基础学科拔尖学生学术论坛

December 11, 2021

概览

1. 引言
   - 数学的思维
   - 公理化发展
2. 范畴论
   - 范畴、函子、自然变换
   - 米田引理、可表函子
   - 范畴论下的数学
3. 聚类算法的研究
   - 什么是聚类算法
   - 公理化刻画
     - 三类算法特征
     - 不可能性定理
   - 聚类算法的性质研究
     - 特征下的分类

数学是什么，什么是数学的思维？

1. 数学不同于科学，科学是研究客观世界的学问，其特征是可验证、可重复，即存在证伪的过程，科学在不断修正错误中前进，而数学的体系是完备的，不能够有内在的逻辑漏洞。现代数学本质上不属于科学，数学本质上是一套基于人类逻辑的逻辑系统。
2. 在不断的抽象化进程中，数学也在不断的修正巩固自己的系统。从最早的结绳计数，到土地面积的丈量;从解方程到十七世纪的微积分；从具体求解一个问题到研究问题背后的数学结构……
3. 数学的思维是将具体问题化为数学问题，从问题各类特征中抽取本质进行研究，以获得相应的有效结果一种思维方法，如递归法、归纳法、不变量法...

数学抽象化进程

1. 欧几里得《几何原本》中的几何系统是公理系统，具有平面几何五大公理；第五条公理称为平行公理；几何学家曾试图证明第五公理可以由前四公理推出；但最终在修改第五公理的条件下，我们可以得到罗巴切夫斯基几何或黎曼椭圆几何。
2. 在数学抽象化的过程中出现过三次数学危机：毕达哥拉斯悖论，发现了无理数；贝克莱悖论引发了微积分的严格化，推动了以极限为基础的微积分；罗素悖论，指出了集合论存在的漏洞，最终修正为朴素集合论体系。
3. 求一元五次方程的根式解曾困扰数学家三百余年，阿贝尔和伽罗瓦的工作证明了一般一元五次方程没有根式解。这一天才的创举将数学由表象研究转为对数学结构与数学体系的研究。

范畴是什么

- 集合论是一种数学基础研究中的基础尝试，但终于在罗素悖论中找到缺陷。
- 在数学研究中，我们发现在数学不同领域中的知识或定理可能有相似的结构，若不加以归纳整理，一方面庞杂的知识会逐渐变得难于理解，另一方面会使得研究者们难以观察到其中的联系与其揭示的本质。
- 范畴的概念比集合论的更加广泛，集合论在于其会出现大小问题(Size Problem)，使得足够大的集合不再能成为集合；而范畴在不同的大小限制下可以对其进行研究。范畴研究的是抽象对象之间的关系，甚至范畴之间的关系。

我们现在给出范畴的定义:

定义

一个范畴$\mathcal{C}$具有以下的结构:

- 一个范畴中的一系列**对象**，可以视作范畴中的元素，构成ob($\mathcal{C}$)
- 对于任意两个对象$X, Y \in \mathrm{ob}\mathcal{C}$，其间有**态射类** $\mathrm{Mor}_{\mathcal{C}}(X, Y)$
- 态射有复合 $\circ$: $\mathrm{Mor}_{\mathcal{C}}(X, Y) \times \mathrm{Mor}_{\mathcal{C}}(Y, Z) \to \mathrm{Mor}_{\mathcal{C}}(X, Z)$，若有$f \in \circ$:$\mathrm{Mor}_{\mathcal{C}}(X, Y)$，$g \in \circ$:$\mathrm{Mor}_{\mathcal{C}}(Y, Z)$，则复合记为$g \circ f$.
- 对于每一个对象X，有一个单位态射id_X，使得在任意左右复合中，结果与原态射相同。

如果有 $f \in \mathrm{Mor}_{\mathcal{C}}(X, Y)$，$g \in \mathrm{Mor}_{\mathcal{C}}(Y, X)$，且$g \circ f = \mathrm{id}_{\mathrm{X}}$, $f \circ g = \mathrm{id}_{\mathrm{Y}}$, 我们称 X, Y 同构(isomorphic)。

范畴

在数学中我们会研究一些常见范畴，如

- 集合范畴 $\mathfrak{Set}$，其态射为集合之间的映射
- 群范畴 $\mathfrak{Grp}$，其态射为群之间的同态
- 环范畴 $\mathfrak{Ring}$，其态射为环之间的同态，一般指含幺环范畴
- 向量空间范畴 $\mathfrak{Vect}$，其态射为向量空间的线性映射
- 序k范畴 k，在 $k \geq 2$ 为有限正整数的时候，定义对象为 $1, 2, ..., k$ ，态射 $f : a \to b$ 定义为 $a \leq b$.

函子

我们定义两个范畴$\mathcal{C}$到$\mathcal{D}$之间的函子F为

定义

$\mathcal{C}$到$\mathcal{D}$之间的函子F为

- 对象之间的映射 $\mathrm{ob}(\mathcal{C}) \to \mathrm{ob}(\mathcal{D})$
- 将态射类映射到对应的态射类，若 $f \in \mathrm{Mor}_{\mathcal{C}}(X, Y)$，则 $Ff \in \mathrm{Mor}_{\mathcal{D}}(FX, FY)$
- $F(\mathrm{id}_X) = \mathrm{id}_{FX}$, $F(g \circ f) = Fg \circ Ff$.

自然变换

两个范畴$\mathcal{C}$到$\mathcal{D}$之间具有函子，可以视作范畴变换到另一个范畴，两个函子关系特殊时，可以定义两者之间的自然变换，我们假设 F, G 都为范畴$\mathcal{C}$到$\mathcal{D}$之间的函子。

定义

F, G之间的自然变换 α

- 对于每个 $\mathcal{C}$ 中的对象 X ，有 $\mathcal{D}$ 中的自然态射 α_X ： $FX \to GX$
- 对于任意 $\mathcal{C}$ 中对象之间的态射 $f : X \to Y$ 下图关于态射的复合是交换的.

$$\begin{tikzcd} FX \arrow[r, "\alpha_X"'] \arrow[d, "Ff"] & GX \arrow[d, "Gf"'] \\ FY \arrow[r, "\alpha_Y"] & GY \end{tikzcd}$$

米田引理(Yoneda Lemma)、可表函子

我们不加证明地给出范畴论中一个比较重要的引理，它刻画了目标为集合范畴的函子。记$\mathcal{C}^\wedge := \mathrm{Fct}(\mathcal{C}^{op}, \mathfrak{Set})$, $\mathcal{C}^\vee := \mathrm{Fct}(\mathcal{C}, \mathfrak{Set})$.
$h_{\mathcal{C}}(X) = h_X := \mathrm{Hom}_{\mathcal{C}}(-, X)$, $k_{\mathcal{C}}(X) = k_X := \mathrm{Hom}_{\mathcal{C}}(X, -)$, 则有

引理 (米田引理)

$$\forall X \in \mathcal{C},\ \ F \in \mathcal{C}^\wedge \quad y' \colon \mathrm{Hom}_{\mathcal{C}^\wedge}(h_X, F) \simeq F(X)$$
$$\forall X \in \mathcal{C},\ \ G \in \mathcal{C}^\vee \quad y \colon \mathrm{Hom}_{\mathcal{C}^\vee}(k_X, G) \simeq G(X)$$

以第二个式子为例，由于 $G \in \mathcal{C}^\vee$ 因此 G 为 $\mathcal{C} \to \mathfrak{Set}$ 的函子，因此 $G(X)$ 为一个集合, 引理指出, 函子 k_X 到 G 的所有自然变换同构于集合 $G(X)$.

定义 (可表函子)

设$\mathcal{C}$为局部小范畴(两个对象间的态射构成集合)，集合范畴 $\mathfrak{Set}$. 协变函子 $F \in \mathrm{Fct}(\mathcal{C}, \mathfrak{Set})$ 为可表函子, 如果存在 $X \in \mathcal{C}$, 使得 $k_X \simeq F$

范畴论的数学

范畴论诞生后，其应用于数学的各个领域，试图将各个领域的概念统一，有时范畴化的语言会将直白的概念抽象化。历史上的术语“抽象的废话”是常常被用于描述范畴论在那些不那么抽象的领域的应用。但在数学发展的进程中，数学家们逐渐发现抽象简洁化语言导出的定理，能够应用于很广泛的场合。
下面我们不加解释地举例范畴论中一个引理，并指出其在具体数学领域中的应用

引理

对于任意的双函子 $F : I \times J \to \mathcal{C}$, 并且图中的极限与余极限都存在, 则存在一个规范映射

$$\kappa : \quad \mathrm{colim}_{i\in I}\lim_{j\in J}F(i,j) \to \lim_{j\in J}\mathrm{colim}_{i\in I}F(i,j)$$

范畴论的数学

推论

对于任意的集合 X 和 Y 和任意的函数 $f : X \times Y \to \mathbb{R}$

$$\sup_{x\in X} \inf_{y\in Y} f(x,y) \le \inf_{y\in Y} \sup_{x\in X} f(x,y)$$

Proof.

我们将全体实数定义为一个范畴 $\mathbb{R}$, 其对象为实数, $f : a \to b$ 为 a 到 b 之间的态射是指具有关系 $a \le b$.
则根据极限与余极限的定义, 函数 $f : X \to \mathbb{R}$ 视为函子的情况下, 有

$$\mathrm{colim}_{x\in X} f(x) = \sup_{x\in X} f(x), \quad \lim_{x\in X} f(x) = \inf_{x\in X} f(x).$$

则根据上述引理可得结果. □

聚类算法的概念

- 我们仅考虑将有限个点分类的情况.
- 对于任意具有 n 个点的空间, 我们的目的是通过某种**算法**得到空间某些点的一个分类，即将这个空间作分划。由于我们要考虑点的一系列特征，因此可以将不同点之间的相似度定量化，即赋予不同点之间的**权重(weight)**，或**距离(distance)**.
- 一个**算法**的本质是, 给定一个空间及其上面的度量，我们可以得到其上的一个分划.
- 或者相反地思考，给定一个分划，我们能否从点的特征中抽象出一个度量，使得在算法下这个度量诱导了给定分划.

三类特征

由于在实践中，构造一个算法、衡量其性质更多地是出自经验，而如何定量控制算法各方面性质是一个重要的问题。于是我们应该抽象出一些算法所可能具有的性质，直接研究抽象性质的关系。已经有许多数据科学家、数学家在这样的公理化思路做出杰出贡献。**J.Kleinberg** 的体系给出了如下三类特征:

定义

一个算法 F 可能具有以下特征.

- **(Scale-Invariance 尺度不变性)** 若对于任意的度量空间 (X, d_X) 和任意 $\lambda > 0$, 有 $F((X, d_X)) = F((X, \lambda d_X))$. 即任意空间缩放度量都得到相同的分划.
- **(Surjectivity 满性)** 对 X 的任意一个分划 P_X, 都存在一个 X 上的度量 d_X 使得 $F((X, d_X)) = P_X$.
- **(Consistency 一致性)** 假设 $F((X, d_X)) = P_X$, 对于 P_X 中同一类中的点减小距离(权重), 不同类之中的点增大距离(权重), 得到新的度量 d'_X, 应有 $F((X, d_X)) = F((X, d'_X))$.

不可能性定理

我们说明，一个算法不可能同时满足三大特征，这一结论是下一定理的推论。首先给出定义：

定义

假设 Γ, Γ' 都是空间 X 的分划, 如果对于任意类 $C \in \Gamma$, 存在 $C' \in \Gamma'$ 使得 $C \subseteq C'$ 则称 Γ 是 Γ' 的**一个加细**.

定义

如果对于一些分划 Γ 构成的集族 $\mathcal{A}$, 其中没有任何两个分划是存在**加细**的, 则称这个分划集合是一个**反链**.

定理

若一个聚类算法满足尺度不变性和一致性，则对于某度量空间 X

$$\mathrm{Range}(F) = \{P_X | \ \exists d_X, \ \mathrm{s.t} \ F((X, d_X)) = P_X\}$$

是一个反链.

不可能性定理

因为一个度量空间 X 所有的分划必然不可能构成一个反链, 因为满性要求所有分划的可能性，所以一个聚类算法不能同时满足三个条件。
这个命题的逆命题也成立，即

定理

假设一些 X 的分划构成的集族 $\mathcal{A}$ 是一个反链, 那么存在一个聚类算法 F, 满足尺度不变性与一致性, 且有 $\mathrm{Range}(F) = \mathcal{A}$.

下面我们说明, 放宽三个条件中的任意一个条件, 都存在一个聚类算法满足剩余两个条件, 为此我们要引入一种算法: Single-linkage 算法.

1. 将每个对象归为一类, 共得到N类, 每类仅包含一个对象. 类与类之间的距离就是它们所包含的对象之间的距离.
2. 找到最接近的两个类并合并成一类, 于是总的类数少了一个.
3. 重新计算新的类与所有旧类之间的距离.
4. 重复第2步和第3步, 直到达到某类终止条件.

不可能性定理

Single-linkage 算法中，两个类之间的距离定义为任意点对的距离的最小值.

1. 当只剩下 k 个类时停止，除满性外, 尺度不变性与一致性都满足.
2. 仅添加距离(权重)至多为 r 的边, 除尺度不变性外, 满性与一致性均满足.
3. α 设 $\rho^* = \max_{i,j} d(i,j)$, 仅添加权重为 $\alpha\rho^*$ 的边，除一致性外, 尺度不变性与满性均满足.

聚类算法函子化

- 我们研究具有函子性质的算法, 首先考虑**标准聚类范畴** $\mathcal{C}$, 其对象为所有的 (X, P_X), 是度量空间及其上所有的分划; 态射 $g: (X, P_X) \to (Y, P_Y)$ 满足若 x, y 属于 P_X 中的同一类, 那么 $g(x), g(y)$ 属于 P_Y 中的同一类. 也就是说, 通过取 g 的原像诱导的 X 的分划被 P_X 加细.
- 定义一个标准的度量空间, 其中有 k 个点, 且任意两点之间的距离为 δ, 我们称为 $k-1$ 维 δ 单形, 记为 $\Delta_k(\delta)$.
- 我们再引入一种特征, **(Excisiveness 可切性)**, 是指若度量空间 (X, d_X) 通过算法 F 得到分划 P_X, 那么对于分划中任意一个类 (X', d_X), 通过算法得到的分划为自身 $\{X'\}$.

公理化描述

我们首先定义三类度量空间的范畴.

定义

- $\mathcal{M}^{gen}$, 其对象为所有的有限点的度量空间, 及其上所有可能的度量 (X, d_X). 对象间的态射是所有的度量不增映射 f, 即 $d(f(x), f(y)) \leq d(x, y)$.
- $\mathcal{M}^{inj}$, 其对象与 $\mathcal{M}^{gen}$ 的相同, 但对象间的态射要求都为单射.
- $\mathcal{M}^{iso}$, 其对象与 $\mathcal{M}^{gen}$ 的相同, 但对象间的态射要求都为同构.

记 $\mathcal{M}$ 为上述为任一范畴, 我们考虑所有的函子 $F : \mathcal{M} \to \mathcal{C}$, 即所有满足**函子性质的聚类算法**.

关于一些特征的分类

定义度量空间 (X, d_X) 上的等价关系 $\sim_\delta$, 对于 $x, y \in X$, 若存在序列 $x = x_0, x_1, \ldots, x_m = y$ 使得 $d(x_{i-1}, x_i) < \delta$, $(\forall i = 1, 2, \ldots, m)$, 则有 $x \sim_\delta y$.

定义

对于某一个 $\delta > 0$, 我们定义 *Vietoris-Rips* 聚类函子

$$\mathfrak{R}_\delta : \mathcal{M}^{gen} \to \mathcal{C}$$

如下, 对于一个有限度量空间 (X, d_X) 我们把具有等价关系 $\sim_\delta$ 的点划为一类. *Vietoris-Rips*算法是可切的.

关于一些特征的分类

设 Ω 由任意一些有限度量空间构成, 定义一个聚类函子

定义

$$\mathfrak{C}^{\Omega} : \mathcal{M} \to \mathcal{C}$$

为, $\mathfrak{C}^{\Omega}(X, d_X) = (X, \{X_\alpha\}_\alpha \in A)$. 若 x, x'满足下列关系则属于一类.

- 一个 X 中的序列 $z_0, z_1, \ldots, z_k \in X$, $z_0 = x$, $z_k = x'$.
- 一个空间序列 $\omega_1, \ldots, \omega_k \in \Omega$, 并且
- 对于任意的 $i = 1, \ldots, k$ 都有一对点 $(\alpha_i, \beta_i) \in \omega_i$ 和态射 $f_i \in \mathrm{Mor}_{\mathcal{M}}(\omega_i, X)$ s.t. $f_i(\alpha_i) = z_{i-1}$ 和 $f_i(\beta_i) = z_i$.

定义

称一个聚类函子 $\mathfrak{C}$ 是**可表的**, 指存在一个有限个度量空间构成的集合 Ω, 使得 $\mathfrak{C} = \mathfrak{C}^{\Omega}$.

关于一些特征的分类

定理

设 $\mathcal{M}$ 为 $\mathcal{M}^{inj}$ 或 $\mathcal{M}^{gen}$, 那么任意 $\mathcal{M}$ 上的聚类函子是**可切的**当且仅当它是**可表的**.

定理 ($\mathcal{M}^{gen}$ 的函子分类)

假设 $\mathfrak{C}:\mathcal{M}^{gen}\to\mathcal{C}$ 是聚类函子, 且存在 $\delta_{\mathfrak{C}}>0$ 使得

- $\mathfrak{C}(\Delta_2(\delta))$ 在一类中, 对任意 $\delta\in[0,\delta_{\mathfrak{C}}]$.
- $\mathfrak{C}(\Delta_2(\delta))$ 在两类中, 对任意 $\delta>\delta_{\mathfrak{C}}$.

则 $\mathfrak{C}=\mathfrak{R}_{\delta_{\mathfrak{C}}}$.

关于一些特征的分类

定理 ($\mathcal{M}^{gen}$上的尺度不变性)

设函子 $\mathfrak{C}: \mathcal{M}^{gen} \to \mathcal{C}$ 是尺度不变的, 即对于任意 $\lambda > 0$, $\mathfrak{C} \circ \sigma_\lambda = \mathfrak{C}$. 那么要么

- $\mathfrak{C}$ 给每个度量空间分成单点类, 或
- $\mathfrak{C}$ 将每个度量空间分成一类, 即它自身.

关于一些特征的分类

定理 ($\mathcal{M}^{inj}$上的尺度不变性)

设函子 $\mathfrak{C}: \mathcal{M}^{inj} \to \mathcal{C}$ 是尺度不变的，即对于任意 $\lambda > 0$, $\mathfrak{C} \circ \sigma_\lambda = \mathfrak{C}$. 令

$$K(\mathfrak{C}) := \{k \geq 2; \mathfrak{C}(\Delta_k(1)) \text{ 是在一类中的}\}$$

- 如果$K(\mathfrak{C}) = \emptyset$, 那么 $\mathfrak{C}$ 将每个有限度量空间 X 分为单点集.
- 否则，令 $k_{\mathfrak{C}=\min K((C))}$. 那么:
  - 对于任意的 $k \geq k_{\mathfrak{C}}$, $\mathfrak{C}$ 把 k 个点的有限度量空间划分成一个类，即它本身.
  - 对于任意的 $2 \leq k \leq k_{\mathfrak{C}}$, $\mathfrak{C}$ 把 k 个点的有限度量空间分为一系列单点集的类.

总结

1. 介绍了范畴的起源、概念；介绍了函子、自然变换。
2. 介绍了范畴论中的一些具体的定理, 并且说明了其在数学上的具体应用。
3. 刻画了聚类算法的数学意义，给出了相应的特征。
4. 共同拥有三类特征的不可能性。
5. 刻画了有限度量空间聚类算法函子的尺度不变性。

感谢您的耐心收听

纰漏难免，欢迎指教

Email:tianzy1019@mails.jlu.cn
QQ:863947348