arXiv:2512.22692v1 [cs.LG] 27 Dec 2025

与 p-进数一起学习

André F. T. Martins[1,2]

[1]Instituto Superior Técnico，里斯本大学，葡萄牙

[2]电信研究所，里斯本，葡萄牙

andre.t.martins@tecnico.ulisboa.pt

摘要

现有的机器学习框架在实数域（$\mathbb{R}$）上运行，并在实数（欧几里得或希尔伯特）向量空间中学习表示（例如，$\mathbb{R}^d$）。其底层几何性质与线性可分性、最小包围球和子空间投影等直观概念高度契合；而基础微积分则为通过基于梯度的优化进行学习提供了工具箱。

但这真的是唯一可能的选择吗？在本文中，我们研究了一种截然不同的域作为 $\mathbb{R}$ 的替代方案—— p进数 $\mathbb{Q}_p$的超度量、非阿基米德空间。p进数的层次结构，以及将其解释为无限字符串的方式，使其成为编码理论和层次表示学习的理想工具。我们的探索性理论工作为使用 p进数进行分类、回归和表示学习奠定了基础，并提供了相应的学习模型和算法。我们展示了如何将简单的 Quillian 语义网络表示为紧致的 p进线性网络；而使用实数域则无法实现这一构造。最后，我们讨论了由这一新框架带来的开放问题和未来研究机遇。

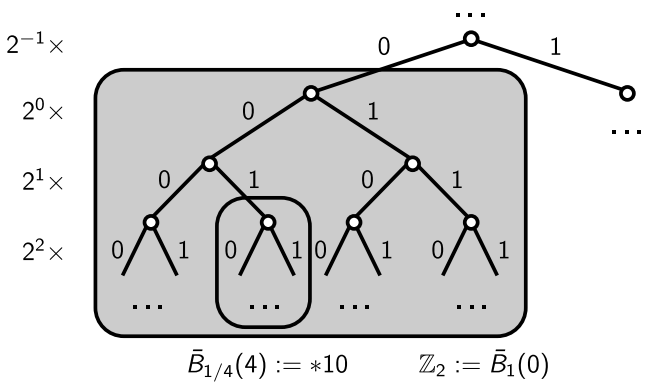

图 1： $\mathbb{Q}_p$ 关于 $p = 2$的层次结构。

目录

1 I引言 3

2 背景 3

2.1 非阿基米德绝对值与超度量 .32.2 $\mathbb{Q}_p$ -进数域与 p-进整数环 $\mathbb{Z}_p$ p . . . . . 42.3 $\mathbb{Q}_p$. 6

3 p-进分类 6

3.1 线性分类器：一维情形 .73.2 非线性分类器：一维情形 .93.3 线性分类器：多维输入 .103.4 学习一个 p- 进线性分类器 .11

4 p-进数回归 12

4.1 线性回归：单维情况 .124.2 线性回归：多维情况 .13

5 p-进数表示 13

6 开放问题 18

6.1 线性多分类 p-进数分类器 .186.2 基于梯度的优化与 p-进数求根 .196.3 多层 p-进数网络 .196.4 使用所有素数：阿代尔预测器 .20

7 结论 21

A 在超度量空间中，每个三角形都是等腰的 23B 关于 p-adic球的附加性质 23

C 有理数域 A p-adic 版 Kraft 不等式的证明 24

D 证明 25

D.1 命题 2 的证明 .25D.2 命题 3 的证明 .25D.3 命题 4 的证明 .26D.4 命题 5 的证明 . 27D.5 命题 6 的证明 .28D.6 命题 7 的证明 .28

1 引言

自Hensel (1897年) 引入以来， p-进数已在数论、代数几何、物理学及其他领域得到广泛应用（Koblitz，1984年；Robert，2000年；Gouvêa，2020年）。它们在许多重要方面都不同于实数，因此产生了许多引人入胜且出人意料的结果，例如等式

$$1+2+4+8+\ldots=-1$$

或者说一个有趣的事实：在 p-进世界中，所有三角形都是等腰三角形，而且球内的任意一点都是该球的中心。

然而，除了少数例外，几乎没有研究探讨 p-进数在机器学习中的潜力。Bradley(2009年)研究了 p-进数据的聚类，并提出了用于最小化簇能量的次优算法。Murtagh (2004年, 2009年)分析了数据中的树状图和超度量性。Chierchia 和 Perret (2019年)以及Cohen-Addad 等人(2020年)开发了将超度量拟合到数据的流程。Khrennikov 和 Tirozzi(1999年)提出了一种“p-进神经网络”（与我们在 §3中的一维线性分类器类似）。Baker 和 Molla-Aliod (2022年)在自然语言处理中的一个小型序列到序列问题上使用了 p-进回归，其做法与我们在 §4中的表述有一定相似之处。

本文尝试通过构建分类（§ p-adic 机器学习的基础模块，来奠定其基础3）和回归问题（§4），并探索 p-adic 表示（§5）的表达能力。本文最后列出一些开放问题（§6），我认为解决这些问题是使这一框架具备实际应用价值的必要条件。（标题有意造成误导：本文主要探讨的是我通过 -adic p所学到的东西，而非如何利用 p-adic进行机器学习。）

这项探索面临若干挑战，因为基于梯度的优化和统计学等经典工具在 p-adic 的世界中并不容易应用：虽然我们可以进行微积分，并用牛顿法求函数的根，但 p-adic 的拓扑性质似乎使导数不那么有用（例如，处处导数为零的函数未必是常数），而且由于 p-adic 不像实数那样构成有序域，如何进行优化也并不显而易见。不过，它们具有一种非常有趣的层次结构，似乎有望用于表示学习，以及某些分类和回归问题。

记号。 我们分别用 $\mathbb{R}$ 和 $\mathbb{Q}$ 表示实数域和有理数域，用 $\mathbb{R}_+$表示非负实数，用 $\mathbb{Z}$ 表示整数环。我们记作 $[n]=\{1,\ldots,n\}$。

2 背景

我们先回顾超度量空间、 p-进数域和 p-进整数环及其基本性质（Gouvêa，2020年）。

2.1 非阿基米德绝对值与超度量

设 $\mathbb{K}$ 是一个域（例如 $\mathbb{Q}$ 或 $\mathbb{R}$），并令 $\mathbb{R}_+$ 表示非负实数。在 $\mathbb{K}$ 上的**绝对值** 是一个函数 $|.|:\mathbb{K}\to\mathbb{R}_+$，满足 (i) $|x|=0$ 当且仅当 $x=0$；(ii) $|xy|=|x||y|$

对所有 $x,\ y \in \mathbb{K}$；(iii) 对所有 $x,\ y \in \mathbb{K},\ |x+y| \leq |x|+|y|$。若绝对值具有以下性质（强于性质 iii），则称其为**非阿基米德**：

$$|x+y| \leq \max\{|x|,|y|\}, \quad \text{for all } x,y \in \mathbb{K}. \tag{1}$$

否则称为"阿基米德"的。[1]

例 1（平凡绝对值）。定义为 $|x|=1$（若 $x \neq 0$，且 $|0|=0$）的函数是一个绝对值，称为平凡绝对值。它是非阿基米德的。

例 2（$\mathbb{Q}$ 和 $\mathbb{R}$上的通常绝对值）。设 $\mathbb{K}=\mathbb{Q}$ 或 $\mathbb{R}$。通常绝对值$|x|:=$最大值$\{x,-x\}$是阿基米德的。

例 3（p进绝对值在 $\mathbb{Q}$上）。设 p 是一个素数。任意非零 $x \in \mathbb{Q}$ 都可以唯一地写成 $x=p^n\frac{a}{b}$，其中 a 和 b 是不被 p 整除的互素整数，并且 $n \in \mathbb{Z}$。$\mathbb{Q}$ 上的 p-进绝对值为

$$|x|_p := \begin{cases} p^{-n} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0. \end{cases}$$

这个绝对值（本文中将广泛使用）是非阿基米德的。

绝对值通过距离函数 $d(x,y):=|x-y|$ 在 $\mathbb{K}$ 上诱导出一个度量（进而诱导出一个拓扑）。由上面的性质 (i–iii) 可知，我们必有 (i) $d(x,y) \geq 0\ \forall x,y \in \mathbb{K}$，等号成立当且仅当 $x=y$；(ii) $d(x,y)=d(y,x)\ \forall x,y \in \mathbb{K}$；(iii)$d(x,z) \leq d(x,y)+d(y,z)\ \forall x,y,z \in \mathbb{K}$。性质 (iii) 称为（弱）三角不等式。如果 $|\cdot|$ 是非阿基米德的，那么诱导出的度量满足强三角不等式：

$$d(x,z) \leq \max\{d(x,y),d(y,z)\} \quad \forall x,y,z \in \mathbb{K}. \tag{2}$$

满足 (2) 的度量称为超度量，相应的 $\mathbb{K}$ 称为超度量空间。超度量空间有一个有趣的性质：每个三角形都是等腰三角形——更具体地说，对于任意 $x,y,z \in \mathbb{K}$, $d(x,y) \neq d(y,z)$，都有 $d(x,z)=\max\{d(x,y)\ d(y,z)\}$。附录 A 中给出了一个简单的证明。

2.2 $\mathbb{Q}_p$ 域：p-进数与 $\mathbb{Z}_p$ 环：p-进整数 ers

让我们回顾一下，实数集 $\mathbb{R}$ 是如何通过完备化有理数集$\mathbb{Q}$构造出来的。给定一个绝对值及其诱导的度量，我们可以定义开球and a收敛的概念。域 $\mathbb{K}$ 是**完备的**，如果 $\mathbb{K}$ 中的任意柯西序列[2] 都收敛到 $\mathbb{K}$中的一个极限点——例如，$\mathbb{Q}$相对于通常绝对值并不完备。[3] 通过填补

[1]这个名称源于阿基米德绝对值满足阿基米德性质：对于任意$x,\ y \in \mathbb{K}$若 $x \neq 0$，则存在一个整数 $k \in \mathbb{Z}$使得 $|kx|>|y|$。这等价于断言存在任意"大"的整数，这一观察可追溯至阿基米德。非阿基米德绝对值则不满足这一点；在那里，(1) 蕴含 $|kx|=|x+\ldots+x| \leq |x|$。[2]元素数列 $x_n \in \mathbb{K}$ 是柯西序列，当且仅当 $\forall \epsilon>0$ 存在 $M \in \mathbb{N}$ 使得 $m,\ n \geq M$ 蕴含 $|x_n-x_m|<\epsilon$。[3]例如，由 $x_{n+1}=\frac{1}{2}x_n+x_n^{-1}$定义的数列 $x_n=1$ 收敛于 $\sqrt{2} \notin \mathbb{Q}$（这是通过对 $x^2-2=0$应用牛顿法求解得到的数列）。

用这些极限点——无理数——填补 $\mathbb{Q}$ 中的“空洞”，我们就得到了更大的完备域实数域 $\mathbb{R}$。

如果我们沿用这个思路，但改用 p-进绝对值 $|\cdot|_p$ 来完备化 $\mathbb{Q}$，会发生什么？[4] $\mathbb{Q}$ 关于度量 $d(x,y)$：仍然是不完备的。$=|x-y|_p$将 $\mathbb{Q}$ 中关于此度量的柯西序列的极限添加进去，我们就得到 **p-进数域**，记作 $\mathbb{Q}_p$。子集 $\mathbb{Z}_p:=\{x\in\mathbb{Q}_p: \ |x|_p\leq 1\}\subset\mathbb{Q}_p$ 称为**p-进整数集合**，并具有环结构。

命题 1(p-进展开)。每个 $x\in\mathbb{Z}_p$ 都可以唯一地写成以 p为底的无限“数字”展开：

$$x=a_0+a_1p+...+a_np^n+...=\sum_{i=0}^{\infty}a_ip^i,$$

其中每个 $a_i\in\{0,...,p-1\}$。当 p 可由上下文清楚得知时，我们将其简写为 $x=\cdots a_2a_1a_0$.

每个 $x\in\mathbb{Q}_p$ 都可以唯一地写成：

$$x=\sum_{i=-m}^{\infty}a_ip^i:=\cdots a_2a_1a_0\,.\,a_{-1}\dots a_{-m}, \tag{3}$$

其中 $m\in\mathbb{Z}$，每个 $a_i\in\{0,...,p-1\}$，且 $a_{-m}\neq 0$。此外，我们有 $|x|_p=p^m$。

.....之间一个关键区别 $\mathbb{Q}_p$ 和 $\mathbb{R}$ 是，在 $\mathbb{Q}_p$ 中，数位展开是向“左”而不是向“右”进行的。此外，p-进展开式（3）是**唯一的**，而实数并非如此（例如，$1.000...=0.999...$ 是同一个数在十进制下的两种不同展开形式）。

在 $\mathbb{Z}_p$ 中，可以利用 p进展开式进行加法和乘法，方法与通常对 $\mathbb{Z}$进行运算时相同，只不过进位会无限地向“左”延伸。在 $\mathbb{Q}_p$ 中也是如此，只需考虑“小数点”即可。

例 4. 取 $p=2$。数 $-1=\cdots111$、$12=1100$ 和 $\frac{1}{3}=\cdots10101011$ 都是2-进整数。数 $\frac{1}{2}=0.1$ 和 $\frac{5}{4}=1.01$ 属于 $\mathbb{Q}_2$，但不是2-进整数。

例 5. 取 $p=5$。$\frac{1}{2}=\cdots2223$ 和 $\frac{1}{4}=\cdots3334$ 都是5-进整数。$5\,2$ 会在$1\,2$的右侧添上一个零，$\frac{5}{2}=\cdots2230$。注意，逐位加法 $\frac{1}{4}+\frac{1}{4}=\frac{1}{2}$ 在向左进位时成立（模 5）。

自然地，我们有 $\mathbb{Q}\subset\mathbb{Q}_p$，$\mathbb{Z}\subset\mathbb{Z}_p\subset\mathbb{Q}_p$，且 $\mathbb{Z}_p\cap\mathbb{Q}=\{\frac{a}{b}\in\mathbb{Q}: \ p\nmid b\}$。从拓扑上看，$\mathbb{Q}_p$ 与 $\mathbb{R}$截然不同。[5] 例如，不同于 $\mathbb{R}$，$\mathbb{Q}_p$ 不是有序域，也就是说，诸如 $x\geq 0$ 或 $x\leq y$ 这样的表达式在 x中 $y\in\mathbb{Q}_p$没有意义。因此，我们需要一种新方法来定义 p-进制二元分类器，详见§3。

[4]事实上，奥斯特罗夫斯基定理（Gouvêa，2020年，定理 3.1.4）出人意料地指出，§2.1中的三个例子完整呈现了 $\mathbb{Q}$中关于绝对值的全貌：$\mathbb{Q}$ 上的每个非平凡绝对值都等价于（即诱导相同的拓扑）某个 p-进绝对值或通常绝对值。这一结果表明，我们应当赋予 p-进绝对值与通常绝对值同等的重要性。将后者记作 $|.|_\infty$，而乘积公式则优美地刻画了所有非平凡绝对值之间的关系：（$\prod_{p\leq\infty}|x|_p=1$（Gouvêa，2020年，命题 3.1.5）。我们将在 §6.4中回到这个优美的公式。[5]形式上，$\mathbb{Q}_p$ 是一个全不连通豪斯多夫拓扑空间。$\mathbb{Z}_p$ 是紧致的，且 $\mathbb{Q}_p$ 是局部紧致的。（Gouvêa，2020年，推论 4.2.6-7）。

2.3 Qp 的性质

有理数域中的球 $\mathbb{Q}_p$。令 $a \in \mathbb{Q}_p$ 与 $r \in \mathbb{R}_+$。以 a 为中心、半径为 r的（闭）球，记作$\bar{B}_r(a)$，是如下集合

$$\bar{B}_r(a) := \{x \in \mathbb{Q}_p : |x-a|_p \le r\}.$$

由于 p-进绝对值函数的像集为集合 $\{0\} \cup \{p^n : n \in \mathbb{Z}\}$，许多半径不同的球实际上相同——例如，对于任意 $r \in [p^n, p^{n+1}[$，$\bar{B}_r(a) = \bar{B}_{p^n}(a)$。因此，在大多数情况下，只需考虑形如 $\bar{B}_{p^n}(a)$ 的球，其中 $n \in \mathbb{Z} \cup \{-\infty\}$。$\mathbb{Q}_p$（更一般地说，超度量空间）有一个有趣的性质：球中的每个点都是该球的中心——也就是说，对于任意 $x \in \bar{B}_r(a)$，$\bar{B}_r(a) = \bar{B}_r(x)$。另一个重要性质是，任意两个球要么不相交，要么一个包含在另一个之中——也就是说，对于满足 $r \le s$ 的任意 a、$b \in \mathbb{Q}_p$ 和 r、$s \in \mathbb{R}_+$，要么 $\bar{B}_r(a) \cap \bar{B}_s(b) = \varnothing$，要么 $\bar{B}_r(a) \subseteq \bar{B}_s(b)$。从拓扑角度看，$\mathbb{Q}_p$ 中的球既闭又开（“闭开集”）。

例6。任何半径为零的球都是一个单点集，$\bar{B}_0(a) = \{a\}$。

例7。p-进整数集是一个单位球，$\mathbb{Z}_p = \bar{B}_1(0)$。选择0 作为中心是任意的：对于任意 $x \in \mathbb{Z}_p$，我们同样有 $\mathbb{Z}_p = \bar{B}_1(x)$ 。我们可以将 $\mathbb{Z}_p$ 分解为 p 较小的球的不交并：

$$\mathbb{Z}_p = \bar{B}_1(0) = \bar{B}_{\frac{1}{p}}(0) \cup \bar{B}_{\frac{1}{p}}(1) \cup \ldots \cup \bar{B}_{\frac{1}{p}}(p-1). \tag{4}$$

这表明 $\mathbb{Z}_p$ 具有层次结构，如图 2所示。注意，任意半径为 p^n的 $\mathbb{Q}_p$ 球也会出现类似情况：它可以分解为 p 个较小的球的不交并，每个球的半径为 p^{n-1}。附录 p展示了更多关于B-进球的性质。

例8 (p-进球是字符串）。任意球 $\bar{B}_r(a) \subseteq \mathbb{Z}_p$，只要 $r > 0$，都可以与一个字符串对应，该字符串取自一个包含 p 个符号的字母表。具体来说，若 $r = p^{-n}$ 且 $n \in \mathbb{N}$，则 $\bar{B}_r(a)$中的任意元素都有一个形如以下形式的 p-进展开式

$$x = *a_{n-1} \ldots a_0,$$

其中，a_i 是 a 的 i^{th} 位数字，而通配符 $*$ 表示左侧出现的任意符号。因此，$\mathbb{Z}_p$ 的元素可视为“无限字符串”，而 $\mathbb{Z}_p$ 的球可视为有限长度的字符串。

在附录 C中，我们利用上文 p-进球与字符串之间的联系，给出一个 **Kraft不等式** (Kraft，1949）的简单证明。Kraft不等式是信息论中的一项关键结果，它确立了一个码成为前缀码的充要条件。（据我们所知，这种基于 p-进球的证明是新的。）

3 p进分类

令 $d \in \mathbb{N}$，并用 $\mathbb{Q}_p^d$ 表示域 $\mathbb{Q}_p$ 上的 d 维向量空间，其加法和标量乘法采用标准定义。本节讨论 p-进制二元分类器，它们以 d 维特征向量 $x \in \mathbb{Q}_p^d$ 作为输入，并预测输出

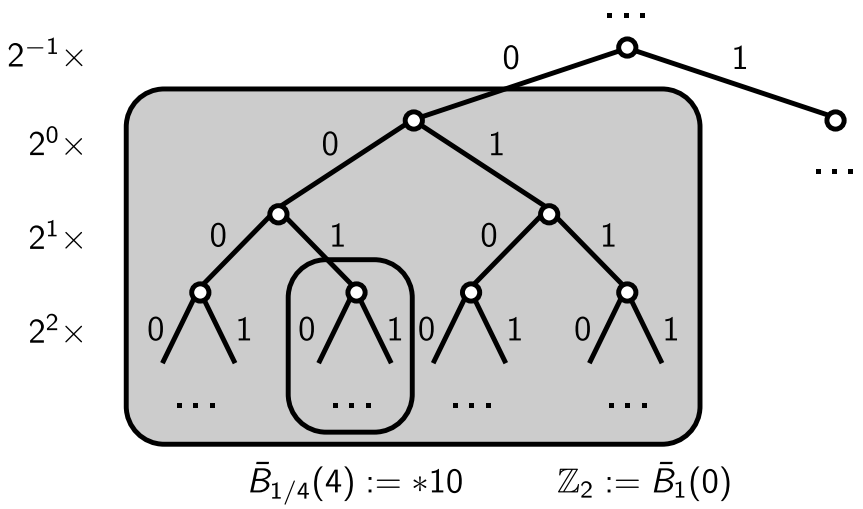

图2：的层次结构 $\mathbb{Q}_p$ 对于 $p = 2$。中的球 $\mathbb{Q}_p$ 要么相互嵌套，要么彼此不相交¯，每个球都对应层级结构中的一个节点。图中显示了两个相互嵌套的球， $B_1(0)$，¯它是以下集合： p-进整数，以及 $B_{1/4}(4)$，它包含于前者之中，其元素为 p形式的-进整数： $\cdots 10$。

$y \in \{\pm 1\}$。我们假设训练数据为 $\mathcal{D} = \left\{(x^{(i)}, y^{(i)})\right\}_{i=1}^{n} \subseteq \mathbb{Q}_p^d \times \{\pm 1\}$，其中 $n \in \mathbb{N}$。我们同时考虑单类分类（也称为异常检测），其训练集仅包含正例；以及二类分类，其训练集同时包含正例（$+1$）和负例（-1）。

在 $\mathbb{R}$ 上运行的二元分类器根据规则 $\hat{y}(x) = \text{sign}(f(x))$ 进行预测，其中f：$\mathbb{R}^d \to \mathbb{R}$ 是判别函数，而 sign(.) 是符号函数。然而，对于 f：$\mathbb{Q}_p^d \to \mathbb{Q}_p$的 p-进分类器，我们无法应用类似的预测规则，因为 $\mathbb{Q}_p$ 不是有序域，因此无法像上面那样使用符号函数。相反，我们定义一条分类规则：如果参数属于 $\mathbb{Z}_p$，预测结果为 $+1$；否则为 -1：

$$\hat{y}(x) := \begin{cases} +1, & \text{if } |f(x)|_p \leq 1 \\ -1, & \text{otherwise.} \end{cases} = \begin{cases} +1, & \text{if } f(x) \in \mathbb{Z}_p \\ -1, & \text{otherwise.} \end{cases} \tag{5}$$

接下来的小节将讨论一维情形（$x \in \mathbb{Q}_p$），然后将其扩展到 $d \geq 1p$-adic 特征（$x \in \mathbb{Q}_p^d$）。

3.1 线性分类器：一维情形

在线性一维情形（$x \in \mathbb{Q}_p$）中，我们使用（5），其中 $f(x) = wx + b$，其中w，$b \in \mathbb{Q}_p$ 是模型参数。如果 $w = 0$，这就成为一个恒为正（$b \in \mathbb{Z}_p$）或恒为负（$b \notin \mathbb{Z}_p$）的平凡分类器，因此我们关注 $w \neq 0$ 的情形。由于 $|wx + b|_p = |w|_p \left|x + \frac{b}{w}\right|_p$，决策规则（5）变为

$$\hat{y}(x) = \begin{cases} +1, & \text{if } x \in \bar{B}_r(a) \\ -1, & \text{otherwise,} \end{cases} \tag{6}$$

其中 $a = -\frac{b}{w}$与 $r = \frac{1}{|w|_p}$。因此，一维 p-adic 二元线性分类器**只需检查输入 x 是否位于一个 p-adic 球内**。要训练这样的分类器，我们需要找到一个球，使其包含尽可能多的正训练样例，同时排除大多数负训练样例。当且仅当存在一个球能够

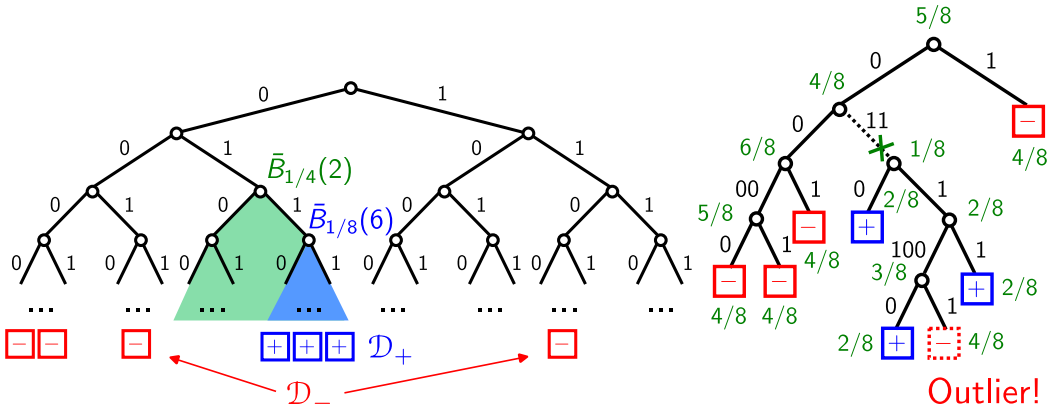

图 3：左：$\mathbb{Q}_2$ 中的一个可分数据集及其包围球。正例的形式为 *110，负例的形式则为 *00 或 *1。我们选择 $x_i = 6 = 110$作为正类的代表。点 $x_j = -2 = \cdots 1110$ 与 x_i的距离最大。利用命题 2，我们得到 $w^\star = 1/8$ 和 $b^\star = -3/4$，分别对应于 $\bar{}$最小包围球 $B_{1/8}(6) = *110$。最大包围球为 $B_{1/4}(6) = *10$。右：加入离群点 $x_- = \cdots 11001110$ 后，数据集变得不可分。设数据集为 $\mathcal{D}_+ = \{*0110, *01001110, *11110\}$ 和 $\mathcal{D}_- = \{*00000, *10000, *100, *11001110, *1\}$（左侧的延续部分并不重要）。我们将这些点表示在一棵树中，其中每个叶节点表示包含单个点的最大球，而节点表示分割点（边标有一个或多个符号）。为了确定误分类误差最小的分类器，我们考虑以每个节点（或叶节点）为根的球，并计算与该球相关的训练误差。在此例中，最优分类器对应$\bar{}$于球 $B_{1/4}(2) = *10$或 $B_{1/8}(6) = *110$中的任一个；切断虚线边后，训练误差为 1/8。

将一类与另一类完美分离——这个问题可以在 $\mathcal{O}(n)$ 时间内解决。在单类分类问题中，没有负例可用，因此一个合理的准则是寻找半径最小且能包含正例的球——这与有理数域 $\mathbb{R}$ (Nolan，1991年；Schölkopf 等人，1999年) 上欧几里得空间／希尔伯特空间中的类似目标相呼应。对于 p-adic 分类器，这等价于寻找正例的最近公共祖先节点。下文给出完整刻画，证明见附录D.1（图3对此作了说明）。

命题 2. 设 $\mathcal{D} = \mathcal{D}_+ \cup \mathcal{D}_-$ 是一个包含正例/负例 $\mathcal{D}_+ = \{x^{(1)}, \ldots, x^{(m)}\}$ 和 $\mathcal{D}_- = \{x^{(m+1)}, \ldots, x^{(n)}\}$的训练集。假设 $m \geq 2$，且 $\mathcal{D}$ 线性可分，也就是说，存在 (w, b) 使得 $|wx_+ + b|_p \leq 1\ \forall x_+ \in \mathcal{D}_+$ 且 $|wx_- + b|_p > 1 \forall x_- \in \mathcal{D}_-$。那么：

1. 任取 $x^{(i)} \in \mathcal{D}_+$，令 $x^{(j)} \in \arg\max_{x_+ \in \mathcal{D}_+} |x_+ - x^{(i)}|_p$ 为距离最远的正例。那么，$w^\star := 1/(x^{(j)} - x^{(i)})$ 和 $b^\star := -x^{(i)}/(x^{(j)} - x^{(i)})$ 参数化了一个分离线性分类器。该分类器满足 $w^\star x^{(i)} + b^\star = 0$ 和 $w^\star x^{(j)} + b^\star = 1$。

2. 上述分类器 $(w^\star, b^\star)$ 对应于包围球 $\bar{B}_r(a)$，其中 $a = x^{(i)}$ 且 $r = |x^{(j)} - x^{(i)}|_p$——这是包含 $\mathcal{D}_+$的（唯一）最小包围球。

3. 令 $x^{(k)} \in \operatorname{argmin}_{x_- \in \mathcal{D}_-} |x_- - x^{(i)}|_p$ 为距离最近的负例（由于¯可分性和超度量性，它也与 $\mathcal{D}_+$中的任意点距离最近）。那么， $B_{r'}(a)$以及 $a = x^{(i)}$ 和 $r' = p^{-1}|x^{(k)} - x^{(i)}|_p$ 构成包含$\mathcal{D}_+$ 的（唯一）最大包围球，并定义了一个分离线性分类器。

4. 任何中心位于 $x^{(i)}$ 且半径属于 $[r, r']$ 的球，都是定义分离线性分类器的包围球。任何分离线性分类器都具有这种形式。

如果问题不可分，那么可以针对某个损失函数（例如误分类率），通过 $\mathcal{O}(n)$ 算法找到最佳包围球（见图3）：

1. 首先，构建一棵树，其叶节点表示 $\mathcal{D}_+$ 中的点，其节点表示包含这些点子集的球（边表示一个或多个数字，且只需考虑具有多个子节点的节点）。

2. 然后，检查每个节点，通过统计相应的球所包围的正类点和负类点的数量，计算损失值。选出最佳节点。

有趣的是，可以将这一过程与实数轴（$\mathbb{R}$）上的情况进行比较；在那里，这个一维问题同样可以处理——只需考察相邻点所界定的区间，而不必对节点进行推理。

3.2 非线性分类器：一维情形

现在考虑具有非线性判别函数 $f\colon \mathbb{Q}_p \to \mathbb{Q}_p$ 的决策规则 (5)，例如次数为 $s \geq 1$ 的多项式， $f(x) = c\prod_{j=1}^{s}(x - a_j)$，其中 $c, a_1, ..., a_s \in \mathbb{Q}_p$。在这种情况下（假设 $c \neq 0$），当且仅当 $\prod_{j=1}^{s} |x - a_j|_p \leq r$ 时，决策为 $\hat{y}(x) = +1$，其中 $r = 1/|c|_p$。我们证明，此时正区域可能是若干 p-进球的并集。[6]

命题 3 (2nd 阶 p-进分类器)**.** 设 $f(x) = c(x - a_1)(x - a_2)$，其中 $\hat{y}(x)$ 定义为 (5)。设 $|a_1 - a_2|_p = p^{-k_{12}}$ 且 $r = 1/|c|_p = p^{-k}$ 对某个 $k_{12}, k \in \mathbb{Z}$。则

1. 如果 $k > 2k_{12}$，$\hat{y}(x) = +1$ 当且仅当 $x \in B_{p^{-k+k_{12}}}(a_1) \cup B_{p^{-k+k_{12}}}(a_2)$。也就是说，正区域是两个半径相同的球的不交并。

2. 如果 $k \leq 2k_{12}$，$\hat{y}(x) = +1$ 当且仅当 $x \in B_{p^{-\lceil k/2 \rceil}}(a_1) = B_{p^{-\lceil k/2 \rceil}}(a_2)$，也就是说，正区域是一个球，且该分类器等价于线性分类器。

证明见附录 D.2。[7]

例 *9*。考虑由 $|f(x)|_p = |\frac{1}{64}(x^2 - 6x + 5)|_p \leq 1$ 在 $\mathbb{Q}_2$ 中定义的分类器。该函数可分解为 $f(x) = \frac{1}{64}(x - 1)(x - 5)$。我们有 $k = 6$ 与 $k_{12} = 2$。由于 $k > 2k_{12}$，且我们有 $-k + k_{12} = -4$，因此正区域由 $B_{\frac{1}{16}}(1) \cup B_{\frac{1}{16}}(5)$ 定义，其中包含形如 $*0001$ 或 $*0101$ 的数。参见图 *4*。

[6]注意，次数足够高的多项式分类器可以通过设置 $s := |\mathcal{D}_+|$，$a_j := x^{(j)}$，对每个 $x^{(j)} \in \mathcal{D}_+$ 进行分类，并选择 $r = 1/|c|_p < \min_{x_- \in \mathcal{D}_-} \prod_{x_+ \in \mathcal{D}_+} |x_- - x_+|_p$，以完美准确率过拟合训练数据。这确保了对于任意 $x_+ \in \mathcal{D}_+$，$f(x_+) = 0$，并且对于任意 $x_+ \in \mathcal{D}_+$，$|f(x_-)|_p > 1$。[7] 对于 $s > 2$，我们得到至多 s 个球的并，但这些球不必大小相同。

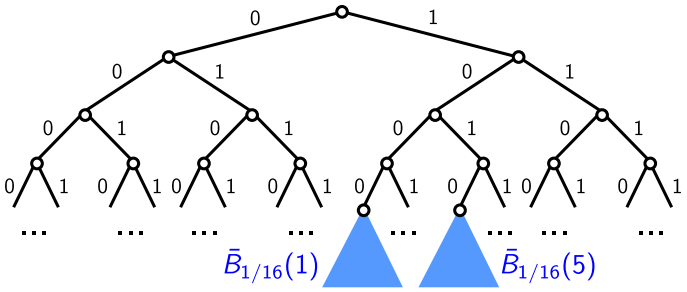

图 4：一个 2nd 阶分类器的示例，其中 $f(x) = \frac{1}{64}(x-1)(x-5)$。正区域 $\bar{\ }$ 是 $B_{\frac{1}{16}}(1)$ 与 $\cup B_{\frac{1}{16}}(5)$ 的并集。

3.3 线性分类器：多维输入

现在令 $d \geq 1$。输入空间现在是一个 d 维向量空间，其上的域为 $\mathbb{Q}_p$，记作 $\mathbb{Q}_p^d$。我们将 §3.1 中的框架扩展到这一情形，方法如下。[8]

我们假设 $\hat{y}(\cdot;\, w, b)$ 按照式 (5) 定义，其中 $f(x) = w^\top x + b$，且 $w \in \mathbb{Q}_p^d$、$b \in \mathbb{Q}_p$ 和 $w^\top x := \sum_{i=1}^d w_i x_i$。令 $\mathcal{H}_{\mathbb{Q}_p^d} := \{\pm\hat{y}(\cdot; w, b) : w \in \mathbb{Q}_p^d, b \in \mathbb{Q}_p\}$ 为 $\mathbb{Q}_p$-线性分类器的假设类，并令 $\mathcal{H}_{\mathbb{R}^d} := \{\hat{y}_{\mathbb{R}}(\cdot;\, w, b) : w \in \mathbb{R}^d, b \in \mathbb{R}\}$ 为$\mathbb{R}$-线性分类器的假设类，其中 $\hat{y}_{\mathbb{R}}(x;\, w, b) = \text{sign}(w^\top x + b)$。

接下来的结果在附录 D.3中得到证明。该结果表明，一些实数线性分类器无法解决的问题——例如异或或奇偶问题；Minsky 和 Papert(1988) 已证明这些问题无法由“有限阶感知机”解决；又如需要“二阶感知机”才能解决的计数问题——对于 p-进线性分类器却很容易解决。反过来，计数阈值问题对实数线性分类器来说微不足道，但无法由 p-进线性分类器直接解决。

命题 4（$\mathbb{Q}_p$-线性分类器的性质）。

1. 对于 $d = 2$ 和任意素数 p，$\mathcal{H}_{\mathbb{Q}_p^d}$ 中的分类器可以计算任意布尔函数。这包括异或（$\mathcal{H}_{\mathbb{R}^d}$ 无法解决异或）。

2. 对于 $d \geq 2$ 和任意素数 p，$\mathcal{H}_{\mathbb{Q}_p^d}$ 中的分类器可以解决模 p^n 同余问题（包括奇偶校验）和计数问题——这些问题是 $\mathcal{H}_{\mathbb{R}^d}$ 中的分类器无法解决的——但它们无法解决计数阈值问题，而该问题可由 $\mathbb{R}$-线性分类器解决。

最后，附录D.4中证明的下一个命题，将我们之前关于一维情形的结果（命题2）推广了。

命题 5。假设 $\hat{y}(\cdot;\ w, b) \in \mathcal{H}_{\mathbb{Q}_p^d}$ 对 $\mathcal{D}_+ := \{x^{(1)}, \ldots, x^{(m)}\} \subseteq \mathbb{Q}_p^d$ 中的所有点都能正确分类，并将此分类器对应的“正区域”记为 $R(w, b) := \{x \in \mathbb{Q}_p^d :\ |w^\top x + b|_p \leq 1\}$。则：

[8]扩展第 §3.1 节的另一种值得考虑的方法，是通过在 $\mathbb{Q}_p^d$中定义球来构造一个类似于（$\mathbb{Q}_p^d$）的球状决策规则，而这可以通过在 $\|x\|_p$ 上定义范数来实现。一种颇具吸引力的选择是上确界范数，$=$ 最大值$_{1\leq i\leq d} |x_i|_p$，它赋予 $\mathbb{Q}_p^d$ 超度量空间的结构，同时保持与 $\mathbb{Q}_p$相同的非阿基米德特性。不幸的是，采用规则 $\|x-a\|_p =$ 最大值 $_i |x_i - a_i|_p \leq r$ 的分类器似乎相当受限——它们要求对每个 $x_i \in \bar{B}_r(a_i)$ 都满足 i，因此类似于将二元分类器的合取应用于每个特征，只是所有分类器都被限制为具有相同的半径。因此，我们选择采用本节中基于线性和的构造。

1. 对任意 $i \in [m]$，参数 (w, b') 满足 $b' = -w^\top x^{(i)}$ 时定义的是同一个分类器，即有 $R(w, b) = R(w, b')$。换言之，对于任意选择的 $i \in [m]$，$R(w, b) = \{x \in \mathbb{Q}_p^d : \ |w^\top (x - x^{(i)})|_p \le 1\}$。

2. 如果分类器在 $\mathcal{D}_+$ 中是“紧的”（这意味着 $\exists j \in [m]$ 使得 $|w^\top x^{(j)} + b|_p = 1$），那么就存在一个参数化 $(w^\star, b^\star) \in \mathbb{Q}_p^d \times \mathbb{Q}_p$，满足 $w^{\star\top} x^{(i)} + b^\star = 0$ 且 $w^{\star\top} x^{(j)} + b^\star = 1$，其中某些 i，$j \in [m]$ 使得 $R(w, b) = R(w^\star, b^\star)$，即 $\hat{}y(\cdot; w^\star, b^\star) = \hat{}y(\cdot; w, b)$。

命题5 中的第二项要求分类器满足紧性这一条件。不过请注意，这实际上并非额外条件——对于任何能正确分类 $\mathcal{D}_+$ 中所有点的分类器，我们总可以将 w 和 b 乘以某个标量 λ，其中 $|\lambda|_p \ge 1$，从而使分类器变得紧。还要注意，这一性质与实数 $d = 1$的任何类似性质都不相仿。

3.4 学习一个 p进制线性分类器

上一节阐明了一个 d 维 $\mathbb{Q}_p$ 进制线性分类器可以和不可以学习什么。现在，我们将介绍一种近似算法，用于从数据中学习分类器。在接下来的两条注释中，我们针对 $x \in \mathbb{Q}_p^d$ 定义 p 进制范数 $\|x\|_p := \max_{i \in [d]} |x_i|_p$。这是向量空间 $\mathbb{Q}_p^d$ 中的一个有效范数；定义 $d(x, z) := \|x - z\|_p$ 后，它将 $\mathbb{Q}_p^d$ 变为超度量空间。

注释 1（化归为整数）。设 $\mathcal{D} = \mathcal{D}_+ \cup \mathcal{D}_- = \{x^{(1)}, \ldots, x^{(n)}\}$ 为训练输入。不失一般性，我们可以假设所有输入都是 p 进制整数——$x^{(j)} \in \mathbb{Z}_p^d$ 对所有 $j \in [n]$ 均成立。为了解释原因，假设 $(w, b) \in \mathbb{Q}_p^{d+1}$；令 i 对应于 $\|x^{(i)}\|_p$ 最大的那个示例，并令 p^m 表示这个量。接着，对所有 $j \in [n]$，定义 $x^{(j)\prime} := p^m x^{(j)} \in \mathbb{Z}_p^d$，并令 $w' := p^{-m} w$。我们有 $|w'^\top x^{(j)\prime} + b|_p = |w^\top x^{(j)} + b|_p$。因此，我们得到一个等价的分类问题，其中所有数据都由 p 进制整数构成，且权重向量 w 已缩放。

注 2 (整数权重)。还有一种与 (5) 等价的表述，可确保 $(w, b) \in \mathbb{Z}_p^{d+1}$，从而可以将学习简化为在 p-进整数上进行搜索：求出 $(w', b') \in \mathbb{Z}_p^{d+1}$ 和 $\ell \in \mathbb{Z}$，使得 $|w'^\top x_+ + b'|_p \le p^{-\ell}$ 对所有正例都成立，x_+ 并且 $|w'^\top x_- + b'|_p > p^{-\ell}$ 对所有负例都成立 x_-。那么，$w = p^{-\ell} w'$ 和 $b = p^{-\ell} b'$将构成一个分离分类器。

不失一般性，我们假设输入是 $(d + 1)$ 维的，并且具有常数特征 $x_{d+1} = 1$，因此可以省略偏置参数。算法 1 展示了一种用于训练线性 p-进分类器的简单束搜索算法。该算法假设所有输入 $x^{(i)}$ 都属于 $\mathbb{Z}_p^{d+1}$，这可以通过注 1 中的预处理来确保。我们将正类错误数和负类错误数分别记为 $\varepsilon_+ := |\{x^{(i)} \in \mathcal{D}_+ : \hat{}y(x^{(i)}; w) = -1\}|$和 $\varepsilon_- := |\{x^{(i)} \in \mathcal{D}_- : \hat{}y(x^{(i)}; w) = +1\}|$。搜索树中深度为 δ 的每个节点 n 都关联一个权重向量 $n.w \in \mathbb{Z}_p^{d+1}$，其中只有 δ 个最高有效位起作用。对于任意后代节点，正类错误数 ε_+ 只能增加，而负类错误数 ε_- 只能减少。因此，ε_+ 是任意后代节点所能达到的错误数下界。算法 1 的最坏情况运行时间复杂度为 $\mathcal{O}(k\delta_{\max} n p^{d+1})$，该复杂度关于数据集大小呈线性增长，但会随特征数 d 指数级快速增长。

在 §5中，我们用此算法来学习一个 $\mathbb{Q}_p$线性分类器，以解决逻辑推理问题。

算法 1 使用束搜索训练 p-进分类器。

输入： 数据集 $\mathcal{D} := \mathcal{D}_+ \cup \mathcal{D}_- \subseteq \mathbb{Z}_p^{d+1}$，束宽 k，最大错误数 $\varepsilon_{\max}$，最大深度 $\delta_{\max}$
$\delta := 0$ ▷ 当前深度
创建节点 n，使用 $n.w := 0$ ▷ 根节点
$\mathcal{A} := \{n\}$ ▷ 波束中的活跃节点
重复
  $\mathcal{N} := \varnothing$ ▷ 可容许节点
  对于 $n \in \mathcal{A}$**执行**
    对于 $a \in \{0, \ldots, p-1\}^{d+1}$ **执行** ▷ 扩展节点
      $w := n.w + ap^{\delta}$
      ▷ 统计正类和负类错误
      $\varepsilon_+, \varepsilon_- :=$ 求值$(w/p^{\delta+1}, \mathcal{D})$
      如果 $\varepsilon_+ \le \varepsilon_{\max}$ **则**
        创建节点 c，包含 $c.w := w$
        $\mathcal{N} := \mathcal{N} \cup \{c\}$
      如果结束
    循环结束
  结束循环
  $\delta := \delta + 1$
  ▷ 保留错误最少的 k 节点 $\varepsilon_+ + \varepsilon_-$
  $\mathcal{A} :=$ BEST-$k(\mathcal{N})$
  如果存在 $n \in \mathcal{A}$且有 $\le \varepsilon_{\max}$ 个错误**那么**
    返回 $w := n.w/p^{\delta}$
  结束 if
直到 $\delta = \delta_{\max}$

4 p-进回归

在线性回归中，目标是估计目标值 $y \in \mathbb{Q}_p$，其形式为$\hat{y}(x) = w^\top x + b$。鉴于 p-进数与字符串之间的关系（例8），我们可以非正式地将p-进回归视为一个序列到序列问题。给定数据集 $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^n$，我们将问题表述为寻找 w 和 b，使残差的上确界范数最小：$\max_{i\in[n]} |w^\top x^{(i)} + b - y^{(i)}|_p$。

4.1 线性回归：一维情形

首先考虑一维情形 $x \in \mathbb{Q}_p$。以下结果表明，这个问题可以高效求解。

命题 6. 假设对于任意 $i \neq j$，都有 $x^{(i)} \neq x^{(j)}$。那么，以下算法可以在 $\mathcal{O}(n^2)$时间内求得$(w^\star, b^\star) \in \arg\min_{w,b} \max_{i\in[n]} |wx^{(i)} + b - y^{(i)}|_p$：

1. 任意选择 $k \in [n]$。*2.* 在 $\mathcal{O}(n^2)$ 时间内求出 $j^\star \in \arg\min_{j\neq k} \max_{i\neq k} |x^{(i)} - x^{(k)}|_p \left| \frac{y^{(j)}-y^{(k)}}{x^{(j)}-x^{(k)}} - \frac{y^{(i)}-y^{(k)}}{x^{(i)}-x^{(k)}} \right|_p$。*3.* 设定 $w^\star = \frac{y^{(j^\star)}-y^{(k)}}{x^{(j^\star)}-x^{(k)}}$ 和 $b^\star = y^{(k)} - w^\star x^{(k)}$。（注意，我们还有 $b^\star = y^{(j)} - w^\star x^{(j)}$。）

证明见附录 D.5。注意，上述解（$w^\star$，$b^\star$）满足 $w^\star x^{(j)} + b = y^{(j)}$ 和 $w^\star x^{(k)} + b = y^{(k)}$，也就是说，**存在一个经过这些点的最优解**

j **以及** k。这与实数上的线性回归截然不同，是 p-进数的超度量性质所致。接下来我们将看到，这一性质在多维情形（$d \geq 1$）中也成立。

4.2 线性回归：多维情形

现在假设 $d \geq 1$，并考虑超定情形 $n \geq d+1$。与 §3.4中一样，我们可以通过附加一个常数特征 $x_{d+1}^{(i)} = 1$来消去偏置参数。定义设计矩阵$X \in \mathbb{Q}_p^{n\times(d+1)}$后，上确界范数可写为 $\|Xw - y\|_p$，我们希望对 $w \in \mathbb{Q}_p^{d+1}$进行最小化。

附录D.6中证明的下一个结果表明，在满足适当的可逆性条件时，[9] 存在一个基于样例的最优解，与 §4.1中的情形一样。

命题 7。 设 $X \in \mathbb{Q}_p^{n\times(d+1)}$为设计矩阵，且满足 $n \geq d+1$；并假设其所有$(d+1)$阶乘$(d+1)$ 子矩阵均可逆。还假设偏置扩增之前的原始设计矩阵$\tilde{X} \in \mathbb{Q}_p^{n\times d}$（即 $\tilde{X} := X_{1:n,1:d}$）的所有 $d\times d$ 子矩阵也均可逆。则存在一个最优解 $w^\star \in \operatorname{argmin}_w \|Xw - y\|_p$，它经过 $d+1$个点。

需要注意的是，命题的条件禁止我们（除其他情况外）将 $\tilde{x}^{(i)} = 0$ 作为输入，因为这会导致出现不可逆子矩阵。命题7 的一个推论是，在上述假设下，我们可以通过从 $w^\star$ 个训练样例中选出所有可能的 $d+1$ 种组合，找到最优的 n 权重：对每种组合，求解一个线性 p-进制系统（由于我们假设$(d+1)\times(d+1)$ 的所有子矩阵 X 均可逆元，因此保证有解），然后选出最优权重 $w^\star$ 使上确界范数最小的组合。该算法的运行时间为 $\mathcal{O}(n^{d+1})$。

5 p-进表示

现在我们来考察 $\mathbb{Q}_p^d$中的表示（“嵌入”）。神经网络之所以如此有效，一个原因在于它们能够学习内部表示，将训练样例映射到 $\mathbb{R}^d$中的点。p-进空间中会发生什么？

先考虑一维情形 $\mathbb{Q}_p$（只有一个“嵌入维度”），并思考如何将 n 输入示例 $\{x^{(i)}\}_{i=1}^n$ 嵌入这个空间。由于 $\mathbb{Q}_p$ 具有层次结构（图2），我们可以为每个 $x^{(i)}$ 关联一个最大的球 $B_{r_i}(x^{(i)})$，其中包含该示例，但不包含其他任何示例。不同的球按层次排列，可以表示为一棵有限树，其叶节点对应每个 n 输入示例。举例来说，如果 $p = 2$，我们会得到一棵二叉树，其中每个球 $B_{r_i}(x^{(i)})$ 对应一个比特串——这与**布朗聚类**(Brown 等人，1992年）类似，是自然语言处理领域一种广受欢迎的表示技术。因此，我们可以将布朗聚类视为 p-进表示，与连续实向量表示形成对比。

Quillian 的语义网络。 在前面的例子中，所有概念都对应树的叶节点。我们不妨认为，树中较高位置的内部节点代表与 $\mathbb{Q}_p$ 中较大的球相关联的“更一般的概念”。这些较大的球包围着较小的球（更具体的概念），形成一种嵌套结构。以.....为例

[9] 矩阵 $A \in \mathbb{Q}_p^{d\times d}$可逆，当且仅当存在 $A^{-1} \in \mathbb{Q}_p^{d\times d}$使得 $AA^{-1} = A^{-1}A = I_d$（即恒等矩阵）。

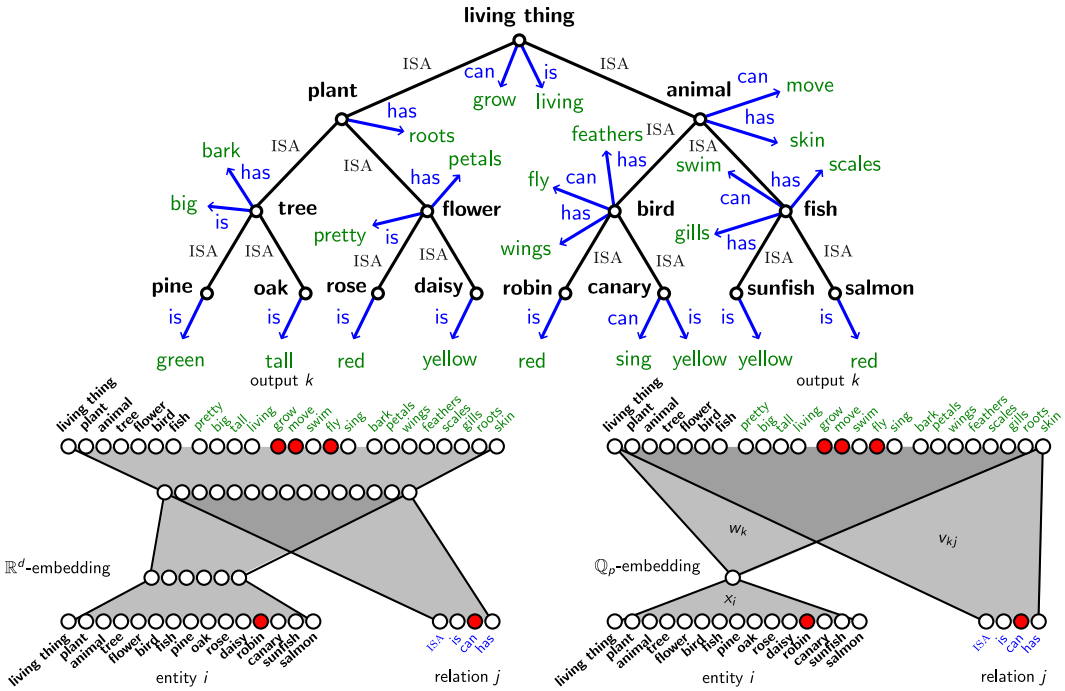

图 5： **上：** 改编自Quillian (1968) 的语义网络。 **左下：** 由Rumelhart (1990) 和McClelland等人 (1995) 开发的神经网络，利用该语义模型回答查询。给定激活输入“知更鸟可以”，网络会生成补全结果“生长 移动 飞行”。注意两个（非线性）隐藏层，其中第一个将输入实体嵌入到 $\mathbb{R}^6$中。 **右下：** 一个单一嵌入维度的线性 p进数网络（$\mathbb{Q}_p$），它解决了同一问题（为简化起见，排除了颜色和叶节点属性）。

图 5（上）中的语义网络对应一个简单的层次命题模型（Quillian，1968）。

Rumelhart (1990) 和 McClelland等人 (1995) 构建了一个具有两个隐藏层的简单神经网络（图 5，左下），通过将概念嵌入 $\mathbb{R}^6$，使其能够回答与该语义网络相关的查询。接下来，我们构建一个更紧致的**线性**分类器，采用 p进数表示（其中 $p = 2$），对该语义网络的所有命题进行编码（图 5，右下）。我们的 p进数分类器与Rumelhart (1990) 的网络结构相似，但只有一个隐藏层（用于编码概念的 p进数表示），而非两个隐藏层，并且不含任何非线性激活。它由线性函数复合而成，因此仍是一个线性分类器。加入隐藏表示层有三方面优势：（i）明确呈现概念的表示；（ii）降低维度；（iii）能够使用共享表示空间来执行多个任务。

我们首先展示网络中不带“是绿色/红色/黄色”属性的叶节点的表示和权重——在这种情况下，一维的 p-adic 表示空间 $\mathbb{Q}_p$ 事实证明已足够。然后，我们通过使用二维的 p-adic 表示空间（$\mathbb{Q}_p^2$）将网络扩展为包含“是绿色/红色/黄色”属性。最后，我们考虑一个新的属性“有叶子”，该属性打破了语义网络的层次结构：所有植物都有叶子除了松树（松树有针叶

表1：与图5所示语义网络相关的实体/概念的表示。不带颜色属性时，只需要一个特征维度（第 x列）。带有颜色属性时，我们需要第二个特征维度（第 x'列）。

实体	x	x（数字）	x'	x'（数字）
生物	1	1	0	0
植物	4	100	0	0
动物	6	110	0	0
tree	16	10000	0	0
花	24	11000	0	0
bird	18	10010	0	0
fish	26	11010	0	0
pine	64	1000000	1	1
oak	96	1100000	0	0
rose	72	1001000	2	10
雏菊	104	1101000	4	100
知更鸟	66	1000010	2	10
金丝雀	98	1100010	4	100
太阳鱼	74	1001010	4	100
鲑鱼	106	1101010	2	10

而不是叶节点）。这一例外可以通过使用三维 p-adic 表示空间（$\mathbb{Q}_p^3$）来处理。在这一构造中，我们始终使用 $p = 2$。

不包括颜色的属性。我们首先构建忽略颜色属性（“是绿色/红色/黄色”）的模型。通过观察图 5 （上）可以看到，除这些属性外，所有其他属性都与语义树中的单个节点相关联。我们还可以看到，每个属性只与单个关系相关联（例如，“生长”与can 相关，而不与是一个、is或has相关）。我们的策略是：（i）将实体/概念表示放置在一个p-adic 树中，使其结构与图 5 中的语义网络相同；（ii）为与树中特定节点相关联的每个属性使用一个二元线性分类器。

我们的网络形式如图 5 （右下）所示。我们有两组输入：15 个实体/概念（生物、植物、动物.....太阳鱼、鲑鱼）和 4 个关系（是一个、is、can、has）。与Rumelhart（1990）一样，我们对实体/概念和关系都使用独热编码。随后，实体/概念被输入到一个嵌入层中 $\mathbb{Q}_p$，且不带偏置参数。这实际上意味着它们获得嵌入 x_1，.....，x_{15}，其中每个$x_i \in \mathbb{Q}_p$。关系 j 直接使用独热编码。实体/概念和关系的表示会被拼接；因此，当实体 i 和关系 j 出现在输入中时，会得到一个 5 维向量

$$[x_i, \underbrace{0, \ldots, 1, \ldots, 0}_{\text{one-hot for relation } j}] \in \mathbb{Q}_p^5.$$

这些表示在所有属性间共享，每个属性 k 都关联一个具有 5 维权重 $[w_k, v_{k1}, \ldots, v_{k4}]$ 和一个偏置参数的二元分类器，我们将偏置参数设为零。输出层预测

$$\hat{y}_k(i,j) = \begin{cases} +1 & \text{if } |w_k x_i + v_{kj}|_p \leq 1 \\ -1 & \text{otherwise,} \end{cases} \tag{7}$$

¯其中（参见式（6））， 顶部条件等价于 $x_i \in B_{1/|w_k|_p}(-v_{kj}/w_k)$。通过考察语义网络的结构，可以直接得出参数 x_i、 w_k和 v_{kj} ，从而为每个属性 k构造出预期的分类器。我们首先选择与图 5（左下）中的实体/概念相容的表示 $x_i \in \mathbb{Q}_p$ 。表1中的表示满足这一要求。接着，对于任意属性 k，我们考虑与该属性关联的（唯一）关系 j 。首先选择 w_k ，使 $1/|w_k|_p$ 处于树中的正确层级，符合式（7）右侧的要求——例如，属性-关系对“ISA 鸟”应涵盖实体/概念鸟、知更鸟和金丝雀，它们的共同前缀是 $\cdots 0010$，因此我们需要 $1/|w_k|_p = 2^{-4}$，而 $w_k = 2^{-4} = 1/16$满足这一条件。然后，我们选择 v_{kj} ，使 $-\frac{v_{kj}}{w_k}$ 成为所需球的中心——例如，对于属性-关系对“ISA 鸟”，任何以 $\cdots 0010$为前缀的中心都可以。一个可行的中心是 2，由此得到 $v_{kj} = -2w_k = -2 \times 2^{-4} = -1/8$。最后，对于彼此不相容的属性-关系对，我们选择一个与所有实体/概念都不相交的球。由此得到表2所示的参数。此参数选择确保对输入中的任意实体/概念和关系都能正确分类。

添加颜色。加入颜色属性后，上述策略便不再适用，因为红色和黄色这两种颜色属性都与多个节点相关联（红色与玫瑰、知更鸟、鲑鱼相关联，黄色与雏菊、金丝雀、太阳鱼相关联）。我们可以通过增加第二个嵌入维度 x'_i 来表示“颜色”特征（表1最右侧的列）。这需要按照与颜色相关的另一种层次结构来组织实体/概念——幸运的是，多维 $\mathbb{Q}_p$线性分类器可以表示多个层次结构。现在，输入表示将是 6 维向量 $[x_i, x'_i, 0, ..., 1, ..., 0] \in \mathbb{Q}_p^6$，且每个属性 k 都对应一个具有 6 维权重 $[w_k, w'_k, v_{k1}, \ldots, v_{k4}]$ 和零偏置的二元分类器。如果属性 k 不是颜色，我们就设定 $w'_k = 0$ ，分类规则与式（7）完全相同。如果属性 k 是颜色（绿色、红色或黄色），我们就设定 $w_k = 0$，由此得到

$$\hat{y}_k(i,j) = \begin{cases} +1 & \text{if } |w'_k x'_i + v_{kj}|_p \leq 1 \\ -1 & \text{otherwise.} \end{cases}$$

同样的逻辑也适用于此处，从而得到权重 w'_k 和 v_{kj} 表2。

处理叶子。 如上所述，几乎所有植物都有叶子，除了松树，它长的是针叶，而不是叶子。我们如何处理这个例外？当然，可以沿用添加颜色时的相同思路：在表示空间中增加一个维度，并将关系/属性对“有叶子”关联到所有有叶子的实体/概念：花、橡树、玫瑰和雏菊。不过，我们将介绍另一种策略，说明如何对规则例外进行建模。首先，我们对“所有植物都有叶子”这一（错误的）规则进行编码，这很容易做到：在表 2中新增一项“有叶子”，其权重与“有根”这一项相同：（$w_k = 1/4$，$w'_k = 0$，以及 $v_{kj} = 0$）。现在我们需要为这条规则设定例外：由于松树没有叶子，我们不能再说它的任何祖先都有叶子；因此，必须为松树、树和植物这些实体/概念设定例外。这可以通过在表示空间中增加第三个维度来实现，记作 x''_i。对于既不是松树、树也不是植物的实体/概念，该维度的值为 0；对于这些实体/概念中的任意一个，该维度的值为 1。接下来，我们选择一个权重 w''_k，用于乘以这个新特征。除“叶子”外，该权重在所有属性上的值均为 0——这不会影响例外列表之外的实体，但会改变判定

表2：每个关系 j 和属性 k 对应的权重为 w_k 和 v_{kj}。每个关系/属性对都对应一个半径为 $1/|w_k|$、中心为 $-v_{kj}/w_k$ 的 2进球。例如，“是一个 鸟”对应球 $B_{2^{-4}}(2)$，它对应于 p-进数位展开

$\cdots 0010$，其中包含鸟、知更鸟和金丝雀的表示（见表1）。属性 k与关系 j (不兼容（例如“ISA 生长”）的权重为 $v_{kj} = -\frac{w_k}{2}$（表中未显示）。因此，它们对应于中心为$1\ 2$、半径为 $1/|w_k|$的球；由于 $1/|w_k| \leq 1$ 对于表中的所有属性 k，这些球都只包含表$\mathbb{Z}_2$ 外的实体，因此不包含表1中的任何实体。颜色属性以蓝色标示，对于这些属性，我们有 $w_k = 0$，显示的值为 w'_k。对于非颜色属性，情况正好相反，$w'_k = 0$显示的值为 w_k。

关系 j	属性 k	w_k （或w'_k）	v_{kj}
ISA	生物	1	0
ISA	植物	$1/4$	0
ISA	动物	$1/4$	$-1/2$
ISA	tree	$1/16$	0
ISA	花	$1/16$	$-1/2$
ISA	bird	$1/16$	$-1/8$
ISA	fish	$1/16$	$-5/8$
is	漂亮	$1/16$	$-1/2$
is	big	$1/16$	0
is	生活	1	0
is	绿色	$1/2$	$-1/2$
is	red	$1/4$	$-1/2$
is	黄色	$1/8$	$-1/2$
is	tall	$1/64$	$-1/2$
can	grow	1	0
can	move	$1/4$	$-1/2$
can	swim	$1/16$	$-5/8$
can	fly	$1/16$	$-1/8$
can	sing	$1/64$	$-17/32$
has	bark	$1/16$	0
has	花瓣	$1/16$	$-1/2$
has	翅膀	$1/16$	$-1/8$
has	羽毛	$1/16$	$-1/8$
has	鳞片	$1/16$	$-5/8$
has	鳃	$1/16$	$-5/8$
has	根	$1/4$	0
has	skin	$1/4$	$-1/2$

松树、树和植物的规则

$$\hat{y}_k(i,j) = \begin{cases} +1 & \text{if } |w_k x_i + w''_k x''_i + v_{kj}|_p \leq 1 \\ -1 & \text{otherwise,} \end{cases}$$

当 j，k 对应“有叶子”时。由于按照设计，所有这些实体 i （松树、树和植物）都满足 $w_k x_i + v_{kj} \in \mathbb{Z}_p$，因此任何 $w''_k \notin \mathbb{Z}_p$ （例如，$w''_k = 1/2$）都会产生所需的例外。

实验。 我们使用上述语义网络运行一个简单的合成实验，具体如下。我们生成结合 15 个实体、4 个关系、

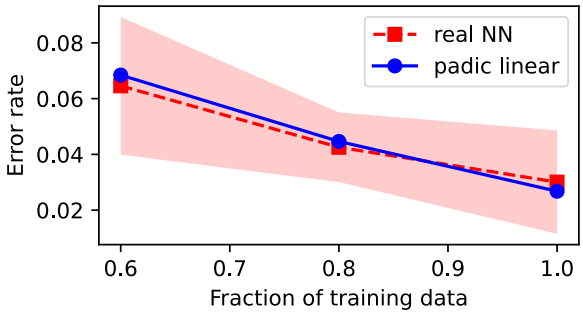

图 6：使用学习得到的神经网络复现 Rumelhart（1990）以及一个线性 p-adic 网络，其嵌入设置如表 1 所示，且其权重 $\{w_k\}$ 和 $\{v_{kj}\}$通过算法 1进行学习。图中显示了使用 {60%, 80%, 100%} 的训练数据时的测试错误率。

以及图 5中所示的 28 个输出的 1,680个命题。我们将这一完整数据集随机划分为 80% 的训练命题和 20% 的测试命题。我们进一步对训练集进行抽样，得到包含 60% 和 80% 示例的缩减版本。我们按照 Rumelhart(1990) 的方法实现了一个具有 6 维嵌入和 15 维隐藏层的神经网络（690 $\mathbb{R}$ 个参数），以及一个具有 3 维嵌入的线性 p-adic 网络（“padic linear”）（241 $\mathbb{Q}_p$个参数），其参数按照表 1中的方式选取；两者均如图 5所示。我们使用梯度反向传播和 Adam 优化器（学习率为 0.1）训练神经网络，并使用算法 1训练线性p-adic 分类器。对于前者，我们使用不同的随机初始化运行 10 次试验；对于后者，由于其具有确定性，因此只运行 1 次。两个模型都成功对完整训练集进行了过拟合（所有试验均如此），错误率为零。图 6 展示了采用不同训练集规模进行训练集/测试集划分的结果。我们观察到，两个模型在测试集上的泛化能力相似。

6 开放问题

本文仅通过建立基本理论结果，初步探讨了如何在 p空间中进行机器学习。要确定这一框架是否具有实际应用价值，还有许多挑战有待克服。下面我总结了一些开放问题，并提出研究这些问题的可能途径。

6.1 线性多类 p分类器

在 §3 中，我们只讨论了单类和二分类问题。在多类问题中，可以有 $K \geq 2$ 个类别。虽然任何多类问题都可以化归为二分类问题的组合（例如，通过一对其余或一对一方案），但尝试为 p-adic 多类分类建立一个原生框架仍然很有意义。

我们可以为每个输出类别 $k \in [K]$ 定义一个权重向量 $w_k \in \mathbb{Q}_p^d$ 和偏置 $b_k \in \mathbb{Q}_p$，并采用如下决策规则

$$\hat{y}(x) = \arg\min_{k \in [K]} |w_k^\top x + b_k|_p.$$

这种方法的一个潜在缺点是，由于 p-进范数具有离散性，出现平局的可能性很高。让我们考虑 $d = 1$ 的一维情形，并假设

$|w_k|_p = 1$ 对所有 k 都成立，以便简化问题。在这种情况下，通过定义“类别质心” $a_k := -b_k/w_k$，问题就转化为：为给定的 x 找到最近的质心。与每个类别相关联的“沃罗诺伊单元”按层级排列成树状图；由于该空间是超度量空间，因此可能出现平局。请注意，并非所有平局都允许：只有当任何与这两个类别中的任意一个更相似的其他类别也处于平局中时，这两个类别之间才可能出现平局。

6.2 基于梯度的优化与 p进制域上的根求解

虽然我们为一维情形提供了高效的分类和回归算法，但对于 $d > 1$ 此处介绍的学习算法（用于1 p进制分类的算法，以及第§4.2 节中针对 p进制回归概述的算法），应仅将其视为概念验证，因为它们并不实用——相对于特征数 d而言，其运行时间呈指数增长。寻找更好的算法是一个开放问题， p进制预测器能否用于任何实际目的，在很大程度上取决于此。

一个开放问题是， p进制微积分的工具能否带来更好的学习算法——事实上，实分析中的许多工具同样适用于 p进制数，包括形式幂级数、连续性和微分（Koblitz，1984年；Robert，2000年）。然而，这需要克服几个障碍——虽然牛顿算法可以在 p进制域中，基于形式导数求出 p进制函数的根，但 p进制微积分与实微积分有根本区别：例如，一个处处导数为零的函数在 $\mathbb{Q}_p$中可能并非常数。此外，机器学习中通过优化损失函数，使模型拟合训练数据的经典思想，在 p进制域中无法直接奏效——由于 $\mathbb{Q}_p$ 不是有序域，“优化”一个取 p 进制值的损失函数没有意义。在第4节的回归表述中，我们通过优化 p进制范数绕过了这一问题，而该范数取值于 $\mathbb{Q}$。许多有趣的 p进制学习问题很可能都可以数学化地表述为 p进制最小范数问题，这一主题值得进一步研究。

不过，我们应当注意， p进分析的基石之一是亨泽尔引理。它类似于用于求根的牛顿法，并具有很强的收敛性质。这表明，学习 p进预测器的一种可行途径，是巧妙地设计一个函数 F，使其导出一个 p进方程 $F(\theta; \mathcal{D}) = 0$；其中 $\theta \in \mathbb{Q}_p^d$ 是模型参数， $\mathcal{D}$ 是训练数据，并且该方程的解在某种意义上对应于一个“最优”的模型配置。对于实数域上的机器学习模型，人们会针对可微损失函数 L选择 $F(\theta; \mathcal{D}) = \nabla L(\theta; \mathcal{D})$；但在 p进情形下，直接使用 F、绕过损失函数会更方便。通过将亨泽尔提升引理（Gouvêa，2020年，§4）推广到 $\mathbb{Q}_p^d$，再利用 p 进牛顿算法求解F，很可能会得到效率极高的算法。

6.3 多层 p-进网络

本文采用了线性预测器（第 §3.2 节讨论的高阶分类器除外）。然而，在实数上的机器学习中，多层网络的表达能力要强得多（Hornik 等人，1989）。我们自然会预期，“p-adic 神经网络”也会如此。例如，虽然我们已在命题 4 中证明，线性 p-adic 分类器无法解决计数阈值问题，但很容易构造一个单隐层 p-adic 神经网络来解决这类问题：我们只需先构建第一层，由 $d-k+1$ p-adic 分类器组成，分别解决关于 精确 计数的问题，其中 $i = k$、 $k+1$、……、 d——使用附录 D.3 中给出的构造——然后注意到

这些 $d-k+1$ 个分类器中至多有一个会返回 $+1$（其余分类器必须返回 -1），接着添加一个顶层，其中包含一个权重为 $w_1 = ... = w_{d-k+1} = p^{-n}$、偏置为$b = (d-k-1)\,p^{-n}$的精确一分类器，且 $n = \lceil \log_p(d-k+1) \rceil$；所得的多层系统在至少有 k 个输入为“真”时返回 $+1$，否则返回 -1。

要在这个研究问题上取得进展，我们需要寻找合适的非线性函数和学习准则。

6.4 使用所有素数：阿代尔预测器

我们的整篇论文都假设固定一个素数 p，并由此构造一个非阿基米德范数 $|\cdot|_p$。然而，应该选择哪个 p 呢？我们能否构建使用多个素数，甚至同时使用所有 素数的预测器?

上述问题与所谓的黑塞局部-整体原理（Gouvêa，2020年，§4.8）有相似之处；这一原理是数论的基石。该原理的核心思想是，尝试通过同时考察所有“局部”完备化来回答 $\mathbb{Q}$ 中的一个“整体”复杂问题（例如，寻找丢番图方程的有理数解）；也就是说，对每个素数 p，考察 $\mathbb{Q}_p$，以及 $\mathbb{R}$。事实上，根据奥斯特罗夫斯基定理（Gouvêa，2020年，定理 3.1.4），有理数域 $\mathbb{Q}$ 的任何完备化都具有以下形式之一：有理数域$\mathbb{Q}$ 上的每个非平凡绝对值（就诱导相同拓扑而言）都等价于某个非阿基米德p-进绝对值 $|\cdot|_p$（例 3），或通常的阿基米德绝对值 $|\cdot|_\infty$（例 2）。数学家常用 $p=\infty$ （“无穷远处的素数”）来为实数集编索引，记作 $\mathbb{Q}_\infty = \mathbb{R}$。因此，我们可以将 $\mathbb{Q}$ 的所有完备化表示为集合$\{\mathbb{Q}_p : \ p \le \infty\}$。局部-整体结果的一个例子是乘积公式（Gouvêa，2020年，命题 3.1.5），它指出，对于任意 $x \in \mathbb{Q} \setminus \{0\}$,

$$\prod_{p \le \infty} |x|_p = 1. \tag{8}$$

（利用算术基本定理很容易证明这个公式。）

接下来，我将勾勒一条受此原则启发的**阿代尔分类**路径。让我们来设想一个由所有素数 p 对应的 p-进分类器组成的“集成”，其中还包括一个“实数分类器”，记作 $p=\infty$，其决策规则——给定一个输入 $x \in \mathbb{Q}^d$——为

$$\hat{y}(x) := \begin{cases} +1, & \text{if } \prod_{p\le\infty} |w_{(p)}^\top x + b_{(p)}|_p \le 1 \\ -1, & \text{otherwise.} \end{cases} \tag{9}$$

我们可以将此决策规则理解为让所有 p-adic 分类器（对于每个 p）通过其对应的绝对值进行投票，使用其特定参数 $w_{(p)} \in \mathbb{Q}_p^d$ 与 $b_{(p)} \in \mathbb{Q}_p$，然后汇总所有投票以产生最终决策。根据乘积公式（8），我们有：如果 $w_{(p)} = w \in \mathbb{Q}^d$ 与 $b_{(p)} = b \in \mathbb{Q}$ （即，如果对于所有 p 而言参数都相同）并且 $w^\top x + b \neq 0$，则有 $\prod_{p\le\infty} |w_{(p)}^\top x + b_{(p)}|_p = \prod_{p\le\infty} |w^\top x + b|_p = 1$，即，所有预测都位于决策边界上。另一方面，如果我们假设只有 $\bar{}$ $\bar{}$有限多个 p 集合中的 $\mathcal{P}$ 参与投票，而其余所有 $p \notin \mathcal{P}$ 都具有 $w_{(p)} = 0$和 $b_{(p)} = 1$，那么我们可以设计算法，逐步添加新的素数 p 以改善模型-数据拟合，直到满足某个停止准则。

为了对分类器（9）进行参数化，我们需要定义 $w = (w_{(p)})_{p\le\infty}$ 和 $b = (b_{(p)})_{p\le\infty}$——这些对象称为**阿代尔**(Goldfeld and Hundley，2011）。形式上，阿代尔环$\mathbb{A}_\mathbb{Q}$ 定义为（按元素进行加法和乘法）

$$\mathbb{A}_\mathbb{Q} := \prod_{p\le\infty}{}' \mathbb{Q}_p = \left\{ z \in \prod_{p\le\infty} \mathbb{Q}_p \ : \ |z_{(p)}|_p \le 1 \text{ for all but finitely many } p \right\},$$

其中 $\prod'$ 表示受限直积（而非笛卡尔积），它要求 $z = (z_{(p)})_{p\leq\infty} \in \mathbb{A}_{\mathbb{Q}}$ 中除有限多个以外的所有项都是 p-阿代尔整数。请注意，对于任意$z \in \mathbb{A}_{\mathbb{Q}}$，受限直积都能确保 $[z]:=\prod_{p\leq\infty}|z|_p$ 收敛，从而使其能够用于式（9）。还要注意，任意有理数 q 都可以视为阿代尔 $z=(q,q,...)$，其中所有项均为常数，即有一个嵌入 $\mathbb{Q}\hookrightarrow\mathbb{A}_{\mathbb{Q}}$。根据乘积公式（8），与这些有理数对应的阿代尔具有 $[z]=1$。此外，任意 $a\in\mathbb{Q}_p$也可以视为阿代尔 z，其中 $z_{(p)}=a$ 且 $z_{(p')}=1$ 适用于 $p'\neq p$，对此我们有$[z]=|z|_p$（对于 $p=\infty$，即实数，这也成立）。由于阿代尔构成一个环，对于每个 $x\in\mathbb{Q}^d$——甚至每个 $x\in\mathbb{A}_{\mathbb{Q}}^d$——$w\in\mathbb{A}_{\mathbb{Q}}^d$ 和 $b\in\mathbb{A}_{\mathbb{Q}}$ 都确保 $z=w^\top x+b\in\mathbb{A}_{\mathbb{Q}}$，因此我们可以将式（9）写为：若 $[w^\top x+b]\leq 1$，则为$\hat{}y=+1$；否则为$\hat{}y=-1$。请注意，§3 中定义的p-阿代尔分类器是此构造的一个特例，只需令$w_{(p)}=0$，并对除某个特定 p之外的所有情况取 $b_{(p)}=1$ 即可。我们还可以通过考察残差 $[w^\top x+b-y]$来定义阿代尔回归，这也是 §4 中所述 p-阿代尔回归框架的一种泛化。

进一步发展上述概念，或许会成为未来研究的一条有趣路径。

7 结论

我们介绍了对“p-进机器学习”的探索性研究，它以$\mathbb{R}$ 取代了 $\mathbb{Q}_p$。我们建立了 p-进分类与回归的主要构成模块，包括预测规则、学习形式和学习算法。我们推导出了一些基础理论性质，其中有些性质颇为出人意料，这些性质源于 $\mathbb{Q}_p$的超度量性：例如，线性回归器是基于样例的（它们经过 $d+1$训练点）；一维线性分类器是包围 p-进球的，而具有多项式判别函数的非线性分类器则是这些球的并集。我们展示了 $\mathbb{Q}_p$ 的拓扑有助于捕捉表示空间中对象之间的层级关系，例如语义网络中出现的关系；此外，我们还为Rumelhart（1990）设计的一个网络提供了概念验证：在 $\mathbb{Q}_p$ 中存在一个具有完美准确率的紧凑线性网络，而在 $\mathbb{R}$中则不存在这种情况。

非欧几里得几何表示（如双曲嵌入）已被用于捕捉层级关系（Nickel 和 Kiela，2017年）。我们的框架与这类方法截然不同，因为后者仍以实数域作为基础。我们认为，我们的工作只是迈向 p-进机器学习的第一步，并确定了许多未来工作的潜在方向。克服这些挑战是未来工作中一个令人兴奋的方向。

参考文献

Baker, G. 和 Molla-Aliod, D.（2022年）。数论与语言学相遇：利用 2进度量对1497种语言中的名词复数化进行建模。*AACL-IJCNLP 2022*，第24页。Bradley, P. E.（2009年）。论 p-adic分类。*p*进数、超度量分析与应用，1：271–285。Brown, P. F.、Della Pietra, V. J.、Desouza, P. V.、Lai, J. C. 和 Mercer, R. L.（1992年）。自然语言的基于类别的n元语法模型。计算语言学，18(4)：467–480。Chierchia, G. 和 Perret, B.（2019年）。通过梯度下降进行超度量拟合。神经信息处理系统进展，32。

Cohen-Addad, V.、Karthik, C. 和 Lagarde, G.（2020年）。论高效的低畸变超度量嵌入。载于国际机器学习会议，第2078–2088页。PMLR。Goldfeld, D. 和 Hundley, J.（2011）。一般线性群的自守表示与L函数：第*1*卷，第129卷。剑桥大学出版社。Gouvêa, F. Q.（2020年）。*p*进数：导论。施普林格，第3版。Hensel, K.（1897年）。论代数数理论的一种新基础。德国数学家协会年报，6：83–88。Hornik, K.、Stinchcombe, M. 和 White, H.（1989年）。多层前馈网络是通用逼近器。神经网络，2(5)：359–366。Khrennikov, A. 和 Tirozzi, B.（1999年）。p进神经网络的学习。预印本。

Koblitz, N.（1984年）。*p* 进数、*p* 进分析与ζ函数。施普林格。

Kraft, L. G.（1949年）。一种用于量化、分组和编码调幅脉冲的装置。博士论文，麻省理工学院。

McClelland, J. L.、McNaughton, B. L. 和 O’Reilly, R. C.（1995年）。海马体和新皮层中为何存在互补的学习系统：从联结主义学习与记忆模型的成功与失败中获得的启示。心理学评论，102(3):419。

Minsky, M. L. 和 Papert, S. A.（1988年）。《感知机：扩展版》。麻省理工学院出版社。

Murtagh, F.（2004年）。论超度量性、数据编码与计算。分类学期刊，21:167–184。

Murtagh, F.（2009年）。从数据到 p 进数或超度量模型。*p*进数、超度量分析与应用，1：58–68。

Nickel, M. 和 Kiela, D.（2017年）。用于学习层级表征的庞加莱嵌入。神经信息处理系统进展，30。Nolan, D.（1991）。超额质量椭球。多元分析杂志，39(2)：348–371。

Quillian, R.（1968）。语义记忆。

Robert, A. M.（2000年）。*p*进分析课程，第198卷。施普林格科学与商业媒体。Rumelhart, D. E.（1990年）。脑式计算：学习与泛化。载于神经网络与电子网络导论，第405–420页。Schölkopf, B., Williamson, R. C., Smola, A., Shawe-Taylor, J. 和 Platt, J.（1999年）。用于新颖性检测的支持向量方法。神经信息处理系统进展，第12卷。

A 在超度量空间中，每个三角形都是等腰三角形

应用强三角不等式（2），并假设 $d(x,y) \neq d(y,z)$。不失一般性，我们可以假设 $d(x,y) > d(y,z)$。由（2）可得 $d(x,z) \leq \max\{d(x,y)\ d(y,z)\} = d(x,y)$，但再次应用（2）可得 $d(x,y) \leq \max\{d(x,z)\ d(y,z)\}$，由于我们假设 $d(x,y) > d(y,z)$，因此必有 $d(x,y) \leq d(x,z)$。因此，结合这两个不等式，必有 $d(x,z) = d(x,y) = \max\{d(x,y)\ d(y,z)\}$。

B 关于 p-进球的其他性质

在接下来的例子中，对于任意集合 A 和 B（它们是某个加法群的子集），我们定义集合 $a+B := \{a+b:\ b \in B\}$, $cB := \{cb:\ b \in B\}$，以及（明可夫斯基）集合和$A+B := \{a+b:\ a \in A,\ b \in B\}$。注意，后两个运算都关于并具有分配性，即，$A+(B \cup C) = (A+B) \cup (A+C)$ 和 $c(A \cup B) = cA \cup cB$。

例 10. 中的任意一个 $\mathbb{Q}_p$ 非零半径的球都可以写成 $B_{p^n}(a) = a + p^{-n}\mathbb{Z}_p$。特别是，分解 (4) 也可以等价地写成自指形式

$$\mathbb{Z}_p = \bigcup_{i=0}^{p-1}(i + p\mathbb{Z}_p).$$

这表明 $\mathbb{Z}_p$ 具有自相似结构。

命题 8. p-进球在闵可夫斯基和运算下封闭，并且我们有

$$\bar{B}_r(a) + \bar{B}_s(b) = \bar{B}_{\max\{r,s\}}(a+b).$$

带有闵可夫斯基和运算的 p-进球构成一个交换幺半群，其中 $\{0\} = B_0(0)$ 是恒等矩阵。此外，p-进球在标量乘法下也封闭，并且对于任意$c \in \mathbb{Q}_p$，我们有，

$$c\bar{B}_r(a) = \bar{B}_{r|c|_p}(ca).$$

证明。首先注意到 $a+B_s(b) = B_0(a)+B_s(b) = \{a+y:\ |y-b|_p \leq s\} = \{u:\ |u-a-b|_p \leq s\} = B_s(a+b)$。现在，假设（不失一般性）$r \geq s$。由于具有相同中心的球必然嵌套，因此我们有 $y \in B_s(b) \implies y \in B_r(b)$，所以对于任意$a \in \mathbb{Q}_p$, $|y-b|_p = |y-b+a-a|_p \leq r$，即 $a+b \in B_r(a+y)$。由于球中的任意点都是该球的中心，因此我们必有 $B_r(a+y) = B_r(a+b)$。因此，我们有 $B_r(a)+B_s(b) = \bigcup_{y\in\bar{B}_s(b)}(B_r(a)+y) = \bigcup_{y\in\bar{B}_s(b)} B_r(a+y) = \bigcup_{y\in\bar{B}_s(b)} B_r(a+b) = B_r(a+b)$。

关于第二部分，请注意，如果 $c=0$，则该命题变为 $\{0\} = B_r(0)$，这是正确的。现在假设 $c \neq 0$，我们有 $c\bar{B}_r(a) = \{cx:\ |x-a|_p \leq r\} = \{u:\ |\frac{u}{c}-a|_p \leq r\} = \{u: |u-ca|_p \leq r|c|_p\} = B_{r|c|_p}(ca)$。 □

于是我们得到以下推论，它推广了（4）。

推论 1。 $\mathbb{Q}_p$ 中的任意球都满足：

$$\bar{B}_{p^n}(a) = \bigcup_{i=0}^{p-1} \bar{B}_{p^{n-1}}(a + p^{-n}i),$$

其中各并集互不相交，同时 $\bar{B}_{p^{n-1}}(a) = p\bar{B}_{p^n}(ap^{-1})$。

证明。 利用 p-进整数球的分解、上述命题，以及闵可夫斯基和与标量乘法对并集的分配性质，我们有 $\bar{B}_{p^n}(a + \bigcup_{i=1}^{p-1} p^{-n}\bar{B}_{p^{-1}}\ i\ = \bigcup_{i=1}^{p-1} \bar{B}_{p^{n-1}}\ a + p^{-n}i\ )$

$= a + p^{-n}\mathbb{Z}_p = a + p^{-n}\bigcup_{i=1}^{p-1}(i + p\mathbb{Z}_p) = a + p^{-n}\bigcup_{i=1}^{p-1}\bar{B}_{p^{-1}}(i)\ (\ )\ (\ )$。显然，这个并集互不相交。第二个结论可由上述命题轻易证明。

□

C p进赋值证明 Kraft不等式

示例中所表达的联系 8 之间 p进赋值球与字符串之间的联系，可用于得出**Kraft不等式**（Kraft，1949），这是信息论中的一项重要结果，它给出了一个码成为前缀码（因而可唯一解码）的充要条件。

命题 9（Kraft不等式）。设字母表中每个信源符号都被编码为字母表大小为 m 的字母表上的一个前缀码，其码字长度为 ℓ_1、ℓ_2、……、ℓ_n。则$\sum_{i=1}^{n} m^{-\ell_i} \leq 1$。反之，对于满足上述不等式的一组自然数 ℓ_1、ℓ_2、……、ℓ_n，存在一个字母表大小为 m的字母表上的前缀码，其码字长度即为这些数。

我们基于p进赋值球的性质，为 $m = p$为素数的情形给出一个简单证明。

我们从以下引理开始。

引理 1。 设 $\{B_{r_i}(a_i)\}_{i=1}^{n}$ 是有理数域 $\mathbb{Q}_p$ 中一个由有限个两两不交的球组成的集合，即满足 $B_{r_i}(a_i) \cap B_{r_j}(a_j) = \varnothing$ 对 $i \neq j$ 均成立。假设 $r_i = p^{n_i}$ 对 $n_i \in \mathbb{Z}$ 均成立。设 $B_r(a)$ 是另一个球，它包含所有球 $\bar{B}_{r_i}(a_i)$，其中 $r = p^n$。那么必有 $\sum_{i=1}^{n} r_i \leq r$，且当且仅当 $\bar{B}_r(a) = \bigcup_{i=1}^{n} \bar{B}_{r_i}(a_i)$ 时等号成立。

证明。 根据推论 1，我们知道 $B_{p^n}(a)$ 可以分解为 p个较小的球的不交并，每个球的半径为 $r_i = p^{n-1}$。于是有 $\sum_{i=0}^{p-1} p^{n-1} = p^n$。我们可以递归地继续分解其中一些较小的球，这样仍能保持等号成立。由于 $\mathbb{Q}_p$ 中的所有球必定彼此包含，或交集为空，因此，将一个球分解为较小的球只有这些可能的方式。集合 $\{B_{r_i}(a_i)\}_{i=1}^{n}$必然是这些分解之一的子集，其中可能缺少一些球，因此，总和 $\sum_{i=1}^{n} r_i$ 不超过 $r = p^n$。

□

现在，我们为命题 9 在 $m := p$ 为素数的情形给出一个简单证明，并用到我们已经了解的关于 p-进球的事实。我们先证明 $\Longrightarrow$ 方向。我们将每个字符串（码）与包含在 $\mathbb{Z}_p$中的一个球对应起来；长度为 ℓ_i 的字符串对应半径为 $p^{-\ell_i}$、中心为某个 $a_i \in \mathbb{Z}_p$ 的球，使得 ℓ_i 位 a_i 的数字与该字符串中的字符相对应。由于这是一个前缀码，与某个码字对应的球都不可能包含在与另一个不同码字对应的球中，也就是说，球集 $\{B_{p^{-\ell_i}}(a_i)\}$ 两两不相交。因此，它满足

¯引理 1 的条件，因此，由于 $\mathbb{Z}_p = B_1(0)$包含所有这些球，我们必有 $\sum_{i=1}^{n} p^{-\ell_i} \leq 1$。为了证明 $\Longleftarrow$ 方向，设 ℓ_1、.....、 ℓ_n 满足 $\sum_{i=1}^{n} m^{-\ell_i} \leq 1$，并且不失一般性地假设 $\ell_1 \leq ... \leq \ell_n$。我们可以通过以下方式构造一个前缀码：¯先选取一个与球 $B_{p^{-\ell_1}}(0)$对应的字符串（码），然后再选取另一个与半径为 $p^{-\ell_2}$ 且不包含在第一个球中的球对应的字符串（码），依此类推。

D 证明

D.1 命题证明2

要证明 1，注意到 $w^\star x^{(i)} + b^\star = x^{(i)}/(x^{(j)} - x^{(i)}) - x^{(i)}/(x^{(j)} - x^{(i)}) = 0$ 与 $w^\star x^{(j)} + b^\star = x^{(j)}/(x^{(j)} - x^{(i)}) - x^{(i)}/(x^{(j)} - x^{(i)}) = 1$。对于每个 $x_+ \in \mathcal{D}_+$，根据 $x^{(j)}$ 的定义，我们有 $|w^\star x_+ + b^\star|_p = |x_+ - x^{(i)}|_p/|x^{(j)} - x^{(i)}|_p \leq 1$。对于每个 $x_- \in \mathcal{D}_-$，我们有 $|w^\star x_- + b^\star|_p = |x_- - x^{(i)}|_p/|x^{(j)} - x^{(i)}|_p > 1$，因为假设 $\mathcal{D}$ 是线性可分的，因此根据超度量性，任意两个正例之间的距离都必须严格小于正例与负例之间的距离。

要证明 2，注意到 $a = -\frac{b^\star}{w^\star} = x^{(i)}$ 与 $r = \frac{1}{|w^\star|_p} = |x^{(j)} - x^{(i)}|_p$。

要证明 3，注意由 $\bar{B}_{r'}(a)$ 定义的分类器对应于某些 w 和 b的 $a = -\frac{b}{w}$ 与 $r' = \frac{1}{|w|_p}$；第二个方程在 $b = -wa = -wx^{(i)}$下成立。对于任意$x_- \in \mathcal{D}_-$，我们有 $|wx_- + b|_p = |w(x_- + x^{(i)})|_p = |w|_p|x_- + x^{(i)}|_p = \frac{1}{r'}|x_- + x^{(i)}|_p \geq pr'/r' = p > 1$。对于任意 $x_+ \in \mathcal{D}_+$，由于线性可分性和超度量性，有 $|wx_+ + b|_p = |w(x_+ + x^{(i)})|_p = |w|_p|x_+ + x^{(i)}|_p = \frac{1}{r'}|x_+ + x^{(i)}|_p < pr'/r' = p$。由于 $|wx_+ + b|_p < p$，我们必有$|wx_+ + b|_p \leq 1$。

第 4 点由第 1 点和第 3 点自动推出。

D.2 命题证明3

我们先证明 1。我们先证明“⇒”方向，也就是说，任何满足 $|f(x)|_p \leq 1$ 的 x 都具有以下形式

$$x \in \bar{B}_{p^{-k+k_{12}}}(a_1) \cup \bar{B}_{p^{-k+k_{12}}}(a_2), \tag{10}$$

我们有 $|f(x)|_p \leq 1$ 等价于 $|x-a_1|_p|x-a_2|_p \leq r = p^{-k}$。令 $p^{-k_1} = |x-a_1|_p$ 和 $p^{-k_2} = |x - a_2|_p$，则当且仅当

$$k_1 + k_2 \geq k. \tag{11}$$

此外，由强三角不等式可得 $p^{-k_{12}} = |a_1 - a_2|_p \leq \max\{|x - a_1|_p, |x - a_2|_p\} = \max\{p^{-k_1}, p^{-k_2}\}$，因此 $k_{12} \geq \min\{k_1, k_2\}$，且当 $k_1 \neq k_2$ 时取等号。现在注意，在情形 1 中，若 $k > 2k_{12}$，则必有 $k_1 \neq k_2$：事实上，若 $k_1 = k_2$，就会有 $k_{12} \geq k_1 = k_2$，从而 $k > 2k_{12} \geq k_1 + k_2$，这与 (11) 矛盾。因此，我们有 $k_{12} = \min\{k_1, k_2\}$。设 $k_1 < k_2$（即 $|x - a_1|_p > |x - a_2|_p$）。那么，$k_{12} = k_1$，结合 (11) 可得 $k_2 \geq k - k_{12}$，这意味着 $|x - a_1|_p = p^{-k_{12}}$ 和

$$|x - a_2|_p \leq p^{-k+k_{12}}. \tag{12}$$

但是，由于 $k > 2k_{12}$ 我们有 $-k + k_{12} < -k_{12}$，因此（12）推出 $|x - a_2|_p < p^{-k_{12}}$；根据强三角不等式，并且由于 $|a_1 - a_2|_p = p^{-k_{12}}$，我们自动得到$|x - a_1|_p = p^{-k_{12}}$，这表明（12）本身始终是一个可行解。若我们在上文假设 $k_1 > k_2$，则由对称性可得 $|x - a_1|_p \leq p^{-k+k_{12}}$，结合（12）可知，析取式（10）在情形 1 中成立。反过来，设

x 具有（10）的形式。若 $x \in \bar{B}_{p^{-k+k_{12}}}(a_1)$，如上所示，我们有 $|x-a_2|_p = p^{-k_{12}}$，因此 $f(x) = |c|_p|x-a_1|_p|x-a_2|_p \le p^k p^{-k+k_{12}} p^{-k_{12}} = 1$，这就证明了“⇐”方向。

现在我们再次证明 2，仍从“⇒”方向开始。设 x 满足 $|f(x)|_p \le 1$，其中 $k \le 2k_{12}$。再次令 $p^{-k_1} = |x-a_1|_p$ 和 $p^{-k_2} = |x-a_2|_p$，我们再次得到 (11)，以及不等式 $k_{12} \ge \min\{k_1, k_2\}$；当且仅当 $k_1 \ne k_2$时取等号。若 $k_1 < k_2$，我们可以用与上文类似的推理得到 (12) 与 $|x-a_1|_p = p^{-k_{12}}$；但现在我们有$-k + k_{12} \ge -k_{12}$，因此强三角不等式推出 $|x-a_2|_p \le p^{-k_{12}} = |x-a_1|_p$。若假设 $k_1 > k_2$，也可得到类似结果 $|x-a_1|_p \le p^{-k_{12}} = |x-a_2|_p$。需要考虑的第三种可能是 $k_1 = k_2$。在这种情况下，强三角不等式推出$k_{12} \ge k_1 = k_2$，因此必有 $|x-a_1|_p = |x-a_2|_p \ge p^{-k_{12}}$；并且由 (11) 还可得 $k_1 = k_2 \ge \lceil k/2 \rceil$，这意味着 $|x-a_1|_p = |x-a_2|_p \le p^{-k/2}$。由于 $\lceil k/2 \rceil \le k_{12}$，，综合以上可得 $x \in B_{p^{-\lceil k/2 \rceil}}(a_1) = B_{p^{-\lceil k/2 \rceil}}(a_2)$，正如所需。最后，为了证明“⇐”方向，注意任何如上所述的 x 都满足 $|x-a_1|_p \le p^{-\lceil k/2 \rceil}$ 和$|x-a_2|_p \le p^{-\lceil k/2 \rceil}$，因此 $f(x) = |c|_p|x-a_1|_p|x-a_2|_p \le p^k p^{-\lceil k/2 \rceil} p^{-\lceil k/2 \rceil} \le 1$。

D.3 命题4的证明

具有 $d = 2$ 的布尔问题。我们首先证明，对于 $d = 2$ 以及任意素数 p，$\mathcal{H}_{\mathbb{Q}_p^d}$ 中的分类器都能计算任意布尔函数。为方便起见，我们假设 $(x_1, x_2) \in \{0,1\}^2$（同样也可以假设 $(x_1, x_2) \in \{-1,+1\}^2$，并通过对权重进行线性变换得到相同的证明）。共有 16 种情况需要考虑 —— 对 4 种输入配置进行2⁴种可能的输出赋值。恒零分类器和恒一分类器可以通过选取 $w = 0$ 并选择合适的 b轻松实现；输出仅取决于 x_1 或仅取决于 x_2 的情况也很简单，因为我们可以分别选择$w_2 = 0$或 $w_1 = 0$，忽略无关输入，从而转化为一维问题。对其中一个输入取反，也可以通过变换 $x_i' = 1 - x_i$ 并相应地重新计算 w 和 b 来处理。剩下唯一需要分析的情况是：

1. 异或函数。这里，目标值是 $y(x) = +1$若 $x_1 \ne x_2$，否则为 -1。我们可以设定 $w_1 = w_2 = \frac{1}{p}$和 $b = -\frac{1}{p}$。于是，当 $x_1 \ne x_2$时，我们得到 $|w^\top x + b|_p = 0 \le 1$；当 $x_1 = x_2$ 时，我们得到 $|w^\top x + b|_p = \left|\frac{1}{p}\right|_p = p > 1$。2. 异或函数的否定。与前一种情况类似；我们可以设定 $w_1 = -w_2 = \frac{1}{p}$ 和 $b = 0$。3. 与函数。这里，目标值是 $y(x) = +1$ 若 $x_1 = x_2 = 1$，否则为 -1。对于$p = 2$，我们可以设定 $w_1 = w_2 = \frac{1}{4}$和 $b = -\frac{1}{2}$。对于 $p > 2$，我们可以设定 $w_1 = w_2 = \frac{1}{p}$和$b = -\frac{2}{p}$。我们将在下文看到，与异或函数一样，这是计数问题的一个特例，而我们将为此提供一个通用解法。

4. 与函数的否定。这个问题无法直接求解。如果可以，我们就需要有 $|b|_p \le 1$、$|w_1 + b|_p \le 1$、$|w_2 + b|_p \le 1$ 和 $|w_1 + w_2 + b|_p > 1$。根据强三角不等式，我们有 $|w_1 + w_2 + b|_p \le \max\{|w_1 + b|_p, |w_2|_p\}$，因此要满足上述不等式，就需要有 $|w_2|_p > 1$。因此 $|w_2|_p \ne |b|_p$，进而必须有 $|w_2 + b|_p = \max\{|w_2|_p, |b|_p\} = |w_2|_p > 1$，这与假设矛盾。我们将在下文看到，这是计数阈值问题的一个特例。然而，由于类 $\mathcal{H}_{\mathbb{Q}_p^d}$被定义为同时包含分类器 $\hat{}y(.;\, w, b)$ 及其否定$-\hat{}y(.; w, b)$，我们可以回到前一个问题。

同余、奇偶校验和计数问题。接下来我们将说明，对于 $d \ge 2$, $\mathcal{H}_{\mathbb{Q}_p^d}$可以求解模 p^n 的同余问题（奇偶校验问题是其中一种特殊情况，

当 $p=2$ 与 $n=1$时），以及计数问题。我们假设输入为二进制（x_1，……，x_d）$\in\{0,1\}^d$。模 p^n 的同余问题对应于以下目标函数：

$$y(x)=\begin{cases}+1 & \text{if } \sum_{i=1}^d x_i \equiv a \bmod p^n \\ -1 & \text{otherwise,}\end{cases} \tag{13}$$

其中 $n\in\mathbb{N}$ 和 $a\in\{0,\ldots,p^n-1\}$。这些问题可通过一个 p进线性分类器求解：对每个 $i\in[d]$，有 $w_i=p^{-n}$，且 $b=-ap^{-n}$。当 c 个输入处于激活状态时，我们得到$|w^\top x+b|_p=|(c-a)p^{-n}|_p=p^n|c-a|_p$；当且仅当 $|c-a|_p\le p^{-n}$时，它为 ≤ 1，也就是说，$c\equiv a$ 模 p^n。

计数问题（其中异或和与是特例）对应于以下目标函数（定义域仍为 $\{0,1\}^d$）：

$$y(x)=\begin{cases}+1 & \text{if } \sum_{i=1}^d x_i = c \\ -1 & \text{otherwise,}\end{cases} \tag{14}$$

其中 $c\in\{0,\ldots,d\}$。这些问题可通过一个 p进线性分类器求解：对每个 $i\in[d]$，有 $w_i=p^{-n}$，且 $b=-cp^{-n}$，其中 n 设为 $n=\lceil\log_p(1+\max\{c,d-c\})\rceil$。为了解这一点，注意到计数问题 (14) 等价于同余问题 (13)，其中 $a=c$ 且 n充分大。更具体地说，对于任意 $c'\in\{0,\ldots,c-1,c+1,\ldots,d\}$，$n$ 应足够大，使得 $|(c'-c)p^{-n}|_p>1$。该不等式等价于 $|c'-c|_p>p^{-n}$；若 $\max\{c,d-c\}<p^n$，或 $n\ge\lceil\log_p(1+\max\{c,d-c\})\rceil$，则可保证这一点。

计数阈值问题。这对应于以下目标函数（定义域仍为 $\{0,1\}^d$）：

$$y(x)=\begin{cases}+1 & \text{if } \sum_{i=1}^d x_i \le c \\ -1 & \text{otherwise,}\end{cases} \tag{15}$$

其中 $c\in\{0,\ldots,d\}$。前面提到的与问题是一个特例，此时 $d=2$ 与 $c=1$，我们已经看到，对于任意 p，它都无法由 p-进制线性分类器解决。事实上，对于一般的 $d\ge 2$ 与 $c\in\{1,\ldots,d-1\}$，情况也相同。我们用反证法证明这一点。假设存在 w与 b，对于某个 $c\in\{1,\ldots,d-1\}$，它们可以解决 (15)。那么（特别是）必须有 $|b|_p\le 1$、$|\sum_{j=1}^c w_j+b|_p\le 1$、$|w_{c+1}+b|_p\le 1$和 $|\sum_{j=1}^c w_j+w_{c+1}+b|_p>1$。根据强三角不等式，有 $|\sum_{j=1}^c w_j+w_{c+1}+b|_p\le\max\{|\sum_{j=1}^c w_j+b|_p,|w_{c+1}|_p\}$，因此若要满足上述不等式，就必须有 $|w_{c+1}|_p>1$。因此，$|w_{c+1}|_p\ne|b|_p$，于是必须有 $|w_{c+1}+b|_p=\max\{|w_{c+1}|_p,|b|_p\}=|w_{c+1}|_p>1$，这与假设矛盾。

D.4 命题5的证明

对于 1，我们需要证明，对于任意满足 $|w^\top x^{(i)}+b|_p\le 1$ 的 $x^{(i)}$，都有 (i) $|w^\top x+b|_p\le 1\Rightarrow|w^\top x-w^\top x^{(i)}|_p\le 1$，以及 (ii) $|w^\top x+b|_p>1\Rightarrow|w^\top x-w^\top x^{(i)}|_p>1$。注意，$|w^\top x-w^\top x^{(i)}|_p=|w^\top x+b-w^\top x^{(i)}-b|_p\le\max\{|w^\top x+b|_p,|w^\top x^{(i)}+b|_p\}$，且当 $|w^\top x+b|_p\ne|w^\top x^{(i)}+b|_p$ 时取等号。对于 (i)，由于 $|w^\top x+b|_p\le 1$ 和 $|w^\top x^{(i)}+b|_p\le 1$，可得 $|w^\top x-w^\top x^{(i)}|_p\le 1$。对于 (ii)，由于 $|w^\top x+b|_p>1$，我们必有 $|w^\top x+b|_p\ne|w^\top x^{(i)}+b|_p$，因此 $|w^\top x-w^\top x^{(i)}|_p=\max\{|w^\top x+b|_p,|w^\top x^{(i)}+b|_p\}=|w^\top x+b|_p>1$。

现在证明 2。令 w和 b满足分类器是紧的，且对某个 i 满足 $w^\top x^{(i)}+b=0$（由第 1 点保证），并选取满足 $|w^\top x^{(j)}+b|_p=1$的 j。那么存在可逆元 $u\in\mathbb{Z}_p^\times$，使得 $u(w^\top x^{(j)}+b)=1$。令 $w^\star=uw$ 和 $b^\star=ub$，并注意到 $w^{\star\top}x^{(i)}+b^\star=u(w^\top x^{(i)}+b)=0$，即可完成证明。

D.5 命题6的证明

我们将用到以下引理。

引理 2. 设 $a_1, \ldots, a_n \in \mathbb{Q}_p$. 那么，对任意 $u \in \mathbb{Q}_p$，存在一个 $j \in [n]$，使得 $|a_j - a_i|_p \leq |u - a_i|_p$ 对所有 $i \in [n]$都成立。具体而言，任意满足此界的 $j \in \operatorname{argmin}_k |u - a_k|_p$ 均满足这一界。

证明。令 $j \in \arg\min_k |u - a_k|_p$ 如上所述，即，对所有 k，都有 $|u - a_j|_p \leq |u - a_k|_p$。那么，对任意 i，都有 $|a_j - a_i|_p = |a_j - u - a_i + u|_p \leq \max\{|u - a_j|_p, |u - a_i|_p\} = |u - a_i|_p$. □

先考察零维情形下会发生什么，这颇有启发性；此时$x^{(i)} = 0$对所有 $i = 1, \ldots, n$。在这种情况下，只有 $y^{(1)}, \ldots, y^{(n)}$才起作用，问题在于找到使 b 最小的“质心”$\max_i |b - y^{(i)}|_p$。引理 2 表明， b 是该问题的解，当且仅当它是由 $y^{(1)}, \ldots, y^{(n)}$张成的球内的一个点（因此，由于超度量性，它也是该球的一个中心）。特别地， $b = y^{(j)}$ （对于任意 $j \in [n]$）都是一个解，这同样是由引理2得出的。

现在我们考虑另一个特殊情形，即不存在偏置参数；也就是说，问题是寻找使 $L(w) := \max_i |wx^{(i)} - y^{(i)}|_p$最小的 $w \in \mathbb{Q}_p$。假设对所有 $i \in [n]$都有 $x^{(i)} \neq 0$。那么我们有 $L(w) = \max_i |x^{(i)}|_p |w - y^{(i)}/x^{(i)}|_p$。根据引理 2，对于任意 w，存在一个 j（即 $j \in \arg\min_k |w - y^{(k)}/x^{(k)}|_p$），使得对所有 i 都有$|x^{(i)}|_p |y^{(j)}/x^{(j)} - y^{(i)}/x^{(i)}|_p \leq |x^{(i)}|_p |w - y^{(i)}/x^{(i)}|_p$；因此，不失一般性，我们可以假设最优的 w 具有 $w = y^{(j)}/x^{(j)}$ 的形式，其中 j是某个值。我们需要找出与极小极大问题相关联的指标 i 和 j

$$\min_j \max_i |x^{(i)}|_p \left| \frac{y^{(j)}}{x^{(j)}} - \frac{y^{(i)}}{x^{(i)}} \right|_p,$$

这可以通过显式枚举在$\mathcal{O}(n^2)$时间内完成。

最后，让我们讨论包含权重和偏置的情形，即我们要解决的问题是

$$\min_{w,b} \max_i |wx^{(i)} + b - y^{(i)}|_p.$$

令 $w^\star$ 为该问题的一个部分解。由零维情形可知，任意形如 $b^\star = y^{(k)} - w^\star x^{(k)}$ （其中 k任意）的 b 都能补全该解。因此，我们可以代入 $b = y^{(k)} - wx^{(k)}$ 并求解 w，从而得到 $\min_w \max_i |w(x^{(i)} - x^{(k)}) - y^{(i)} + y^{(k)}|_p$。由于当 $i = k$ 时，我们有 $|w(x^{(i)} - x^{(k)}) - y^{(i)} + y^{(k)}|_p = |0|_p = 0$，该问题等价于 $\min_w \max_{i \neq k} |w(x^{(i)} - x^{(k)}) - y^{(i)} + y^{(k)}|_p$。这就回到了不含偏置的问题，由此我们得到命题 6 中所述的算法。

D.6 命题证明7

设 w 是一个经过 $r < d + 1$ 个点的回归器。我们证明，存在另一个经过 $r + 1$ 个点的回归器 $\hat{}w$，使得 $\|X\hat{}w - y\|_p \leq \|Xw - y\|_p$。该结论可由归纳法得到。

假设 $w^\top x^{(i)} = y^{(i)}$对于 $i \in \{j_1, \ldots, j_r\}$成立，且 $w^\top x^{(i)} \neq y^{(i)}$对于 $i \in [n] \setminus \{j_1, \ldots, j_r\}$成立。令$j_{r+1}, \ldots, j_d$为任意互不相同的输入索引。我们将证明存在 $k \in [n] \setminus \{j_1, \ldots, j_d\}$和一个分类器 $\hat{}w$，使得:

- $\hat{w}^\top x^{(i)} = w^\top x^{(i)} = y^{(i)}, \quad i \in \{j_1, \ldots, j_r\}$;
- $\hat{w}^\top x^{(i)} = w^\top x^{(i)} \notin y^{(i)}, \quad i \in \{j_{r+1}, \ldots, j_d\}$;

- $\hat{w}^\top x^{(k)} = y^{(k)}$;
- $|\hat{w}^\top x^{(i)} - y^{(i)}|_p \le$ 最大值$_{j\in[n]} |w^\top x^{(j)} - y^{(j)}|_p, \quad i \notin \{j_{r+1}, ..., j_d, k\}$。

这些条件意味着 $\|X\hat{w} - y\|_p \le \|Xw - y\|_p$。

定义 $\bar{X} \in \mathbb{Q}_p^{(d+1)\times(d+1)}$为 X中由行 $\{j_1, ..., j_d, k\}$ 构成的子矩阵——根据假设，该矩阵可逆。类似地，令$\hat{y} \in \mathbb{Q}^{d+1}$ 定义为当 $i \in \{j_1, ..., j_d\}$ 时的$\hat{y}^{(i)} = w^\top x^{(i)} = \hat{w}^\top x^{(i)}$ 以及$\hat{y}^{(d+1)} = y^{(k)}$。我们有 $\bar{X}\hat{w} = \hat{y}$，因此$\hat{w} = \bar{X}^{-1}\hat{y}$。此外，对于 $i \in [n]$，我们有：

$$\begin{aligned}
\hat{w}^\top x^{(i)} - y^{(i)} &= w^\top x^{(i)} - y^{(i)} + (\bar{X}^{-1}\hat{y} - w)^\top x^{(i)} \\
&= w^\top x^{(i)} - y^{(i)} + (\bar{X}^{-1}(\hat{y} - \bar{X}w))^\top x^{(i)} \\
&= w^\top x^{(i)} - y^{(i)} - (w^\top x^{(k)} - y^{(k)})[\bar{X}^{-1}]_{:,d+1}^\top x^{(i)}. \qquad (16)
\end{aligned}$$

现在注意到，矩阵 X 可分解为

$$\bar{X} = \begin{bmatrix} \tilde{X}_{1:d} & 1_d \\ \tilde{x}^{(k)^\top} & 1 \end{bmatrix},$$

其中，我们用 $\tilde{x}^{(i)} \in \mathbb{Q}_p^d$ 表示对 $i \in [n]$ 进行偏置扩增之前的输入，也就是说，$x^{(i)} = [\tilde{x}^{(i)^\top}, 1]^\top$；并用 $\tilde{X}_{1:d}$ 表示行向量为 $\tilde{x}^{(j_1)^\top}$，……，$\tilde{x}^{(j_d)^\top}$ 的矩阵。根据分块矩阵求逆公式——该公式适用，因为 $X_{1:d}$ 可逆且 $\bar{X}$ 也可逆，后者意味着舒尔补 $(1 - \tilde{x}^{(k)^\top}\tilde{X}_1^{-1}{}_{:d}1_d)$ 可逆——我们得到

$$\begin{aligned}
[\bar{X}^{-1}]_{:,d+1}^\top x^{(i)} &= \begin{bmatrix} -\tilde{X}_{1:d}^{-1}1_d(1 - \tilde{x}^{(k)^\top}\tilde{X}_{1:d}^{-1}1_d)^{-1} \\ (1 - \tilde{x}^{(k)^\top}\tilde{X}_{1:d}^{-1}1_d)^{-1} \end{bmatrix}^\top \begin{bmatrix} \tilde{x}^{(i)} \\ 1 \end{bmatrix} \\
&= \frac{1 - \tilde{x}^{(i)^\top}\tilde{X}_{1:d}^{-1}1_d}{1 - \tilde{x}^{(k)}\tilde{X}_{1:d}^{-1}1_d}.
\end{aligned}$$

现在我们为 k 定义如下选择规则：

$$k := \underset{i\in[n]\setminus\{j_1,\ldots,j_d,k\}}{\arg\max} \left|1 - \tilde{x}^{(i)}\tilde{X}_{1:d}^{-1}1_d\right|_p,$$

这确保了$\left|[\bar{X}^{-1}]_{:,\ d+1}^\top x^{(i)}\right|_p \le 1$。

最后，我们对（16）应用强三角不等式：

$$\begin{aligned}
|\hat{w}^\top x^{(i)} - y^{(i)}|_p &= \left|w^\top x^{(i)} - y^{(i)} - (w^\top x^{(k)} - y^{(k)})[\bar{X}^{-1}]_{:,d+1}^\top x^{(i)}\right|_p \\
&\le \max\left\{|w^\top x^{(i)} - y^{(i)}|_p,\ |w^\top x^{(k)} - y^{(k)}|_p \underbrace{\left|[\bar{X}^{-1}]_{:,d+1}^\top x^{(i)}\right|_p}_{\le 1}\right\} \\
&\le \max\left\{|w^\top x^{(i)} - y^{(i)}|_p,\ |w^\top x^{(k)} - y^{(k)}|_p\right\} \\
&\le \max_{j\in[n]} |w^\top x^{(j)} - y^{(j)}|_p,
\end{aligned}$$

对所有 $i \in [n] \setminus \{j_1, ..., j_d, k\}$ 均成立，证毕。