1 介绍

1.1 预测

预测：在给出有关过去元素和可能的其他可用信息的一些知识的情况下，预测未知序列的下一个元素。

在最基本的预测问题中，预测器观察一个又一个序列中的元素 $y_1, y_2...$，在每个时刻 t,t 之前所有的元素 y 被揭示，预测器利用前 $t-1$ 个观测来预测第 t 个元素。在顺序预测的经典统计理论中，元素序列（我们称之为结果）被假设为平稳随机过程的实现。在这种假设下，可以根据过去观察的序列来估计过程的统计特性，并且可以从这些估计中导出有效的预测规则。在这样的设置中，预测规则的风险可以定义为一些损失函数的预期值，该损失函数测量预测值和真实结果之间的差异，并且根据其风险的行为来比较不同的规则。

这本书放弃了结果是由潜在随机过程生成的基本假设，并将序列 $y_1, y_2, ...$ 视为某种未知和未指定机制的产物（这可能是确定性的、随机的，甚至与我们自己的机制相违背）。为了与随机建模进行对比，这种方法通常被称为单个序列（*individual sequences*）的预测。

如果没有概率模型，就无法定义风险的概念，并且如何正式设定预测目标也不是显而易见的。事实上，存在多种可能性，本书中讨论了其中许多可能性。在我们的基本模型中，预测器的性能是通过多轮预测期间累积的损失来衡量的，其中损失是通过某些固定损失函数进行评分的。由于我们希望避免对要预测的序列的生成方式进行任何假设，因此没有明显的基线来衡量预测者的表现。为了提供这样的基线，我们引入了一类参考预测者，也称为专家。这些专家在下一个结果揭晓之前向预测者提供他们的预测。然后，预测者可以根据专家的“建议”做出自己的预测，以使自己的累积损失接近同类最佳专家的水平。

预测者的累积损失与专家的累积损失之间的差异称为遗憾（*regret*），因为它衡量的是事后预测者对没有遵循该特定专家的建议感到后悔的程度。遗憾是本书的一个基本概念，并且非常注重构建预测策略，以保证班上所有专家都不会感到遗憾。事实证明，将遗憾保持在较小水平的可能性很大程度上取决于专家类别的规模和结构以及损失函数。

“专家”的抽象概念可以用不同的方式解释，这也取决于正在考虑的具体应用。在某些情况下，可以将专家视为具有未知计算能力的黑匣子，可能可以访问私人信息来源。在其他应用中，专家类别被统称为统计模型，其中该类别中的每个专家代表某些给定“自然状态”的最佳预测者。对于最后的解释，最小化任意序列的遗憾的目标可以被认为是鲁棒性要求。事实上，一个小的遗憾保证了，即使模型没有完美地描述自然状态，预测者也能几乎和最佳模型表现地一样好。

在计算机科学中，顺序接收输入的算法被称为以在线方式运行。在线算法的典型应用领域包括涉及决策序列的任务，例如选择如何服务流中的每个传入请求。决策问题和预测问题之间的相似性，以及在线算法通常在任意输入序列上进行分析的事实，导致了两个领域之间富有成效的思想和技术交流。然而，预测框架中缺少的顺序决策问题的一些关键特征（例如对决策者与生成请求流的机制之间的交互进行建模的状态的存在）迄今为止阻碍了允许对两类问题进行统一分析的通用理论的推导。

1.2 学习

单个序列的预测也一直是机器学习理论研究的主要课题，更具体地说是在线学习领域。机器学习和信息论的研究人员也对预测的计算方面感兴趣。当考虑非常大的参考预测器类别时，这成为一个特别重

要的问题，并且需要发明各种技巧以使预测器在实际应用中可行。

1.3 博弈

本书研究的在线预测模型与博弈论有着密切的联系。首先该模型是最自然地根据预测者和生成结果序列的“环境”之间进行的重复游戏来定义的，从而提供了一种描述基本主题的变体的便捷方法。这种联系甚至更深。事实证明，如果重复正常形式博弈中的所有玩家都根据某些简单的后悔最小化预测策略进行游戏，那么所诱发的动态在某种意义上会导致均衡。

1.4 一个例子

为了帮助读者了解本书中的研究成果，我们首先详细描述一个简单的预测过程，并分析它在任意结果序列上的表现。

假设我们要预测一个未知的二进制序列 $y_1, y_2, \ldots$，其中每个 y_t 取值于 $\{0,1\}$。在每个时间点 t，预测者会先给出一个预测 $\widehat{p}_t \in \{0,1\}$ 来预测 y_t。随后，真实的 y_t 会被揭示，预测者就能知道自己的预测是否正确。为了计算 $\widehat{p}_t$，预测者会参考 N 位专家的建议。这些建议用一个二进制向量表示 $(f_{1,t}, \ldots, f_{N,t})$，其中 $f_{i,t} \in \{0,1\}$ 是第 i 位专家对下一位 y_t 的预测。我们的目标是给出一个界限，限制预测者在 $\widehat{p}_t \neq y_t$ 的时间步数，也就是预测者犯错误的次数。

为了从一个更简单的情况入手，假设我们事先知道，在这个特定的结果序列中，有某位专家 i 是从不出错的。也就是说，存在某个 i 满足对所有 t 都有 $f_{i,t} = y_t$，但我们不知道具体是哪个专家满足这个条件。基于这个信息，我们不难设计出一种预测策略，使预测者在整个序列上最多犯 $\lfloor \log_2 N \rfloor$ 次错误。为了解释这个结论，考虑这样一个预测者：他一开始给每位专家 $j = 1, \ldots, N$ 分配权重 $w_j = 1$。在每个时间点 t，预测者会在满足以下条件时预测 $\widehat{p}_t = 1$：即当前权重 $w_j = 1$ 的专家中，预测 $f_{j,t} = 1$ 的人数多于预测 $f_{j,t} = 0$ 的人数。当真实值 y_t 被揭示后，如果 $\widehat{p}_t \neq y_t$，预测者会将所有在 t 时刻预测错误的专家 k 的权重置为 0，也就是执行 $w_k \leftarrow 0$。简单来说，这个预测者会追踪哪些专家犯错，并根据至今为止从未犯错的专家中的大多数意见来进行预测。

这个分析非常直观。设 W_m 为预测者在犯了 m 次错误后所有专家权重的总和。初始时，$m = 0$，$W_0 = N$。当预测者犯了第 m 次错误时，至少有一半之前一直正确的专家会首次犯错。这意味着 $W_m \leq W_{m-1}/2$，因为那些首次预测错误的专家的权重被置为 0。由于这个不等式对所有 $m \geq 1$ 都成立，我们可以得到 $W_m \leq W_0/2^m$。再结合第 i 位专家从不犯错的事实，即 $w_i = 1$，我们知道 $W_m \geq 1$。同时，$W_0 = N$，因此我们有 $1 \leq N/2^m$。解得 m（必须是整数）满足不等式 $m \leq \lfloor \log_2 N \rfloor$。

现在我们来分析一般情况，即预测者在序列上没有任何关于专家错误次数的预先信息。我们的目标是将预测者犯的错误次数与表现最好的专家犯的错误次数联系起来，而不论是哪一个序列被预测。

回顾之前的预测策略，可以看到，只有在确定某位专家永远不会犯错的情况下，将出错专家的权重归零才是合理的。如果没有这种保证，一个更安全的选择是每次专家 k 犯错时，将权重更新为 $w_k \leftarrow \beta w_k$，其中 $0 < \beta < 1$ 是一个自由参数。换句话说，每当某位专家出错时，我们不再把他的权重清零，而是将其缩小一个固定的比例。这是我们对之前预测者策略唯一的修改，而它的分析过程也和之前一样简单。更具体地，新预测者会将建议预测 1 的专家的总权重与建议预测 0 的专家的总权重进行比较，并根据加权多数的意见来进行预测。和之前一样，当预测者犯下第 m 次错误时，出错专家的总权重至少是 $W_{m-1}/2$。这些出错专家的权重被乘以 β，而其他专家的权重（最多为 $W_{m-1}/2$）保持不变。因此，我们有 $W_m \leq W_{m-1}/2 + \beta W_{m-1}/2$。由于这个不等式对所有 $m \geq 1$ 都成立，我们得到 $W_m \leq W_0(1+\beta)^m/2^m$。现在，设 k 为在预测者犯第

m 次错误时，错误次数最少的专家，记这个最小错误次数为 m^*。那么这个专家当前的权重是 $w_k = \beta^{m^*}$，因此我们有 $W_m \geq \beta^{m^*}$。于是可以得到不等式 $\beta^{m^*} \leq W_0(1+\beta)^m/2^m$。结合 $W_0 = N$，我们得出最终的界限：

$$m \leq \left\lfloor \frac{\log_2 N + m^* \log_2(1/\beta)}{\log_2 \frac{2}{1+\beta}} \right\rfloor .$$

对于任何固定的 β 值，上述不等式表明预测者在任意预测次数后的错误与当时表现最好的专家的错误次数 m^* 之间存在线性关系。注意，这个界限对结果序列的选择是无关的。

m 和 m^* 之间的线性关系意味着，在某种程度上，随着专家与结果序列之间的“差异” m^* 的增加，预测者的表现逐渐下降。而这个界限对专家数量的依赖性较弱：$\log_2 N$ 项表明，除了计算方面的考虑外，专家数量翻倍只会导致界限增加一个较小的加法项。

尽管这个例子非常简单，但它包含了本书中发展的主要思想之一，例如利用与专家过去表现相关的权重来计算预测。在接下来的章节中，我们会对这一思想以及其他许多概念进行严格而系统的展开，目的是提供关于这个迷人主题的全面视角。

1.5 给读者的说明

本书主要面向计算机科学、数学、工程学和经济学领域的研究人员和学生，他们对预测和学习的各个方面感兴趣。尽管我们尽量使文本尽可能自成一体，但读者应熟悉一些基本的概率、分析和线性代数概念。为帮助读者，我们在附录中汇总了本书中用到的一些技术工具。这些材料中有些是相当常见的，但可能并非所有潜在读者都熟悉。

为了尽量减少对文本流畅性的干扰，我们将书目的参考文献集中在每章末尾。这些参考文献旨在追溯书中描述结果的来源，并指向一些相关的文献。我们为可能的遗漏致歉。其中一些材料是首次发表的，但这些结果并未特别标注。每章末尾还有一组练习题，难度差异很大。部分练习题可以通过对正文材料的简单改编来解决，这些练习题应该能帮助读者掌握内容。而其他练习则涉及一些困难的研究成果。在某些情况下，我们提供了解题思路，但没有解答手册。

图 1.1 描述了本书各章之间的依赖关系结构。它应该能帮助读者集中精力于特定主题，也为教师组织不同课程的材料提供帮助。

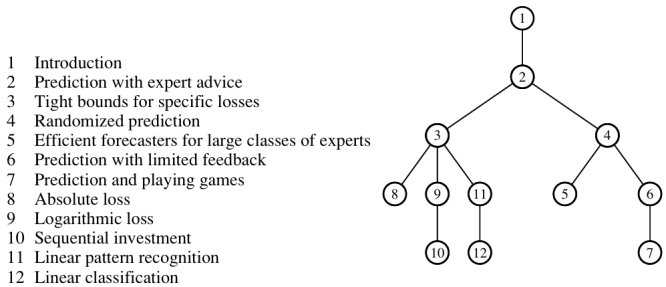

图 1.1: 章节结构