NOISE

第三部分

预测性判断中的噪声

哪里有预测，哪里就有客观无知

很多判断都是预测性的，而预测可以被验证和评估，所以我们可以通过考察预测性判断来更好地了解噪声和偏差。在这一部分中，我们主要关注预测性判断。

在第 9 章，我们比较了专业人士、机器和简单规则预测的准确性，结论是：专业人士预测的准确性是三者中最差的——你可能并不会对此感到惊讶。在第 10 章，我们探究了上述现象背后的成因，并发现噪声是导致人类判断力降低的主要因素。

为了得出这些结论，我们需要评估预测的品质，这就需要一种测量预测准确性的方法，从而回答“预测与结果的协变关系”（covariation）这样的问题。例如，假设人事部门定期评估新员工的潜力，那么我们就可以在几年后考察这些员工的绩效表现，并对比其潜力评估与绩效评估之间的相似度。如果某位员工在入职时被评价为高潜力员工，并且他在随后的工作中也获得了较高的评价，那就说明针对他的潜力评估在一定程度上是准确的。

符合这一直觉的测量指标叫作“一致性比率”（Percent Concordant，PC）。它回答了一个更具体的问题：假设你随机挑选了两名新员工，

那么在潜力评估中得分较高的新员工，在随后的工作中也表现较好的概率是多少？如果早期评估完全准确，则 PC 应为 100%，即对两名新员工的潜力评估完美地预测到了两人的绩效排名。如果早期预测完全没用，那么一致性只会随机发生，所谓的“高潜力”员工表现好和表现差的概率相当，即 PC 为 50%。我们将在第 9 章继续讨论这个已被广泛研究的有关招聘的例子。再举个简单点的例子，成年男性脚的尺码与身高的 PC 值为 71%，也就是说，如果对两个人“评头论足”，你会发现长得高的那个人同时脚也大的可能性是 71%。

PC 是一个衡量协变关系的直观指标，这是它的优点，但它并非社会科学家所使用的标准度量指标。标准度量指标是相关系数（correlation cofficient，r），当两个变量正相关时，其值在 0 ～ 1 的范围内变化。在前面的例子中，身高和脚的尺码之间的相关系数约为 0.6。

我们可以使用很多方法来审视相关系数。这里有一种很直观的方法：两个变量之间的相关系数就是指决定它们的因素中共有成分所占的百分比。例如，如果某个特征完全由遗传所决定，那么我们可以推测亲兄弟姐妹在该特征上的相关系数为 0.5，因为他们拥有 50%的共同基因，而堂兄弟姐妹之间的相关系数为 0.25，因为他们拥有 25%的共同基因。对于身高和脚的尺码之间的相关系数为 0.6，我们也可以理解为，决定身高的因素中有 60%也同时决定了脚的尺码。

以上两种协变关系的测量指标是直接相关的。表 1 列出了一系列相关系数的 PC 值。在本书的其余部分，当讨论人和模型的预测表现时，我们通常会同时应用这两个指标。

表 1　相关系数和一致性比率（PC）的对应关系

相关系数	PC
0	50%
0.1	53%
0.2	56%
0.3	60%
0.4	63%
0.6	71%
0.8	79%
1	100%

在第 11 章，我们会讨论预测准确性的一个重要局限：影响未来的很多事件是无法预知的，因而大多数判断都是在我们所谓的客观无知的状态下做出的。然而令人惊讶的是，在大多数情况下人们往往会忽视这一局限，并满怀信心或过度自信地进行预测。最后，在第 12 章中，我们发现客观无知不仅会影响我们对事件的预测能力，甚至会影响我们对事件的理解能力，这也是“为什么噪声会隐而不见”谜题的一个重要答案。

第 9 章

判断与模型，简单的模型普遍优于人类判断

很多人都对预测未来的工作绩效感兴趣，不只是自己的，还有别人的。因此，绩效预测是用来考察预测性判断的实用例子。例如，一家大公司在招聘高管时，聘请了一家专业咨询公司对两名候选人莫妮卡和娜塔莉进行评估，并以取值为 1 ～ 10 分的量表对两人的领导力、沟通能力、人际交往能力、职业技能、应聘动机等维度进行打分（见表 9-1）。你的任务是：预测她们在两年后的工作绩效，并用 1 ～ 10 分来评分。

表 9-1　两名高管候选人的能力评估得分

	领导力	沟通能力	人际交往能力	职业技能	应聘动机	你的预测
莫妮卡	4	6	4	8	8	
娜塔莉	8	10	6	7	6	

大多数人在面对此类问题时，只是简单地盯着每一行数字并心算出平均分，然后快速做出判断。如果你也是这样，那么你可能会得出这一结论：娜塔莉是更优人选，因为莫妮卡的平均分比她差一两分。

判断还是公式

针对此问题，你采取的这种方法被称为“诊断性判断”（clinical judgment）。在此过程中，你会考虑相关信息，或许再快速计算一下，然后利用直觉做出判断。事实上，诊断性判断就是我们在这本书中简单描述的判断过程。

现在假设你以实验参与者的身份完成了上述预测工作。莫妮卡和娜塔莉的数据来自一个信息数据库，其中记录了此前聘用的数百名经理的信息，以及这些经理在 5 个维度上的得分。你可以使用那些评分来预测两人的工作绩效，而现在你还获得了两人的实际工作绩效数据。那么，想一想你对这两人的诊断性判断有多接近她们的实际绩效呢？

这个例子大致来源于一项关于绩效预测的真实研究。如果你曾参加过该项研究，你可能会对自己的预测结果非常不满意。一家国际咨询公司聘请了拥有博士学位的心理学家来做预测，结果发现，预测与绩效评估的相关系数仅为 0.15（PC = 55%）。也就是说，当他们评估一名候选人优于另一名候选人时，他们所偏爱的候选人最终获得更高绩效的可能性仅为 55%，比随机选择的结果高不了多少。显然，这不是一个令人满意的结果。

也许你会认为，预测准确性之所以这么差，是因为评分信息对预测没有用。因此，我们不禁要问：对候选人的评分到底包含了多少有用的预测信息？如何将它们进行整合才能获得与实际表现相关性最高的预测分数？

有一种标准的统计方法可以回答上述问题。在上述研究中使用这种方法，可以使相关系数达到 0.3（PC=60%）。结果虽然仍不尽如人意，但至少优于诊断性预测。

这种方法叫作“多元回归”（multiple regression），它是对各种预测因素的平均值进行加权后获得预测分数的方法。多元回归可以找到一组最佳权重，使整合后的预测分数与目标变量之间的相关性最大。最佳权重可以使预测的均方误差最小——这就是最小平方法在统计学中举足轻重的有力证明。你可能认为，与目标变量相关性越密切的预测因素，其权重也应该越大；而无用的预测因素，其权重应该为 0。然而事实上，权重也可能是负数，例如候选人乘公交的逃票次数在预测其工作绩效上的权重就可能是负的。

多元回归是一个“机械性预测”（mechanical prediction）的例子。机械性预测种类繁多，从简单规则（如雇用完成高中学业的人）到复杂的人工智能模型不等。“线性回归”（linear regression）模型是最为常见的一种，因此该模型也被称为“判断和决策研究的主力军”。为方便起见，我们将线性回归模型称为“简单模型”（simple models）。

上文提到的莫妮卡和娜塔莉的案例，可以帮助我们对诊断性预测和机械性预测进行比较。二者都具有如下一些简单的结构：

- 用一组预测因素（如案例中对候选人的评分）来预测目标结果（如候选人的工作绩效）。

- 利用人类的判断做出诊断性预测。

- 基于某项规则（如多元回归），使用同一组预测因素来生成机械性预测的结果。

- 比较诊断性预测与机械性预测的整体准确性。

梅尔：最优模型击败了你

在了解诊断性预测和机械性预测之后，人们往往想知道两者之间的区别，即相比于公式，人类的判断会更优吗？

这个问题早已有人提出过，但是直到 1954 年，明尼苏达大学心理学教授保罗·梅尔（Paul Meehl）出版了《临床与统计预测：理论分析和证据综述》（*Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*）一书后，该问题才引起了人们的广泛关注。梅尔回顾了 20 项研究，并根据一系列研究结果（如学业成就和精神病预后效果）分析了诊断性判断和机械性判断的优劣。他得出了一个强有力的结论：简单的机械性规则普遍优于人类的判断。梅尔发现，临床医生和其他专业人员在整合信息的能力上表现极差，虽然他们自认为在这方面具有优势。

为了更好地理解上述发现为何如此令人惊讶，以及它与噪声的关系，我们需要明白简单的机械性预测模型是如何工作的。机械性预测最关键的特征是：它的预测规则适用于所有情况。每个预测因素都有特定的权重，这个权重不会因个案的不同而发生变化。你可能会认为，这种严格的约束会使模型比不上人类的判断，比如，在上文的例子里，也许你会认为莫妮卡的应聘动机和职业技能相结合是一项重要优势，能弥补她在其他方面的不足；也许你还认为，考虑到娜塔莉的其他长处，她在这两方面的劣势没什么大不了。也就是说，你会不由自主地设想她们两人不同的成功途径。凭借这些看似合理的诊断性推测，你针对两人的情况，对不同的预测因素赋予了不同的权重，而简单模型不存在这样的问题。

简单模型的另一个限制是，预测因素每增加 1 个单位，总是会产生相同大小的效果，即如果增加 2 个单位，那么所产生的效果是前者的 2 倍，而诊断性直觉经常与这一原则相悖。例如，娜塔莉的沟通能力是满分 10 分，如果你对此印象深刻，认为此分数值得你提高对其沟通能力的预测权重，那么你所做的就是简单模型所不能做到的。在加权平均公式中，得分 10 和 9 之间的差异与得分 7 和 6 之间的差异是相同的，但诊断性判断往往不遵循这一原则，相反，它反映了一种普遍性的直觉，即相同的差异在一种情况下可能无关紧要，在另一种情况下却可能非常重要。因此，我们认为没有哪个简单模型可以完整地描述你对莫妮卡和娜塔莉所做出的判断。

本文使用的例子就是梅尔模式的一个典型案例。正如我们所指出的那样，诊断性预测与工作绩效之间的相关系数仅为 0.15（PC = 55%），而机械性预测获得的相关系数为 0.3（PC = 60%）。再回忆一

下你在莫妮卡和娜塔莉的例子中对她们的优点所持有的信心。梅尔的结果强烈表明，你对自己判断品质的满意感只是一种错觉，即“效度错觉”（illusion of validity）。

人们做出预测性判断时总会出现效度错觉，因为我们无法区分预测性任务的两个不同阶段：对当前证据的评估阶段和对实际结果的预测阶段。如果要评估两名候选人哪个看起来更好，你通常会充满信心，但是这跟猜测哪名候选人实际上更好完全是两码事。比如，你可以胸有成竹地说“娜塔莉看起来是比莫妮卡更优秀的候选人”，但是，如果要断言娜塔莉将成为一位比莫妮卡更成功的经理，则要冒很大的风险，原因很明显：评估两名候选人所需要的大部分信息你都是知道的，但要想预测未来，就存在很大的不确定性。

然而，我们的思维对上述差别的感受是模糊不清的，事实上，几乎每个人都对二者的差别感到困惑。但是，如果你做出预测时表现出的自信与你进行案例评估时一样，那么你就产生了效度错觉。

即使是临床医生也无法避免效度错觉。梅尔的研究发现，最简单的公式，只要持续应用，竟然可以胜过医学专家的临床判断。你肯定能想象到临床心理学家对此做何反应，他们会感到震惊、怀疑，甚至会鄙视这种对神奇的临床直觉假装进行的肤浅研究。这种反应很容易理解，梅尔的发现与人类判断的主观经验相矛盾，大多数人都更相信自己的经验而非学者的主张。

梅尔本人对自己的发现也持模棱两可的态度，因为一提到他的名字，我们就会想起“统计优于诊断性判断”这一论断，我们可能把他

想象成人类洞察力的无情批判者，或者是“量化分析之父”，但事实并非如此。梅尔不仅是一位学术研究人员，还是一位有着丰富临床经验的精神分析师，他的办公室里挂着心理学家弗洛伊德的照片。同时，他也是一个多才多艺的人，不仅教授心理学课程，还教授哲学和法律学课程，他还撰写了一些有关形而上学、宗教学、政治学甚至超心理学（parapsychology）的文章。这些特征都不符合一个刻薄的数据狂人的形象。梅尔对临床医生并没有恶意，但是正如他所说，存在“大量且一致”的证据表明：采用机械性的方法来整合意见会更具优势。

“大量且一致”是个不偏不倚的表述。一篇发表于 2000 年的论文对 136 项研究进行了回顾，清晰地表明机械性整合确实优于诊断性判断。这篇论文涵盖的研究主题广泛，包括对黄疸的诊断、军人的身体素质测评和婚姻满意度调查等，其中，63 项研究表明机械性预测更准确；65 项研究表明两者难分伯仲；8 项研究表明诊断性预测更好。以上结果可能还低估了机械性预测的优势，因为机械性预测比诊断性预测速度更快、成本更低。此外，在许多此类研究中，人类在判断时还具有不对等的优势，因为他们可以获取未提供给计算机模型的“私人”信息。这些发现都支持了一个显而易见的结论：**简单模型的决策优于人类判断。**

戈德堡：你的判断模型击败了你

梅尔的发现引出了一些重要的问题：公式到底为什么会更优？模型在哪些方面可以做得更好？事实上，一个更好的问题是：为什么人

类做出的判断很差？答案是：**人类在许多方面都不如统计模型，其中一个主要弱点在于人类的判断过程存在噪声。**

为了支持这一结论，我们来看另一项关于简单模型的研究，该研究始于美国俄勒冈州的小城市尤金（Eugene）。保罗·霍夫曼（Paul Hoffman）是一位富有且有远见的心理学家，他对当时的学术环境颇为不满，因此，他成立了一家研究所，招募了一批非常得力的研究人员，这使尤金市成为著名的人类判断行为研究重镇。

其中有一位名叫刘易斯·戈德堡（Lewis Goldberg）的研究人员，他因在“大五人格模型”的基础上发展出了“领导力角色模型”而闻名于世。在 20 世纪 60 年代后期，戈德堡基于霍夫曼的早期工作，开始研究用于描述个体判断行为的统计模型。

建立这样一个判断模型和建立一个“现实模型”（model of reality）一样简单，因为两者所使用的预测因素完全相同。与我们最初的例子一样，预测因素是高管在工作绩效的 5 个维度上的得分，使用的工具也是多元回归。不同的是，该公式并非用于预测候选人的实际绩效，而是用于预测人的判断，比如你对莫妮卡、娜塔莉和其他高管候选人的判断。

用加权平均的方式对你的判断进行建模，可能看起来有些奇怪，因为你的判断并不是这样形成的。当你评价莫妮卡和娜塔莉的工作绩效时，你并没有采用这种规则，事实上，你可能没有采用任何规则。总之，判断模型并非描述实际判断过程的模型。

然而，即使你在实际判断过程中并未基于线性公式去运算，你的判断结果仍可能像是使用了线性公式一般。比如，台球专家们在描述某一杆如何进球时，表现得就好像他们解开了复杂的方程一样，然而实际上他们并未真的那样做。同理，你做出的预测就好像使用了简单公式一样，然而实际上你所做的要复杂得多。对于一个假设模型来说，即使它对过程的描述存在明显的错误，但只要该模型可以合理准确地预测人们的行为，那么它也是很有用的。简单模型就是这样的假设模型。一项针对判断研究的报告全面回顾了 237 项研究，发现判断模型和诊断性判断的平均相关系数为 0.8（PC = 79%），尽管不是完全相关，但这种相关性已经足以支持所谓的“假设”理论了。

戈德堡的研究想要解决的问题是：简单的判断模型预测实际结果的效果究竟如何？由于该模型只是对判断者的一个粗略的模拟，因此我们可以合理地假定它的预测效果不佳。那么，用模型替代判断者时，会损失多少准确性呢？答案可能会让你大吃一惊。当我们依据模型做出预测时，预测的准确性并没有降低，相反，在大多数情况下，判断模型反而表现更优，该模型甚至优于专业人士的预测。我们或许可以这样来理解：**替代品竟然比真品更好用。**

这一结论已被许多领域的研究所证实。早期一项关于预测学生毕业成绩的研究证实了戈德堡的结论。研究人员要求 98 名参与者基于 10 条线索预测 90 名学生的 GPA。研究人员根据这些预测，为每名参与者做出的判断建立了一个线性模型，并比较了参与者本人和模型预测的准确性。结果发现，对于这 98 名参与者来说，模型都比他们本人的预测更准确！几十年后，一项对近 50 年研究成果的综述性研究也得出了同样的结论：**判断模型的表现一如既往地胜过判断者本人。**

我们不知道这些研究中的参与者是否收到了有关个人表现的反馈，但是，如果有人告诉你，对你的判断进行粗略建模后的模型实际上比你本人预测得更准确（这极具讽刺性），想必你会感到非常沮丧。对于大多数人来说，判断活动是复杂、丰富且有趣的，这也恰恰是因为它不符合简单规则。当我们发明并应用一些复杂规则来做判断或对某些案例有了不同于其他案例的见解时，即当我们做出了无法用简单的加权求和模型去简化的判断时，我们会自我感觉更加良好，对自己的判断能力更加信心十足。但关于判断模型的研究进一步证实了梅尔的结论——很多细节都是无用的，复杂性和丰富性并不会使预测更准确。

为什么会这样呢？要了解戈德堡的发现，我们需要了解是什么导致你的实际判断与预测这些判断的简单模型之间有了差异。

基于你的判断建立起来的统计模型，不可能将所有用于判断的信息都纳入其中，模型能做的只是抽象和简化。尤其是，你的简单模型不会将你一直遵循的任何复杂规则表征出来。比如，你可能会认为沟通能力评分为10分和9分之间的差别要比7分和6分之间的差别更大，或认为在所有维度上得分均为7分的候选人比平均分相同但优势和劣势都更加明显的候选人更优秀，然而你的模型并不会表征这些复杂规则，即使你经常使用这些规则。

如果你的复杂规则行之有效，那么简单模型会因为不能重复你的规则而导致自身的预测力下降。例如，假设你必须从一个人的技能和动机两个方面来预测他成功完成一项困难任务的可能性，那么加权平均并非好方法，因为动机再强，也无法弥补能力的不足，反之亦然。

如果你使用复杂的预测规则，那么你的预测准确性将比无法获取复杂规则的简单模型更高。但复杂规则通常只会给你带来效度错觉，这实际上会降低你的判断品质。也就是说，少数复杂规则是有效的，但大多数是无效的。

此外，你的简单模型并不会表征你在判断中的噪声，它不能重现你在特定案例中由于随机反应而产生的正误差或负误差。同理，你在做出特定判断时会受到当时的环境和心理状态的影响，而模型并不会。这些判断的噪声带来的误差很可能与任何事物都不相关，这意味着在大多数情况下，我们可以将其视为随机误差。

从你的判断中消除噪声通常会提高你的预测准确性。例如，假设你的预测与结果的相关系数是 0.5（PC = 67%），此时你的判断中包含了 50%由噪声导致的变异，而如果你的判断没有噪声，那么它们与结果的相关系数将提升至 0.71（PC = 75%）。由此可见，**用机器减少噪声可以提高预测判断的有效性。**

简而言之，用模型代替人类的判断意味着两件事：消除了人类的复杂规则，消除了噪声。判断模型比判断更有效这一强有力的发现说明：从人类判断的复杂规则中获得的好处（如果存在的话）不足以补偿噪声所带来的损失。**你可能会认为自己比一般人更擅长思考、更有洞察力，但实际上只是你的噪声更多而已。**

为什么我们以为复杂的规则更有效，实际上它们却损害了判断的准确性呢？一方面，人们发明的许多复杂规则并不正确；另一方面，即使复杂规则在原则上是有效的，它们也不可避免地仅适用于少

数能被观察到的情况。例如，假设你已经得出结论：对于一个独创性极高的候选人，即使他在其他方面得分一般，也值得被雇用。可问题在于，从定义上看，具有独创性的候选人总是很稀缺。既然对独创性的评估可能不可靠，在这一指标上得了高分的人就可能是侥幸，而真正具有独创性的人才往往无法被发现。即使在绩效评估中具有较高独创性的候选人最终真的表现得特别优秀，绩效评估本身也存在很多问题。两端的测量误差会不可避免地削弱预测的有效性，一些小概率事件尤其可能被忽略，复杂模型的优势很快就会被测量误差所掩盖。

马丁·于（Martin Yu）和内森·昆塞尔（Nathan Kuncel）报告了一项比戈德堡所做的更激进的研究。该研究基于莫妮卡和娜塔莉的案例，使用了一家跨国咨询公司的数据，这家跨国咨询公司聘请专家评估了 3 个独立样本中共 847 名高管职位的候选人。专家们在 7 个不同的评估维度上对这些候选人进行了评分，并使用他们的诊断性判断为每位候选人生成了一个预测总分，然而结果令人大吃一惊。

马丁·于和昆塞尔将判断结果与随机线性模型进行比较，而非与他们的最佳简单模型进行比较。他们为 7 个预测因素生成了 10 000 套随机权重，并应用了这 10 000 个随机公式来预测工作绩效。他们吃惊地发现，用任何线性模型来对所有案例进行预测，其结果均优于人类基于相同信息所做出的判断。在其中一个样本中，10 000 个随机加权线性模型中有 77%优于人类专家；在另外两个样本中，随机模型 100%胜过人类专家。换句话说，该研究表明，所有简单模型的表现都比人类专家好。

这项研究得出的结论比我们从戈德堡的判断模型中得出的结论更

有力。事实上，这是个非常极端的例子。在这种情况下，人类的判断确实非常糟糕，这就解释了为什么即使是不尽如人意的线性模型，其表现也超越了人类判断。当然，我们并不能因此下结论说机器绝对比人强，尽管如此，机械地遵守简单规则（马丁·于和昆塞尔称其为“无意识的一致性”（mindless consistency））可以显著提高针对困难问题所做判断的品质，这一事实说明了噪声对诊断性预测的巨大影响。

本章简要地说明了噪声对诊断性判断造成的负面影响。在预测性判断中，人类专家很容易被简单的公式所击败，其中包括真实模型、判断模型甚至随机生成的模型。这一发现支持我们使用无噪声的方法——规则和算法，这也是下一章的主题。

消除噪声

判断中存在噪声，但判断模型中没有

- 人们往往相信自己的判断能更好地考虑问题的复杂性和微妙的细节，但复杂性和微妙的细节基本上没什么用，因为它们并不会提升简单模型的准确性。

- 在保罗·梅尔的书出版 60 多年后的今天，机械性预测优于人类的判断这一观点仍然令人震惊。

- 判断中有很多噪声，因此无噪声的判断模型会做出更准确的预测。

第 10 章

无噪声的规则

近年来，人工智能（Artificial Intelligence）特别是机器学习技术让机器能够执行许多以前只有人类才能执行的任务。机器学习算法可以承担人脸识别、语言翻译、分析医学影像等任务，并且可以以惊人的速度和准确性来处理计算问题，例如为成千上万名驾驶员迅速规划行车路线。它们还可以执行困难的预测任务：预测美国最高法院的判决；识别哪些嫌疑人更可能在保释期逃脱；评估儿童保护部门接到的哪些电话更紧急，并需要工作人员上门访视。

尽管如今我们一听到“算法”一词，首先想到的是上面这些应用，但这个词还有更广泛的含义。在词典中，算法的定义是：在解决计算或其他问题时（尤其是借助计算机）所遵循的步骤或规则。根据这一定义，我们在上一章中所描述的简单模型和其他形式的机械性判断也属于算法。

事实上，从简单的规则到最复杂且难以理解的机器学习算法，许多机械性方法都可以胜过人类的判断。机械性方法之所以有这种出色表现，一个关键原因可能是所有机械性方法均无噪声，尽管这不是唯一的原因。

为了研究不同类型的基于规则的方法，并了解每种方法为何以及在何种条件下更有价值，我们从第 9 章的基于多元回归的简单模型（即线性回归模型）开始我们的旅程。由此出发，我们将在复杂性频谱上朝着两个相反的方向前进，首先从极端简捷的一端开始，然后朝着逐渐复杂的方向前进（见图 10-1）。

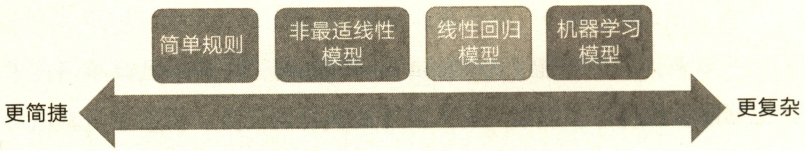

图 10-1　4 类规则和算法的相对复杂性

简捷：稳定之美

罗宾·道斯（Robyn Dawes）是 20 世纪六七十年代美国俄勒冈州尤金市研究人类判断行为的团队中的另一位明星成员。1974 年，道斯在简化预测任务方面取得了突破。他的研究思路令人惊讶：他建议不要使用多元回归模型来确定每个预测因素的精确权重，而应给所有预测因素分配均等的权重。

道斯将均等权重的公式定义为“非最适线性模型（improper linear

model）。他出人意料地发现，这些均等权重模型（equal-weight models）的准确性与合适的回归模型差不多，且远胜于诊断性判断。

连“并非最合适的模型”的支持者也承认，这种说法是不可信的，并且与统计直觉相悖。的确，道斯及其助手伯纳德·科里根（Bernard Corrigan）最初曾努力将论文发表在学术期刊上，但是编辑们根本不认同。如果回顾一下上一章中的莫妮卡和娜塔莉的例子，你就会相信某些预测因素比其他预测因素更重要。例如，相比于职业技能，大多数人会给予领导力更高的权重。因此，简单的未加权平均值怎么可能比精细加权的平均值或专家判断更好地预测一个人的表现呢？

在道斯取得研究突破多年后的今天，人们已经很熟悉这种令其同时代人惊讶的统计现象。正如前文所解释的那样，多元回归模型计算出了最佳权重，从而使均方误差最小化，而多元回归使原始数据中的误差最小化，因此，公式会自行调整以便预测数据中的每个偶然因素。例如，如果样本中包含一些具有较高职业技能但是由于不相关原因而表现异常出色的经理，该模型就将增加职业技能的权重。

这其中的挑战是：当将公式应用于样本之外时，也就是用它预测不同数据集的结果时，这些权重将不再是最优的。原始样本中的偶然因素不再存在，因为它们是“偶然因素”。在新样本中，具有较高职业技能的经理并不会都表现出色，而且新样本中具有原公式无法预测的新因素。要衡量模型预测的准确性，正确的做法是观察它在新样本中的表现，也就是观察它的“交叉验证相关性”（cross-validated correlation）。事实上，回归模型在原始样本上过于出色，因此交叉验证相关性的表现几乎总是比它在原始样本中的表现差。道斯和科里根

在几种情况下对均等权重模型和多元回归模型（交叉验证后）进行了比较。他们采用的一个案例就是预测伊利诺伊大学 90 名心理学研究生第一年的 GPA，使用的是与学业成就相关的 10 个变量，如能力测试分数、大学成绩、各种同龄人评分（peer ratings，如外向性）以及各种自评（如责任心）等。标准多元回归模型的预测相关系数为 0.69，经过交叉验证后降至 0.57（PC = 69%）；均等权重模型与第一年 GPA 预测的相关系数与此大致相同，为 0.6（PC = 71%）。许多其他研究也得到了相似的结果。

当原始样本较小时，经过交叉验证后，准确性会降低更多，因为小样本的偶然性多，变异性较大。道斯指出，社会科学研究中通常使用小样本，以致所谓的最佳权重的优势消失殆尽。正如统计学家霍华德·怀纳（Howard Wainer）在一篇研究最适当权重估值的学术论文中所使用的副标题：它并不重要。用道斯的话说，“我们不需要比我们的测量更精确的模型”。均等权重模型之所以表现出色，是因为它不容易受样本中偶然因素的影响。

道斯的研究的直接理论成果值得广为人知：即使你缺少有关结果先前的数据，你也可以进行有效的统计预测，只须收集一些你认为与预测结果相关的预测因素即可。

假设你必须对已经在多个维度上获得评分的高管的绩效做出预测，如第 9 章中高管的例子所示。你相信这些评分有很强的预测力，但是你没有每个评分预测的准确性数据。你也不可能花费几年的时间来追踪大量管理人员的绩效情况，但是，你可以基于这 7 个评分的均等权重模型来做预测。那么，这个均等权重模型的预测效果如何呢？

它与结果的相关系数将为 0.25（PC = 58%），远优于诊断性预测——相关系数为 0.15（PC = 55%），并且肯定与交叉验证后的回归模型非常相似，也不需要任何你没有的数据或任何复杂的计算。

用道斯的话说，均等权重模型具有“稳定之美”。研究判断的学生已对这句话达成共识。介绍这一观点的开创性文章的最后一句话给出了另一个精妙的总结：“应用均等权重模型所需的全部技巧是决定要关注哪些变量，并知道如何将这些变量进行叠加。”

更简捷：简约模型

另一种简化的方式是采用简约模型（frugal models）或简单规则。简约模型是对现实进行极端简化并无须复杂计算的模型，但在某些情况下，它们可以产生令人惊叹的预测效果。

令很多人感到惊讶的是，这些模型是基于多元回归的一个特征建立的。假设你使用了两个准确性很高的预测因素，它们与结果的相关系数分别为 0.6（PC = 71%）和 0.55（PC = 69%），且这两个预测因素彼此相关，相关系数为 0.5。当将这两个预测因素进行最佳组合时，预测的准确性会有多好呢？答案令人失望，相关系数是 0.67（PC = 73%），这个结果比之前好，但并没有好太多。

该示例说明了一条一般性规则：将两个或多个相关预测因素组合后，预测效果相比于单个预测因素并不会好多少。因为在现实生活中，预测因素几乎总是相关的，所以这一统计事实支持使用包含少量

预测因素的简约模型进行预测。与使用很多预测因素的模型相比，简单规则只需少量计算或根本无须计算，就能在某些情况下达到令人吃惊的预测效果。

一个研究团队于 2020 年发表了一项研究成果。他们将简约模型应用于一系列现实问题，内容包括在案件待审期间法官是否该批准被告的保释申请。这项决策隐含着对被告行为的预测，如果错误地拒绝保释，被告将被不必要地拘押，从而对个人和社会造成巨大损失；如果错误地批准保释，则被告可能在受审前逃脱，甚至犯下其他罪行。

研究人员仅使用两个可高度预测被告在保释期逃脱可能性的已知变量来建立模型：被告的年龄（年龄越大，逃脱风险较低）和未按时出庭受审的次数（有未按时出庭受审记录的人，更可能逃脱）。该模型将这两个变量转换为一系列分数，并针对风险进行评分，在计算被告保释期逃脱的风险时无须使用计算机，甚至不需要计算器。

当用真实数据来测试时，该简约模型的表现与那些使用众多变量的统计模型一样好，而在预测逃脱风险方面，简约模型比几乎所有法官的判断都要好。同样的简约模型采用少数几个整数（−3 ～ +3）对最多 5 个特征进行评分，并以此来对各种任务进行预测，如基于乳房 X 线片判断肿瘤的严重程度、诊断心脏病、预测信用风险等。在所有这些任务中，简约模型的表现都与复杂回归模型一样好，只不过它通常不如机器学习模型的表现好。

另一项研究也证明了简约模型的有效性。另外一个研究小组研究了一个与上述案例相似但有所不同的司法问题：预测惯犯。研究人员

在评估被告再次犯罪的风险时，使用的模型只有两个输入变量，但该模型的预测效果与使用 137 个变量的模型相同。毫无疑问，这两个预测因素（年龄和先前被定罪的次数）与保释模型中使用的两个因素密切相关，而大量证据表明，它们与犯罪行为也是紧密相关的。

简约模型的吸引力在于其透明性和易用性，而且相比于其他复杂模型，它只需略微牺牲一点准确性就能获得这些优势。

更复杂：机器学习

在旅程的第二部分，让我们在复杂性频谱上朝相反的方向前进。如果我们可以使用更多预测因素，收集更多数据，发现前人未发现的关系模式，并对这些模式进行建模以实现更好的预测效果，那会如何呢？从本质上讲，这就是人工智能的目的。

海量数据集对复杂分析至关重要。而获得此类数据集越来越容易，是近年来人工智能快速发展的主要原因之一。例如，大型数据集可以机械地处理“断腿的例外”（broken-leg exceptions）这种情况。这个有点神秘的短语可以追溯到前文中梅尔假想的一个示例：设想有这样一个模型，它可以预测人们今晚去看电影的可能性，无论你对该模型有多大信心，如果你碰巧知道某人刚摔断了腿，你都可能会比模型更准确地预测他今晚是否会去看电影。

在使用简单模型时，“断腿原则”给决策者提供了重要启示：它告诉人们何时需推翻模型，何时则不需要这样做。如果你掌握了模型

未考虑的如“断腿”这样的决定性信息，你就应该推翻模型的建议。此外，即使你缺少此类信息，有时你也不会同意模型的建议。在这种情况下，你试图推翻模型的行为，反映了你对相同预测因素做出反应的个人模式。这种个人模式很可能是无效的，你的干预可能会降低预测的准确性，因此你应该避免推翻模型。

机器学习模型之所以能够在预测方面表现出色，其中一个原因就是，它们能够发现人类所无法想象的各种“断腿”情况。在具有大量案例、海量数据的条件下，追踪观影行为的模型真的会学习，例如在固定观影日去了医院的人当晚不太可能去看电影。可以说，以这种方式改进对不常见事件的预测，可减少对人工监督的需求。

人工智能不是魔法，也不需要理解什么，它仅仅是在识别模式。虽然我们必须佩服机器学习的力量，但我们也要明白，人工智能可能要花很长时间才能理解为什么断腿之人会错过电影之夜。

更明智的保释决策

在前面提到的研究团队将简单规则应用于保释决策问题的同时，由塞德希尔·穆来纳森（Sendhil Mullainathan）[①] 领导的另一个团队

① 哈佛大学终身教授、“麦克阿瑟天才奖”获得者穆来纳森和普林斯顿大学心理学教授埃尔德·沙菲尔（Eldar Shafir）推出了行为经济学的重磅著作《稀缺》，首度提出“带宽 = 认知能力 + 执行控制力”，并指出处于稀缺状态中的人的大脑会被稀缺心态俘获，认知能力与执行控制力会变得低下。同时，书中还提出了一些改善稀缺心态的方法。该书中文简体字版由湛庐引进、浙江人民出版社 2018 年出版。——编者注

训练了复杂的人工智能模型来执行相同的任务。研究团队获得了更大的数据集——包含 758 027 个保释裁定的案例库。对于每种情况，研究团队可以获得和法官一样的信息：被告的罪行、犯罪记录、未按时出庭受审的次数等。除年龄外，参与训练的算法没有其他任何人口统计学信息适合使用。对于每一起案件，研究人员还知道关于被告是否被释放，以及他如果被释放，之后是否会按时出庭或被重新逮捕（被告中有 74%的人获得保释，其中 15%的人在那之后没有按时出庭，26%的人则被重新逮捕）的信息。研究人员利用这一数据来训练一个机器学习算法，并评估了该算法的表现。该模型是通过机器学习构建的，因此并不限于线性组合。如果它在数据中检测到更复杂的规律，它就会使用此模式来改进预测。

该模型用于预测嫌疑人在保释期逃脱的风险，因此将风险量化为数字，而非只产生是否准予保释的决定。这种方法确定了最大可接受风险的阈值，即如果风险高于该阈值，就应该拒绝保释。然而，研究人员发现，无论如何设置风险阈值，使用该模型的预测得分都高于法官的预测。穆来纳森的团队计算得出，如果将风险阈值设置为一个值，使模型预测的拒绝保释人数与法官判决的拒绝保释人数相同，则犯罪率最多可降低 24%，个中原因在于，被关押的人最有可能再次犯罪。相反，如果将风险阈值设置为使该模型在不提高犯罪率的情况下，尽可能减少被拒绝保释的人数，则研究人员计算得出，被羁押的人数最多可再减少 42%。换句话说，机器学习模型在预测哪些被告属于犯罪高风险人群方面，表现要比法官好得多。

利用机器学习建立的模型，也比使用相同信息的线性模型成功得多，原因很有趣：机器学习算法在变量组合中发现了一些会被线性模

型遗漏的重要信息。算法能对风险最高的被告进行归类，就证明它有能力找到很容易被其他模型忽略的模式。换句话说，数据中的某些模式尽管很少见，却非常准确地预测出了高风险人群。利用算法找到罕见但具有决定性作用的模式，让我们想起了“断腿”的概念。

研究人员还使用该算法为每位法官构建了模型，类似于我们在第 9 章中描述的判断模型（但不限于简单线性组合）。他们将这些模型应用于整个数据集，使团队能够模拟法官在遇到相同案件时可能做出的判决，并比较这些判决。结果表明，保释裁定中存在相当大的系统噪声，其中一些是水平噪声：根据宽容程度对法官进行分类时，20% 最宽容的法官（即保释率最高的前 20% 的法官）准予保释的概率为 83%，而 20% 最严厉的法官准予保释的概率为 61%。法官对于哪些被告具有较高逃脱风险的判断方式也大不相同，被一位法官视为具有低逃脱风险的被告，可能被另一位更严厉的法官视为具有高逃脱风险。这些结果为模式噪声提供了清晰的证据。更详细的分析表明，案例之间的变异占总变异的 67%，系统噪声占 33%。系统噪声包括一些水平噪声，即平均严厉程度之间的差异，但其中大多数（79%）是模式噪声。

幸好，机器学习程序的高准确性并不以牺牲法官追求的其他目标，如种族平等为代价。从理论上讲，尽管该算法不使用种族相关数据，但它也可能会无意间加剧种族歧视。如果模型使用与种族信息高度相关的预测因素（如邮政编码），或是用于算法训练的数据源暗含偏见，则可能会出现种族歧视。例如，如果将过去的逮捕次数作为预测因素，而过去的逮捕次数受到种族歧视的影响，那么得到的算法也会存在歧视问题。

尽管从原则上讲，这种歧视无疑是一种风险，但在一些重要层面，该算法所做出的决策比法官群体中存在的种族歧视要轻微。例如，如果通过设置风险阈值使犯罪率与法官判决的犯罪率相同，则该算法可将有色人种被判入狱的概率减少 41%。在其他情况下，算法也得出了类似的结果，即提高准确性不必以加剧种族歧视为代价。正如研究小组所指出的：通过训练，该算法很容易用于减少种族歧视。

另一项不同领域的研究阐述了算法如何在提高准确性的同时减少歧视。哥伦比亚商学院教授博·考吉尔（Bo Cowgill）考察了一家大型科技公司招聘软件工程师的情况。考吉尔并未使用人工筛选简历的方式来筛选可进入面试流程的人，而是基于该公司收到并评估过的超过 30 万份简历，来训练机器学习算法进行筛选。该算法选出的候选人比人工筛选的候选人被录取的可能性要高 14%。当候选人收到录取通知后，算法组筛选出来的候选人，比人工组筛选出的候选人接受工作机会的可能性要高 18%。该算法还根据种族、性别和其他指标选择了一组更加多样化的候选人，而它更有可能选择“非传统”候选人，例如非名校毕业生、缺乏相关工作经验以及没有推荐信的候选人。在筛选软件工程师的简历时，人们通常倾向于选择符合这一群体所有典型特征的人，而该算法则为每个相关预测因素赋予了适当的权重。

需要明确的是，这些例子并不能证明算法始终是公平、无偏见和非歧视的。大家比较熟悉的一个例子是：一个用于预测求职者能否通过面试的算法，实际上是根据过去的晋升决策数据训练出来的，因此，这种算法必然会重蹈过去晋升决策中人类所有偏差的覆辙。

构建一个使种族或性别不平等持续存在的算法，不仅是可能的，

而且十分容易做到。许多算法已做到了这一点。这些例子表明，人们越来越关注算法决策中的偏见，但是，在得出关于算法的一般性结论之前，我们应当记住：**某些算法不仅比人类判断更准确，而且也更公平。**

为什么我们不更多地利用规则

通过这一简短的机械性决策之旅，我们总结出，各种规则之所以会超越人类判断，有两个原因。首先，如第 9 章所述，不仅仅是最新的和更为复杂的技术，所有机械性预测技术都能显著改善人类的判断。个性化的模式和情境噪声的结合会极大地影响人类判断的品质，因为简单的规则和无噪声是提高决策品质的关键。**明智的简单规则比人类的判断要好很多。**

其次，当数据足够丰富时，我们可以用复杂的人工智能技术找出有效的模式，并使其预测力远超简单模型。这些模型相对于人类判断的优势在于，它们不仅没有噪声，而且还具有利用更多信息的能力。

既然算法具有如此多的优点，得到大量证据的支持，那么为什么我们在本书中讨论的各种类型的专业判断，没有广泛地使用算法呢？尽管对算法和机器学习的讨论很热烈，但人们对它们的使用仍然很有限（一些特定领域除外）。许多专家不关心诊断性判断与机械性判断孰优孰劣，而是相信自己的判断。他们对自己的直觉充满信心，并对机器能否做得更好持怀疑的态度。他们将算法决策视为不人道的，认为使用算法是一种放弃责任的表现。

尽管算法决策已取得了令人瞩目的进步，但是在诸如医学诊断等领域，使用算法仍然不是惯常的做法，也很少有企业在招聘和晋升决策中使用算法。好莱坞电影制作公司的高管们是根据自己的经验判断而非公式来选择拍摄哪部电影的；图书出版商也在做同样的事情。而且，正如迈克尔·刘易斯（Michael Lewis）的畅销书《点球成金》（*Moneyball*）所讲述的那样，人们之所以对痴迷于统计的奥克兰田径队的故事印象深刻，恰恰是因为算法在运动团体中的运用是一种例外而非常规。即使在今天，教练、经理人以及与他们共事的其他人通常也更相信自己的直觉，并坚持认为统计分析不可能取代人类自身良好的判断力。

梅尔和他的合著者在 1996 年的一篇论文中，列举了精神科医生、医师、法官和其他专业人士反对机械性判断的至少 17 种理由，并对这些理由进行了驳斥。他们得出的结论是，需要结合社会心理因素来解释临床医生对这类判断的排斥，这些因素包括“对技术性失业的恐惧”“了解不足”和“对计算机的普遍厌恶”。

从那时起，研究人员已经确定了导致这种排斥的其他因素。我们不打算在这里对该研究进行完整的回顾，本书的目标是为改善人类判断提供建议，而不是像弗兰克尔法官那样，主张“用机器取代人类”。

但是，关于哪些因素会导致人类抵触机械性预测，其中的一些发现与我们对人类判断的讨论有关。最近的一项研究得出了一个重要观点：人们对算法并非全盘否定。例如，当从人类的建议和算法的建议之间进行选择时，人们通常会选择后者。对算法的抵制或厌恶并不代表一味地拒绝采用新的决策支持工具，人们愿意给算法机会，而一旦

发现它会犯错误，就不会再信任它。

这种反应似乎是明智的：为什么要在你不信任的算法上浪费精力呢？作为人类，我们敏锐地意识到自己会犯错误，但这是我们不准备分享的特权，我们希望机器是完美的，如果机器不完美，那就丢弃它。

由于存在这种直觉性的期望，人们仍可能不信任算法，而继续相信自己的判断力，即使自己的判断明显不尽如人意。这种态度是根深蒂固的，除非算法能够达到近乎完美的预测准确性，否则这是不可能改变的。

幸好，可改进规则和算法的相关因素同样可用于改善人类的判断。我们不能奢望能够像人工智能模型一样有效地利用信息，但是至少可以努力模仿简单模型的简单性和无噪声性。在一定程度上，我们可以采用减少系统噪声的方法来提高预测判断的品质。如何改善我们的判断力是本书第五部分的主题。

没有噪声的规则和算法

- 当有大量数据时，机器学习算法比人和简单模型的预测要好。相比于人类的判断，即使是最简单的规则和算法也具有很大的优势，因为它们没有噪声，也不会尝试将复杂而无效的因素用于做预测。

- 既然缺少预测结果所需的数据，那么为什么不使用均等权重模型？它几乎和最合适的回归模型一样好，且比人类视情况而定的判断更胜一筹。

- 你不认同该模型的预测？是由于这里有“断腿”的例外情况，还是单纯因为不喜欢这个预测？

- 算法当然也会犯错，但是如果人类犯的错更多，那么我们应该相信谁？