HZ BOOKS
华章教育

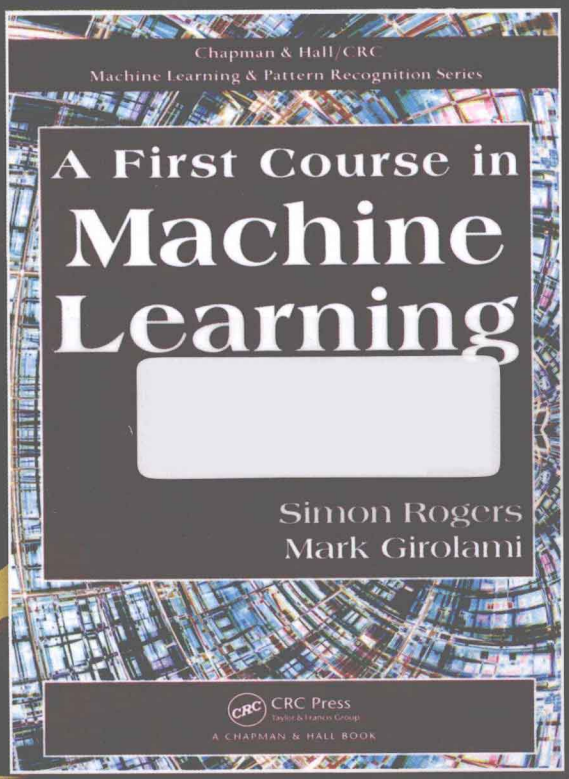
计 算 机 科 学 丛 书

CRC Press
Taylor & Francis Group

机器学习基础教程

（英） Simon Rogers Mark Girolami 著
郭茂祖 王春宇、刘扬 刘晓燕 译

A First Course in Machine Learning



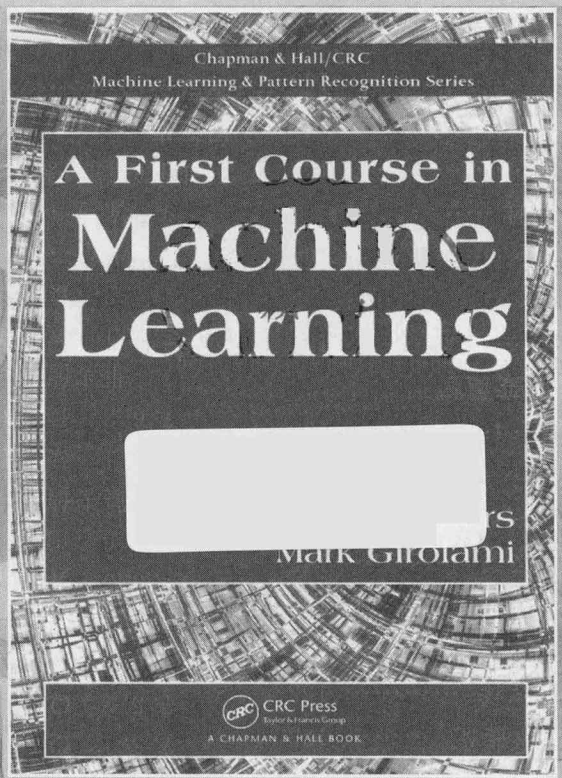
机械工业出版社
China Machine Press

计 算 机 科 学 丛 书

机器学习基础教程

（英） Simon Rogers Mark Girolami 著
郭茂祖 王春宇 刘扬 刘晓燕 译

A First Course in Machine Learning



机械工业出版社
China Machine Press

图书在版编目（CIP）数据

机器学习基础教程/（英）罗杰斯（Rogers，S.），（英）吉罗拉米（Girolami，M.）著；郭茂祖等译．—北京：机械工业出版社，2013.10
（计算机科学丛书）
书名原文：A First Course in Machine Learning

ISBN 978-7-111-40702-7

Ⅰ.机… Ⅱ.①罗… ②吉… ③郭… Ⅲ.机器学习-教材 Ⅳ.TP181

中国版本图书馆 CIP 数据核字（2013）第 109878 号

本书介绍机器学习技术及应用的主要算法，重点讲述理解主流的机器学习算法所需的核心数学和统计知识。书中介绍的算法涵盖机器学习的主要问题：分类、聚类和投影。由于本书是机器学习基础课程的教材，所以尽量减少了数学难度，仅对一小部分重要算法给出详细的描述和推导，而对大部分算法仅给出简单介绍，目的在于使学生打好基础，增强信心和兴趣，鼓励他们进一步学习该领域的高级主题或从事相关研究工作。

本书是机器学习导论课程教材，适合作为计算机、自动化及相关专业高年级本科生或研究生的教材，也可供研究人员和工程技术人员参考。

机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码 100037）
责任编辑：盛思源
藁城市京瑞印刷有限公司印刷
2014 年 1 月第 1 版第 1 次印刷
185mm×260mm·12.5 印张
标准书号：ISBN 978-7-111-40702-7
定 价：45.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991 88361066 投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259 读者信箱：hzjsj@hzbook.com

文艺复兴以降，源远流长的科学精神和逐步形成的学术规范，使西方国家在自然科学的各个领域取得了垄断性的优势；也正是这样的传统，使美国在信息技术发展的六十多年间名家辈出、独领风骚。在商业化的进程中，美国的产业界与教育界越来越紧密地结合，计算机学科中的许多泰山北斗同时身处科研和教学的最前线，由此而产生的经典科学著作，不仅擘划了研究的范畴，还揭示了学术的源变，既遵循学术规范，又自有学者个性，其价值并不会因年月的流逝而减退。

近年，在全球信息化大潮的推动下，我国的计算机产业发展迅猛，对专业人才的需求日益迫切。这对计算机教育界和出版界都既是机遇，也是挑战；而专业教材的建设在教育战略上显得举足轻重。在我国信息技术发展时间较短的现状下，美国等发达国家在其计算机科学发展的几十年间积淀和发展的经典教材仍有许多值得借鉴之处。因此，引进一批国外优秀计算机教材将对我国计算机教育事业的发展起到积极的推动作用，也是与世界接轨、建设真正的世界一流大学的必由之路。

机械工业出版社华章公司较早意识到“出版要为教育服务”。自 1998 年开始，我们就将工作重点放在了遴选、移译国外优秀教材上。经过多年的不懈努力，我们与 Pearson，McGraw-Hill，Elsevier，MIT，John Wiley & Sons，Cengage 等世界著名出版公司建立了良好的合作关系，从他们现有的数百种教材中甄选出 Andrew S. Tanenbaum，Bjarne Stroustrup，Brain W. Kernighan，Dennis Ritchie，Jim Gray，Afred V. Aho，John E. Hopcroft，Jeffrey D. Ullman，Abraham Silberschatz，William Stallings，Donald E. Knuth，John L. Hennessy，Larry L. Peterson 等大师名家的一批经典作品，以“计算机科学丛书”为总称出版，供读者学习、研究及珍藏。大理石纹理的封面，也正体现了这套丛书的品位和格调。

“计算机科学丛书”的出版工作得到了国内外学者的鼎力襄助，国内的专家不仅提供了中肯的选题指导，还不辞劳苦地担任了翻译和审校的工作；而原书的作者也相当关注其作品在中国的传播，有的还专程为其书的中译本作序。迄今，“计算机科学丛书”已经出版了近两百个品种，这些书籍在读者中树立了良好的口碑，并被许多高校采用为正式教材和参考书籍。其影印版“经典原版书库”作为姊妹篇也被越来越多实施双语教学的学校所采用。

权威的作者、经典的教材、一流的译者、严格的审校、精细的编辑，这些因素使我们的图书有了质量的保证。随着计算机科学与技术专业学科建设的不断完善和教材改革的逐渐深化，教育界对国外计算机教材的需求和应用都将步入一个新的阶段，我们的目标是尽善尽美，而反馈的意见正是我们达到这一终极目标的重要帮助。华章公司欢迎老师和读者对我们的工作提出建议或给予指正，我们的联系方法如下：

华章网站：www. hzbook. com
电子邮件：hzjsj@hzbook. com
联系电话：(010) 88379604
联系地址：北京市西城区百万庄南街 1 号
邮政编码：100037

华章教育
华章科技图书出版中心

译者序

A First Course in Machine Learning

机器学习起初只是人工智能（AI）的一个研究分支，随着其他研究分支的成熟发展或逐步淡化，目前机器学习发展成为 AI 中最具活力的研究方向。一方面它源于机器学习，已经成为人工智能理论研究与应用研究的桥梁；另一方面，随着计算机技术的发展，机器学习也日益成为计算机科学的重要研究领域之一。此外，模式识别与数据挖掘的核心算法大多也与机器学习有关。

机器学习作为人工智能理论研究的一部分，需要一定的数学知识作为基础。本书就是为计算机等信息类专业的学生理解最流行的机器学习算法提供核心数学知识和统计技术。本书并没有面面俱到地介绍所有的机器学习算法，而是给出部分代表性算法的核心思想及详细描述。最后，本书主要涉及基于示例的归纳学习，至于神经网络等进化学习以及关于 agent 与环境交互的强化学习这两大类机器学习的相关内容，请读者参阅其他书籍。

本书共 7 章。第 1、2 章介绍如何选择线性模型参数以对观测数据做出预测。第 1 章给出通过最小化损失函数来学习模型参数的方法。第 2 章介绍最大似然函数的方法。第 3 章介绍机器学习中的贝叶斯方法。第 4 章介绍计算后验的三种近似方法。第 5 章及后续各章涉及机器学习领域分类、聚类和预测方面的主要算法，其中第 5 章关注监督学习；第 6、7 章介绍无监督学习，第 6 章研究 K 均值和混合模型两种聚类方法，第 7 章介绍通过将高维数据投影到一个低维空间，对数据进行可视化或特征选择的方法。本书还包括词汇表和索引。

本书适合作为高等院校计算机、自动化等专业本科生及研究生的机器学习教材。同时，本书也是机器学习领域的研究者或者那些想了解和应用当前机器学习技术的工作人员的一本宝贵的参考资料。

本书的翻译工作由郭茂祖主持，郭茂祖审校了全部译稿，邢林林负责校对。其中，郭茂祖翻译了前言和第 1 章，王春宇翻译了第 2、3 章，刘扬翻译了第 4、5 章和词汇表、索引，刘晓燕翻译了第 6、7 章。在本书的翻译过程中，王娟、刘茹、徐云刚、滕志霞、李艳娟、车凯、程爽、史文丽、孟宪伟、代启国、李晋、吴伟宁、徐立秋给予了很多帮助，对他们表示由衷的感谢。

目前机器学习日益成为计算机科学重要的实践、研究与开发领域之一，一方面这反映在它的学术研究规模上，另一方面反映在新的机器学习从业人员遍布于主要的国际银行和金融机构，以及微软、谷歌、雅虎和亚马逊等公司。

从某种角度来讲，这种发展源于人们对世界认知方式的数量和种类的增加。一个特别显著的例子是，在首个基因组测序完成之前，不断涌现出了各种生物检测新技术。不久前，检测生物体的复杂分子状态是难以想象的，因为这已经远远超出了我们的认识能力。现在，机器学习方法在生物体中有用分子结构提取方面的广泛应用，使其成为可能。

本书改编自英国格拉斯哥大学计算机科学学院机器学习课程的讲义，该课程包括 20 学时的授课和 10 学时的实验，面向高年级本科生开设并由研究生讲授。如此少的教学时数不可能涵盖机器学习所有的内容，所以该课的目的是为理解流行的机器学习算法提供核心数学知识和统计技术，并描述其中部分算法，这些算法涵盖了机器学习中的分类、聚类和投影等主要问题。通过本课程的学习，学生应该具备通过考察机器学习相关文献来寻求适合他们所需方法的知识和能力，希望本书的读者也能做到这一点。

鉴于选学该课学生的数学水平参差不齐，我们只假定需要很少的数学知识，计算机科学、工程类、物理学（或其他数值处理类学科）的本科生阅读本书应该没有问题，没有以上经历的读者也可以阅读本书，因为穿插在文中的注解框内给出了相应的数学解释。此外，突出强调了重要公式（公式加阴影），在继续阅读前，花些时间理解这些公式是值得的。

选学该课的学生通常会发现其中的实践环节非常有用，实验有助于将涉及的各种算法和概念由抽象的等式转化为解决实际问题的工具。我们已通过大量的 MATLAB® /Octave[⊖] 软件脚本完成以上转化，这些脚本可通过相关的网页并参考本书正文获得，利用它们读者能够重新绘制书中的插图，并研究如何改变模型说明和参数取值。

最后，本书选择的机器学习方法是我们认为学生应该掌握的，在有限的篇幅和时间内，更有必要给出一小部分算法的详细描述和研究进展，而不是泛泛地描述许多算法，因而多数读者在本书中可能找不到他们最喜欢的算法！

Simon Rogers
Mark Girolami

⊖ 免费数学软件环境，源于 www. gnu. org/software/octave/。

目 录

A First Course in Machine Learning

第 1 章

A First Course in Machine Learning

线性建模：最小二乘法

在有着广泛应用的机器学习中，一个重要且普遍的问题是学习或者推断属性变量与相应的响应变量或目标变量之间的函数关系，使得对任何一个属性集合，我们可以预测其响应。例如，我们可能想要建立一个能够执行疾病诊断的模型。为了构建这个模型，需要使用一个数据集，这个数据集是从已知疾病状态（响应，健康或患病）的患者中得到的测量（属性，如血压、心率、体重等）的集合。在完全不同的例子中，我们希望给顾客提出建议。在这种情况下，我们能够建立一个关于某个顾客以前买过物品的描述（属性）和该顾客最终是否喜欢该产品（响应）的**模型**。这个模型可以帮助我们预测顾客可能喜欢的物品，并因此进行推荐。这一章将涉及许多更重要的应用领域。

1.1 线性建模

首先，通过一个实际例子来考虑机器学习最直接的学习问题——线性建模[⊖]：在属性与响应之间学习**线性**关系。图 1-1 显示了从 1896 年开始，每次奥林匹克运动会（简称奥运会）男子 100 米比赛赢得金牌所需的比赛时间。我们的目标是用这些数据学习一个函数模型，此模型依赖于奥运会举办年份和 100 米获胜时间，并且用这个模型预测将来比赛中的获胜时间。显然，年份并不是影响获胜时间的唯一因素，如果我们认真对待这个预测，可能还会考虑其他因素（例如，主要参赛者的最近情况）。然而，通过图 1-1 可以看出，年份和获胜时间之间至少存在一个统计关系（它不可能是因果关系——时间的流逝并不是获胜时间下降的直接原因），并且这个例子足以帮助我们引入和发展线性建模的主要思想。

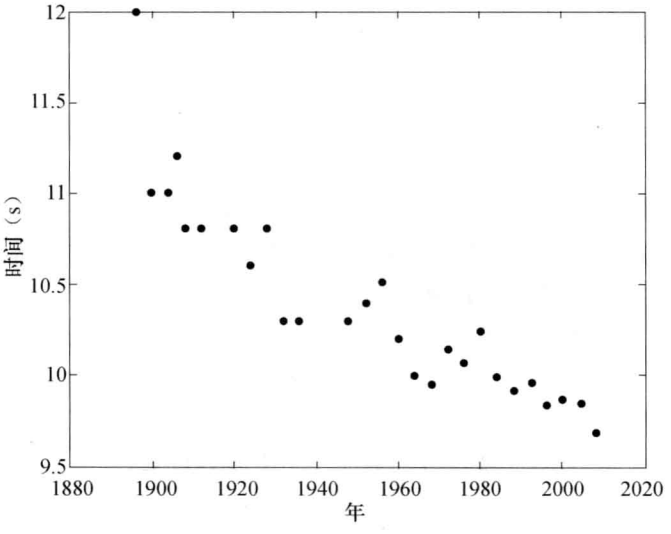


图 1-1 从 1896 年开始，夏季奥运会男子 100 米的获胜时间。注意：在 1914 年、1940 年和 1944 年，由于两次世界大战而中断了这个比赛

⊖ 这里将要考虑的模型类型称为回归，它最初被 Francls Galton（1877 年）用在遗传学方面。当时 Francls Galton 研究智力如何从一代传到一代（或者不是这样，由于这种情况也是可能的）。此术语后来被在统计背景下发展 Galton 工作的统计学家所采用。

1.1.1 定义模型

首先将模型定义为一个将输入属性（在这个例子中，是举办奥运会的年份）映射到输出或者目标值（获胜时间）的函数。对于属性，我们用年份的数值（如1980），尽管还有另外一个公式（例如，从第一届运动年开始，1980－1896＝84），这对潜在的假设没有实质差别。

有许多**函数**可以定义这个映射。一般地，这个函数将以 x（奥运会年份）为输入，并且将返回 t（用秒表示的获胜时间）。也就是说，t 是 x 的函数。数学上，把这个记为 $t=f(x)$。在有些情况下，我们需要知道的是用来评估函数的 x。例如，如果 $f(x)=\sin(x)$，或者 $f(x)=x$，那么对任何 x，我们可以计算 t。一般地，我们需要更灵活并且我们的模型可能有一个相关**参数**的集合。例如，$t=ax$ 有一个参数 a，此参数需要用某种方法定义。在机器学习中，从一个合适的数据集中学习模型参数是一个普遍的问题。我们将用 $t=f(x;a)$ 来
2 表示 x 与参数 a 之间的函数 $f(\cdot)$。

1.1.2 模型假设

为了便于选择特定的模型来使用，我们需要做一些假设。在这个阶段的初始假设是：

x 与 t 之间的关系是线性的（参见注解1.1）。

> **注解1.1（线性关系）：** 等式
>
> $$y=mx+c$$
>
> 这里 m 和 c 是常量，在 x 和 y 之间定义了一个线性关系。它称为是线性的，因为从直观上看，在 x 与 y 之间的关系呈一条直线。下面的等式是非线性的，因为变量 x 和 y 的形式更复杂：
>
> $$y=mx^2+c^2, y=\sin(x), \sqrt{y}=mx+c$$
>
> m 和 c 的值不影响关系的线性性。例如，如下都表示 x 和 y 之间的线性关系：
>
> $$y=mx+c^2, y=x\sin(m)+c$$

或者可以表述为：

图1-1中的数据可以用一条直线模拟。

或者：

每 M 年，获胜时间下降相同数量。

观察图1-1，我们可以看到这个假设并不是完全满足。然而，我们希望它是一个可用的模型，并且它可以对将来的获胜时间做出预测。

满足我们假设的最简单模型是

$$t=f(x)=x$$

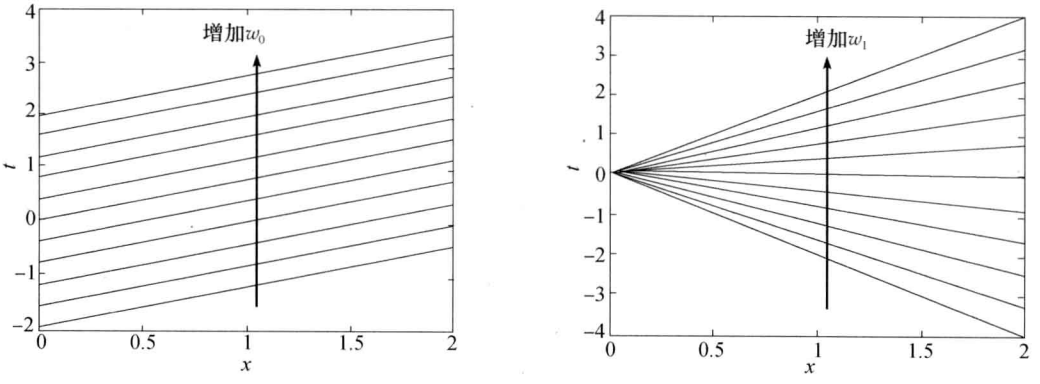
获胜时间等于奥运会年份。x 大于等于1880，t 小于等于12，随着年份的增长获胜时间在下降，这个事实说明这个模型是不适当的。添加一个单参数得到：

$$t=f(x;w)=wx$$

这里 w 为正或者负。这个改进的模型产生了一条直线，通过选择 w，可以使这条直线有任何梯度。这个模型在灵活性（flexibility）方面有所提升，但是它仍然是受限制的，因为在奥运会年份0年时，模型预测的获胜时间是 $w\times0=0$。通过这个数据可以看出，这是不现实的——按照数据的一般趋势，在0年时，获胜时间实际上应该是一个相当大的数。通过对模
3 型添加多个参数，可以克服这个限制：

$$t = f(x; w_0, w_1) = w_0 + w_1 x \tag{1-1}$$

这是直线的标准等式，这个等式许多读者以前都遇到过。现在学习任务是用图 1-1 的数据为两个参数 w_0 和 w_1 选择合适的值。这两个参数常常认为是截距（w_0，直线与 t 轴的截距）和梯度（w_1，直线的梯度），以及改变它们的影响（effect），如图 1-2 所示（MATLAB 脚本：`plotlinear.m`）（参见练习 EX 1.1）。



a）w_0的增加改变了直线与t轴的相交点　　b）w_1的增加改变了直线的梯度

图 1-2　在式（1-1）定义的线性模型中，改变 w_0 和 w_1 带来的影响

1.1.3　定义什么是好的模型

为了选择在某种方式下最好的 w_0 和 w_1 值，我们需要定义最好的意义是什么。常识表明所谓最好的解是由 w_0 和 w_1 的一些值组成，这些值可以产生一条能尽可能与所有数据点接近的直线。衡量一个特定模型与数据点接近程度的普遍方法是真正的获胜时间与模型预测的获胜时间之间的平方差。用 x_n、t_n 分别表示第 n 次的奥运会年份和获胜时间，平方差定义为：

$$(t_n - f(x_n; w_0, w_1))^2$$

这个数值越小，模型在 x_n 处越接近 t_n。对差值取平方是很重要的。如果不这样做，就可以通过连续增加 $f(x_n; w_0, w_1)$ 来无限减小这个量。

这个表达称为平方损失函数（squard loss function），因为它描述了使用 $f(x_n; w_0, w_1)$ 模拟 t_n 所损失的精度。在本章中，我们用 $\mathcal{L}_n()$ 表示损失函数。在这种情况下，

$$\mathcal{L}_n(t_n, f(x_n; w_0, w_1)) = (t_n - f(x_n; w_0, w_1))^2 \tag{1-2}$$

是 n 年的损失。损失总是正的，并且损失越小，函数描述这个数据就越好。由于对于所有的 N 年，我们想有一个低的损失，所以考虑在整个数据集上的平均损失，即

$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N} \mathcal{L}_n(t_n, f(x_n; w_0, w_1)) \tag{1-3}$$

这是每 N 年的平均损失值。它越低越好。因此我们将调整 w_0 和 w_1 值来产生一个模型，此模型得到平均损失的最低值 $\mathcal{L}$。寻找 w_0 和 w_1 的最好值，用数学表达式可以表示为

$$\arg\min_{w_0, w_1} \frac{1}{N}\sum_{n=1}^{N} \mathcal{L}_n(t_n, f(x_n; w_0, w_1))$$

argmin 项是数学上“找到最小化参数”的缩写。在这个例子中，参数是 w_0 和 w_1 的值同时

最小化的表达式是平均损失。图 1-3 显示了一个假设的损失，它是单参数 w 的函数。使 $\mathcal{L}$ 最小的参数 w 的值是 $w=5$。历史上，平方损失的最小化是函数估计的最小二乘误差法的基础，它是由 Gauss 和 Legendre（1809 年）在预测行星运动时发展的方法。

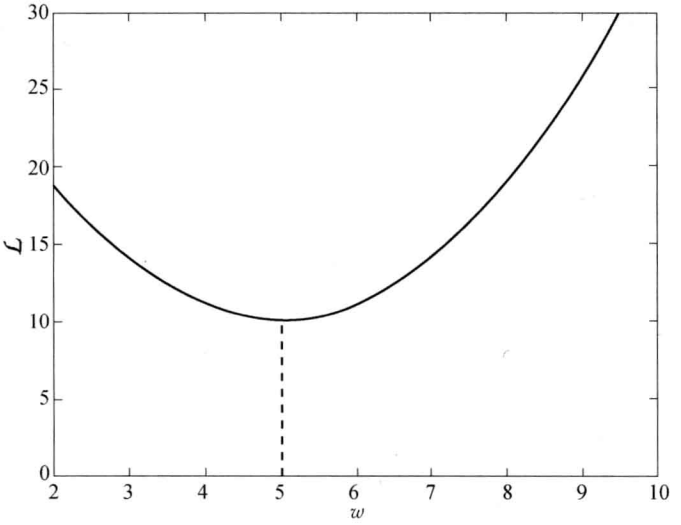


图 1-3 单参数（w）损失函数的例子。虚线表明了 $w=5$ 时损失最小

其他的损失函数适合回归。例如，另一个常用的是绝对损失：

$$\mathcal{L}_n = | t_n - f(x_n; w_0, w_1) |$$

平方损失是非常常见的选择，部分上由于它找到 w_0 和 w_1 的最好值相对直接这一事实——我们可以得到一个**分析**解。然而，现代计算能力已经降低了数学方便的重要性——在多个适合的数据上选择一个方便的损失函数不再有任何借口。显然，我们的目标是介绍对平方损失合适的通用模型概念。值得注意的是，在许多情况下，还有其他一些模型是可行的并且可能是更适合的。

5

1.1.4 最小二乘解：一个有效的例子

简要说明我们的数据集由 $n=1, \cdots, N$ 观测值构成，它们中的每一个由一个年 x_n 和时间（秒）t_n 构成。

我们继续尽力寻找一个函数关系，此函数关系用一个线性模型定义为

$$f(x; w_0, w_1) = w_0 + w_1 x \tag{1-4}$$

我们决定将用最小二乘损失函数来选择适合的 w_0 和 w_1。用表达式中的线性模型替代平均损失，在括号外面相乘结果为

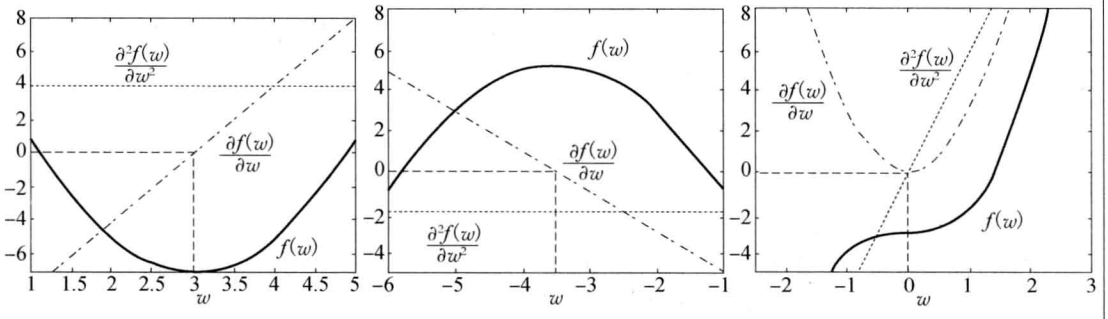
$$\begin{aligned}
\mathcal{L} &= \frac{1}{N}\sum_{n=1}^{N} \mathcal{L}_n(t_n, f(x_n; w_0, w_1)) \\
&= \frac{1}{N}\sum_{n=1}^{N} (t_n - f(x_n; w_0, w_1))^2 \\
&= \frac{1}{N}\sum_{n=1}^{N} (t_n - (w_0 + w_1 x_n))^2 \\
&= \frac{1}{N}\sum_{n=1}^{N} (w_1^2 x_n^2 + 2w_1 x_n w_0 - 2w_1 x_n t_n + w_0^2 - 2w_0 t_n + t_n^2)
\end{aligned}$$

$$= \frac{1}{N}\sum_{n=1}^{N}(w_1^2 x_n^2 + 2w_1 x_n(w_0 - t_n) + w_0^2 - 2w_0 t_n + t_n^2) \tag{1-5}$$

对损失函数求导数：在 $\mathcal{L}$ 的最小值点处，其关于 w_0 和 w_1 的**偏导数**一定是 0（参见注解 1.2）。因此，求出偏导数，使其等于 0 并对 w_0 和 w_1 求解，解 w_0 和 w_1 可以使我们得到最小值。从 w_1 开始，我们知道在式（1-5）中不包含 w_1 的项可以被忽略（由于这些项关于 w_1 的偏导数为 0）。去掉这些项得到

$$\frac{1}{N}\sum_{n=1}^{N}[w_1^2 x_n^2 + 2w_1 x_n w_0 - 2w_1 x_n t_n]$$

注解 1.2（拐点）：通过搜索使函数梯度 $\dfrac{\partial f(w)}{\partial w}$ 为 0 的点，可以找到函数 $f(w)$ 的拐点（可能对应于最小值）。为了确定一个拐点是最大值、最小值还是鞍点，需要检验其 2 阶导数 $-\dfrac{\partial^2 f(w)}{\partial w^2}$。在拐点 $\hat{w}$，如果其 2 阶导数是正的，那么这个拐点是最小值点。下面三个图显示了三个例子函数及其 1 阶和 2 阶导数：



一般地，一个函数可能有多个拐点。一个有趣的特殊情况是，如果函数的 2 阶导数是正的常量，那么这个函数仅有一个最小值点。

在求偏导数之前，我们重新排列这个表达式，使它更简单。尤其是，把没有下标 n 的项放在和的外面并重新排列得到的结果

$$w_1^2 \frac{1}{N}\Big(\sum_{n=1}^{N} x_n^2\Big) + 2w_1 \frac{1}{N}\Big(\sum_{n=1}^{N} x_n(w_0 - t_n)\Big)$$

如下表达式给出了其关于 w_1 的偏导数

$$\frac{\partial \mathcal{L}}{\partial w_1} = 2w_1 \frac{1}{N}\Big(\sum_{n=1}^{N} x_n^2\Big) + \frac{2}{N}\Big(\sum_{n=1}^{N} x_n(w_0 - t_n)\Big) \tag{1-6}$$

现在对 w_0 做相同的操作。去掉不含 w_0 的项后，得到

$$\frac{1}{N}\sum_{n=1}^{N}[w_0^2 + 2w_1 x_n w_0 - 2w_0 t_n]$$

另外，我们在求导之前重新排列它。将没有下标 n 的项移到和的外面（注意 $\sum_{n=1}^{N} w_0^2 = Nw_0^2$），结果为

$$w_0^2 + 2w_0 w_1 \frac{1}{N}\Big(\sum_{n=1}^{N} x_n\Big) - 2w_0 \frac{1}{N}\Big(\sum_{n=1}^{N} t_n\Big)$$

对 w_0 求偏导数得

$$\frac{\partial \mathcal{L}}{\partial w_0} = 2w_0 + 2w_1 \frac{1}{N}\Big(\sum_{n=1}^{N} x_n\Big) - \frac{2}{N}\Big(\sum_{n=1}^{N} t_n\Big) \tag{1-7}$$

导数等于 0：现在我们有了损失函数关于 w_0 和 w_1 的偏导数表达式。为了找到对应于拐点（希望是最小值点）的 w_0 和 w_1 值，必须使这些表达式为 0 并且对 w_0 和 w_1 求解。从关于 w_0 的表达式开始。将式（1-7）设置为 0 并且对 w_0 求解：

$$2w_0 + 2w_1 \frac{1}{N}\Big(\sum_{n=1}^{N} x_n\Big) - \frac{2}{N}\Big(\sum_{n=1}^{N} t_n\Big) = 0$$

$$2w_0 = \frac{2}{N}\Big(\sum_{n=1}^{N} t_n\Big) - w_1 \frac{2}{N}\Big(\sum_{n=1}^{N} x_n\Big)$$

$$w_0 = \frac{1}{N}\Big(\sum_{n=1}^{N} t_n\Big) - w_1 \frac{1}{N}\Big(\sum_{n=1}^{N} x_n\Big)$$

6 ~ 7

将平均获胜时间表示为 $\bar{t} = \frac{1}{N}\sum_{n=1}^{N} t_n$ 以及平均奥运会年份为 $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$，在拐点 $\hat{w}_0$ 处，可以重写 w_0 值的表达式为

$$\hat{w}_0 = \bar{t} - w_1 \bar{x} \tag{1-8}$$

我们从这个表达式可以洞悉到什么？这个新的表达式是初始表示（$t_n = w_0 + w_1 x_n$）的重新排列，这里 t_n 和 x_n 已经被平均值 $\bar{t}$ 和 $\bar{x}$ 取代。考虑在 N 个数据点上的平均函数值，表达式如下：

$$\frac{1}{N}\sum_{n=1}^{N} f(x_n; w_0, w_1) = \frac{1}{N}\sum_{n=1}^{N} (w_0 + w_1 x_n) = w_0 + w_1 \bar{x}$$

平均获胜时间通过 $\bar{t}$ 给出，因此在式（1-8）中，选择 $\hat{w}_0$ 来确保函数的平均值等于平均获胜时间。直观地，用这种方式匹配的平均值似乎是非常有意义的。

在我们用式（1-6）得到关于 $\hat{w}_1$（w_1 在拐点处的值，见注解 1.2）的表达式之前，值得简要地检验它的 2 阶导数，以确保它是最小值点。再一次对式（1-6）关于 w_1 求导并对式（1-7）关于 w_0 求导，结果为：

$$\frac{\partial^2 \mathcal{L}}{\partial w_1^2} = \frac{2}{N}\sum_{n=1}^{N} x_n^2$$

$$\frac{\partial^2 \mathcal{L}}{\partial w_0^2} = 2 \tag{1-9}$$

这两个量一定都是正的。这说明它仅有一个拐点并且此拐点对应于损失函数的最小值。

我们将此过程应用于关于 $\hat{w}_0$（最小化损失函数的 w_0 的值）的表达式中。这个表达式依赖于 w_1，这暗示了对于特定的 w_1，我们知道最好的 w_0。式（1-6）用我们的表达式替代最好的 w_0（式（1-8））并重新排列，我们得到仅含 w_1 项的表达式：

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial w_1} &= w_1 \frac{2}{N}\Big(\sum_{n=1}^{N} x_n^2\Big) + \frac{2}{N}\Big(\sum_{n=1}^{N} x_n(\hat{w}_0 - t_n)\Big) \\
&= w_1 \frac{2}{N}\Big(\sum_{n=1}^{N} x_n^2\Big) + \frac{2}{N}\Big(\sum_{n=1}^{N} x_n(\bar{t} - w_1\bar{x} - t_n)\Big) \\
&= w_1 \frac{2}{N}\Big(\sum_{n=1}^{N} x_n^2\Big) + \bar{t}\,\frac{2}{N}\Big(\sum_{n=1}^{N} x_n\Big) - w_1\bar{x}\,\frac{2}{N}\Big(\sum_{n=1}^{N} x_n\Big) - \frac{2}{N}\Big(\sum_{n=1}^{N} x_n t_n\Big)
\end{aligned}$$

依然用 $\bar{x} = (1/N)\sum_{n=1}^{N} x_n$ 来简化这个表达式并合并包含 w_1 的项：

$$\frac{\partial \mathcal{L}}{\partial w_1} = 2w_1\Big[\Big(\frac{1}{N}\Big(\sum_{n=1}^{N} x_n^2\Big)\Big) - \bar{x}\,\bar{x}\Big] + 2\bar{t}\,\bar{x} - 2\,\frac{1}{N}\Big(\sum_{n=1}^{N} x_n t_n\Big)$$

8

最后，通过将这个偏导数设置为 0，我们能得到关于 $\hat{w}_1$ 的表达式并且对 w_1 求解：

$$2w_1\left[\left(\frac{1}{N}\left(\sum_{n=1}^{N}x_n^2\right)\right)-\overline{x}\,\overline{x}\right]+2\overline{t}\,\overline{x}-2\frac{1}{N}\left(\sum_{n=1}^{N}x_nt_n\right)=0$$

$$2w_1\left[\left(\frac{1}{N}\left(\sum_{n=1}^{N}x_n^2\right)\right)-\overline{x}\,\overline{x}\right]=2\frac{1}{N}\left(\sum_{n=1}^{N}x_nt_n\right)-2\overline{t}\,\overline{x}$$

$$\widehat{w}_1=\frac{\frac{1}{N}\left(\sum_{n=1}^{N}x_nt_n\right)-\overline{t}\,\overline{x}}{\frac{1}{N}\left(\sum_{n=1}^{N}x_n^2\right)-\overline{x}\,\overline{x}}$$

现在定义一些新的平均量是非常有用的。第一个，$(1/N)\sum_{n=1}^{N}x_n^2$ 是数据的平均平方值并且我们把它记为 $\overline{x^2}$。注意，这个量与 $(\overline{x})^2$ 不同。第二个是 $(1/N)\sum_{n=1}^{N}x_nt_n$（同样，它与 $\overline{x}\,\overline{t}$ 不同）。我们将它记为 $\overline{xt}$。将这些在关于 w_1 的表达式中替换，得到：

$$\widehat{w}_1=\frac{\overline{xt}-\overline{x}\,\overline{t}}{\overline{x^2}-(\overline{x})^2} \tag{1-10}$$

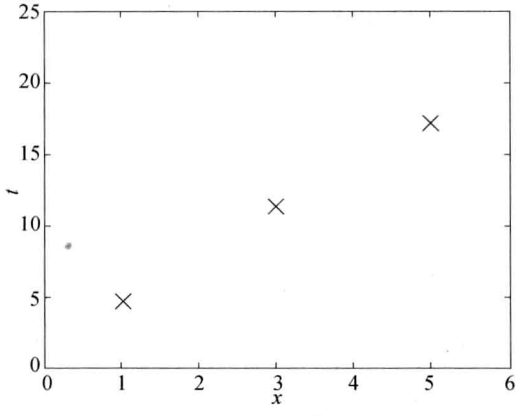
式（1-10）和式（1-8）为计算最好的参数值提供了全部所需的一切。首先用式（1-10）的 $\widehat{w}_1$ 替换式（1-8）来计算 $\widehat{w}_0$（MATLAB 脚本：`fitlinear.m`）。

1.1.5　有效的例子

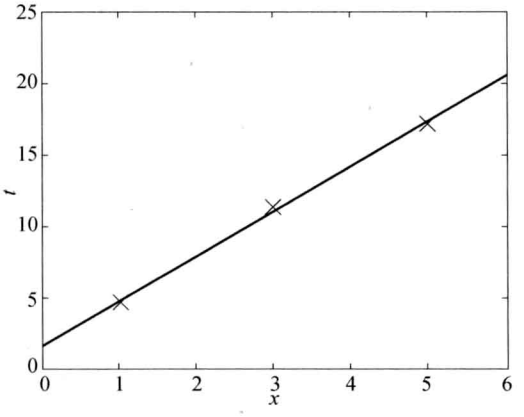
在用线性模型拟合奥运会数据之前，在一个更小数据集上提供一个有效的例子是非常有用的。假设我们观察到 $N=3$ 个数据点，如表 1-1 所示。最后一行给出了计算 $\widehat{w}_0$ 和 $\widehat{w}_1$ 所需的各种平均值：$\overline{x}$、$\overline{t}$、$\overline{xt}$ 和 $\overline{x^2}$。图 1-4 画出了这 3 个数据点。

表 1-1　线性回归例子的合成数据集

n	x_n	t_n	x_nt_n	x_n^2
1	1	4.8	4.8	1
2	3	11.3	33.9	9
3	5	17.2	86	25
$(1/N)\sum_{n=1}^{N}$	3	11.1	41.57	11.67



a）在表1-1中描述的3个合成数据点



b）由$f(x;w_0,w_1)=1.8+3.1x$定义的最小二乘拟合

图 1-4　1.1.5 节中有效的例子中的数据和函数

将这些值代入式（1-10），得到：

$$w_1 = \frac{41.57 - 3 \times 11.1}{11.67 - 3 \times 3} = \frac{8.27}{2.67} = 3.1$$

和

$$w_0 = 11.1 - 3.1 \times 3 = 1.8$$

因此最好的线性函数是：

$$f(x; w_0, w_1) = 1.8 + 3.1x$$

并且如图 1-4b 所示。

1.1.6 奥运会数据的最小二乘拟合

表 1-2 总结了奥运会 100 米数据集的数据（见图 1-1）。

表 1-2 奥运会男子 100 米数据

n	x_n	t_n	$x_n t_n$	x_n^2
1	1896	12.00	22 752.0	3.5948×10^6
2	1900	11.00	20 900.0	3.6100×10^6
3	1904	11.00	20 944.0	3.6252×10^6
4	1906	11.20	21 347.2	3.6328×10^6
5	1908	10.80	20 606.4	3.6405×10^6
6	1912	10.80	20 649.6	3.6557×10^6
7	1920	10.80	20 736.0	3.6864×10^6
8	1924	10.60	20 394.4	3.7018×10^6
9	1928	10.80	20 822.4	3.7172×10^6
10	1932	10.30	19 899.6	3.7326×10^6
11	1936	10.30	19 940.8	3.7481×10^6
12	1948	10.30	20 064.4	3.7947×10^6
13	1952	10.40	20 300.8	3.8103×10^6
14	1956	10.50	20 538.0	3.8259×10^6
15	1960	10.20	19 992.0	3.8416×10^6
16	1964	10.00	19 640.0	3.8573×10^6
17	1968	9.95	19 581.6	3.8730×10^6
18	1972	10.14	19 996.1	3.8888×10^6
19	1976	10.06	19 878.6	3.9046×10^6
20	1980	10.25	20 295.0	3.9204×10^6
21	1984	9.99	19 820.2	3.9363×10^6
22	1988	9.92	19 721.0	3.9521×10^6
23	1992	9.96	19 840.3	3.9681×10^6
24	1996	9.84	19 640.6	3.9840×10^6
25	2000	9.87	19 740.0	4.0000×10^6
26	2004	9.85	19 739.4	4.0160×10^6
27	2008	9.69	19 457.5	4.0321×10^6
$(1/N)\sum_{n=1}^{N}$	1952.37	10.39	20 268.1	3.8130×10^6