机器学习导论
第一章

王小航

简介

LIM Advisors

- 办公室：C1-1417
- 电邮：anita.xhwang@yahoo.com
- 办公时间：周三上午，周四下午

参考书

- 推荐教材(Recommended Teaching Materials)：

机器学习实战：基于Scikit-Learn、Keras和TensorFlow，Aurélien Géron著，宋能辉、李娴译，机械工业出版社，2020年

- 参考教材(Additional Reading Materials)：

统计学习方法（第2版），李航，清华大学出版社，2019年

教学大纲

章节	内容	时间
第一章	机器学习概览	Week 1
第二章	端到端的机器学习项目	Week 2-3
第三章	分类	Week 4-5
第四章	训练模型	Week 6-7
第五章	支持向量机	Week 8-9
第六章	决策树	Week 10-11
第七章	集成学习和随机森林	Week 12-13
第八章	降维	Week 14-15
第九章	无监督学习技术	Week 16
Project presentation	第一到九章内容	Week 17

考核标准

课程总评成绩 Grade	满分100分 Full mark: 100			
课程总评成绩构成 The proportion of grade	考勤 Attendance	个人作业 Homework	PPT展示 Presentation	期末考试 Final exam
	10%	3次，10%/次 合计30%	1次，20%	40%

编程语言

- 使用Python进行编程（Jupyter Notebook），其中机器学习的模型大多使用Scikit-Learn。
- 教材中所有章节的详细代码都发布在GitHub上。项目地址为：

https://github.com/ageron/handson-ml2

- 每两周一次上机课，时长为一个半小时。

什么是机器学习

- 机器学习其实可以被定义为人工智能的一个分支，或者也可以认为是人工智能的一些具体应用。
- 机器学习是让机器具有独立学习的能力，可以直接通过大量的事实数据进行独立学习，让应用程序根据实时场景中的数据进行调整、纠正，并根据分析自我调整，最终得出结果。
- 机器学习一般最基本的做法就是利用算法来解析数据并从中学习，最终对真实世界中的事件作出决策和预测。
- 从 20 世纪 80 年代末到 21 世纪，人们研究了多种机器学习方法，包括神经网络、生物学和进化技术以及数学建模。早期最成功的结果是通过机器学习的统计方法获得的。线性和逻辑回归、分类、决策树、支持向量机等算法大受欢迎。

什么是机器学习

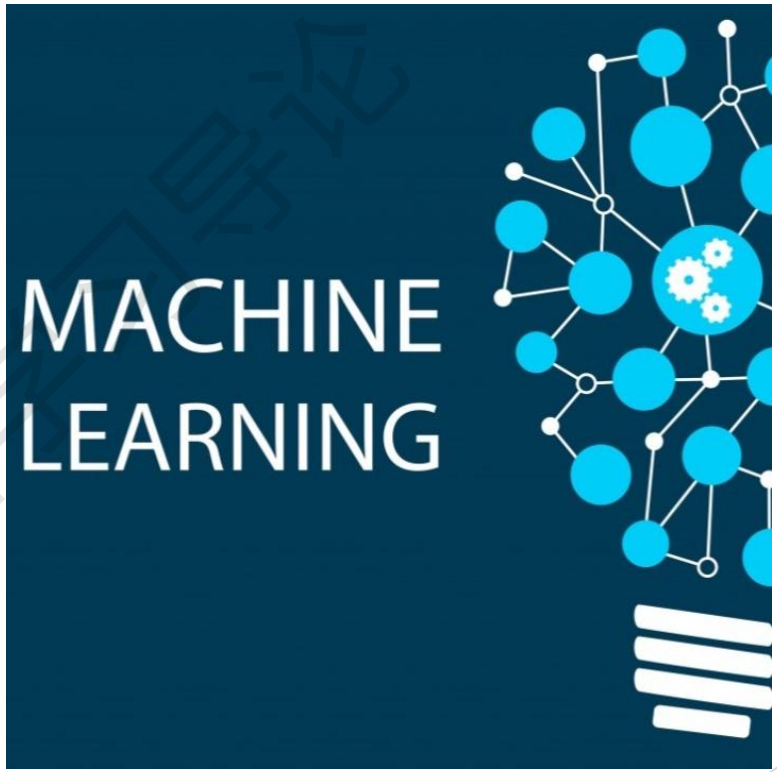
- 机器学习是一门通过编程让计算机从数据中进行学习的科学（和艺术）。

- 机器学习是一个研究领域，让计算机无须进行明确编程就具备学习能力。

——亚瑟·萨缪尔（Arthur Samuel），1959

- 一个计算机程序利用经验E来学习任务T，性能是P，如果针对任务T的性能P随着经验E不断增长，则称为机器学习。

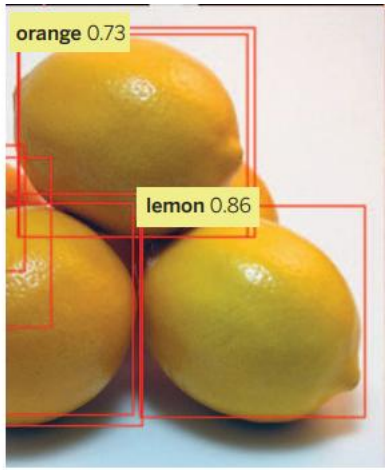
——汤姆·米切尔（Tom Mitchell），1997

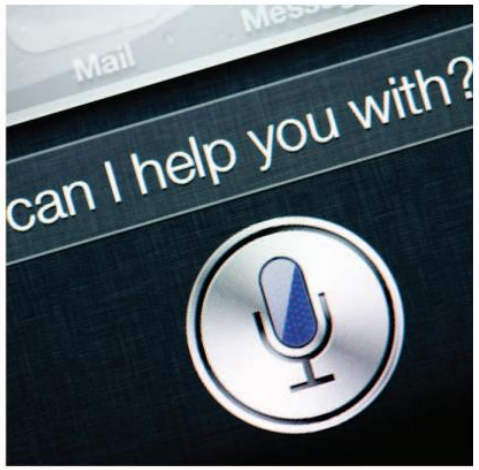


举例

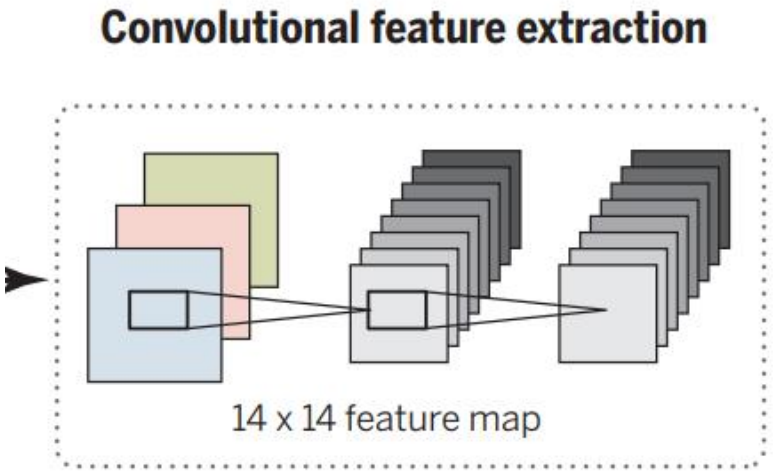
- 垃圾邮件过滤器就是一个机器学习程序，它可以根据垃圾邮件（比如，用户标记的垃圾邮件）和普通邮件（非垃圾邮件，也称作ham）学习标记垃圾邮件。
- 系统用来进行学习的样例称作训练集。
- 每个训练样例称作训练实例（或样本）。
- 在这个示例中，任务T就是标记新邮件是否是垃圾邮件，经验E是训练数据，性能P需要定义。例如，可以使用正确分类邮件的比例。这个性能指标称为准确率，通常用在分类任务中。

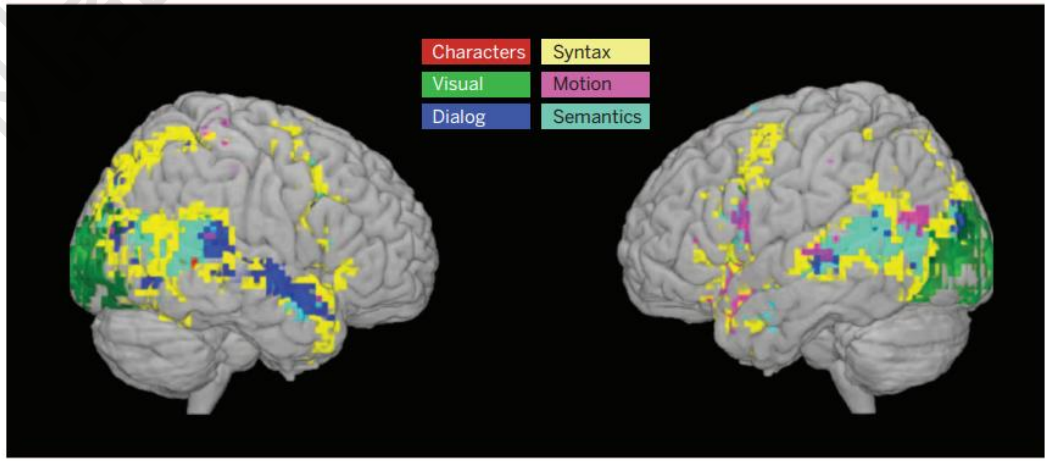
为什么学习机器学习





现今机器学习在日常生活中无处不见，如优酷，淘宝，今日头条的推荐系统;百度和必应等搜索引擎;微博和微信这样的社交媒体;Siri和天猫精灵这样的语音助理

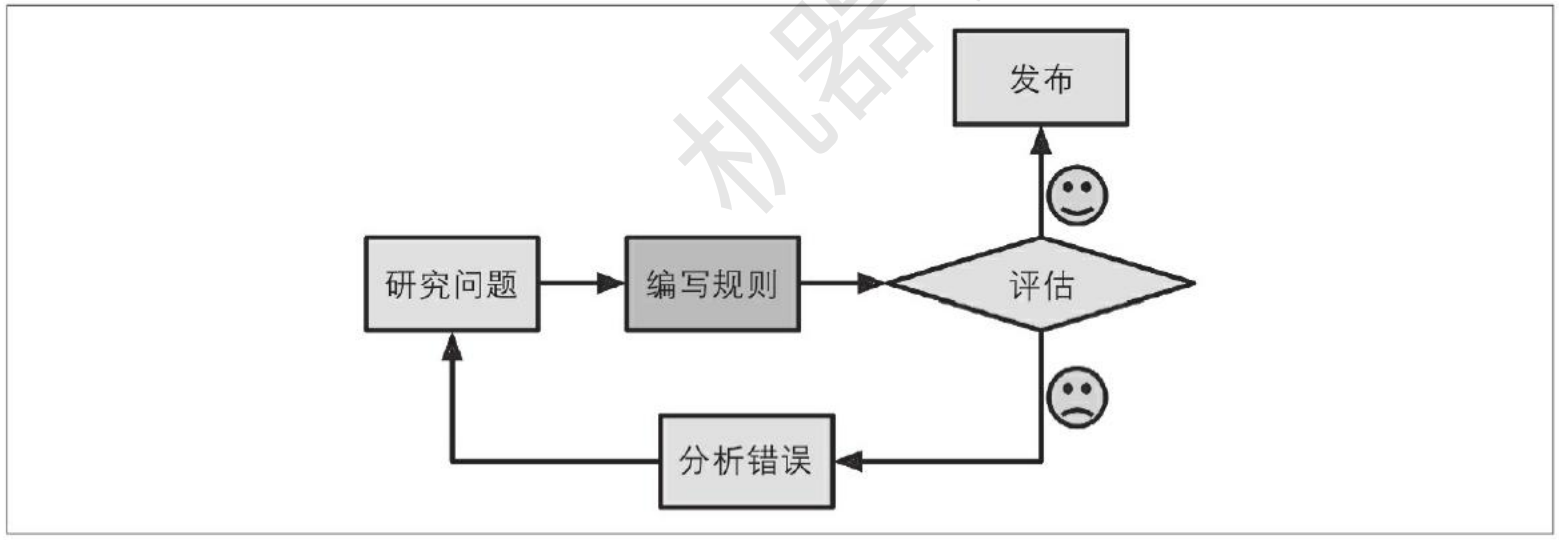




为什么学习机器学习

- 传统方法：

1. 你会先看一下垃圾邮件一般都是什么样子。你可能注意到一些词或短语（比如4U、credit card、free、amazing）在邮件主题中频繁出现等等。
2. 你会为观察到的每个模式各写一个检测算法，如果检测到了某个规律，程序就会将邮件标记为垃圾邮件。
3. 你会测试程序，重复第1步和第2步，直到足够好可以发布

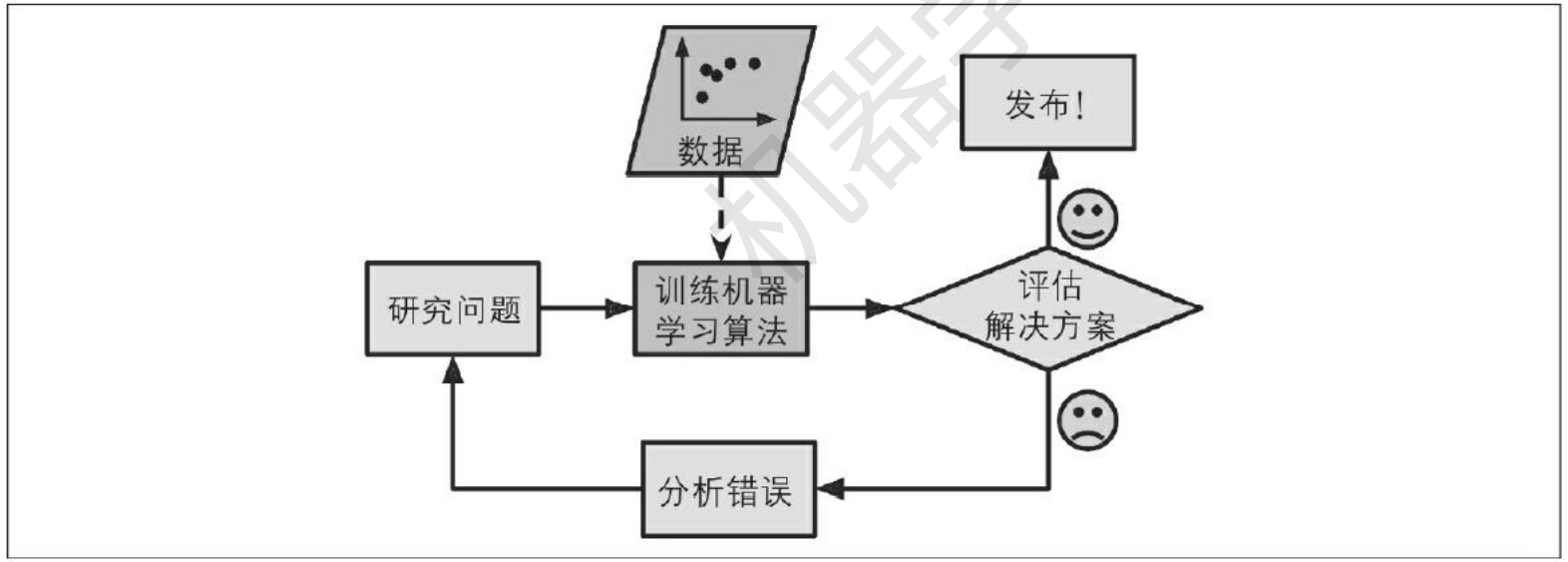


传统方法的缺点

- 你的程序很可能会变成一长串复杂的规则——很难维护。
- 如果垃圾邮件的发送者发现所有包含“4U”的邮件都被屏蔽了，他们会转而使用“For U”。使用传统方法的垃圾邮件过滤器需要更新来标记“For U”。如果垃圾邮件的发送者持续更改，你就需要一直不停地写入新规则

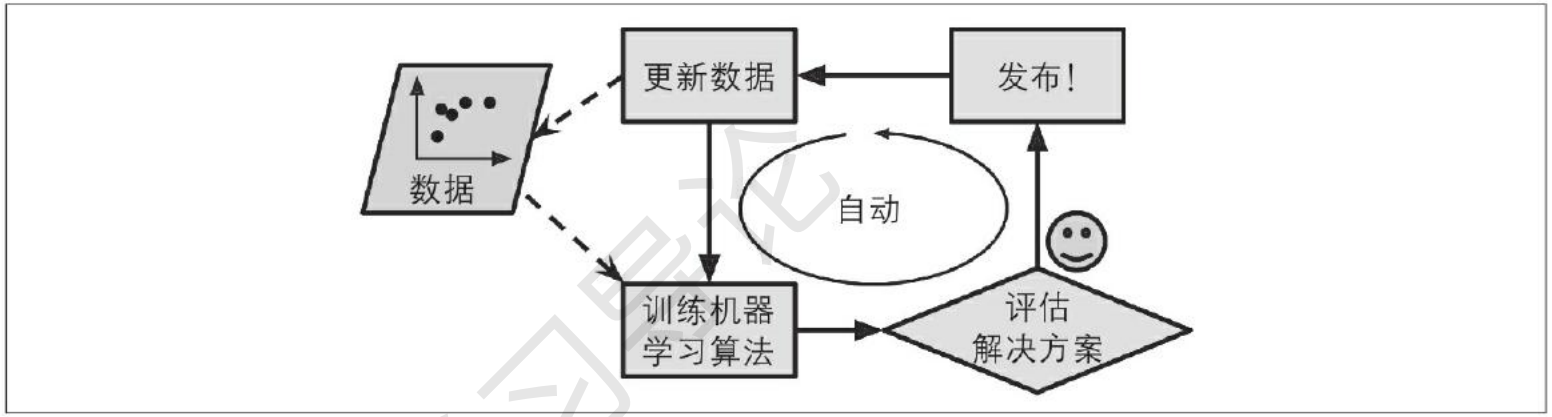
为什么学习机器学习

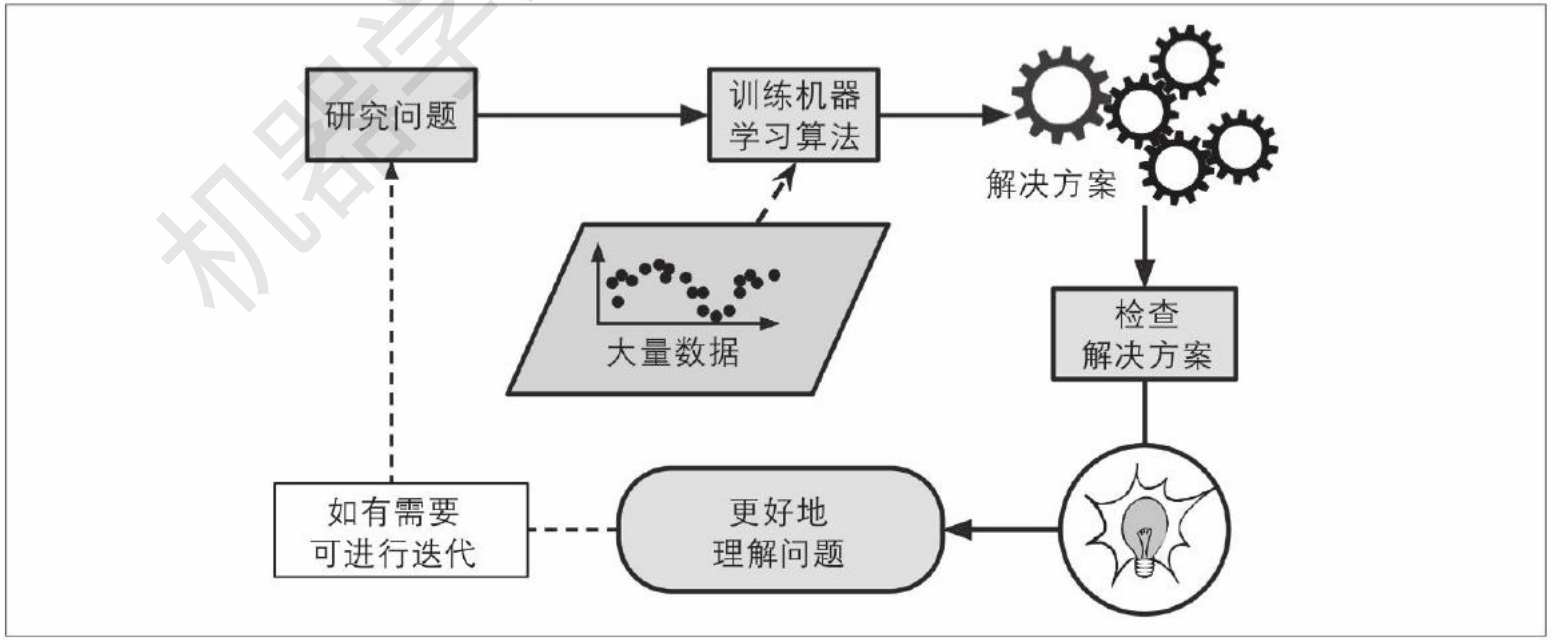
- 基于机器学习技术的垃圾邮件过滤器会自动学习词和短语，这些词和短语是垃圾邮件的预测因素，通过与非垃圾邮件比较，检测垃圾邮件中反复出现的词语模式。
- 这个程序更短，更易维护，也更精确。



机器学习的优点

- 机器学习方法可以自我学习。对于数据的更新，机器学习可以自动发现新规律。
- 机器学习可以帮助人类进行学习。有时可能会发现不引人关注的关联或新趋势，这有助于更好地理解问题。





机器学习的应用示例

- 基于很多性能指标来预测公司下一年的收入

这是一个回归问题（如预测值），需要使用回归模型进行处理，例如线性回归或多项式回归（见第4章）、SVM回归（见第5章）、随机森林回归（见第7章）

- 检测信用卡欺诈

这是异常检测（见第9章）

- 基于客户的购买记录来对客户进行分类，对每一类客户设计不同的市场策略

这是聚类问题（见第9章）

- 用清晰而有洞察力的图表来表示复杂的高维数据集

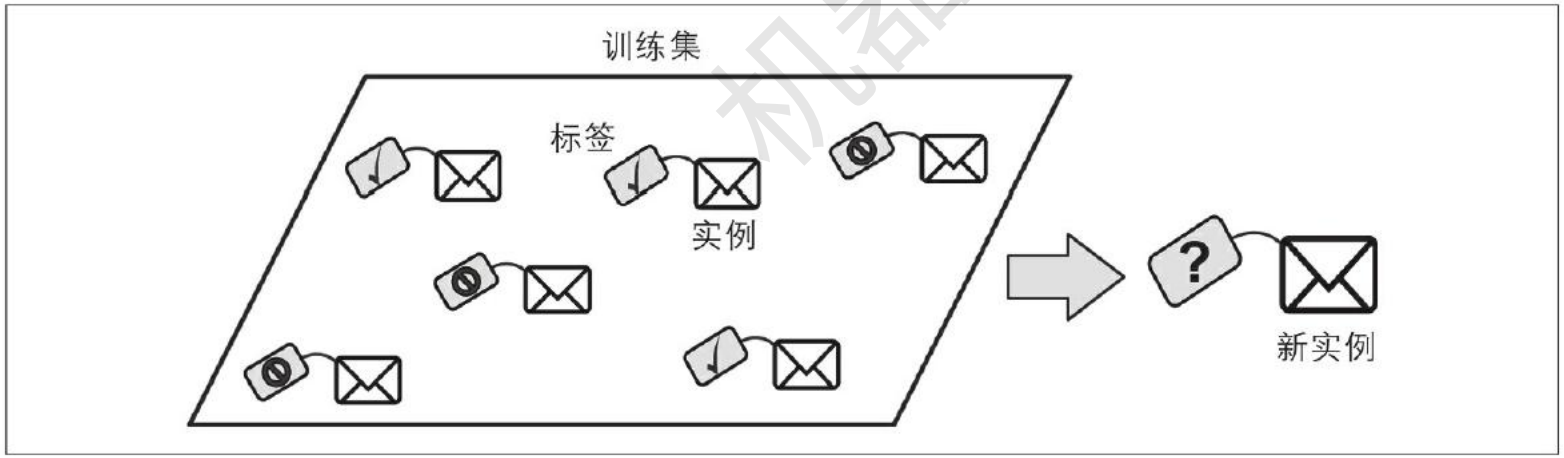
这是数据可视化，经常涉及降维技术（见第8章）

机器学习系统的类型

- 是否在人类监督下训练
  - 有监督学习
  - 无监督学习
  - 半监督学习
- 是否可以动态地进行增量学习
  - 在线学习
  - 批量学习
- 是简单地将新的数据点和已知的数据点进行匹配，还是像科学家那样，对训练数据进行模式检测然后建立一个预测模型
  - 基于实例的学习
  - 基于模型的学习

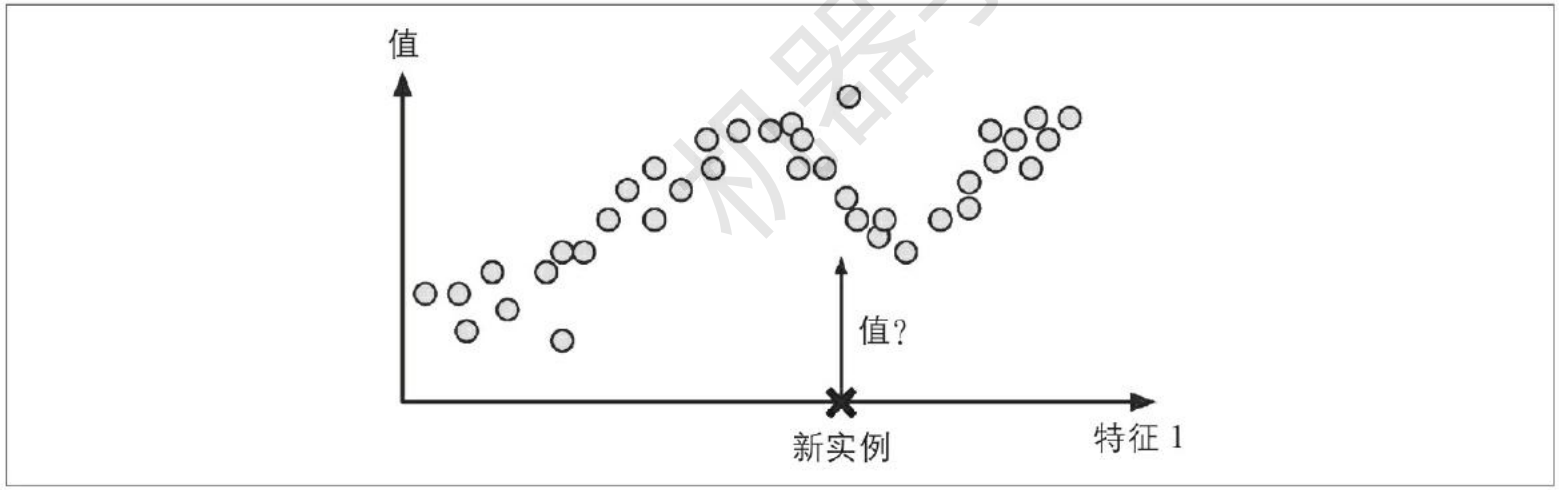
有监督学习

- 在有监督学习中，提供给算法的包含所需解决方案的训练集称为**标签**。
- **分类**任务是一个典型的有监督学习任务。垃圾邮件过滤器就是一个很好的示例：通过大量的电子邮件示例及其所属的类别（垃圾邮件还是常规邮件）进行训练，然后学习如何对新邮件进行分类。



有监督学习

- 另一个典型的任务是通过给定一组称为预测器的特征（里程、使用年限、品牌等）来预测一个目标数值（例如汽车的价格）。这种类型的任务称为**回归**。要训练这样一个系统，需要提供大量的汽车示例，包括它们的预测器和标签（即价格）。

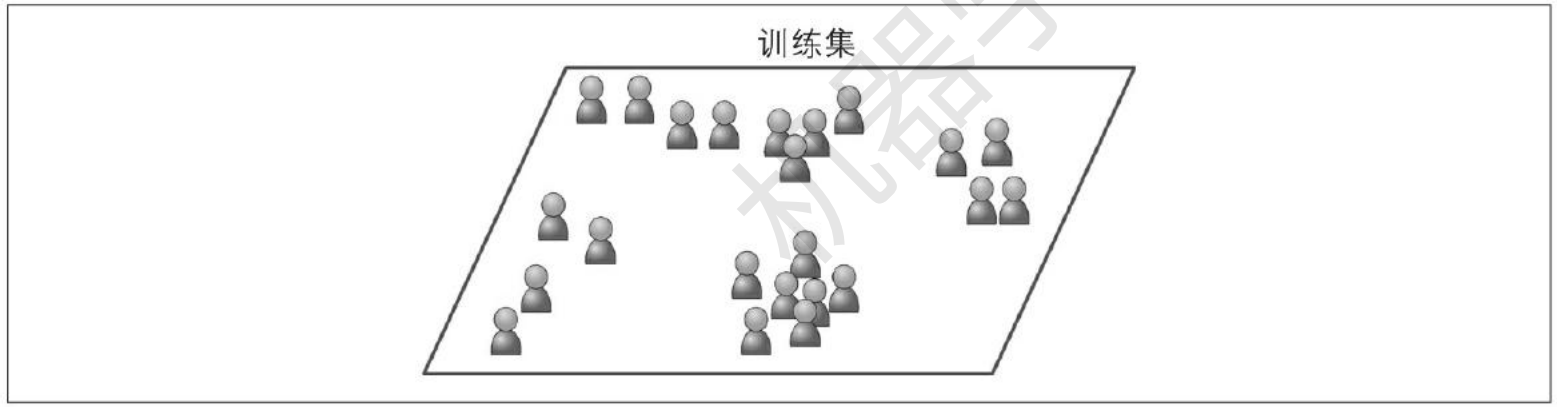


有监督学习

- 一些重要的有监督学习算法：
  - k-近邻算法
  - 线性回归
  - 逻辑回归
  - 支持向量机（SVM）
  - 决策树和随机森林

无监督学习

- 无监督学习的训练数据都是未经标记的。系统会在没有“老师”的情况下进行学习。

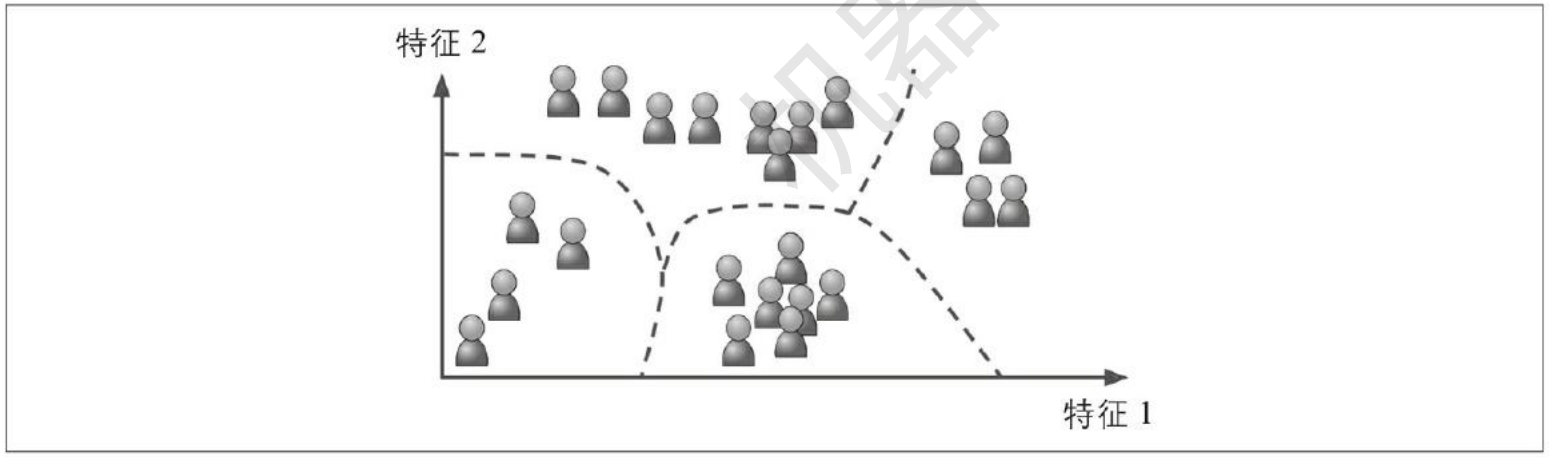


无监督学习

- 一些重要的无监督学习算法：
  - 聚类算法（k-均值算法，DBSCAN，分层聚类分析（HCA））
  - 异常检测和新颖性检测（单类SVM，孤立森林）
  - 可视化和降维（主成分分析（PCA），核主成分分析，局部线性嵌入（LLE），t-分布随机近邻嵌入（t-SNE））
  - 关联规则学习（Apriori，Eclat）

无监督学习

- 例如，假设你现在拥有大量关于自己博客访客的数据。你想通过一个**聚类**算法来检测相似访客的分组）。你不大可能告诉这个算法每个访客属于哪个分组——算法会自行寻找这种关联。例如，它可能会注意到40%的访客是喜欢漫画的男性，并且通常在夜晚阅读你的博客；20%的访客是年轻的科幻爱好者，通常在周末访问；等等。



无监督学习

- **可视化算法**也是无监督学习算法的一个不错的示例：你提供大量复杂的、未标记的数据，算法轻松绘制输出2D或3D的数据表示。这些算法会尽其所能地保留尽量多的结构（例如，尝试保持输入的单独集群在可视化中不会被重叠），以便于你理解这些数据是怎么组织的，甚至识别出一些未知的模式。
- 与之相关的一个任务是**降维**，降维的目的是在不丢失太多信息的前提下简化数据。方法之一是将多个相关特征合并为一个。例如，汽车里程与其使用年限存在很大的相关性，所以降维算法会将它们合并成一个代表汽车磨损的特征。这个过程叫作**特征提取**。

无监督学习

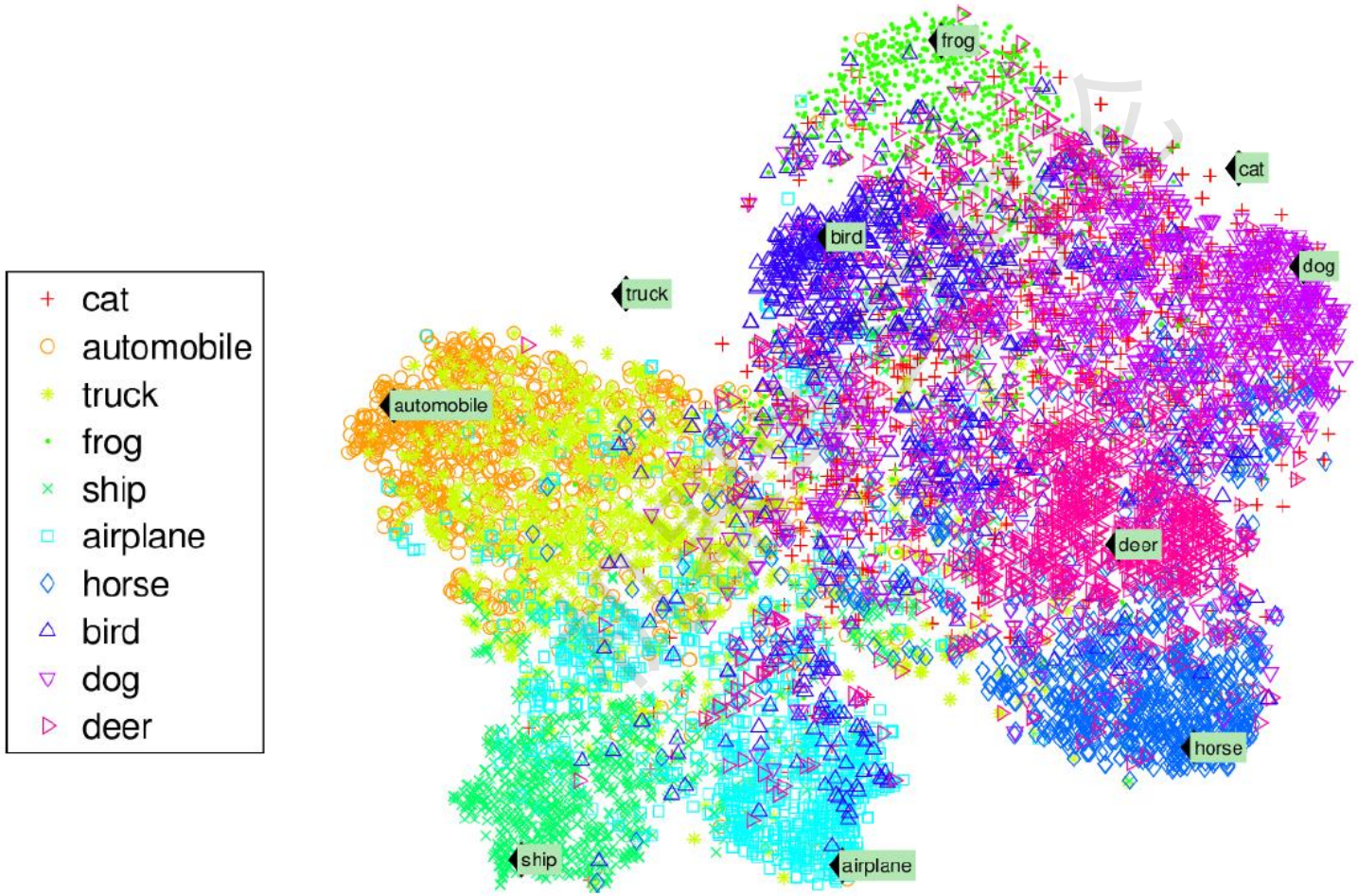
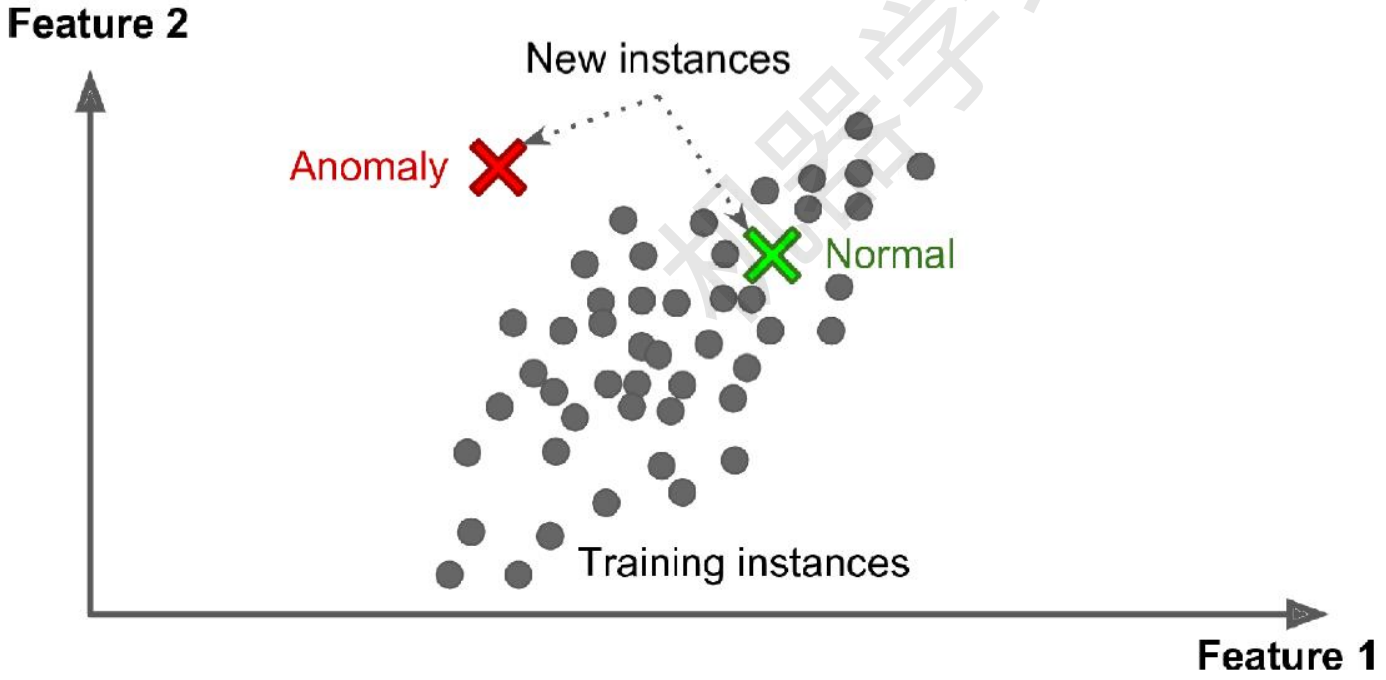


Figure 1-9. Example of a t-SNE visualization highlighting semantic clusters[3]

无监督学习

- 另一个很重要的无监督任务是**异常检测**——例如，检测异常信用卡交易以防止欺诈，捕捉制造缺陷，或者在给另一种机器学习算法提供数据之前自动从数据集中移除异常值。系统用正常实例进行训练，然后当看到新的实例时，它就可以判断出这个新实例看上去是正常还是异常。

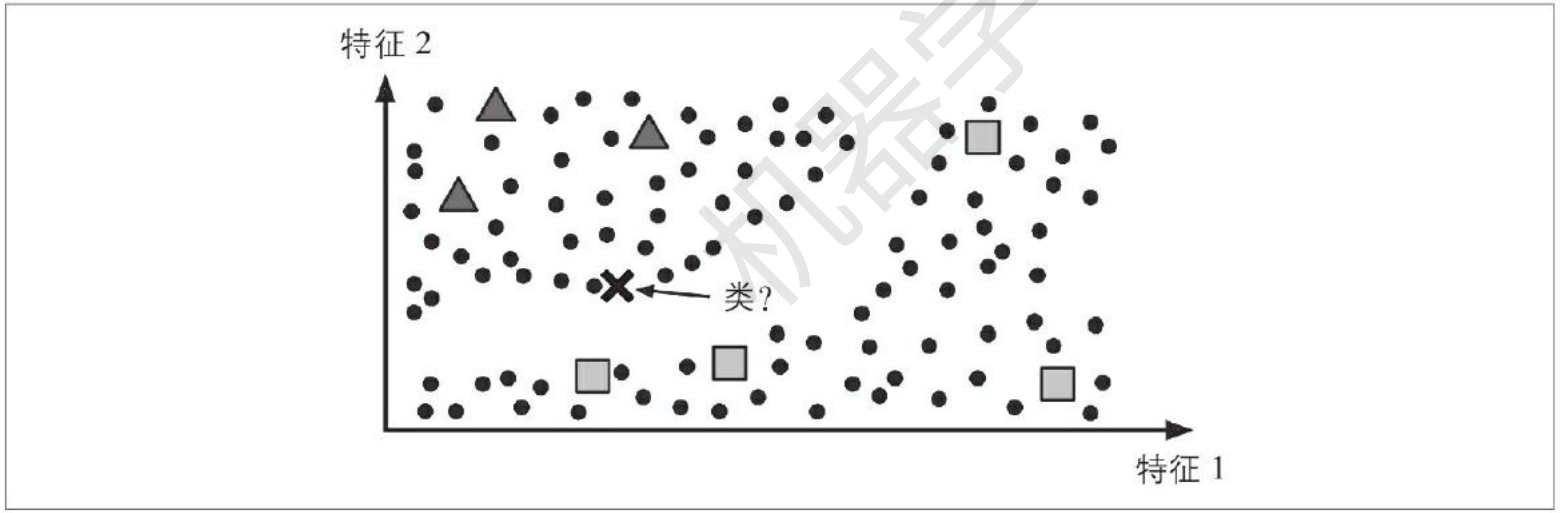


无监督学习

- 一个常见的无监督任务是**关联规则学习**，其目的是挖掘大量数据，发现属性之间的有趣联系。
- 例如，假设你开了一家超市，在销售日志上运行关联规则之后发现买烧烤酱和薯片的人也倾向于购买牛排。那么，你可能会将这几样商品摆放得更近一些。

半监督学习

- 由于通常给数据做标记是非常耗时和昂贵的，你往往会有很多未标记的数据而很少有已标记的数据。有些算法可以处理部分已标记的数据。这被称为**半监督学习**。
- 大多数半监督学习算法是无监督算法和有监督算法的结合。

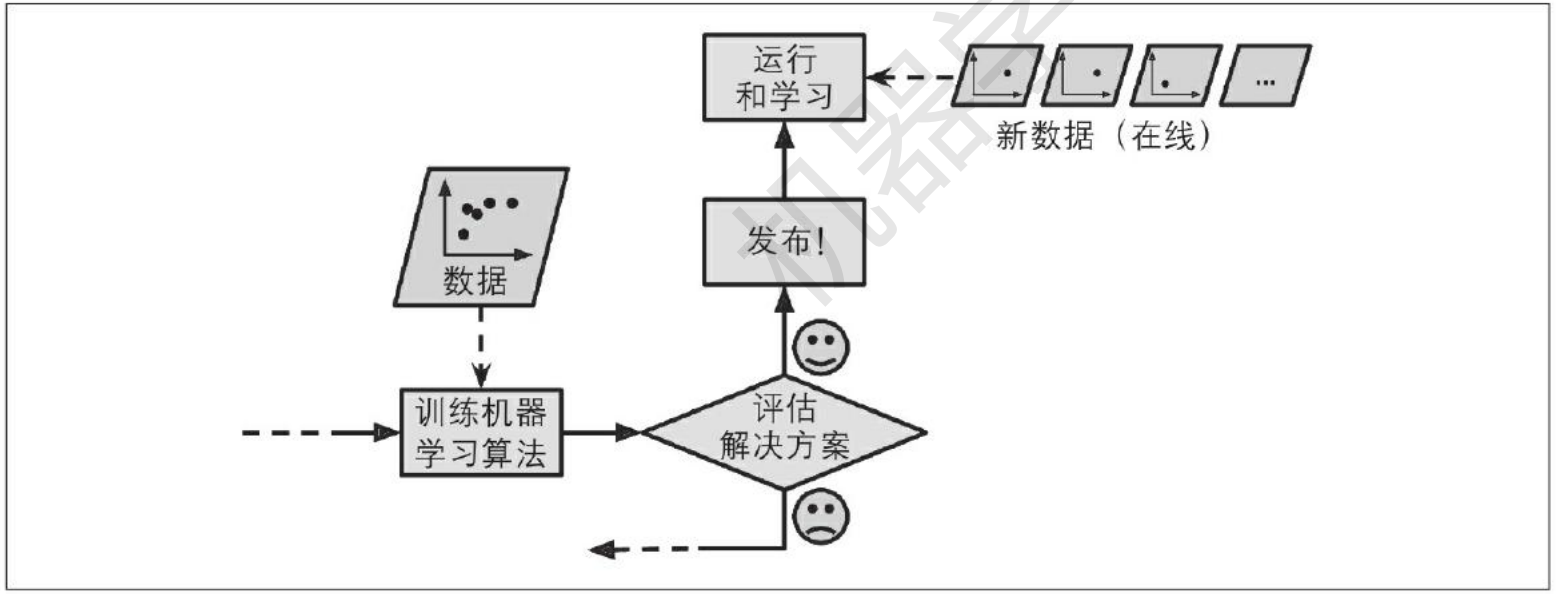


批量学习

- 在批量学习中，系统无法进行增量学习——即必须使用所有可用数据进行训练。
- 这需要大量时间和计算资源，所以通常都是离线完成的。
- 离线学习就是先训练系统，然后将其投入生产环境，这时学习过程停止，它只是将其所学到的应用出来。
- 如果希望批量学习系统学习新数据（例如新型垃圾邮件），需要在完整数据集（包括新数据和旧数据）的基础上重新训练系统的新版本，然后停用旧系统，用新系统取而代之。

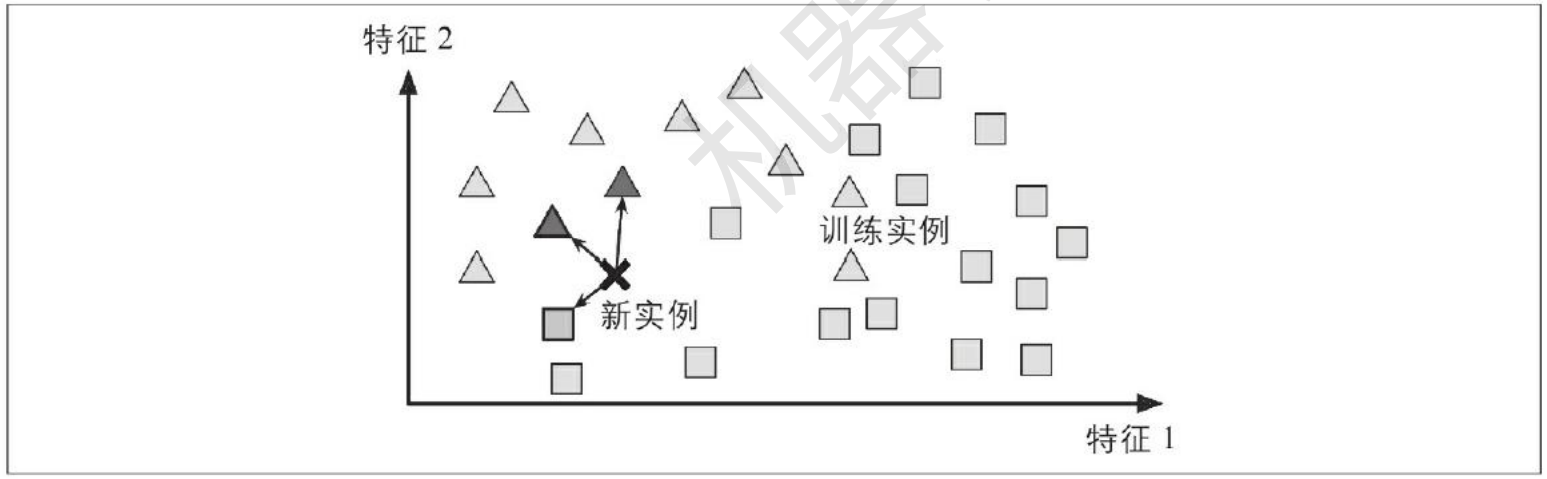
在线学习

- 在在线学习中，你可以循序渐进地给系统提供训练数据，逐步积累学习成果。这种提供数据的方式可以是单独的，也可以采用小批量的小组数据来进行训练。每一步学习都很快速并且便宜，这样系统就可以根据飞速写入的最新数据进行学习。



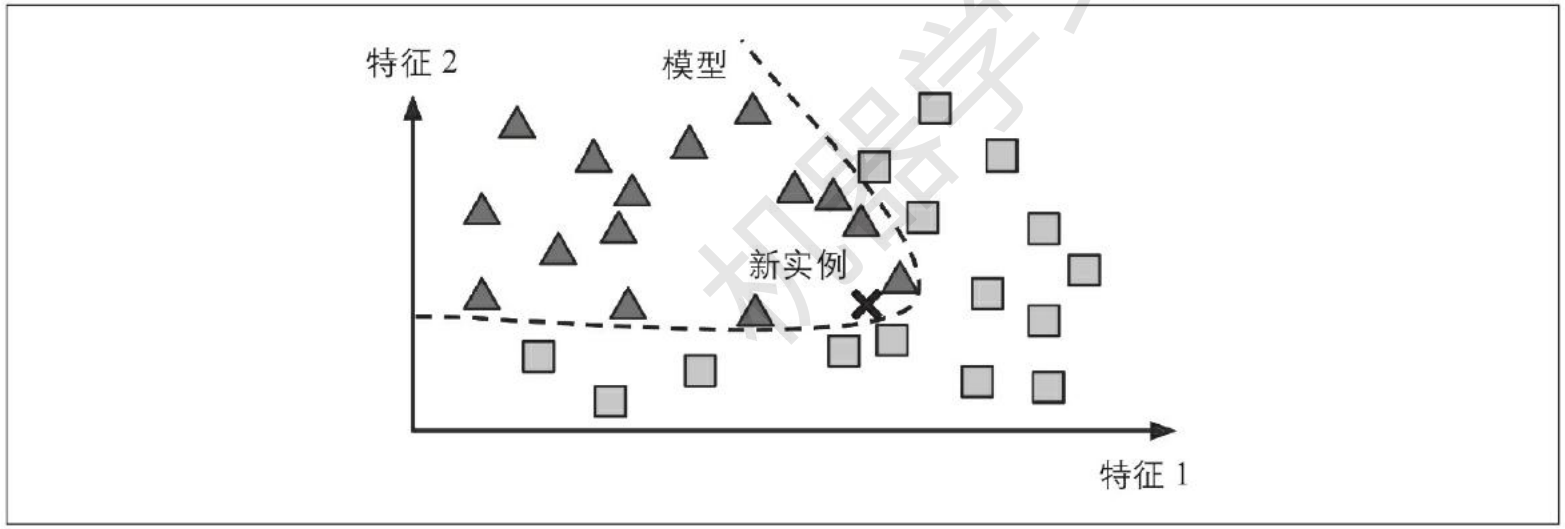
基于实例的学习

- 你可以通过编程让系统标记与已知的垃圾邮件非常相似的邮件。这里需要两封邮件之间的相似度度量。一种相似度度量方式是计算它们之间相同的单词数目。如果一封新邮件与一封已知的垃圾邮件有许多单词相同，系统就可以将其标记为垃圾邮件。



基于模型的学习

- 从一组示例集中实现泛化的另一种方法是构建这些示例的模型，然后使用该模型进行预测。这称为基于模型的学习



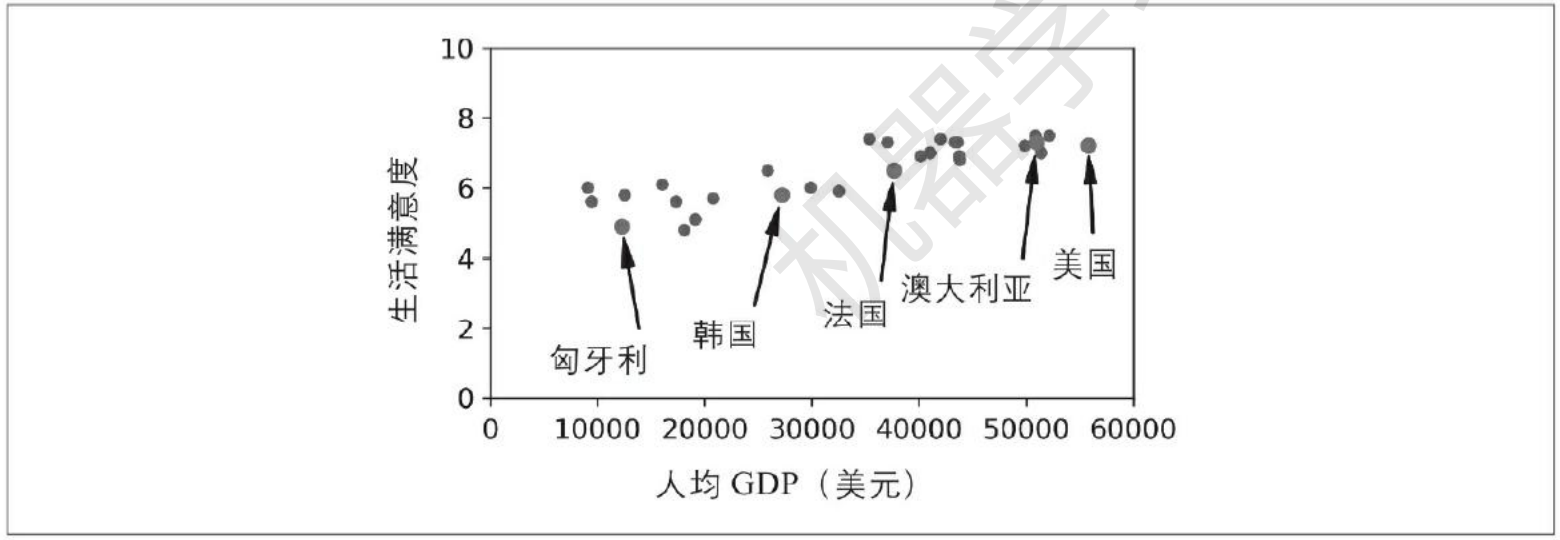
基于模型的学习：举例

- 假设你想知道金钱是否让人感到快乐，你可以从经合组织（OECD）的网站上下载“幸福指数”的数据，再从国际货币基金组织（IMF）的网站上找到人均GDP的统计数据，将数据并入表格，按照人均GDP排序，你会得到如表所示的摘要。

国家	人均 GDP（美元）	生活满意度
匈牙利	12 240	4.9
韩国	27 195	5.8
法国	37 675	6.5
澳大利亚	50 962	7.3
美国	55 805	7.2

基于模型的学习：举例

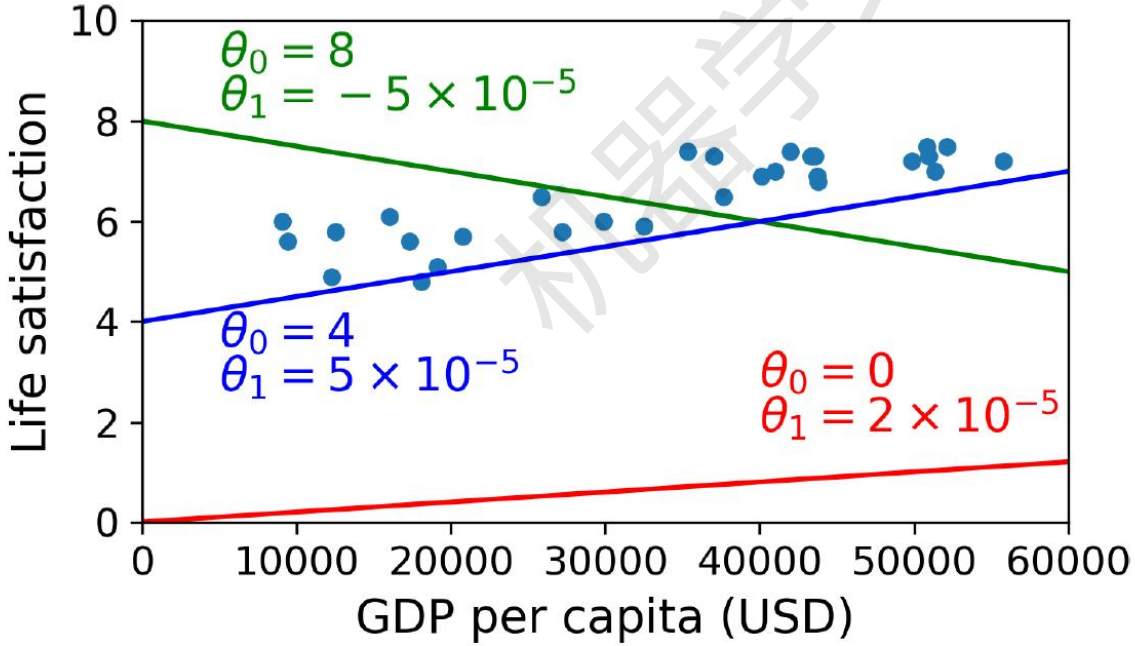
- 这里似乎有一个趋势！虽然数据包含噪声（即部分随机），但是仍然可以看出随着该国人均GDP的增加，生活满意度或多或少呈线性上升的趋势。



基于模型的学习：举例

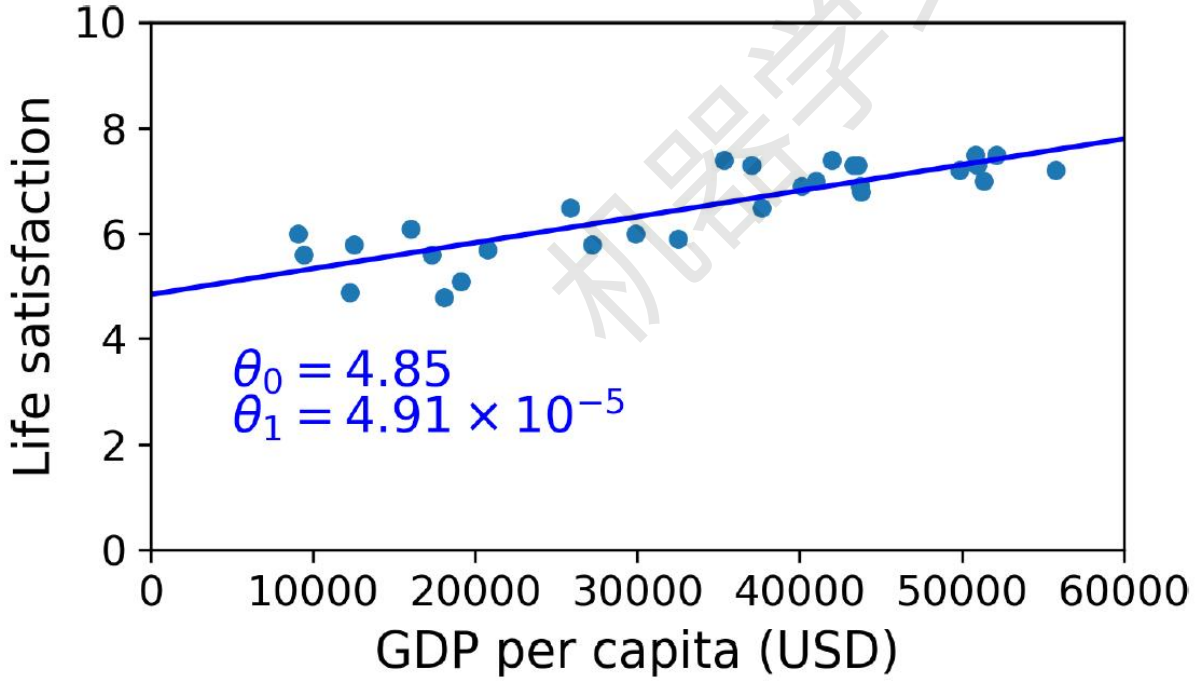
- 你可以把生活满意度建模成一个关于人均GDP的**线性函数**。这个过程叫作**模型选择**。你为生活满意度选择了一个线性模型，该模型只有一个属性，就是人均GDP。

$$\text{生活满意度} = \theta_0 + \theta_1 \times \text{人均GDP}$$



基于模型的学习：举例

- 在使用模型之前，需要先定义参数θ_0和θ_1的值。怎么才能知道什么值可以使模型表现最佳呢？对于线性回归问题，通常的选择是使用成本函数来衡量线性模型的预测与训练实例之间的差距，目的在于尽量使这个差距最小化。



基于模型的学习：举例

- 可以运行模型来进行预测。
- 例如，你想知道塞浦路斯人有多幸福，但是经合组织的数据没有提供答案。
- 先查查塞浦路斯的人均GDP是多少，发现是22587美元，然后应用到模型中，发现生活满意度大约是$4.85 + 22587 \times 4.91 \times 10^{-5} = 5.96$。

基于实例的学习：举例

- 如果使用**基于实例的学习**算法，你会发现斯洛文尼亚的人均GDP最接近塞浦路斯（20 732美元），而经合组织的数据告诉我们，斯洛文尼亚人的生活满意度是5.7，因此你很可能会预测塞浦路斯的生活满意度为5.7。
- 如果稍微拉远一些，看看两个与之最接近的国家——葡萄牙和西班牙的生活满意度分别为5.1和6.5。取这三个数值的平均值得到5.77，这也非常接近基于模型预测所得的值。这个简单的算法被称为**k-近邻回归**（在本例中，k=3）。

机器学习的主要挑战

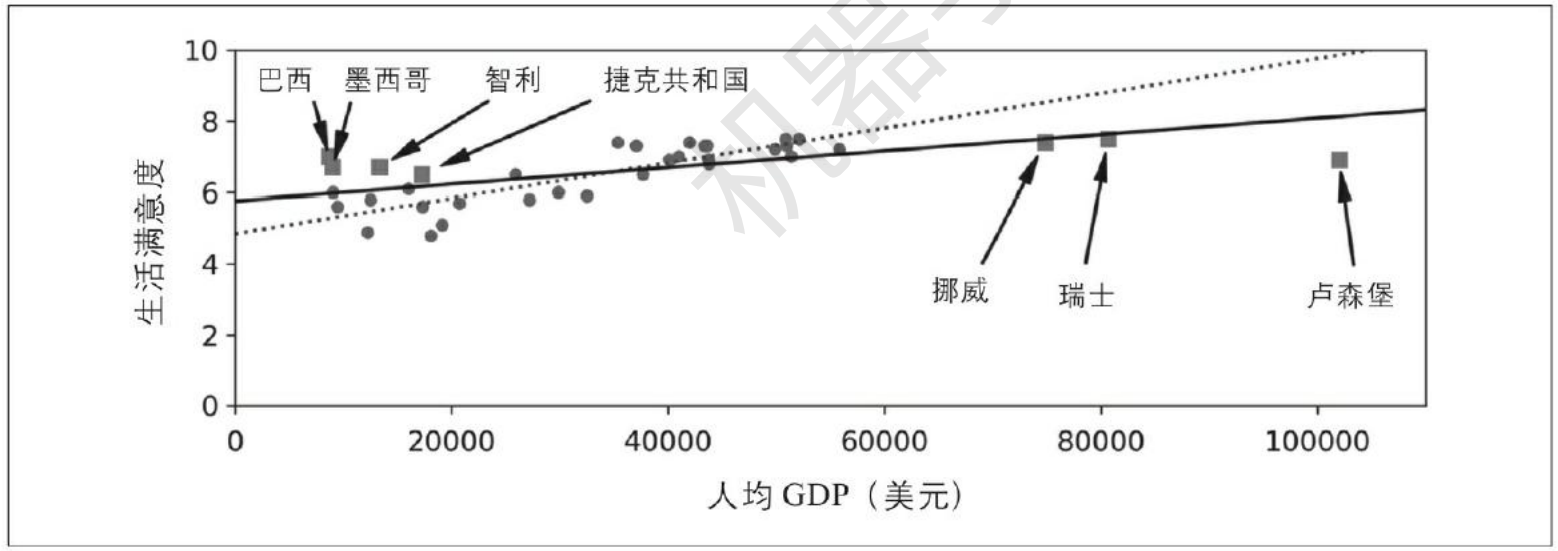
- 训练数据的数量不足
- 训练数据不具代表性
- 无关特征
- 过拟合训练数据
- 欠拟合训练数据

训练数据的数量不足

- 大部分机器学习算法需要大量的数据才能正常工作。
- 即使是最简单的问题，很可能也需要成千上万个示例，而对于诸如图像或语音识别等复杂问题，则可能需要数百万个示例（除非你可以重用现有模型的某些部分）。

训练数据不具代表性

- 例如，前面用来训练线性模型的国家数据集并不具备完全的代表性，有部分国家的数据缺失。
- 如果你用这个数据集训练线性模型，将会得到图中的实线，而虚线表示旧模型。
- 使用不具代表性的训练集训练出来的模型不可能做出准确的预估，尤其是针对那些特别贫穷或特别富裕的国家。



低质量数据

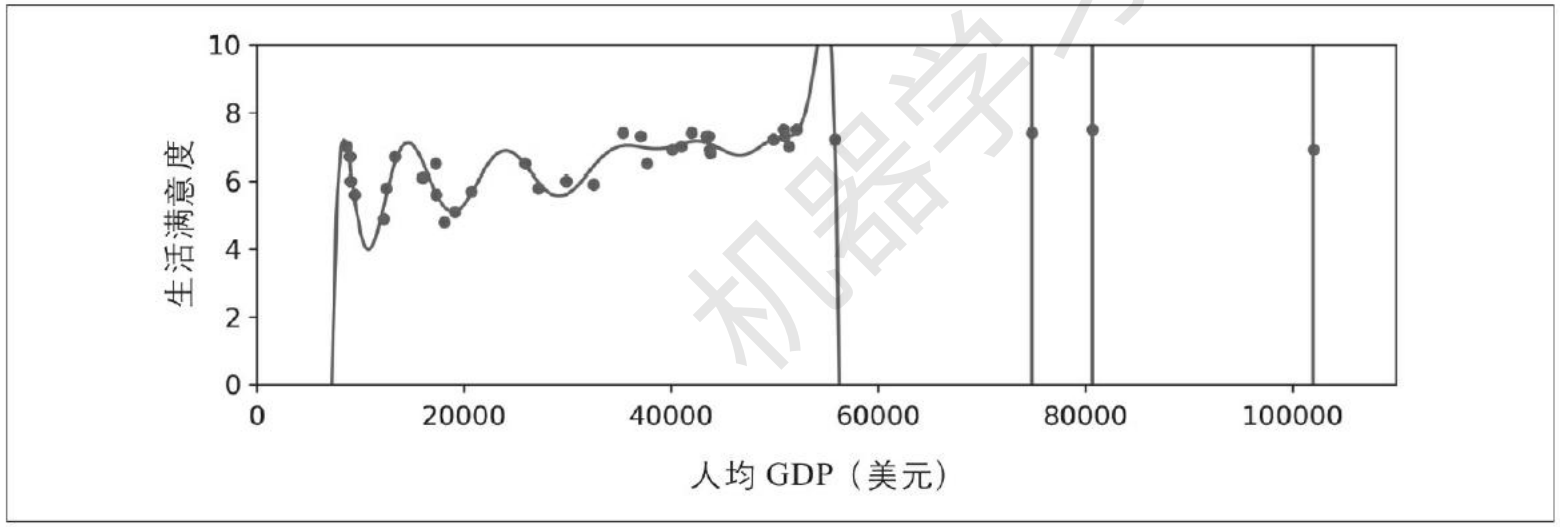
- 如果训练集满是错误、异常值和噪声（例如，低质量的测量产生的数据），系统将更难检测到底层模式，更不太可能表现良好。
- 所以，清理训练数据成为必要，例如：
  - 如果某些实例明显是异常情况，那么直接将其丢弃，或者尝试手动修复错误，都会大有帮助。
  - 如果某些实例缺少部分特征（例如，5%的顾客没有指定年龄），你必须决定是整体忽略这些特征、忽略这部分有缺失的实例、将缺失的值补充完整（例如，填写年龄值的中位数），还是训练一个带这个特征的模型，再训练一个不带这个特征的模型。

无关特征

- 只有训练数据里包含足够多的相关特征以及较少的无关特征，系统才能够完成学习。
- 一个成功的机器学习项目，其关键部分是提取出一组好的用来训练的特征集。这个过程叫作特征工程，包括以下几点：
  - 特征选择（从现有特征中选择最有用的特征进行训练）。
  - 特征提取（将现有特征进行整合，产生更有用的特征——正如前文提到的，降维算法可以提供帮助）。
  - 通过收集新数据创建新特征。

过拟合训练数据

- 过拟合指模型在训练数据上表现良好，但是泛化时却效果不好。

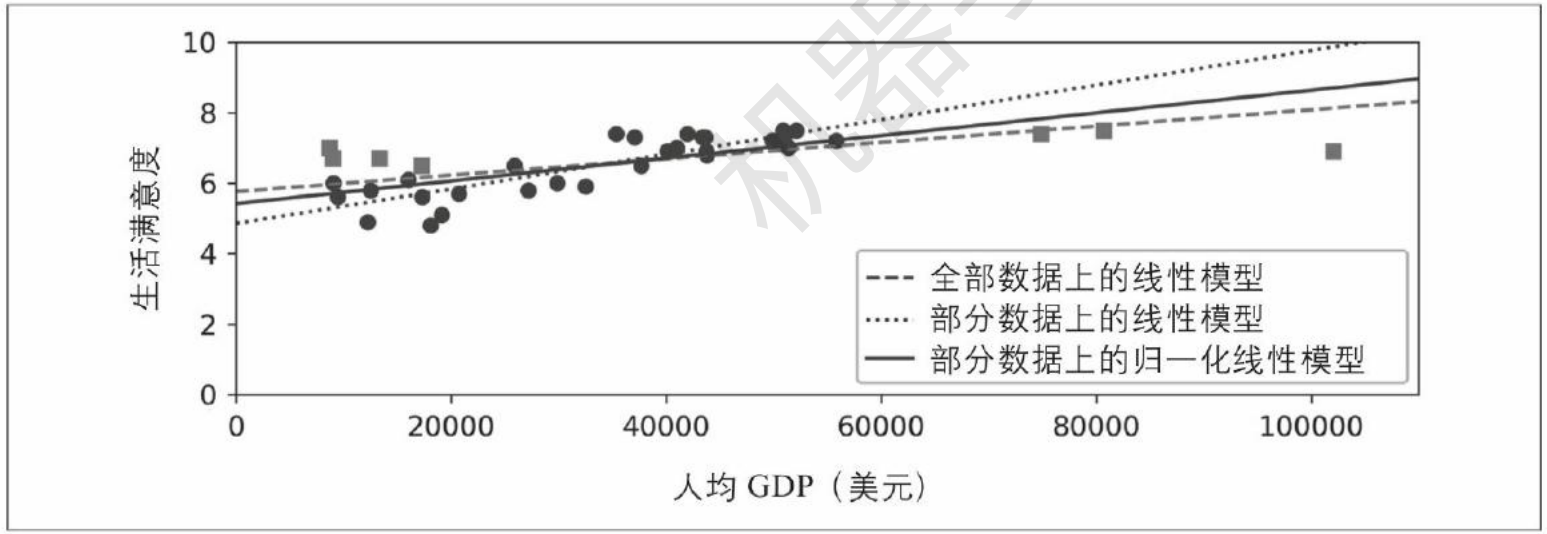


过拟合训练数据

- 当模型相对于训练数据的数量和噪度都过于复杂时，会发生过拟合。可能的解决方案如下：
  - 简化模型：可以选择较少参数的模型（例如，选择线性模型而不是高阶多项式模型）也可以减少训练数据中的属性数量，或者是约束模型。
  - 收集更多的训练数据。
  - 减少训练数据中的噪声（例如，修复数据错误和消除异常值）。

过拟合训练数据

- 通过约束模型使其更简单，并降低过拟合的风险，这个过程称为正则化。
- 下图显示了三个模型。可以看到，正则化强制了模型的斜率较小：该模型与训练数据（圆圈）的拟合不如第一个模型，但它实际上更好地泛化了它没有在训练时看到的新实例（方形）。



欠拟合训练数据

- 欠拟合和过拟合正好相反。它的产生通常是因为对于底层的数据 结构来说，你的模型太过简单。
- 例如，用线性模型来描述生活满意度就属于欠拟合。现实情况远比模型复杂得多。
- 解决这个问题的主要方式有：
  - 选择一个带有更多参数、更强大的模型。
  - 给学习算法提供更好的特征集（特征工程）。
  - 减少模型中的约束（例如，减少正则化超参数）。