
MathWorks®
使用 MATLAB
进行机器学习

目录

第 1 部分: 简介

机器学习是什么？

机器学习教计算机执行人和动物与生俱来的活动：从经验中学习。机器学习算法使用计算方法直接从数据中“学习”信息，而不依赖于预定方程模型。当可用于学习的样本数量增加时，这些算法可自适应提高性能。

更多数据、更多问题、更多解答

机器学习算法可从能够带来洞察力的数据中发现自然模式，帮助您更好地制定决策和做出预测。医疗诊断、股票交易、能量负荷预测及更多行业每天都在使用这些算法制定关键决策。媒体网站依靠机器学习算法从数百万种选择中筛选出为您推荐的歌曲或影片。零售商利用这些算法深入了解客户的购买行为。

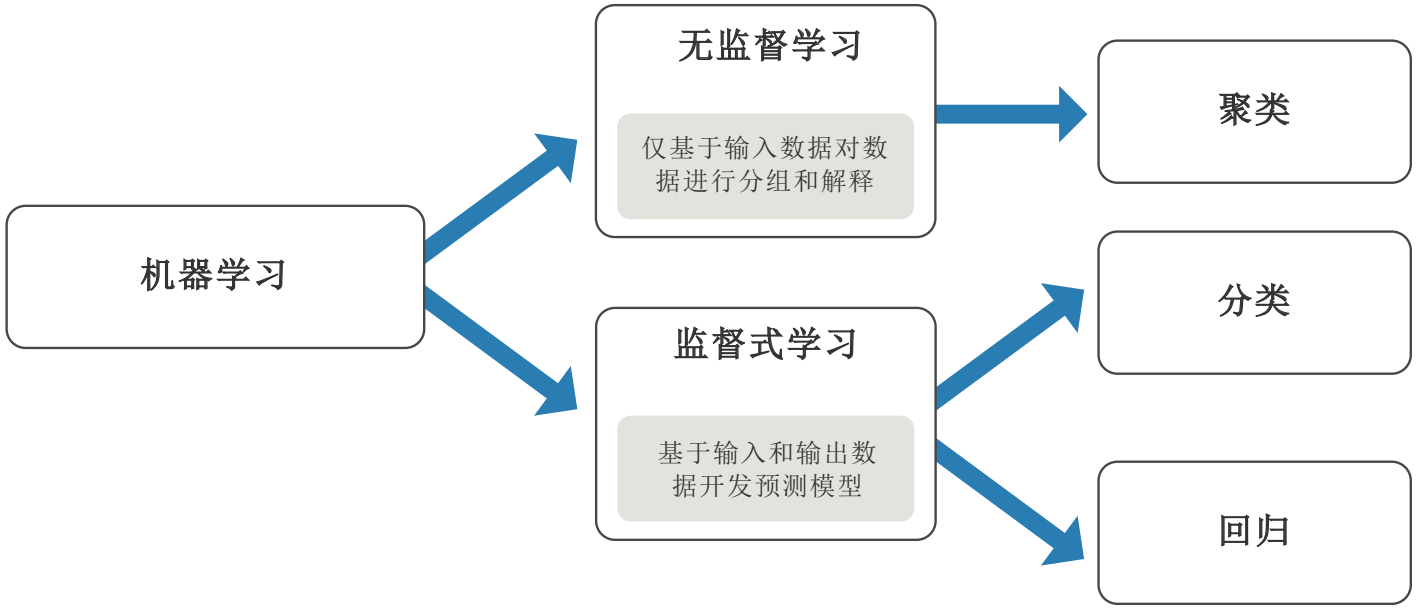
实际环境中的应用程序

随着大数据的增加，机器学习对于解决以下领域的问题变得尤为重要：

- 计算金融学，用于信用评估和算法交易
- 图像处理和计算机视觉，用于人脸识别、运动检测和对象检测
- 计算生物学，用于肿瘤检测、药物发现和 DNA 顺序分析
- 能源生产，用于预测价格和负荷
- 汽车、航空航天和制造业，用于预见性维护
- 自然语言处理

机器学习的工作原理

机器学习采用两种类型的技术：监督式学习和无监督学习。监督式学习根据已知的输入和输出训练模型，让模型能够预测未来输出；无监督学习从输入数据中找出隐藏模式或内在结构。



监督式学习

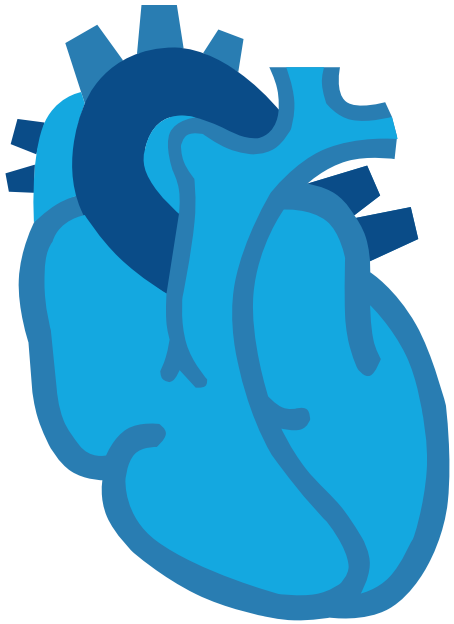
监督式机器学习旨在构建能够根据存在不确定性的证据做出预测的模型。监督式学习算法接受已知的输入数据集和对数据的已知响应（输出），然后训练模型，让模型能够为新输入数据的响应生成合理的预测。

监督式学习采用分类和回归技术开发预测模型。

- **分类技术可预测离散的响应** — 例如，电子邮件是真正邮件还是垃圾邮件，肿瘤是恶性还是良性的。分类模型可将输入数据划分成不同类别。典型的应用包括医学成像、语音识别和信用评估。
- **回归技术可预测连续的响应** — 例如，电力需求中温度或波动的变化。典型的应用包括电力系统负荷预测和算法交易。

使用监督式学习预测心脏病发作

假设临床医生希望预测某位患者在一年内是否会心脏病发作。他们有以前就医的患者的相关数据，包括年龄、体重、身高和血压。他们知道以前的患者在一年内是否出现过心脏病发作。因此，问题在于如何将现有数据合并到模型中，让该模型能够预测新患者在一年内是否会出现心脏病发作。

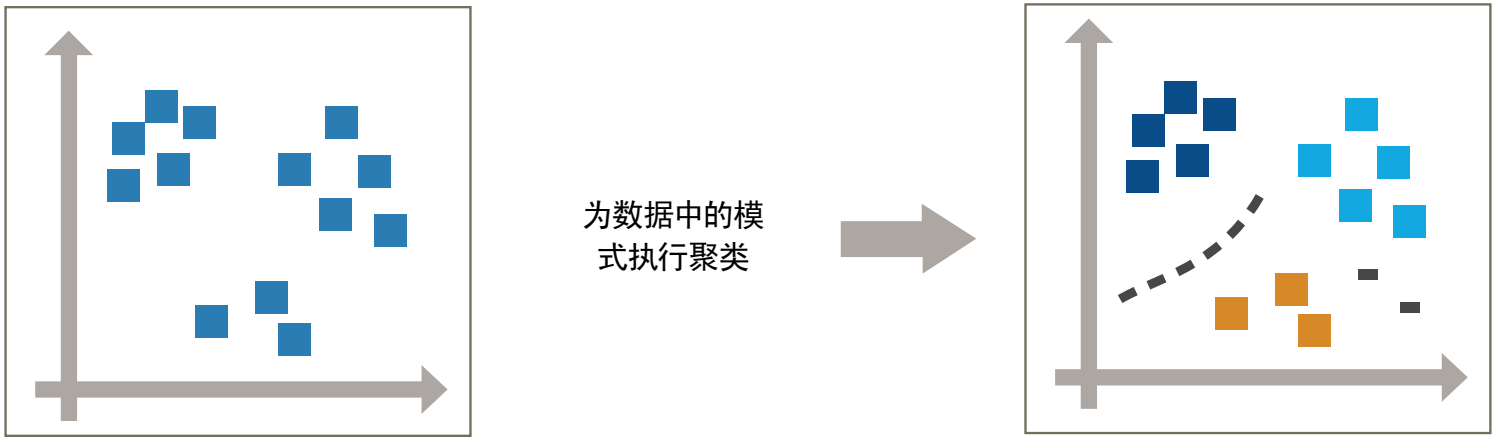


无监督学习

无监督学习可发现数据中隐藏的模式或内在结构。这种技术可根据包含未标记响应的输入数据的数据集执行推理。

聚类是一种最常用的无监督学习技术。这种技术可通过探索性数据分析发现数据中隐藏的模式或分组。

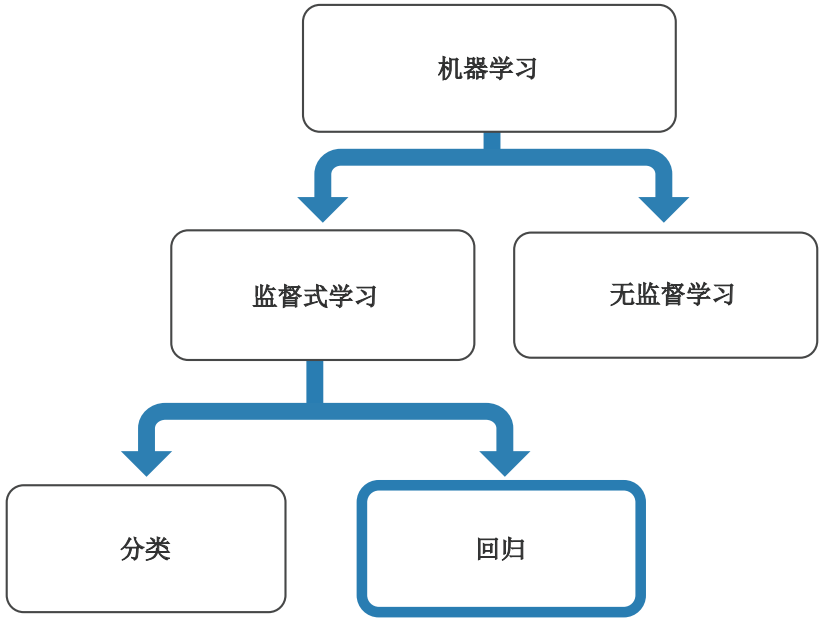
聚类的应用包括基因序列分析、市场调查和对象识别。



如何确定使用哪种算法？

选择正确的算法看似难以驾驭——需要从几十种监督式和无监督机器学习算法中选择，每种算法又包含不同的学习方法。

没有最佳方法或万全之策。找到正确的算法只是试错过程的一部分——即使是经验丰富的数据科学家，也无法说出某种算法是否无需试错即可使用。但算法的选择还取决于您要处理的数据的大小和类型、您要从数据中获得的洞察力以及如何运用这些洞察力。



何时应该使用机器学习？

当您遇到涉及大量数据和许多变量的复杂任务或问题，但没有现成的处理公式或方程式时，可以考虑使用机器学习。例如，如果您需要处理以下情况，使用机器学习是一个很好的选择：

- 手写规则和方程式太过复杂——例如人脸识别和语音识别。
- 任务的规则始终在变化——例如事务处理记录的欺诈检测。
- 数据本身在不断变化，程序也必须适应这种变化——例如自动交易、能量需求预测和购物趋势预测等。

实际环境中的示例

创建可分析艺术作品的算法

美国罗格斯大学艺术与人工智能实验室的研究人员曾经想知道计算机算法能否像人类一样根据风格、流派和艺术家将绘画作品轻松归类。开始时，他们通过识别视觉特征来对绘画作品的风格分类。他们开发的绘画风格分类算法在数据库中的准确度达到 60%，远超过普通非专业人士。

研究人员假定可用于对风格分类（监督式学习问题）的视觉特征也能用于确定艺术影响力（无监督学习问题）。

他们将经过训练的分类算法应用到 Google 图像，用于确定具体对象。他们对跨度长达 550 年的 66 位不同艺术家的 1,700 幅绘画作品测试了此算法。此算法可以可靠地识别出相关的作品，包括迭戈•委拉斯开兹的《教皇英诺森十世肖像》对弗朗西斯•培根的《教皇英诺森十世肖像的习作》产生的影响。

实际环境中的示例

优化大型建筑中的 HVAC 能耗

在办公大楼、医院及其他大型商业楼宇内使用的暖通空调系统 (HVAC) 通常效率低下，原因在于这些系统未考虑不断变化的气候模式、多变的能耗或建筑物的热性能。

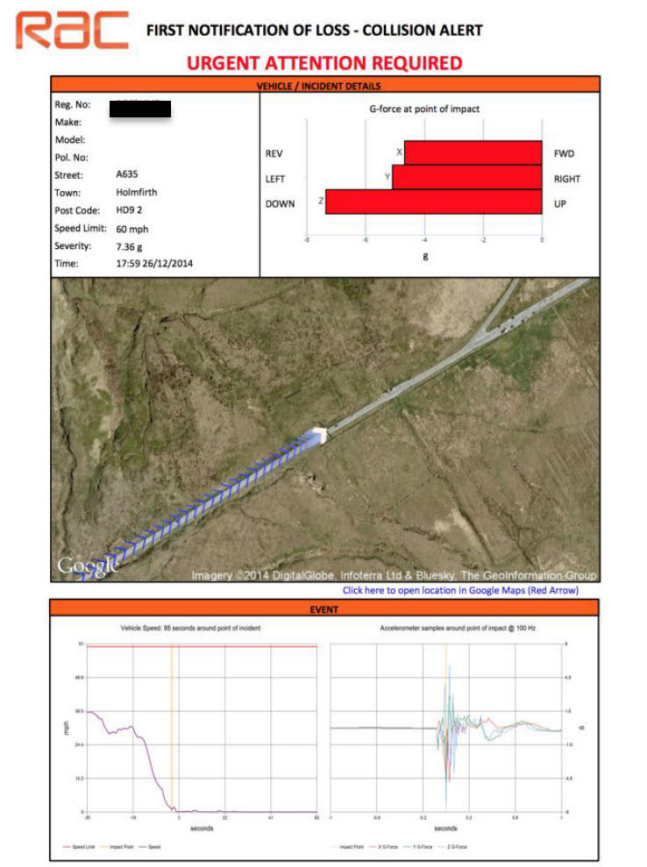
Building IQ 的基于云的软件平台可解决这个问题。该平台采用先进的算法和机器学习方法连续处理来自功率计、温度计和 HVAC 压力传感器的数千兆字节信息以及天气和能耗。更为特殊的是，机器学习可用于对数据分段和确定天然气、电力、蒸汽和太阳能对加热和冷却流程的相对贡献量。Building IQ 平台将大型商业楼宇内使用的 HVAC 在正常运行期间的能耗降低了 10% - 25%。

实际环境中的示例

检测低速撞车

RAC 是英国最大的汽车公司之一，拥有超过 8 百万名成员，为私人和商业汽车司机提供道路救援、保险和其他服务。

为了能够快速响应道路事故、减少碰撞事故和保险费用，RAC 开发了车载撞击感应系统，该系统采用先进的机器学习算法检测低速碰撞，并且可将这些行为与更常见的驾驶行为（例如驶过路面减速带或路面凹坑）区分开。独立测试数据显示 RAC 系统在碰撞检测测试中获得的准确度达到 92%。



了解更多

准备更深入地钻研？深入了解这些资源以了解有关机器学习方法、示例和工具的更多信息。

观看

机器学习一点通 34:34

使用信号处理和机器学习功能进行传感器数据分析 42:45

阅读

机器学习博客：社会网络分析、文本挖掘、贝叶斯推理及更多

机器学习的挑战：选择最佳模型并避免过度拟合

深入了解

应用 MATLAB 的机器学习示例

机器学习解决方案

使用分类学习器应用程序进行数据分类

第 2 部分: 快速入门

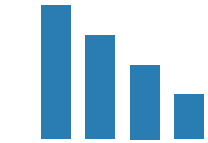
极少一帆风顺

在机器学习中，极少能够自始至终一帆风顺——您将会发现自己始终在改变和尝试各种不同思路和方法。本章介绍系统化机器学习工作流程，重点介绍整个流程中的一些关键决策点。

机器学习的挑战

大多数机器学习挑战都与数据处理和查找正确的模型相关。

数据会以各种形式和大小出现。真实数据集可能比较混乱、不完整，并且采用各种不同格式提供。您可能只有简单的数值型数据。但有时您要合并多种不同类型的数据，例如传感器信号、文本，以及来自于相机的图像数据流。



预处理数据可能需要掌握专业知识和工具。例如，对象检测算法训练中的特征选取，需要掌握图像处理领域的专业知识。不同类型的数据需要采用不同的预处理方法。

找到拟合数据的最佳模型需要时间。如何选择正确的模型是一项平衡过程。高度灵活的模型由于拟合了噪声的细微变化而造成了过度拟合。另一方面，简单的模型可能要有更多的假设条件。这些始终是在模型速度、准确性和复杂性之间权衡取舍。

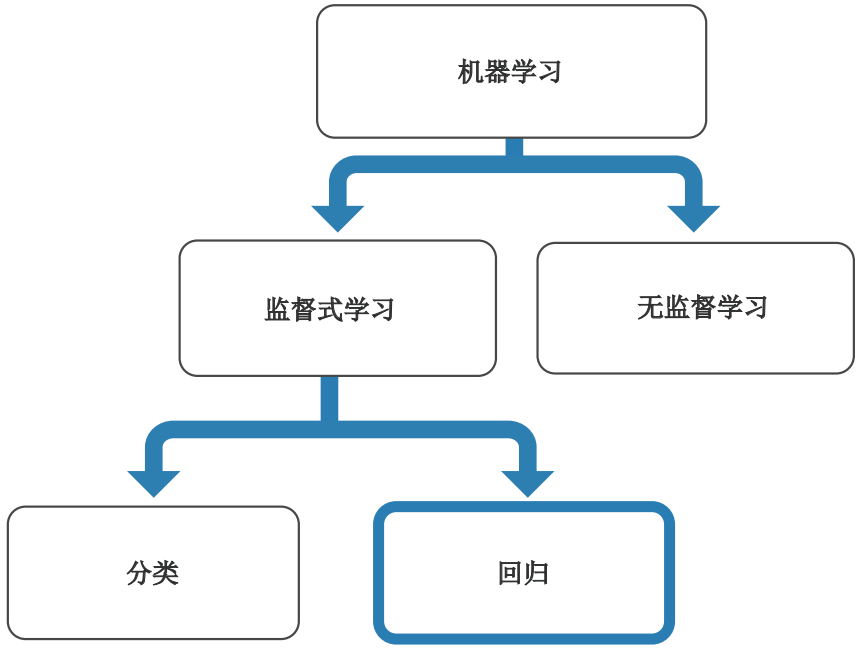
听起来很让人望而生畏？不要泄气。要记住，反复尝试和出错才是机器学习的核心——如果一个方法或算法不起作用，只需尝试另一个。但系统化工作流程有助于创造一个顺利的开端。

开始之前需要考虑的问题

每个机器学习工作流程都从以下三个问题开始：

- 您要处理哪种类型的数据？
- 您想要从中获得哪些洞察力？
- 这些洞察力将如何应用以及在哪里应用？

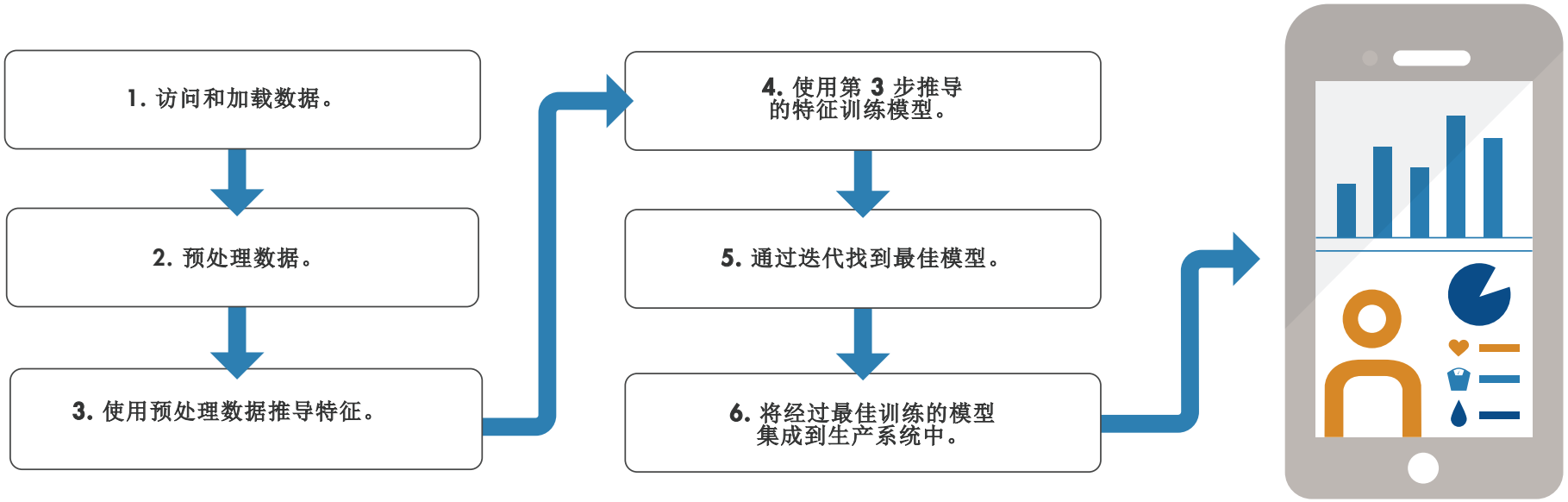
回答这些问题有助于确定您采用监督式学习还是无监督学习。



在以下情况下选择监督式学习：您需要训练模型进行预测（例如温度和股价等连续变量的未来值）或者分类（例如根据网络摄像头的录像片段确定汽车的技术细节）。

在以下情况下选择无监督学习：您需要深入了解数据并希望训练模型找到好的内部表示形式，例如将数据拆分到集群中。

工作流程概览



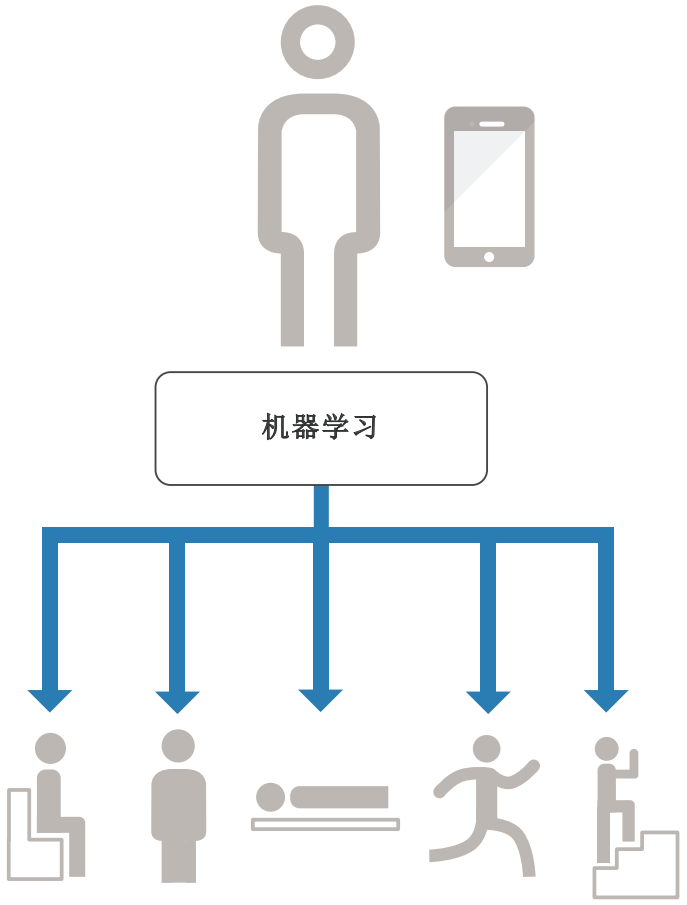
在接下来的章节中，我们将以健康监控应用程序为例更详细地介绍具体步骤。整个工作流程将在 MATLAB® 中完成。

训练模型对身体活动进行分类

本示例基于手机的健康监控应用程序。输入数据包含通过手机的加速计和陀螺仪提供的三轴传感器数据。获得的响应（或输出）为日常的身体活动，例如步行、站立、跑步、爬楼梯或平躺。

我们希望使用输入数据训练分类模型来识别这些活动。由于我们的目标是分类，因此我们将应用监督式学习。

经过训练的模型（或分类器）将被集成到应用程序中，帮助用户跟踪记录全天的身体运动水平。



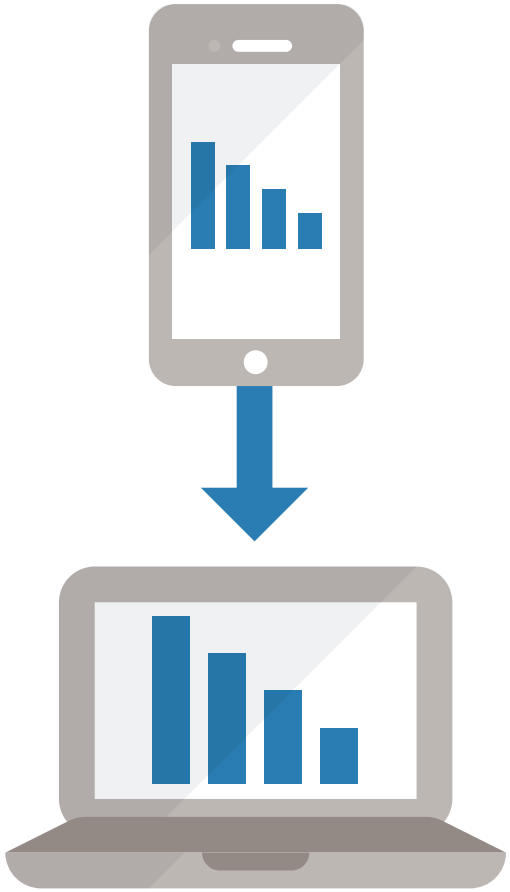
1 步骤 1：加载数据

要加载加速计和陀螺仪的数据，我们要执行以下操作：

1. 手持手机坐下，记录手机的数据，然后将其存储在标记为“坐”的文本文件中。
2. 手持手机站着，记录手机的数据，然后将其存储在第二个标记为“站立”的文本文件中。
3. 重复上述步骤，直到我们获得希望分类的每个活动的数据。

我们将标记的数据集存储在文本文件中。诸如文本或 CSV 等平面文件格式更易于处理，可以直接导入数据。

机器学习算法还不够智能，无法辨别噪声和有价值的信息之间的差异。使用数据进行训练之前，我们需要确保数据简洁和完整。



2 步骤 2：预处理数据

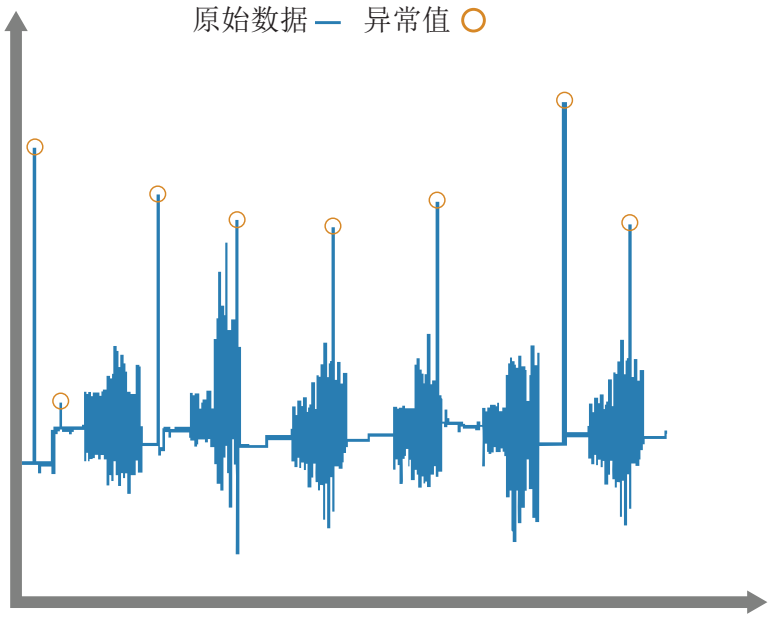
我们将数据导入 MATLAB，然后为每个带有标签的数据集绘图。
要预处理数据，我们可以执行以下操作：

1. 查找位于绝大多数数据所在范围之外的异常值数据点。

我们必须确定异常值能否忽略或者它们是否表示模型应该考虑的现象。
在我们的示例中，可以安全地将其忽略掉（这些异常值是我们记录数据时无意中移动所产生的结果）。

2. 检查是否有缺失值（在记录期间我们可能会因为断开连接而丢失数据）

我们可以简单地忽略这些缺失值，但这会减少数据集的大小。
或者，我们可以通过插值或使用其他示例的参照数据来作为缺失值的近似。



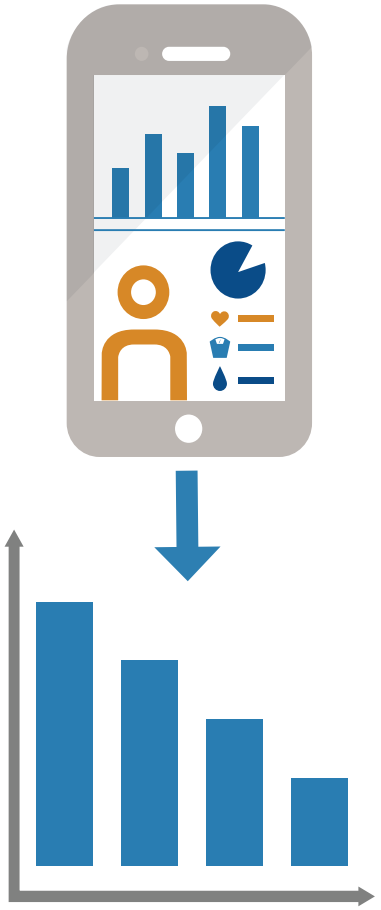
活动跟踪记录数据中的异常值。

在许多应用程序中，异常值提供了关键信息。例如，在信用卡欺诈检测应用程序中，它们表示超出客户常规购买模式的购买行为。

2 步骤 2：预处理数据（续）

3. 从加速计数据中删除重力效应数据，这样我们的算法就能专注处理物体的移动情况，而非手机的移动情况。我们通常使用简单的高通滤波器（例如双二阶滤波器）来处理此问题。
4. 将数据分为两组。我们保存部分数据用于测试（测试组），将其余数据（训练组）用于构建模型。这种方法被称为保留方法，是一种有用的交叉验证技术。

使用建模过程中未使用过的数据测试模型，您就能了解模型如何处理未知数据。



3 步骤 3：推导特征

推导特征（也称为特征工程或特征提取）是机器学习中最为重要的部分之一。此过程可将原始数据转换为机器学习算法可以使用的信息。

作为活动跟踪记录者，我们希望提取那些捕获了加速计数据的频谱的特征。这些特征将会帮助算法区分步行（低频）和跑步（高频）。我们创建了一个包含选定特征的新表。

使用特征选择执行以下操作：

- 提高机器学习算法的准确性
- 提升高维数据集的模型性能
- 提高模型的可解释性
- 防止过度拟合

3 步骤 3：推导特征（续）

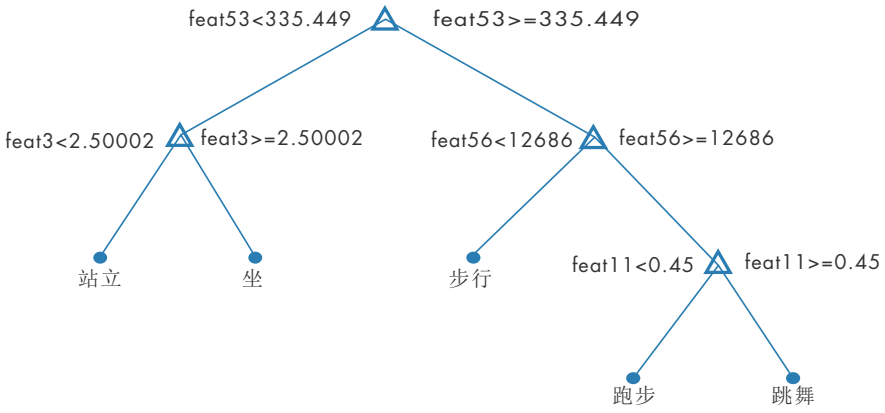
您可以推导出的特征数量只会受您的想象力限制。然而，我们通常可以采用许多技术来处理不同类型的数据。

数据类型	特征选择任务	技术
传感器数据	从原始传感器数据中提取信号属性以生成更高级别的信息	**峰值分析** – 执行 FFT（快速傅立叶变换），然后确定主导频率 **脉冲和转换指标** – 推导出信号特征，例如上升时间、下降时间和稳定时间 **频谱测量** – 绘图信号功率、带宽、平均频率和中值频率
图像和视频数据	提取边缘位置、分辨率和颜色等特征。	**视觉关键词袋** – 为诸如边缘、角和斑点等局部图像特征创建直方图 **方向梯度直方图 (HOG)** – 为局部梯度方向创建直方图 **最小值特征值算法** – 检测图像上的角位置 **边缘检测** – 识别亮度发生急剧变化的点
事务处理数据	计算增强数据信息的派生值	**时间戳分解** – 将时间戳分解为诸如天和月等分量 **汇总值计算结果** – 创建更高级别的特征，例如特殊事件发生的总次数

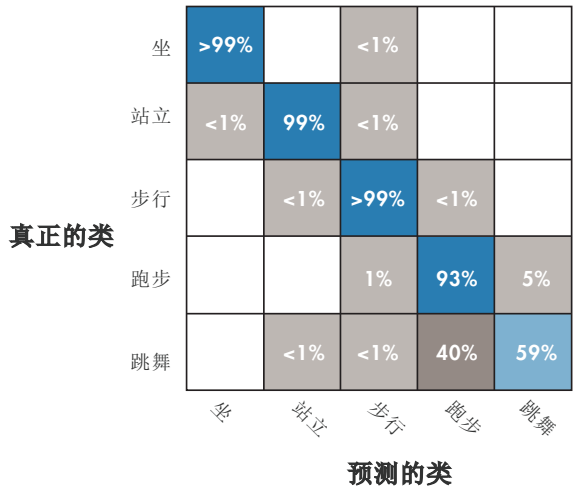
4 步骤 4：构建和训练模型

构建模型时，最好先从构建简单模型开始；这样可以更快的运行并且更易于解释。

我们从构建基本决策树开始。



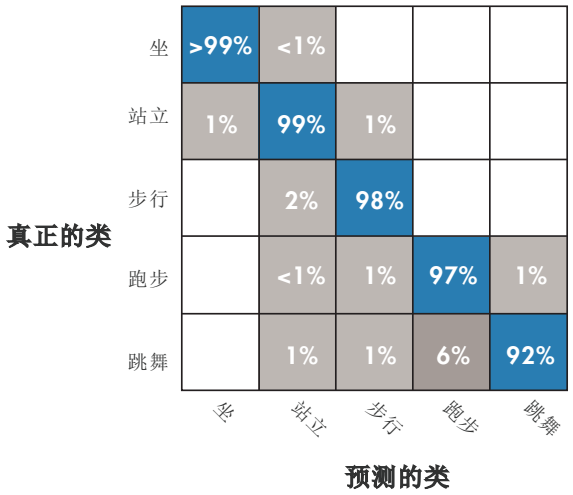
为了了解决策树的执行情况，我们绘制了混淆矩阵，该表将模型产生的分类与我们在步骤 1 中创建的实际分类标签进行了比较。



此混淆矩阵显示我们的模型难以区分步行和跑步。决策树可能无法处理这种类型的数据。我们尝试一些不同的算法。

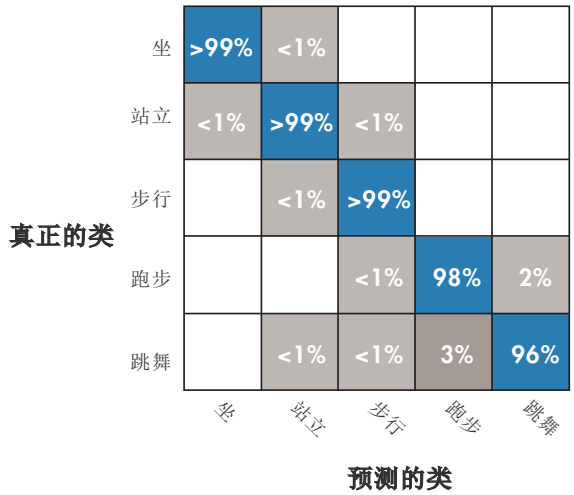
4 步骤 4：构建和训练模型（续）

我们尝试使用 K-近邻算法 (KNN)，这种简单的算法可以存储所有训练数据，将新点与训练数据进行比较，然后返回最近的“K”个点的多数类别。。相比于简单决策树提供的 94.1% 的准确度，此算法的准确度能达到 98%。混淆矩阵也更易于查看：



然而，KNN 需要占用大量内存才能运行，因为该算法需要使用所有训练数据来进行预测。

我们尝试了线性判别模型，但也无法改进结果。最后，我们尝试了多类支持向量机 (SVM)。SVM 处理的结果非常好，我们现在获得的准确度为 99%：



我们通过对模型的更换和不断尝试不同算法实现了目标。如果我们的分类器仍无法可靠地区分步行和跑步，我们将寻找方法来改进这个模型。

5 步骤 5：改进模型

可通过两种不同方式改进模型：简化模型或增加模型的复杂度。

简化

首先，我们要找机会减少特征的数量。热门的特征减少技术包括：

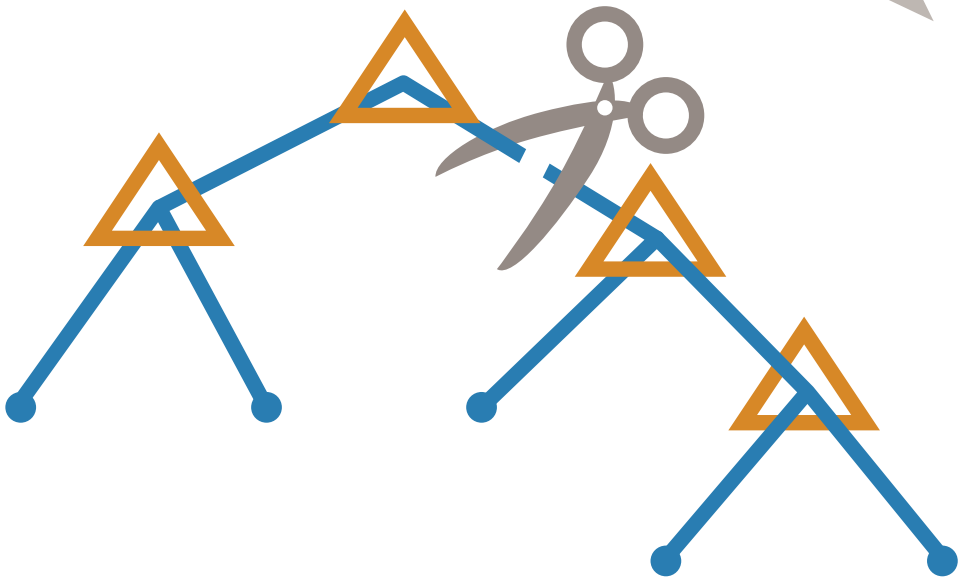
- 相关矩阵 – 可显示变量之间的关系，因此可以删除并非高度相关的变量（或特征）。
- 主分量分析 (PCA) – 可消除冗余，具体方法是找到一组捕获了原始特征的关键区别的特征，并推导出数据集中存在的强模式。
- 序列特征减少 – 采用迭代的方式减少模型的特征，直到无法改进模型性能为止。

接下来，我们寻找方法来简化模型本身。我们可以通过以下方式实现：

- 修剪决策树的分支
- 从集成结构中删除学习器

一个好的模型应该只包含预测能力最强的特征。具有很好泛化能力的简单模型要由于泛化能力较弱或未能完善训练处理新数据的复杂模型。

在机器学习中，和许多其他计算流程一样，经过简化的模型更易于理解、更稳健、计算效率更高。



5 步骤 5：改进模型（续）

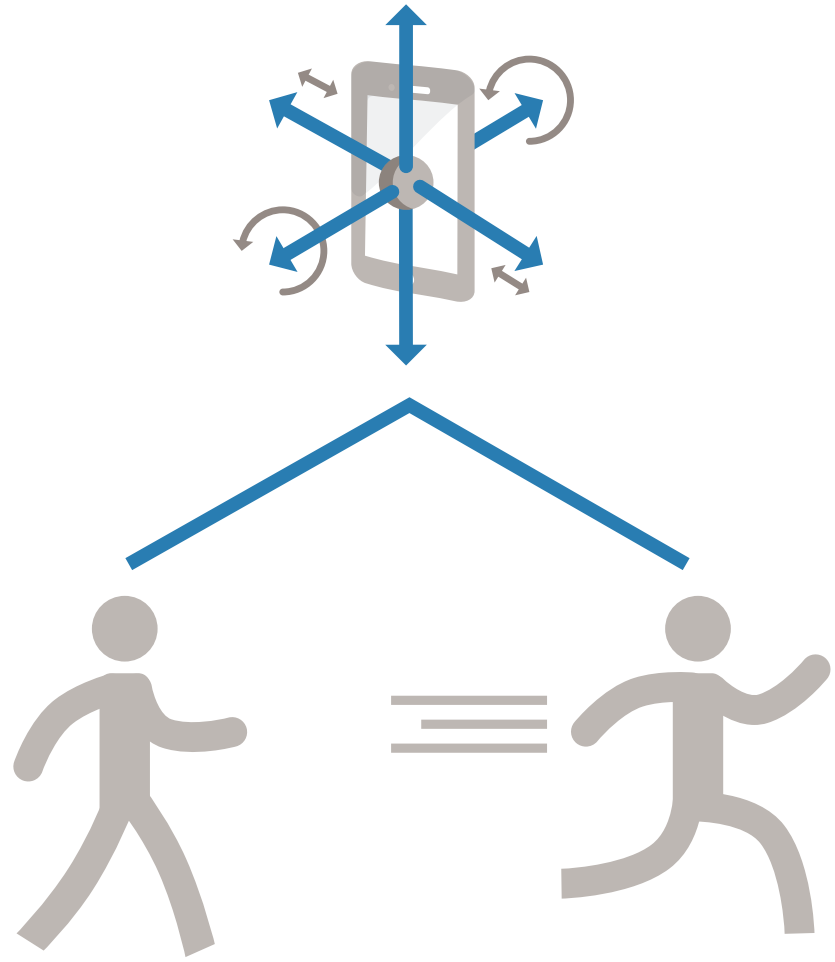
增加复杂度

如果我们的模型由于过度泛化而无法区分步行和跑步，我们就需要寻找新方法来进一步完善该模型。我们可以通过以下方式实现：

- 使用模型组合 – 将多个简单的模型组合成强模型，这样提供的数据趋势要优于其中任何一个简单模型单独提供的趋势。
- 添加更多数据源 – 查看陀螺仪和加速计的数据。陀螺仪记录活动期间手机所处的方向。此数据可提供不同活动的唯一标志，例如，可能存在一个跑步所独有的加速度和旋转的组合。

对模型进行调整后，我们使用预处理期间保留的测试数据验证其性能。

如果模型能够在测试数据集上对活动实现可靠的分类，我们就能将其应用到手机上，开始跟踪记录。



了解更多

准备更深入地钻研？查看这些资源以深入了解有关机器学习方法、示例和工具的更多信息。

观看

机器学习一点通 34:34

使用信号处理和机器学习进行传感器数据分析 42:45

阅读

监督式学习工作流程和算法

运用 MATLAB 分析而获得的数据驱动洞察力：能量负荷预测案例研究

深入了解

应用 MATLAB 的机器学习示例

使用分类学习器应用程序进行数据分类

第 3 部分: 应用无监督学习

何时考虑无监督学习

无监督学习适用的场景是，您想要探查数据，但还没有特定目标或不确定数据包含什么信息。这也是减少数据维度的好方法。

无监督学习技术

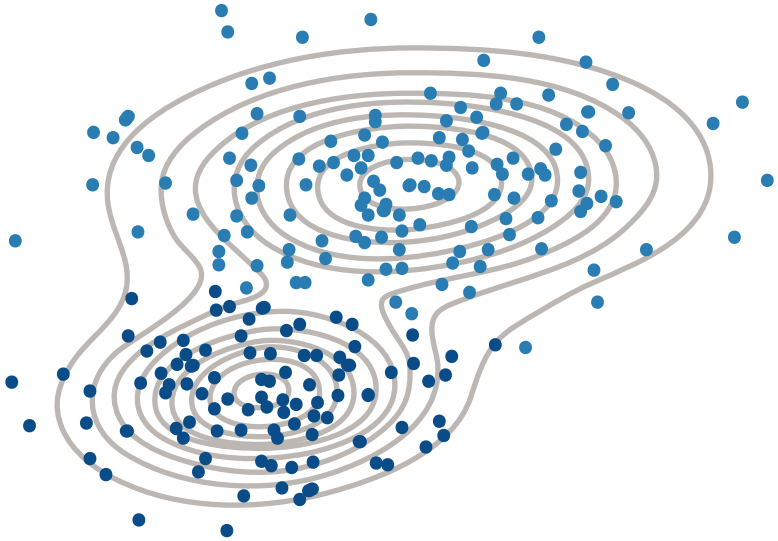
如我们在第 1 部分所见，绝大多数无监督学习技术是聚类分析的形式。

在聚类分析中，根据某些相似性的量度或共有特征把数据划分成组。采用聚类的组织形式，同一类（或簇）中的对象非常相似，不同类中的对象截然不同。

聚类算法分为两大类：

- 硬聚类，其中每个数据点只属于一类
- 软聚类，其中每个数据点可属于多类

如果您已经知道可能的数据分组，则可以使用硬聚类或软聚类技术。



高斯混合模型可用来将数据分成两类。

如果您不知道数据可能如何分组：

- 使用自我组织的特征图或层次聚类，查找数据中可能的结构。
- 使用聚类评估，查找给定聚类算法的“最佳”组数。

常见硬聚类算法

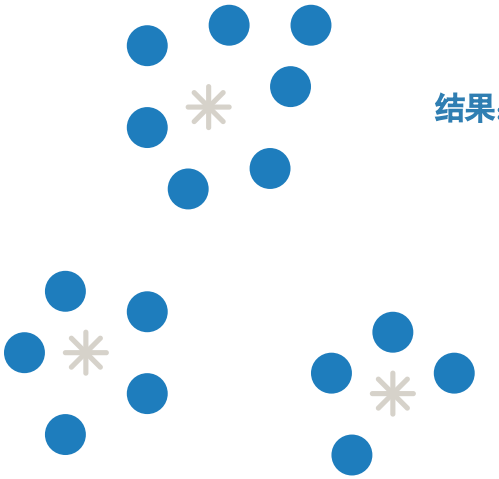
k-均值

工作原理

将数据分割为 k 个相互排斥的类。一个点在多大程度上适合划入一个类由该点到类中心的距离来决定。

最佳使用时机...

- 当聚类的数量已知时
- 适用于大型数据集的快速聚类



结果: 聚类中心

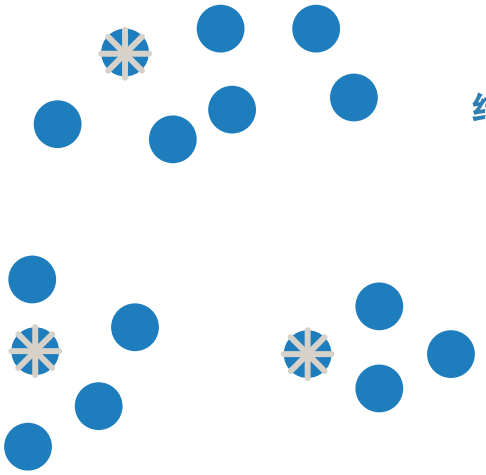
k-中心点

工作原理

与 k-均值 类似，但要求类中心与数据中的点契合。

最佳使用时机...

- 当聚类的数量已知时
- 适用于分类数据的快速聚类
- 扩展至大型数据集



结果: 与数据点契合的聚类中心

常见硬聚类算法（续）

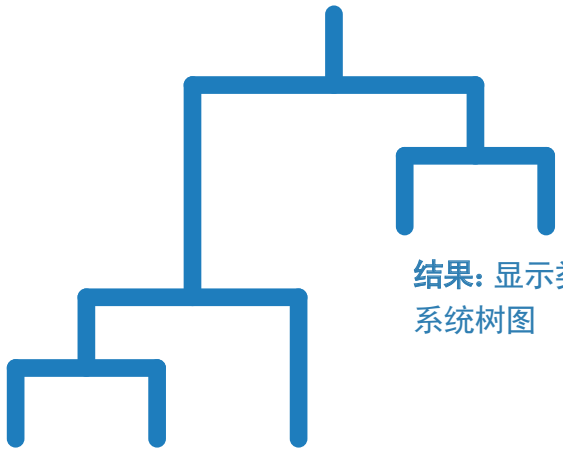
层次聚类

工作原理

通过分析成对点之间的相似度并将对象分组到一个二进制的层次结构树，产生聚类的嵌套集。

最佳使用时机...

- 当您事先不知道您的数据中有多少类时
- 您想要可视化地指导您的选择



结果：显示类之间层次关系的系统树图

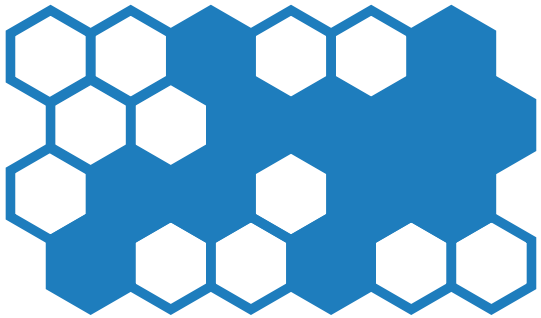
自组织映射

工作原理

基于神经网络的聚类，将数据集变换为保留拓扑结构的 2D 图。

最佳使用时机...

- 采用 2D 或 3D 方式可视化高维数据
- 通过保留数据的拓扑结构（形状）降低数据维度



结果：
低维度（通常 2D）表现形式

常见硬聚类算法（续）

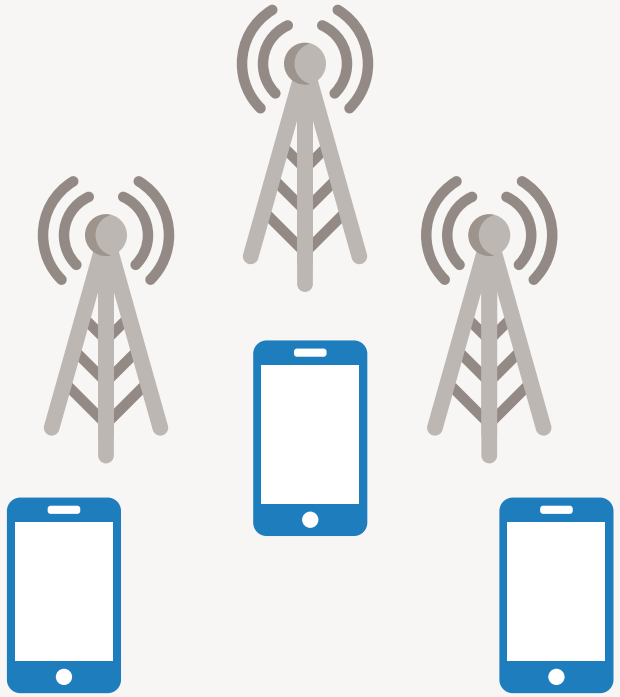
示例：使用 k-均值 聚类为手机信号塔选址

移动电话公司想知道手机信号塔的数量和位置，以便提供最可靠的服务。为实现最佳信号接收，这些塔必须位于人群聚集的地方。

工作流程从最初猜想需要划分多少个人群开始。为了评估这个猜想，工程师采用三个塔和四个塔比较服务效果，查看每种情形下的聚类有多好（换句话说，信号塔提供服务的效果如何）。

一部电话一次只能与一个塔通信，所以这是硬聚类问题。该团队使用 k-均值聚类，因为 k-均值将数据中的每个观察点视为空间中的一个点。找到了一种分割方法，每个类中的对象尽可能地相互靠近，并且尽可能远离其他类中的对象。

在运行算法之后，该团队能够准确地确定将数据分割成三个和四个类的结果。



常见软聚类算法

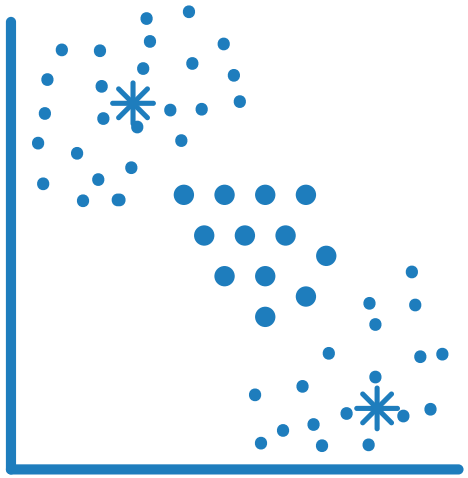
模糊 c-均值

工作原理

当数据点可能属于多个类时进行基于分割的聚类。

最佳使用时机...

- 当聚类的数量已知时
- 适用于模式识别
- 当聚类重叠时



结果： 聚类中心（类似于 k-均值），但有模糊性，所以点可能属于多个类

高斯混合模型

工作原理

基于分割的聚类，数据点来自具有一定概率的不同的多元正态分布。

最佳使用时机...

- 当数据点可能属于多个类时
- 当聚集的类具有不同的大小且含有相关结构时

结果： 一个高斯分布的模型，给出一个点在一个类中的概率

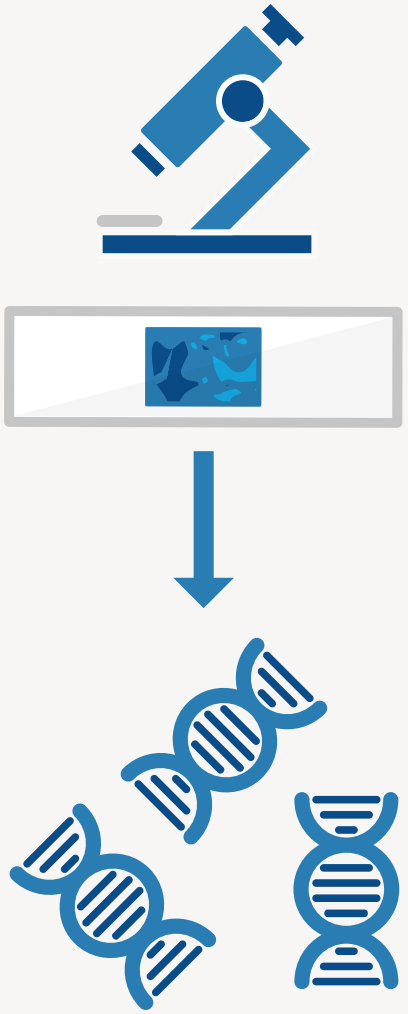
常见软聚类算法（续）

示例：使用模糊 c-均值聚类法分析基因表达数据

一个生物家团队正在通过微阵列分析基因表达数据，更好地了解涉及正常和异常细胞分裂的基因。（如果某个基因积极参与蛋白质生产之类的细胞功能，则称该基因为“已表达”。）

微阵列包含两个组织检体的表达数据。研究人员想要比较检体，确定某些基因表达模式是否与癌细胞增生有牵连。

在对数据进行预处理以消除噪声之后，他们对数据进行聚类。因为相同的基因可能涉及多个生物学过程，没有单个基因可能只属于一类。研究人员对数据运用模糊 c-均值算法。然后，他们对聚集生成的类进行可视化，识别具有类似行为方式的基因组。



用降维的方法改进模型

机器学习是一种发现大数据集内部规律的有效方法。但较大的数据增加了复杂度。

随着数据集越来越大，您经常需要减少特征或**维度**的数量。

示例：EEG 数据减缩

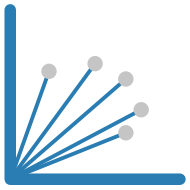
假设您有捕获脑电活动的脑电图 (EEG) 数据，您想使用此数据预测未来的癫痫发作。使用许多导线捕获数据，每根导线对应原始数据集中的一个变量。每个变量都包含噪声。为使您的预测算法更稳健，您使用降维技术生成数量较少的特征。由于这些特征是从多个传感器计算出来的，所以不太容易受单个传感器中的噪声影响，如果您直接使用原始数据，则噪声的影响会非常明显。

常见降维技术

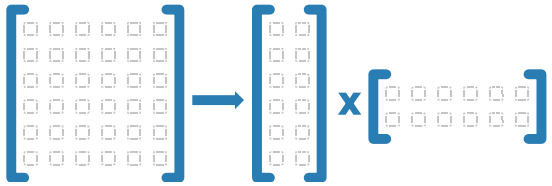
三个最常用的降维技术是：

主成分分析 (PCA) — 对数据执行线性变换，使您的高维数据集中的绝大多数方差或信息被前几个主成分捕获。第一个主成分将会捕获大部分方差，然后是第二个主成分，以此类推。

因子分析 — 识别您的数据集中各变量之间潜在的相关性，提供数量较少的未被发现的潜在因子或公共因子的一种表现方式。



非负矩阵分解 — 当模型项必须代表非负数（比如物理量）时使用。



使用主成分分析

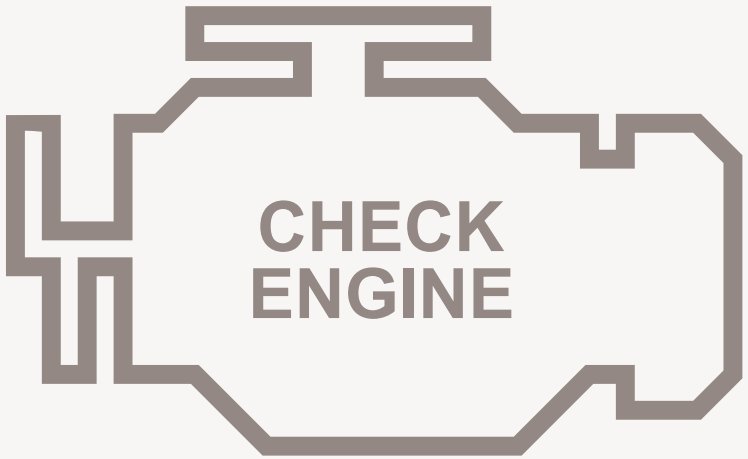
在有许多变量的数据集中，变量组经常一起移动。PCA 充分利用这种信息冗余，通过原始变量的线性组合生成新变量，使少数新变量能够捕获大多数信息。

每个主成分都是原始变量的线性组合。因为所有主成分互不相关，所以没有冗余信息。

示例：发动机健康状况监测

您有一个数据集，包括了对发动机上不同传感器的测量（温度、压力、排放等）。尽管大量数据来自健康的发动机，传感器也会捕获来自需要维护的发动机的数据。

查看任何一个传感器，可能看不出任何明显的异常。发动机异常，通过应用 PCA，您可以变换此数据，使传感器测量中的大部分变动被少数的主成分捕获。与观察原始传感器数据相比，通过检查这些主成分来区别健康和不健康的发动机比较容易。



使用因子分析

您的数据集可能包含重叠的已测变量，意味着这些变量相互依赖。通过因子分析，可将模型拟合到多元数据来评估这种相互依存关系。

在因子分析模型中，已测变量依赖数量较少的未发现（潜在）因子。因为每个因子都可能影响多个变量，所以称为公因子。假定每个变量都取决于公因子的线性组合。

示例：跟踪股价变动

在 100 个星期的时间里，对十家公司记录了股价的百分比变化。这十家公司，有四家是科技公司，三家从事金融业，还有三家从事零售业。假设相同行业的公司股价将随经济环境的变化而一同变化，这似乎很合理。因子分析可以提供数量证据来支持这一假定。

使用非负矩阵因式分解

此降维技术基于特征空间的低秩逼近。除了减少特征数量以外，还保证特征为非负数，从而产生遵守诸如物理量非负等特征的模型。

示例：文本挖掘

假设您想要探查多个网页间词汇和风格的变化。您创建一个矩阵，其中每行对应一个网页，每列对应一个单词（“the”、“a”、“we”等）。数据将是一个特定词出现在特定页面上的次数。

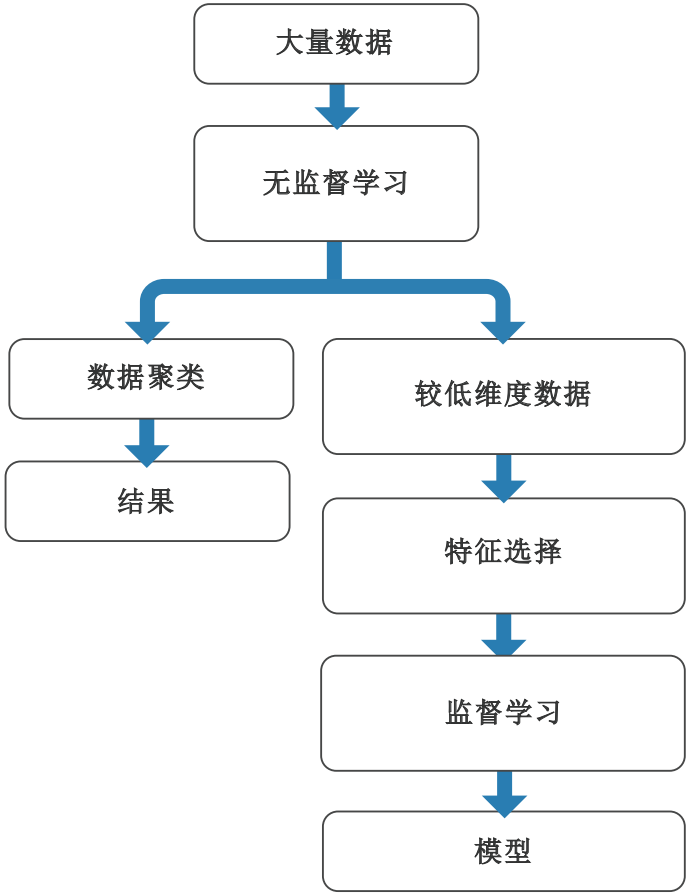
由于英语有一百多万个单词，所以您应用非负矩阵因式分解，创建任意数量的特征，表示高级别概念，而不是一个个单词。运用这些概念，更容易区分新闻、教育内容和在线零售内容。

后续步骤

在本部分中，我们详细介绍了无监督学习的硬聚类和软聚类算法，提供了一些为您的数据选择合适算法的技巧，展示了减少数据集内的特征数量如何改进模型性能。至于后续步骤：

- 无监督学习可能是您的最终目标。例如，如果您做市场研究并想根据网站行为有针对性地划分消费群体，那么，聚类算法几乎肯定能给您想要寻求的结果。
- 另一方面，您可能想使用无监督学习，作为监督式学习的预处理步骤。例如，应用聚类技术得出数量较少的特征，然后使用这些特征作为训练分类器的输入。

在第 4 部分，我们将探索监督学习算法和技术，了解如何通过特征选择、特征减缩和参数调节来改进模型。



了解更多

准备更深入地钻研？浏览以下无监督学习资源。

聚类算法和技术

k-均值

使用 K-均值和层次聚类来发现数据中的自然模式

使用 K-均值和自组织映射进行基因聚类

使用 K-均值聚类实现基于颜色的分割

分层聚类

基于连接的聚类

鸢尾花聚类

自组织映射

使用自组织映射进行数据聚类

模糊 c-均值

使用模糊 C-均值聚类法对拟随机数据进行聚类

高斯混合模型

高斯过程回归模型

对来自高斯分布混合的数据进行聚类

使用软聚类法对高斯混合数据进行聚类

调节高斯混合模型

图像处理示例：使用高斯混合模型检测汽车

降维

使用 PCA 分析美国城市的生活质量

使用因子分析法分析股价

非负矩阵因式分解

执行非负矩阵因式分解

使用减法聚类对郊区通勤进行建模

第 4 部分: 应用监督式学习

MathWorks®

何时考虑监督式学习

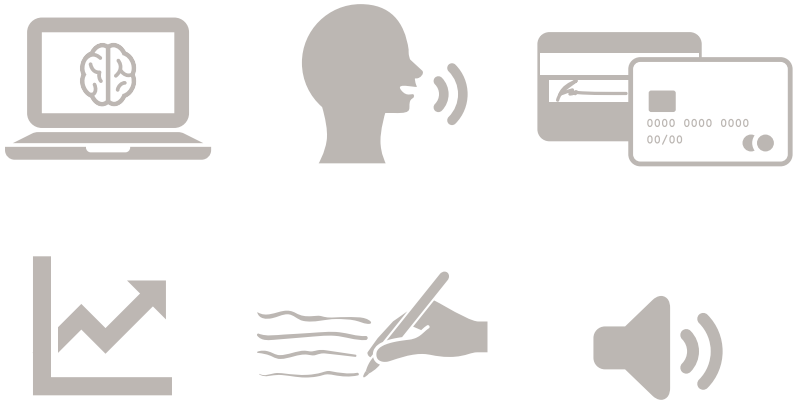
监督式学习算法接受已知的输入数据集合（训练集）和已知的对数据的响应（输出），然后训练一个模型，为新输入数据的响应生成合理的预测。如果您尝试去预测现有数据的输出，则使用监督式学习。

所有的 “监督式学习”改成“监督学习”技术

监督学习技术可分成分类或者回归的形式。

分类技术预测离散的响应 — 例如，电子邮件是真正邮件还是垃圾邮件，肿瘤是小块、中等还是大块。分类模型经过训练后，将数据划分成类别。应用软件包括医学成像、语音识别和信用评分。

回归技术预测连续的响应 — 例如，电力需求中温度或波动的变化。应用软件包括预测股价、笔迹识别和声信号处理。



- 您的数据能否进行标记或分类？如果您的数据能分为特定的组或类，则使用分类算法。
- 处理数据范围？如果您的响应性质是一个实数（比如温度，或一件设备发生故障前的运行时间），则使用回归方法。

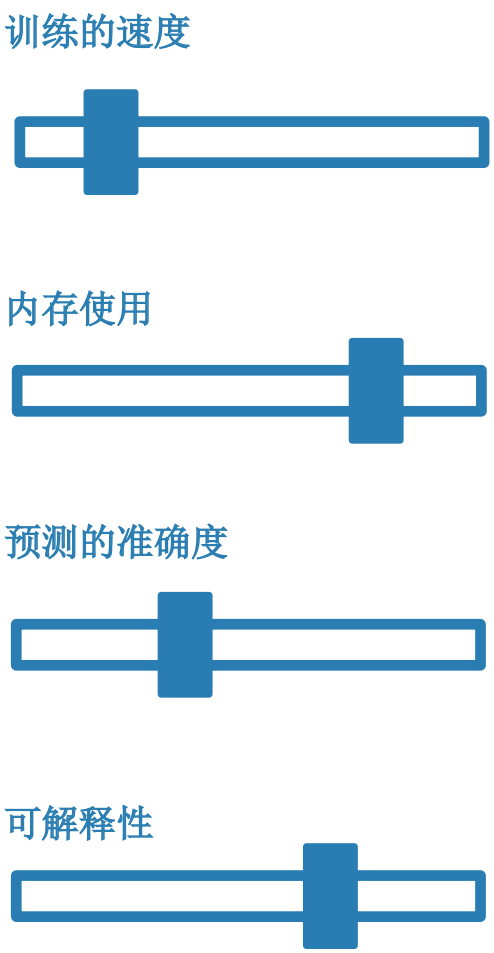
选择合适的算法

如我们在第 1 部分所见，选择机器学习算法是一个试错过程。
同时也是算法具体特性的一种权衡，比如：

- 训练的速度
- 内存使用
- 对新数据预测的准确度
- 透明度或可解释性（您对算法做出预测的理由的理解难易程度）

我们详细介绍最常用的分类和回归算法。

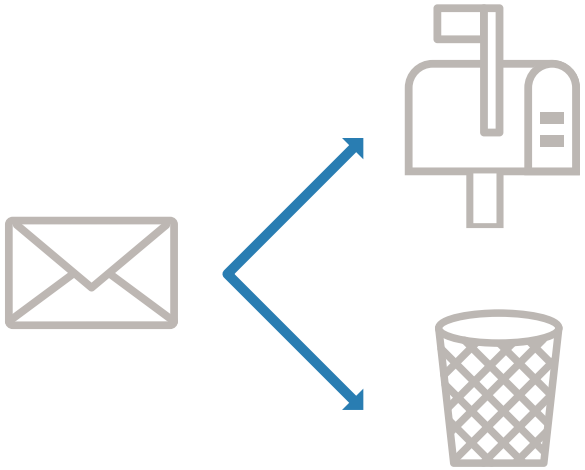
使用较大的训练数据集生成的模型通常对新数据归纳得比较完善。



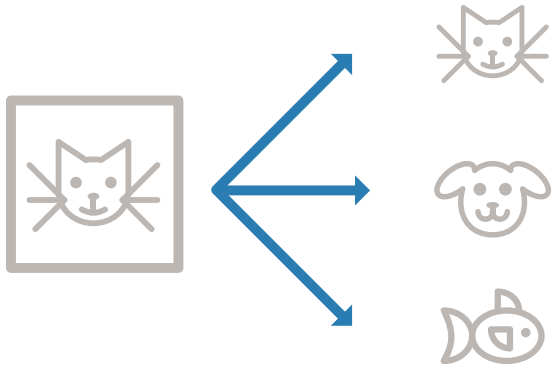
二分类与所有的“多类分类”改为“多分类”

在处理分类问题时，一开始就要确定该问题是二元问题还是多类问题。对于二元分类问题，单个训练或测试项目（实例）只能分成两类 — 例如，如果您想确定电子邮件是真正邮件，还是垃圾邮件。对于多类分类问题，可以分成多个类 — 例如，如果您想训练一个模型，将图像分类为狗、猫或其它动物。

请记住，多类分类问题一般更具挑战性，因为需要比较复杂的模型。



某些算法（例如逻辑回归）是专门为二分类问题设计的。在训练过程中，这些算法往往比多类算法更高效。



常见分类算法

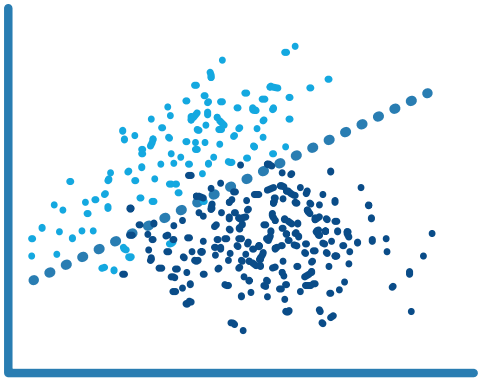
逻辑回归

工作原理

适合可以预测属于一个类或另一个类的二元响应概率的模型。因为逻辑回归比较简单，所以常用作二分类问题的起点。

最佳使用时机...

- 当数据能由一个线性边界清晰划分时
- 作为评估更复杂分类方法的基准



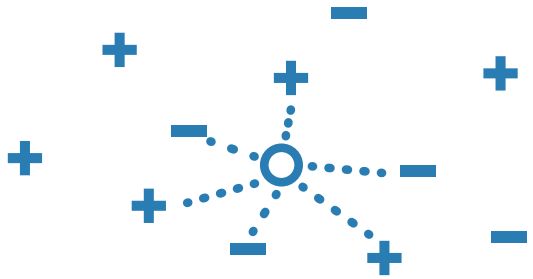
k 最近邻 (kNN)

工作原理

kNN 根据数据集内类的最近邻关系划分对象的类别。kNN 预测假定相互靠近的对象是相似的。距离量度（如欧氏距离、绝对值距离、夹角余弦和 Chebychev 距离）用来查找最近邻。

最佳使用时机...

- 当您需要简单算法来设立基准学习规则时
- 当无需太关注 训练模型的内存使用时
- 当无需太关注 训练模型的预测速度时



常见分类算法（续）

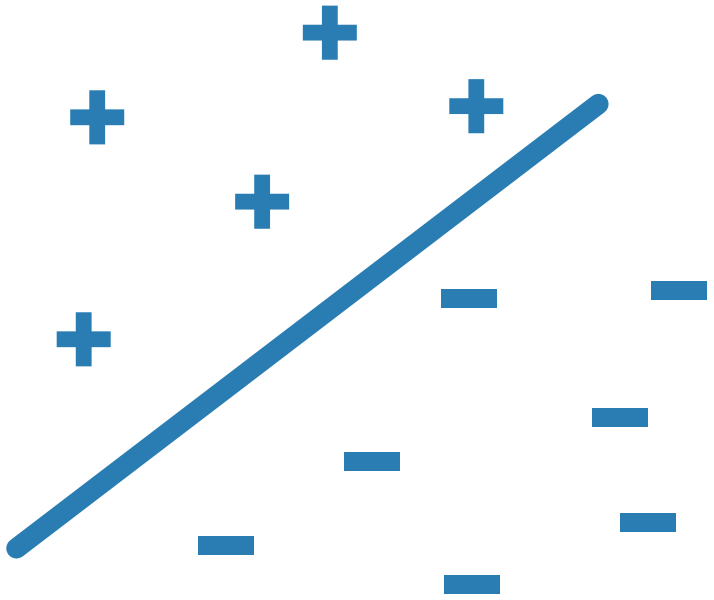
支持向量机 (SVM)

工作原理

通过搜索能将全部数据点分割开的判别边界（超平面）对数据进行分类。当数据为线性可分离时，SVM 的最佳超平面是在两个类之间具有最大边距的超平面。如果数据不是线性可分离，则使用损失函数对处于超平面错误一边的点进行惩罚。SVM 有时使用核变换，将非线性可分离的数据变换为可找到线性判定边界的更高维度。

最佳使用时机...

- 适用于正好有两个类的数据（借助所谓的纠错输出码技术，也可以将其用于多类分类）
- 适用于高维、非线性可分离的数据
- 当您需要一个简单、易于解释、准确的分类器时



常见分类算法（续）

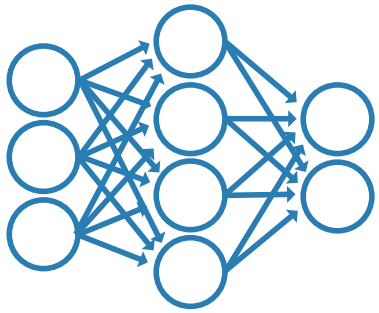
神经网络

工作原理

受人脑的启发，神经网络由高度互连的神经元网络组成，这些神经元将输入与所需输出相关联。通过反复修改联系的强度，对网络进行训练，使给定的输入映射到正确的响应。

最佳使用时机...

- 适用于高度非线性系统建模
- 当数据逐渐增多，而您希望不断更新模型时
- 当您的输入数据可能有意外变动时
- 当模型可解释性不是主要考虑因素时



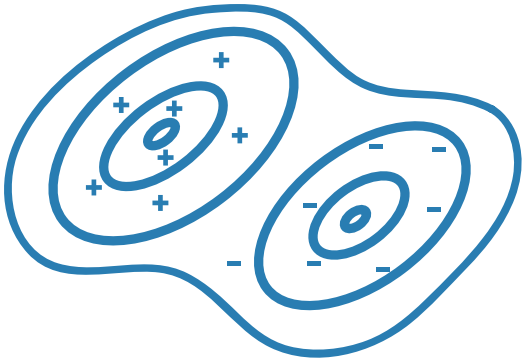
朴素贝叶斯

工作原理

朴素贝叶斯分类器假设类中某一具体特征的存在与任何其他特征的存在不相关。根据数据属于某个特定类的最高概率对新数据进行分类。

最佳使用时机...

- 适用于包含许多参数的小数据集
- 当您需要易于解释的分类器时
- 当模型会遇到不在训练数据中的情形时，许多金融和医学应用就属于这种情况



常见分类算法（续）

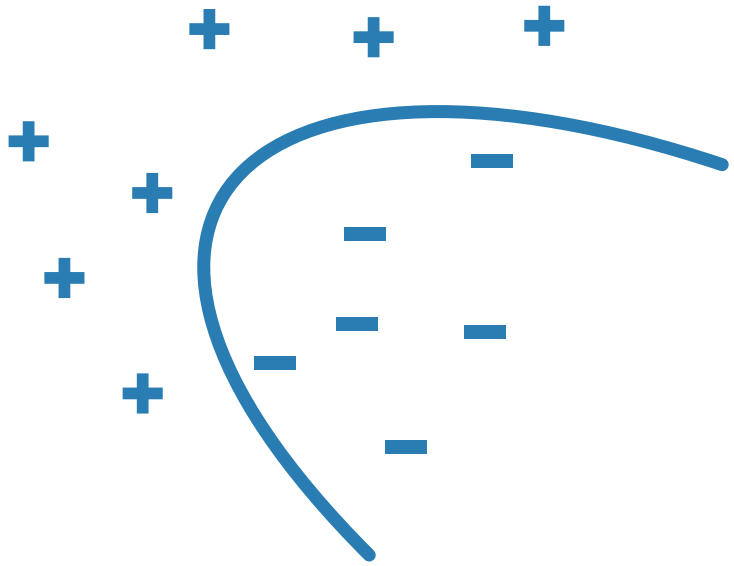
判别分析

工作原理

判别分析通过发现特征的线性组合来对数据分类。判别分析假定不同的类根据高斯分布生成数据。训练判别分析模型涉及查找每个类的高斯分布的参数。分布参数用来计算边界，边界可能为线性函数或二次函数。这些边界用来确定新数据的类。

最佳使用时机...

- 当您需要易于解释的简单模型时
- 当训练过程中的内存使用是需要关注的问题时
- 当您需要快速预测的模型时



常见分类算法（续）

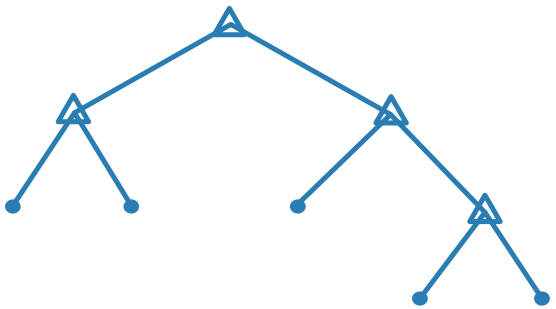
决策树

工作原理

利用决策树预测对数据响应的方法是，按照树中根节点（起始）到叶节点的顺序自上而下地决策。树由分支条件组成，在这些条件中，预测元的值与训练的权重进行比较。分支的数量和权重的值在训练过程中确定。附加修改或剪枝可用来简化模型。

最佳使用时机...

- 当您需要易于解释和快速拟合的算法时
- 最小化内存使用
- 当不要求很高的预测准确性时



保留英文，很少 Bagged 和 Boosted 决策树翻译这两个词

工作原理

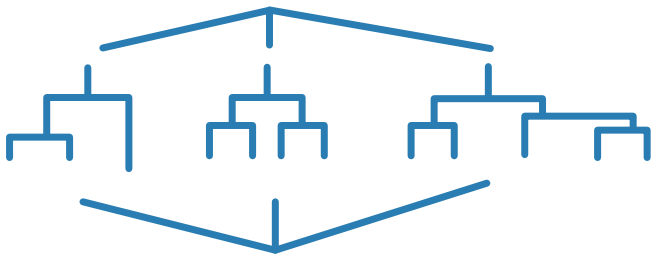
在这些集成方法中，几个“较弱”的决策树组合成一个“较强”的整体。

袋装决策树由根据从输入数据中自举的数据进行独立训练的树组成。

促进决策树涉及创建一个强学习器，具体方法是，迭代地添加“弱”学习器并调节每个弱学习器的权重，从而将重点放在错误分类的样本。

最佳使用时机...

- 当预测元为无序类别（离散）或表现非线性时
- 当无需太关注训练一个模型所用时间时



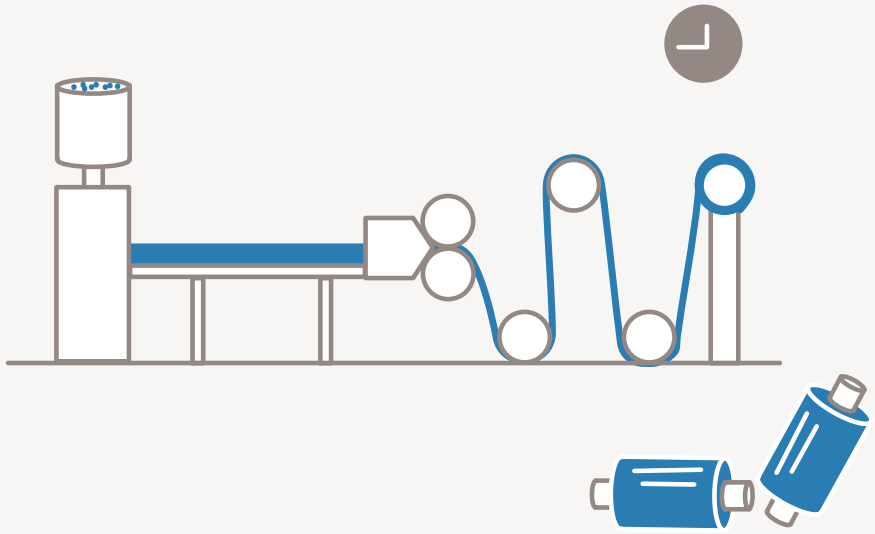
常见分类算法（续）

示例：生产设备的预测性维护

一家塑料加工厂每年生产大约 1800 万吨的塑料和薄膜产品。工厂的 900 名工人一年 365 天、一天 24 小时保证机器运转。

为达到机器故障率最小化，工厂效率最大化，工程人员开发运行状况监测和预测性维护应用软件，使用先进的统计和机器学习算法，找出机器的潜在问题，以便操作人员能够采取正确措施，防止发生严重问题。

在收集、清理和记录工厂中所有机器的数据后，工程人员评估几项机器学习技术，包括神经网络、k-最近邻、袋装决策树和支持向量机 (SVM)。对于每项技术，他们使用记录的机器数据训练一个分类模型，然后测试该模型预测机器问题的能力。测试表明，袋装决策树的整体集成是预测生产质量的最精确模型。



常见回归算法

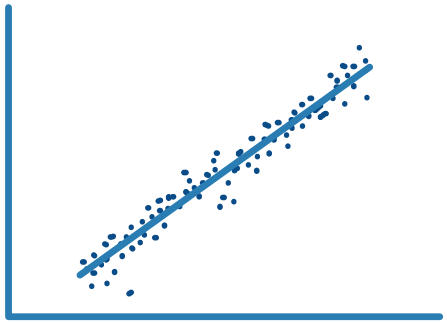
线性回归

工作原理

线性回归是一项统计建模技术，用来描述作为一个或多个预测元变量的线性函数的连续应变量。因为线性回归模型解释简单，
易于训练，所以通常是第一个要与新数据集拟合的模型。

最佳使用时机...

- 当您需要易于解释和快速拟合的算法时
- 作为评估其他更复杂回归模型的基准



非线性回归

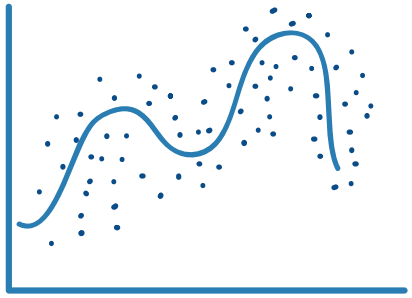
工作原理

非线性回归是一种有助于描述实验数据中非线性关系的统计建模技术。通常将非线性回归模型假设为参数模型，将该模型称为非线性方程。

“非线性”是指一个拟合函数，它是多个参数的非线性函数。例如，如果拟合参数为 b_0、b_1 和 b_2：方程式 $y = b_0+b_1x+b_2x^2$ 是拟合参数的线性函数，而 $y = (b_0x^{b1})/(x+b_2)$ 是拟合参数的非线性函数。

最佳使用时机...

- 当数据有很强的非线性趋势，不容易转化成线性空间时
- 适用于自定义模型与数据拟合



常见回归算法（续）

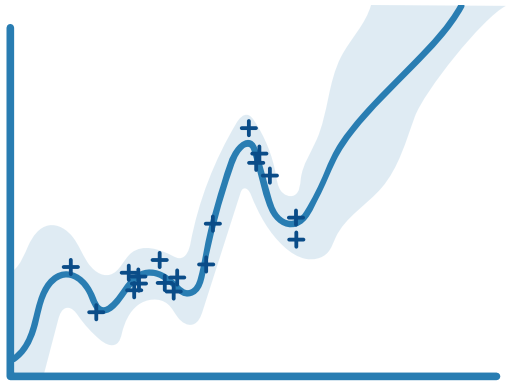
高斯过程回归模型

工作原理

高斯过程回归 (GPR) 模型是非参数模型，用于预测连续应变量的值。这些模型广泛用于对存在不确定情况下的插值进行空间分析的领域。GPR 也称为克里格法 (Kriging)。

最佳使用时机...

- 适用于对空间数据插值，如针对地下水分布的水文地质学数据
- 作为有助于优化汽车发动机等复杂设计的替代模型



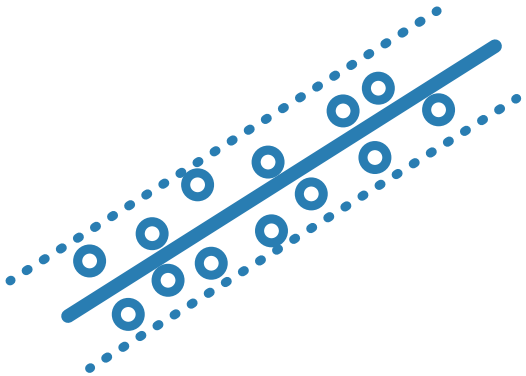
SVM 回归

工作原理

SVM 回归算法类似于 SVM 分类算法，但经过改良，能够预测连续响应。不同于查找一个分离数据的超平面，SVM 回归算法查找一个偏离测量数据的模型，偏离的值不大于一个小数额，采用尽可能小的参数值（使对误差的敏感度最小）。

最佳使用时机...

- 适用于高维数据（将会有大量的预测元变量）



常见回归算法（续）

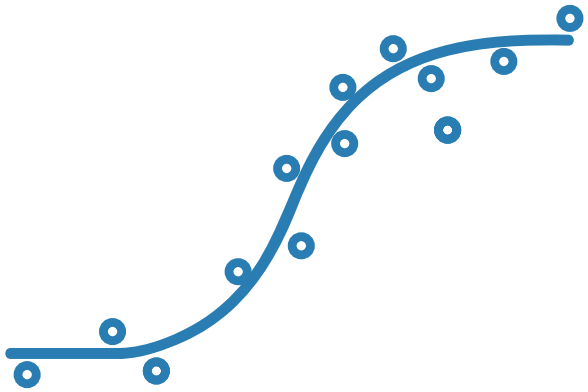
广义线性模型

工作原理

广义线性模型是使用线性方法的非线性模型的一种特殊情况。它涉及输入的线性组合与输出的非线性函数（连接函数）拟合。

最佳使用时机...

- 当应变量有非正态分布时，比如始终预期为正值的应变量



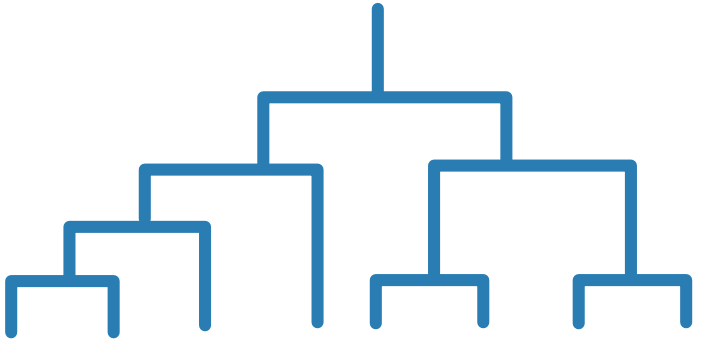
回归树

工作原理

回归的决策树类似于分类的决策树，但经过改良，能够预测连续响应。

最佳使用时机...

- 当预测元为无序类别（离散）或表现非线性时

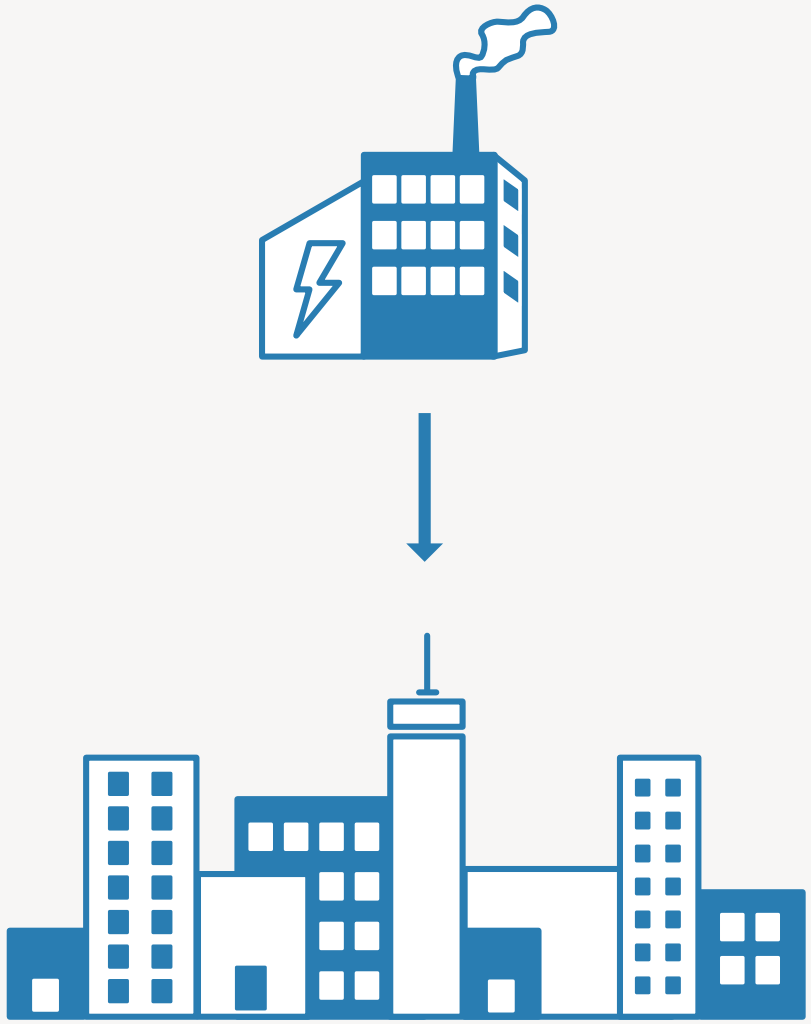


常见回归算法（续）

示例：预测能量负荷

一家大型煤气和电力公司的公用事业分析师开发了能够预测第二天能量需求的模型。电网操作人员使用这些模型能够优化资源，安排电厂发电。每个模型均可访问中央数据库中的历史电力消耗记录和价格数据、天气预报以及各发电厂的参数，包括最大功率输出、效率、成本和所有影响工厂调度的运营约束。

分析师寻找一个模型，对测试数据集提供较低的平均绝对百分比误差(MAPE)。在尝试几个不同类型的回归模型后，最后确定了神经网络，由于其能够捕获系统的非线性行为，所以提供最低的 MAPE。



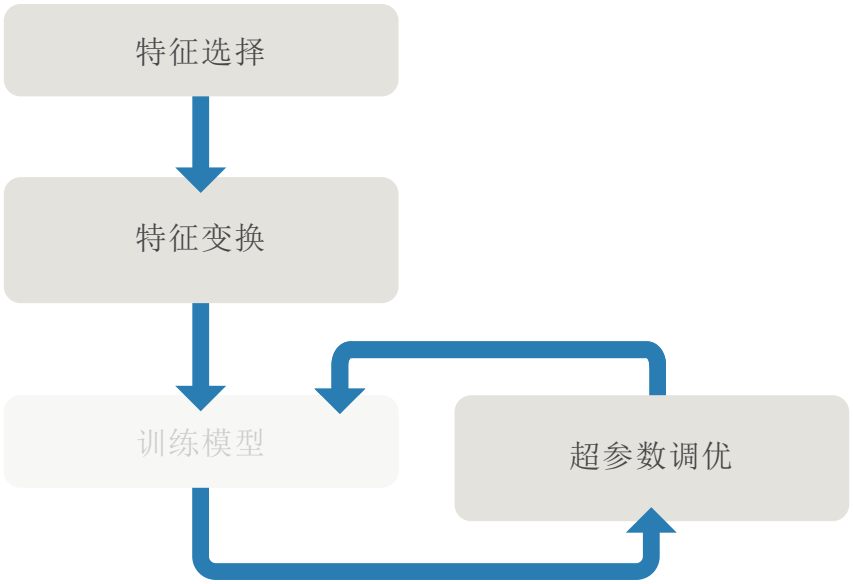
改进模型

改进模型意味着提高其准确性和预测能力，防止过拟合（当模型无法区分数据和噪声时）。模型改进涉及特征工程（特征选择和变换）和超参数调优。

特征选择：识别最相关的特征或变量，在对您的数据建模中提供最佳预测能力。这可能意味着向模型添加变量，或移除不能改进模型性能的变量。

特征变换：使用主成分分析、非负矩阵因式分解和因子分析等技术，将现有特征转变为新特征。

超参数调优：识别能提供最佳模型的参数集的过程。超参数控制机器学习算法如何实现模型与数据拟合。



特征选择

特征选择是机器学习中最重要的任务之一。当您在处理高维数据时，或您的数据集包含大量特征和有限的观察值时，特征选择特别有用。减少特征还节省存储空间和计算时间，使您的结果更容易理解。

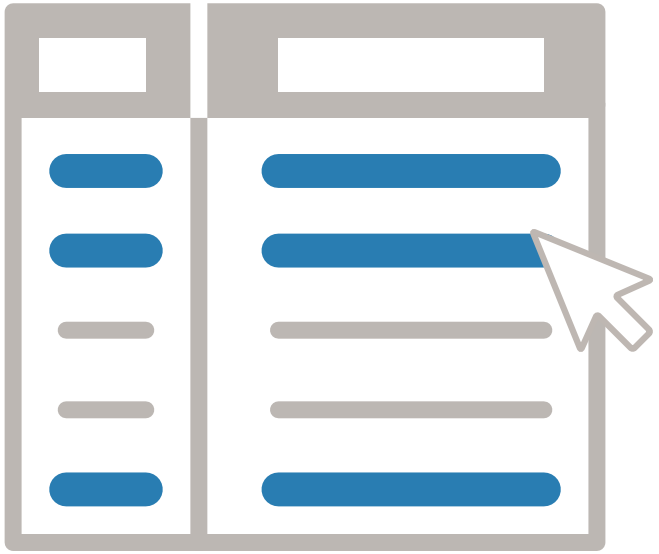
常用特征选择技术包括：

逐步回归：依次添加或移除特征，直到预测精度没有改进为止。

顺序特征选择：迭代地添加或移除预测元变量并评估每次变动对模型性能的影响。

正则化：使用收缩估计量，通过将冗余特征权重（系数）减至零消除冗余特征。

近邻元分析 (NCA)：查找每个特征在预测输出中的权重，以便能够丢弃权重较低的特征。



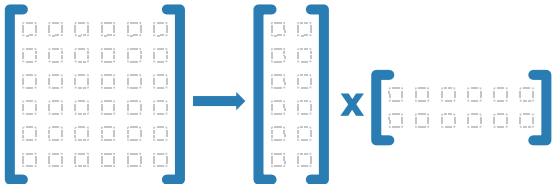
模型的优劣取决于您选择用来训练它的特征。

特征变换

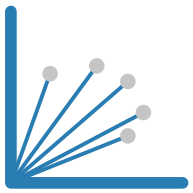
特征变换是一种降维的形式。如我们在第 3 部分所见，三个最常用的降维技术是：

主成分分析 (PCA)：对数据执行线性变换，使您的高维数据集中的绝大多数方差或信息被前几个主成分捕获。第一个主成分将会捕获大部分方差，然后是第二个主成分，以此类推。

非负矩阵因式分解：当模型术语必须代表非负数量（比如物理量）时使用。



因子分析：识别您的数据集中各变量之间潜在的相关性，提供数量较少的未发现潜在因子或公共因子的一种表现方式。

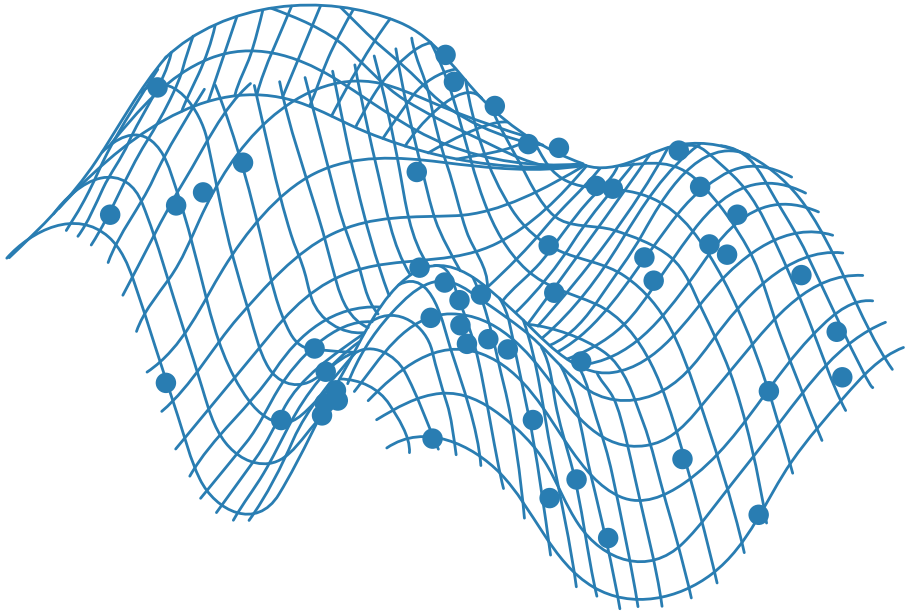


超参数调优

与许多机器学习任务一样，参数调优也是一个迭代过程。一开始设置参数是根据对结果的“最佳猜测”。您的目标是找到“最佳可能”值 — 这些值生成最佳模型。随着您调整参数，模型性能开始改进，您会看到哪些参数设置有效，哪些仍需调优。

三个常用的参数调优方法是：

- 贝叶斯优化
- 网格搜索
- 基于梯度的优化



采用适当调优参数的简单算法通常比调优不充分的复杂算法能够生成更好的模型。

了解更多

准备更深入地钻研？深入了解这些机器学习方法、示例和工具。

监督式学习快速入门

分类

MATLAB 机器学习：分类入门

初步分类示例

贝叶斯解题

以交互方式探讨决策树

支持向量机

K 最近邻点的分类

训练分类集成

使用袋装决策树通过基因表达数据预测肿瘤类

回归

线性回归

什么是广义线性模型？

回归树

训练一个回归集成来预测汽车的燃油经济性

特征选择

选择特征对高维数据进行分类