MathWorks®

机器学习快速入门

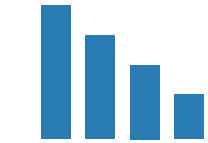
极少一帆风顺

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。

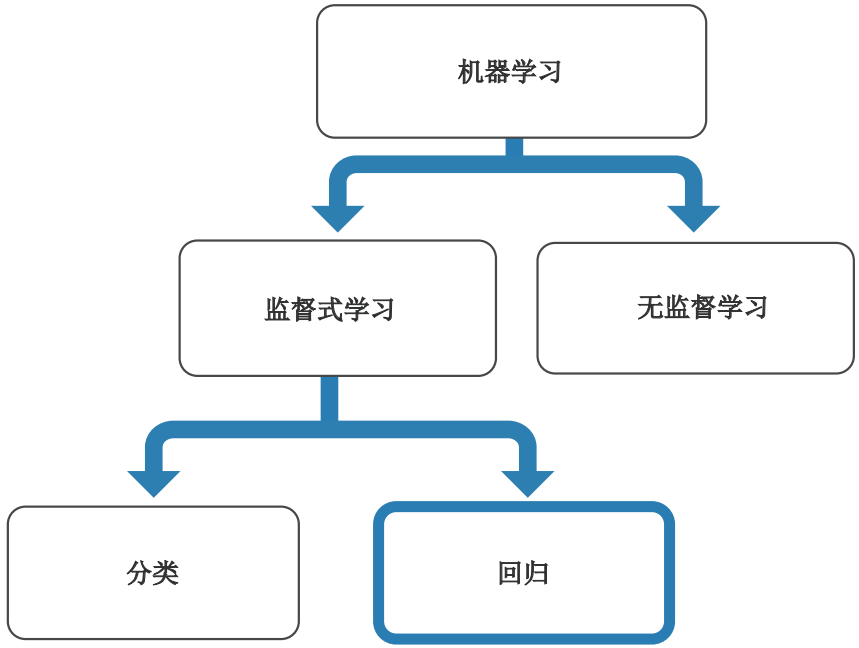
听起来很让人望而生畏？不要泄气。要记住，反复尝试和出错才是机器学习的核心——如果一个方法或算法不起作用，只需尝试另一个。但系统化工作流程有助于创造一个顺利的开端。

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？

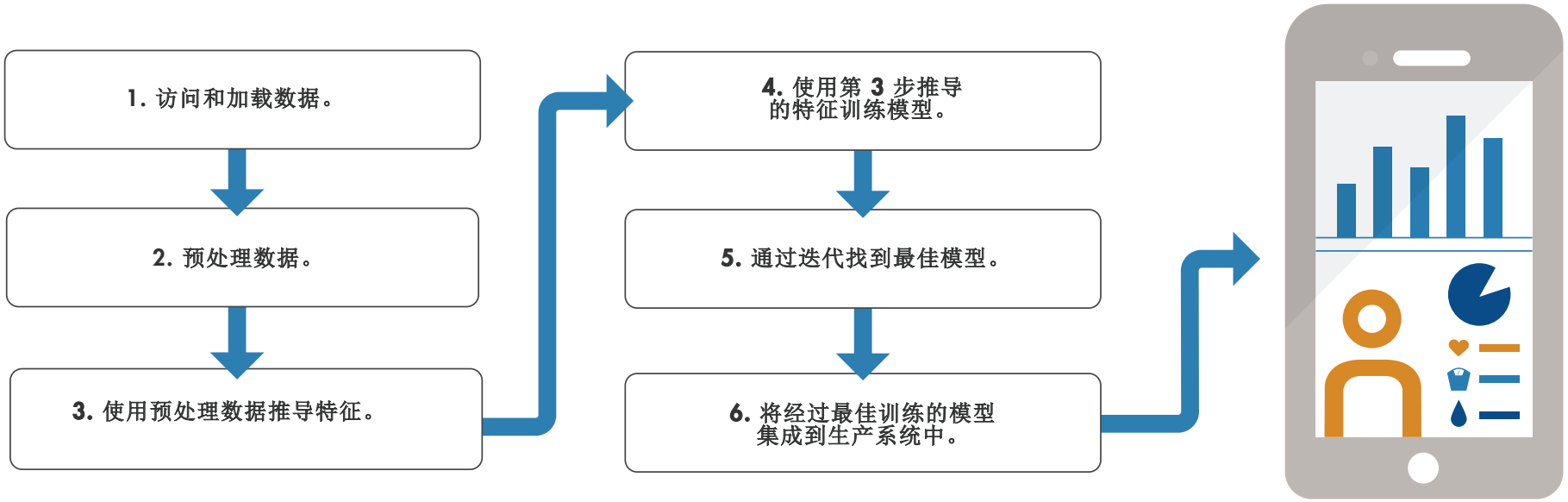
回答这些问题有助于确定您采用监督式学习还是无监督学习。



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



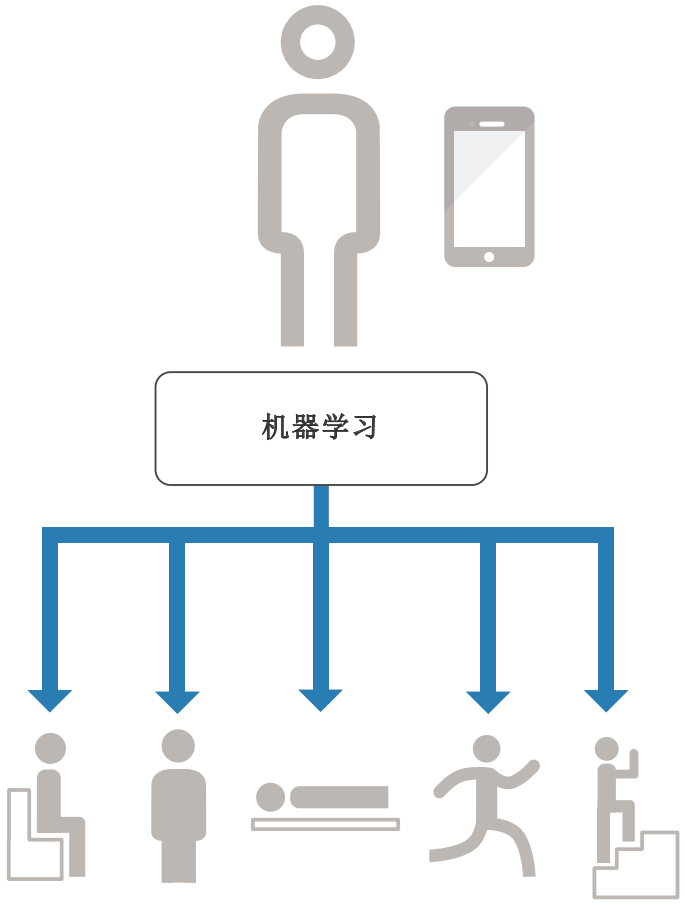
在接下来的章节中，我们将以健康监控应用程序为例更详细地介绍具体步骤。整个工作流程将在 MATLAB® 中完成。

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。



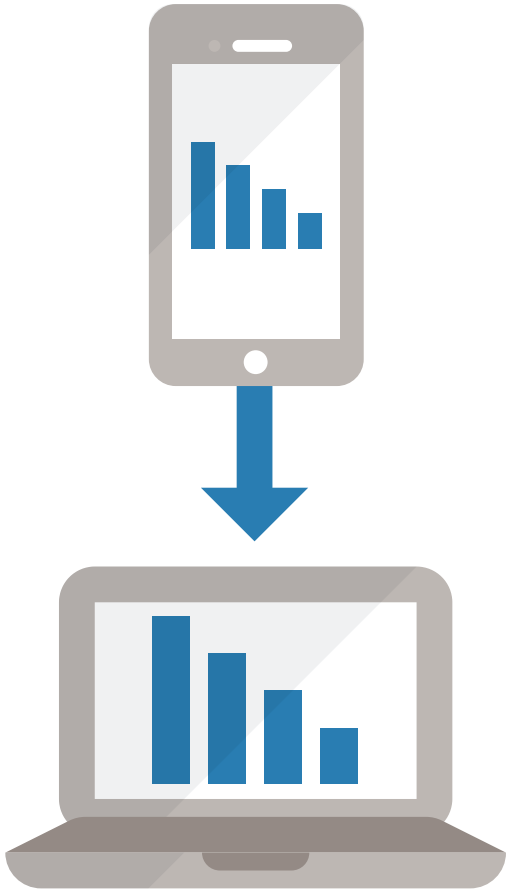
1 步骤 1：加载数据

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

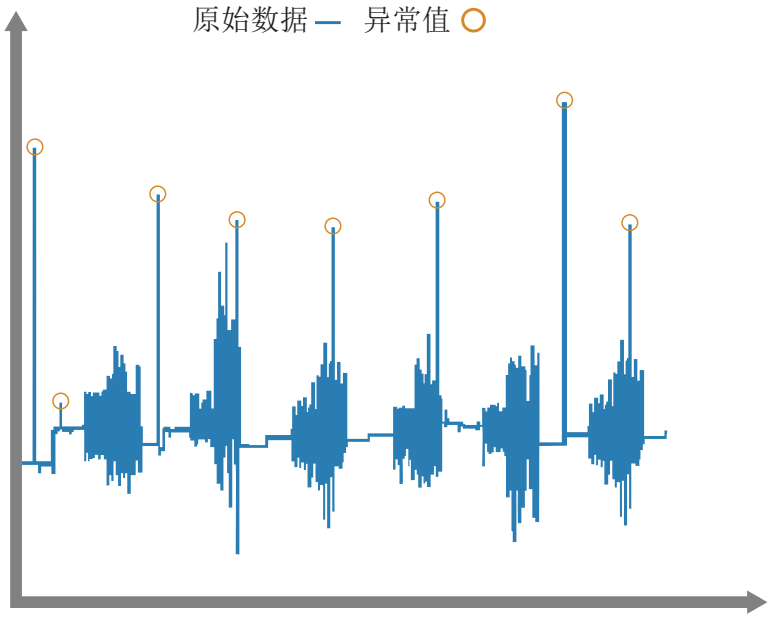
我们将数据导入 MATLAB，然后为每个带有标签的数据集绘图。
要预处理数据，我们可以执行以下操作：

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。



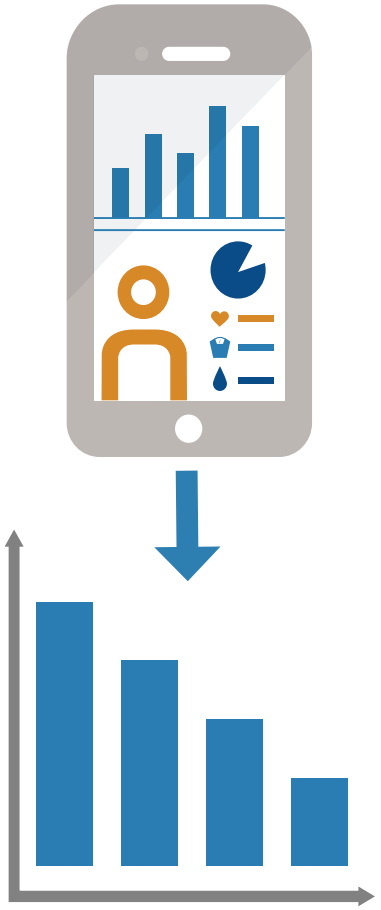
活动跟踪记录数据中的异常值。

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）

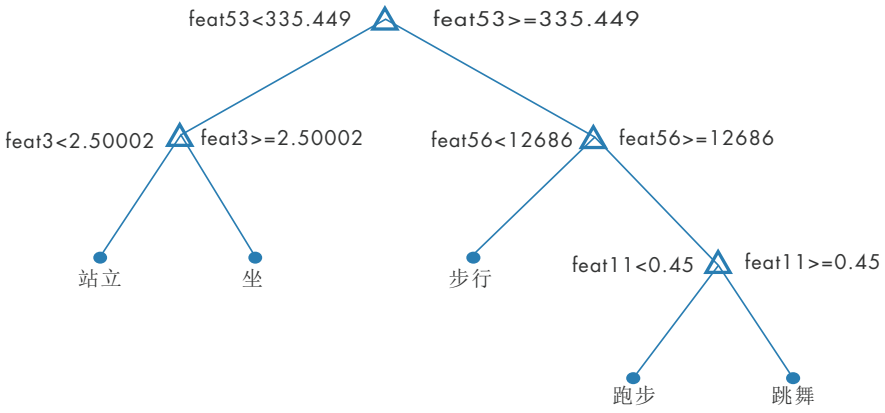
您可以推导出的特征数量只会受您的想象力限制。然而，我们通常可以采用许多技术来处理不同类型的数据。

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

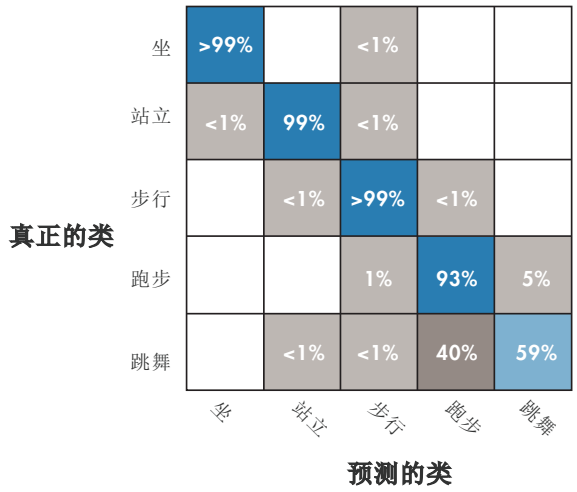
4 步骤 4：构建和训练模型

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。



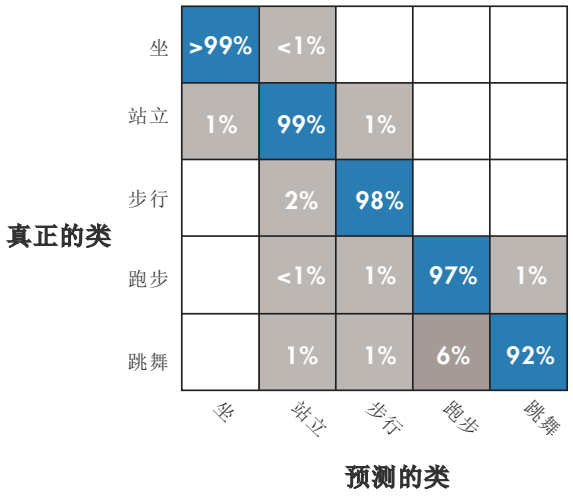
为了了解决策树的执行情况，我们绘制了混淆矩阵，该表将模型产生的分类与我们在步骤 1 中创建的实际分类标签进行了比较。



此混淆矩阵显示我们的模型难以区分跳舞和跑步。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。

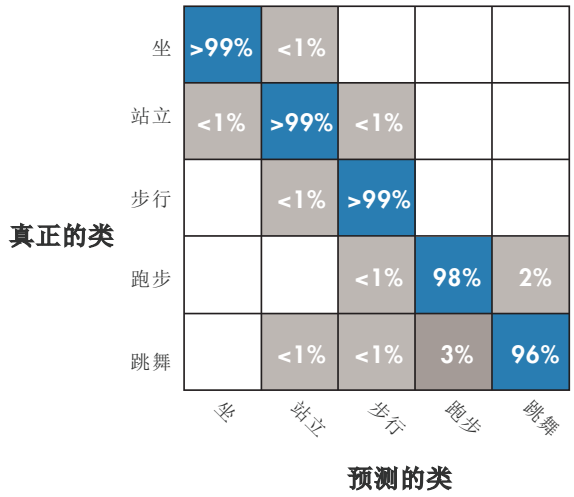
4 步骤 4：构建和训练模型（续）

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：

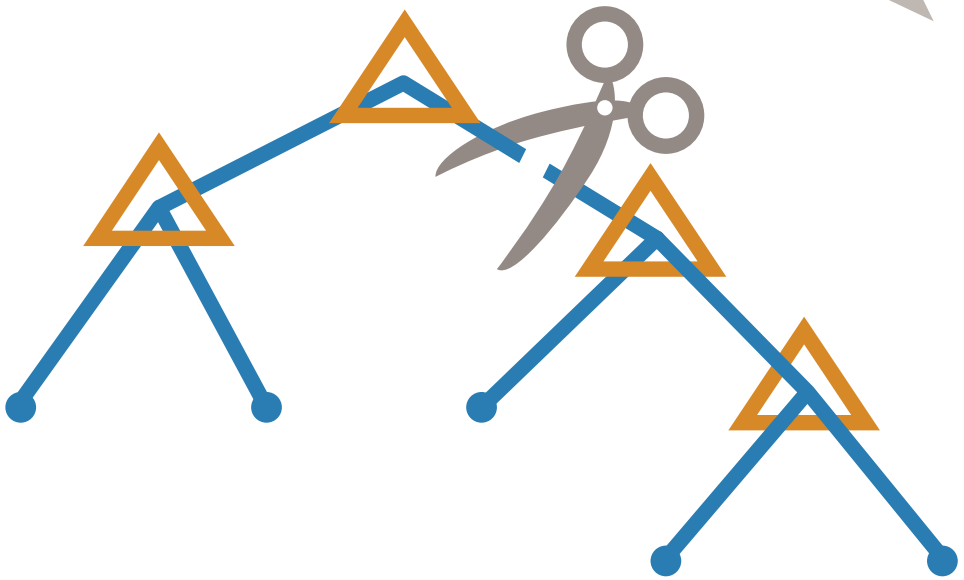
- 相关矩阵 – 可显示变量之间的关系，因此可以删除并非高度相关的变量（或特征）。
- 主分量分析 (PCA) – 可消除冗余，具体方法是找到一组捕获了原始特征的关键区别的特征，并推导出数据集中存在的强模式。
- 序列特征减少 – 采用迭代的方式减少模型的特征，直到无法改进模型性能为止。

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）

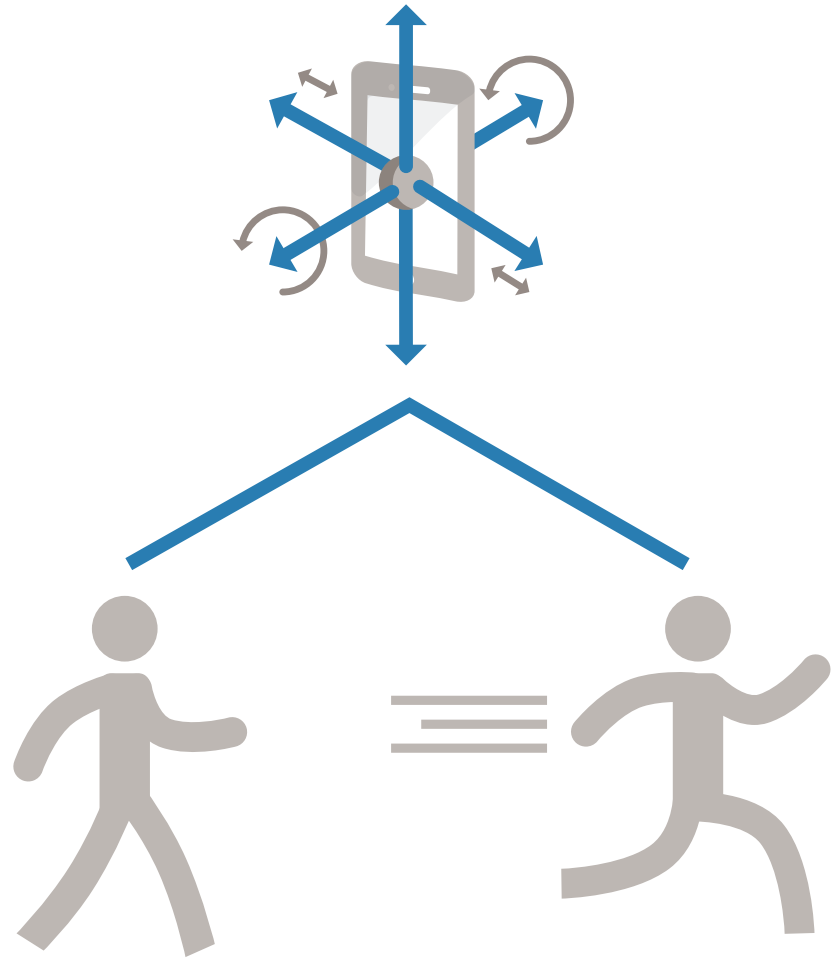
增加复杂度

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

93115v00