MathWorks®

应用监督式学习

MathWorks®

何时考虑监督式学习

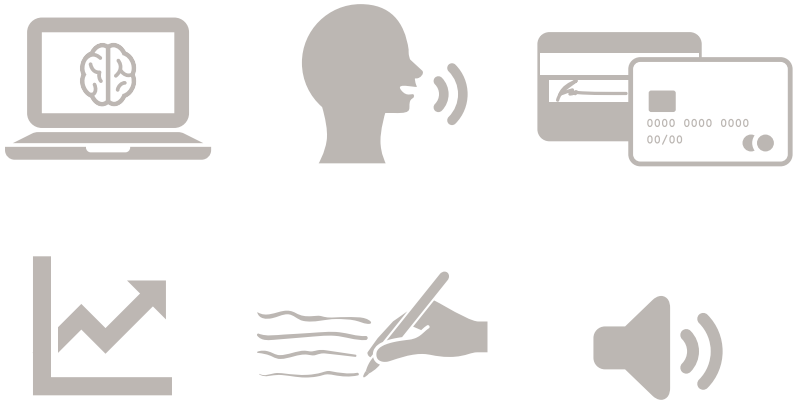
监督式学习算法接受已知的输入数据集合（训练集）和已知的对数据的响应（输出），然后训练一个模型，为新输入数据的响应生成合理的预测。如果您尝试去预测现有数据的输出，则使用监督式学习。

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。

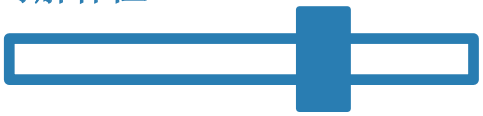
使用较大的训练数据集生成的模型通常对新数据归纳得比较完善。

训练的速度

内存使用

预测的准确度

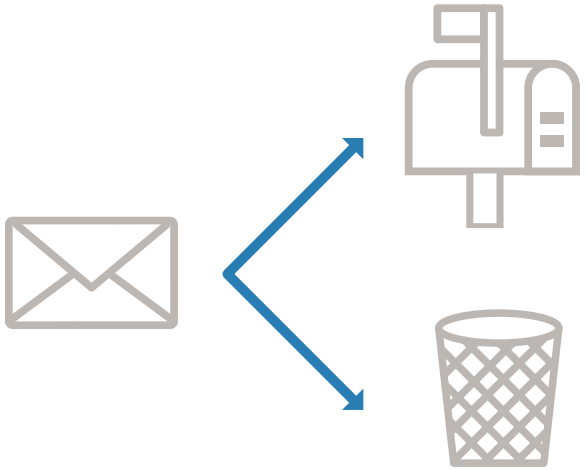
可解释性



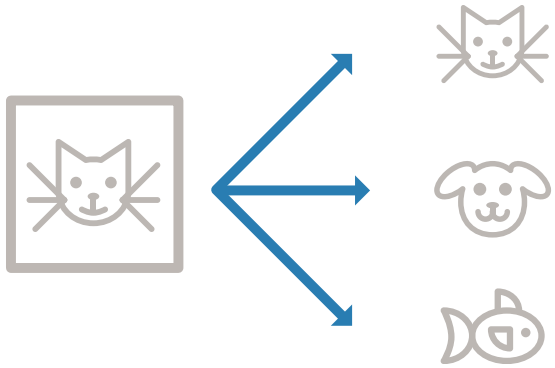
二分类与所有的“多类分类”改为“多分类”

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

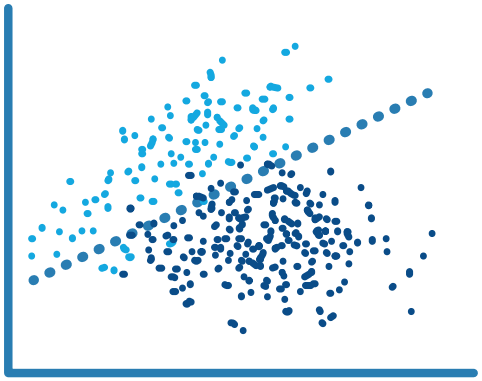
逻辑回归

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



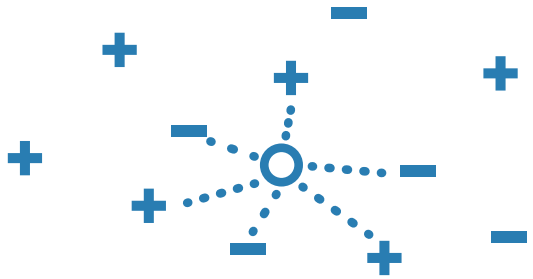
k 最近邻 (kNN)

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）

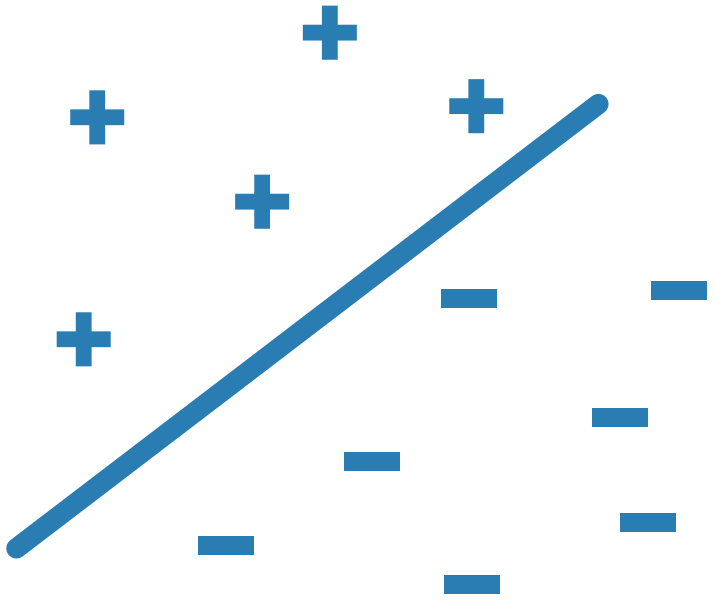
支持向量机 (SVM)

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）

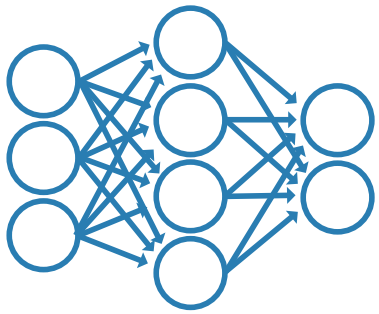
神经网络

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



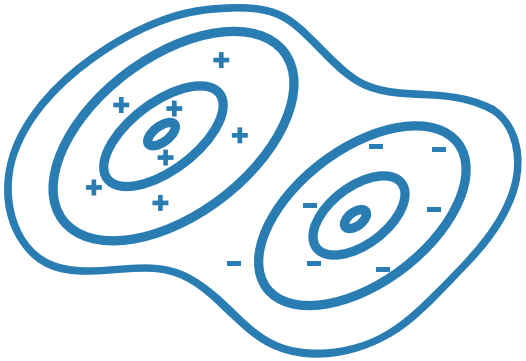
朴素贝叶斯

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）

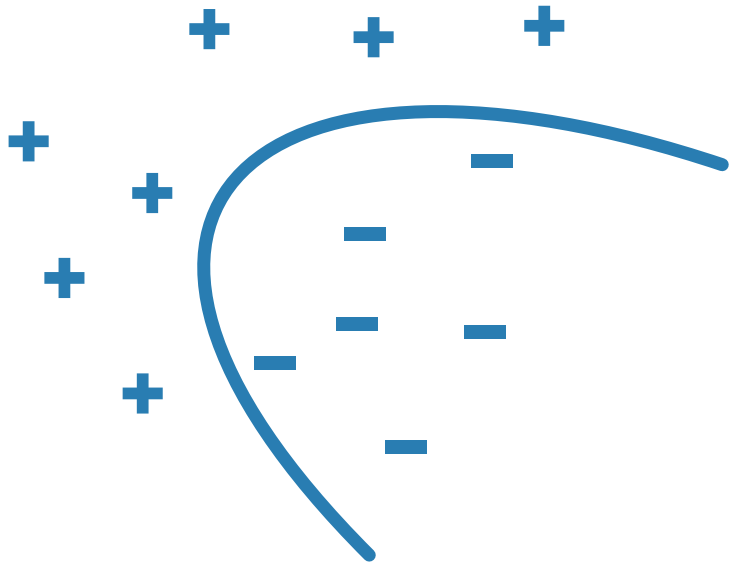
判别分析

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）

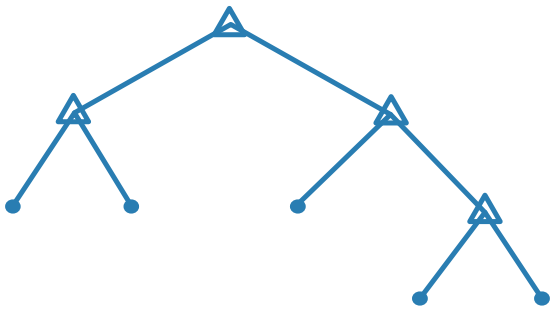
决策树

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

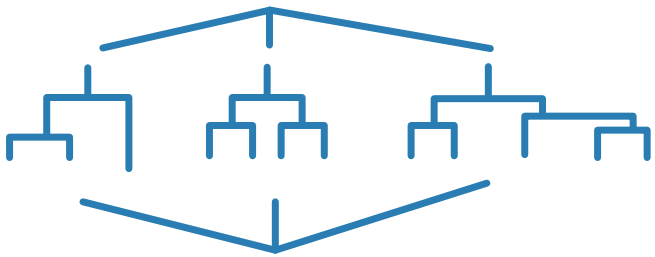
在这些集成方法中，几个“较弱”的决策树组合成一个“较强”的整体。

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



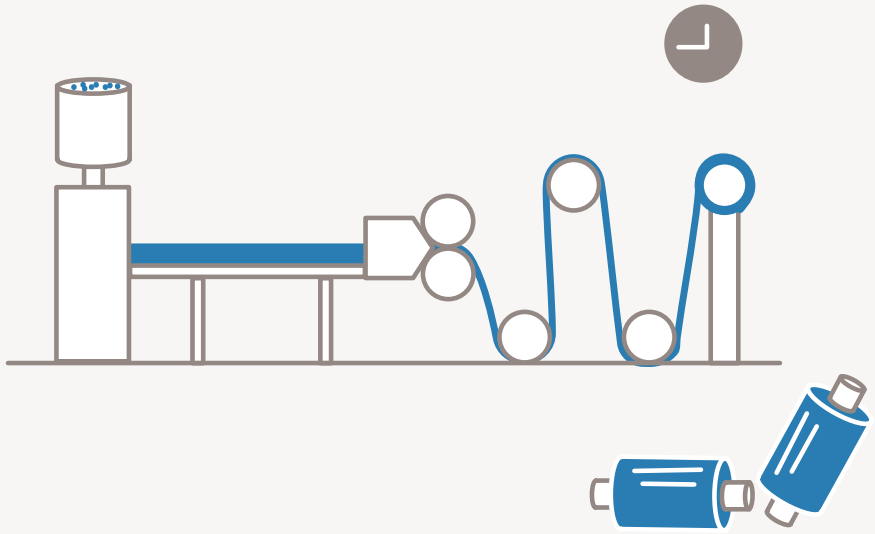
常见分类算法（续）

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

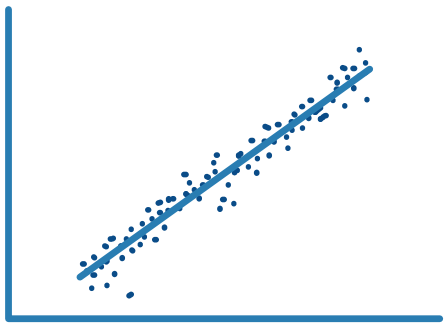
线性回归

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

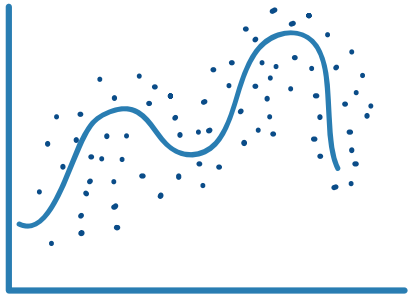
工作原理

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）

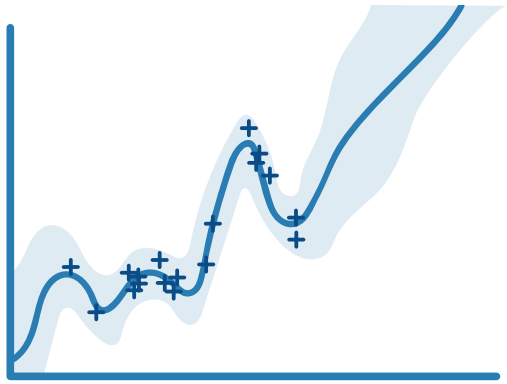
高斯过程回归模型

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



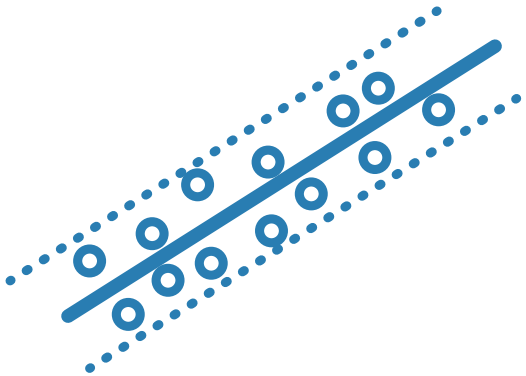
SVM 回归

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）

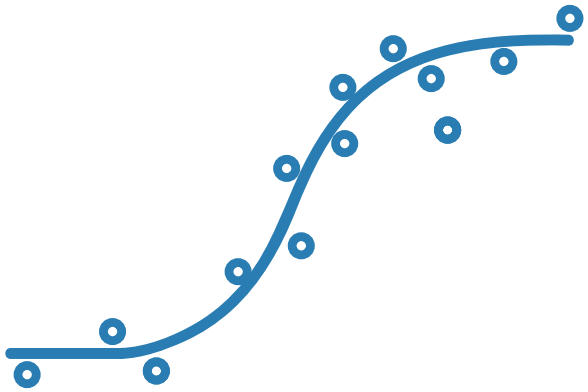
广义线性模型

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



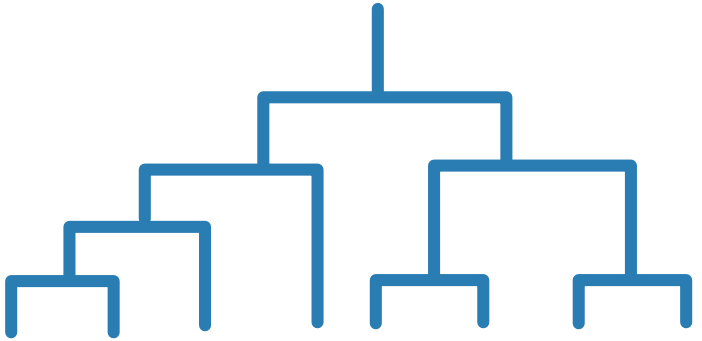
回归树

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

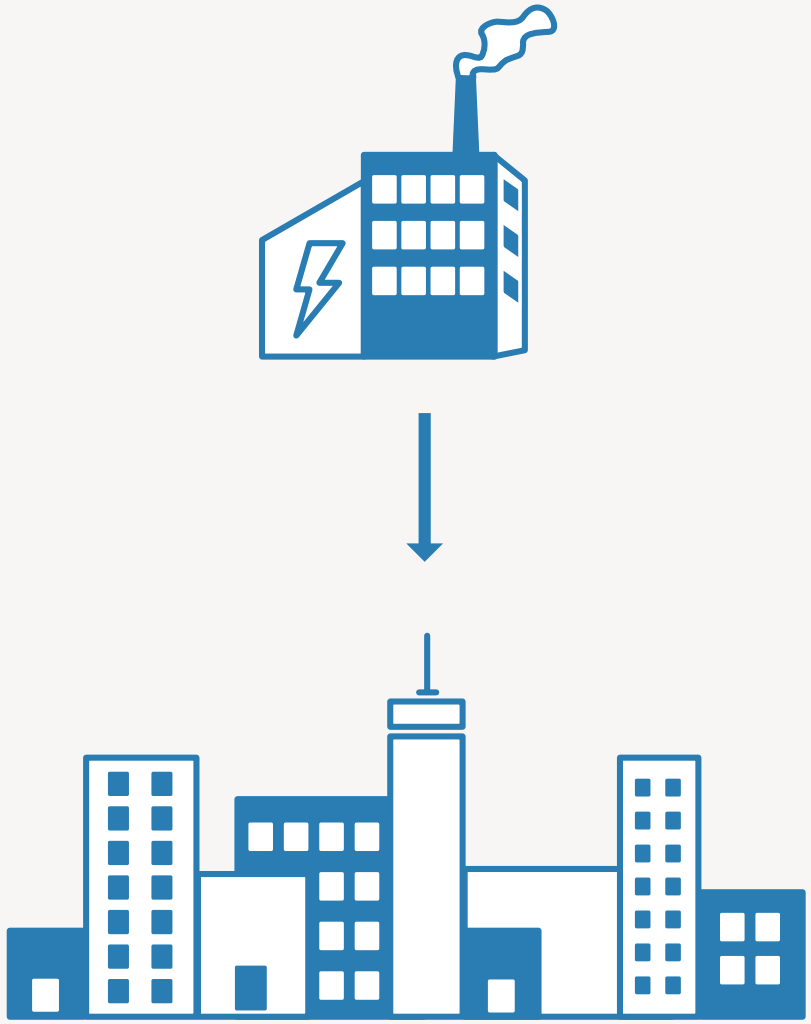


常见回归算法（续）

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。



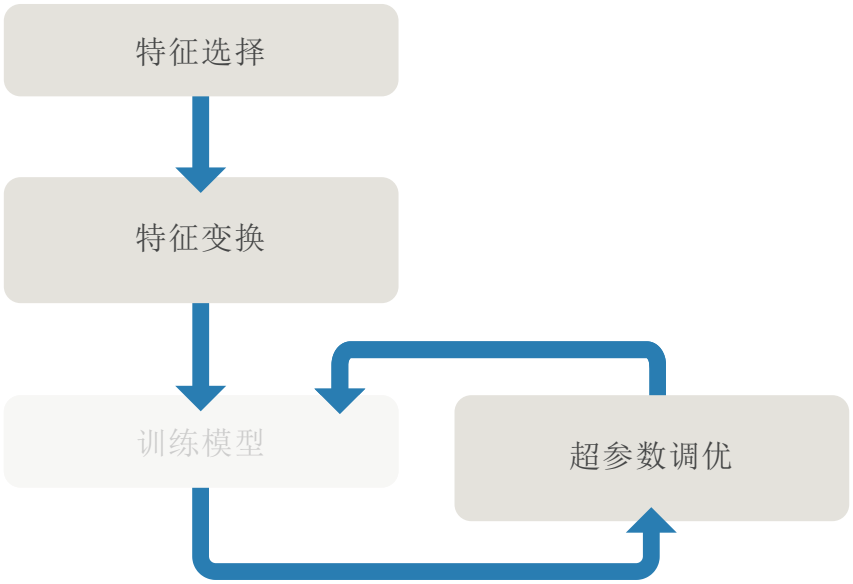
改进模型

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。

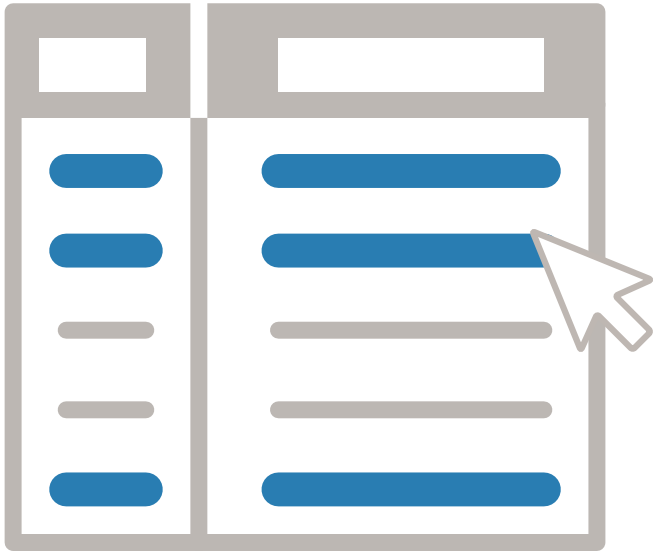
常用特征选择技术包括：

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。



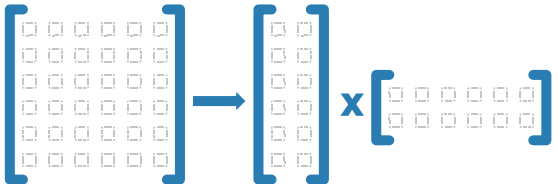
模型的优劣取决于您选择用来训练它的特征。

特征变换

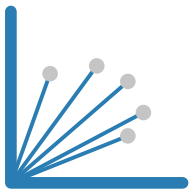
特征变换是一种降维的形式。如我们在第 3 部分所见，三个最常用的降维技术是：

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。

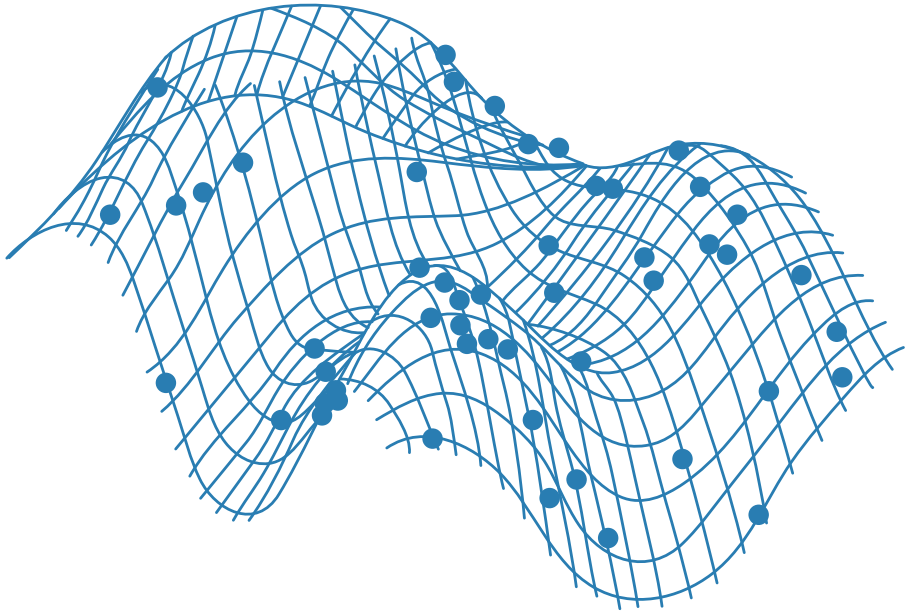


超参数调优

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

93117v00