# 单元 7 机器学习模型

欢迎回来！在这个关于物种分布模型的在线开放课程中的前两个单元中，我们已经研究了物种分布的地理分布、框架和统计回归模型。在这个单元中，我们将研究另一组物种分布模型：机器学习模型。

机器学习模型使用物种分布有/无数据。通常使用一部分数据来学习和描述数据集，另一部分数据来评估模型精度。机器学习模型有多种类型，在这个单元中我们研究其中四个。我们先来看看基于最基本的分类树-决策树的三种算法，然后探讨更复杂的随机森林和增强回归树算法。最后学习人工神经网络算法。

## 分类树

根据数据和预测变量的关系，基于树的模型将数据分为逐渐同类分布有和分布无两类数据。这些是你输入模型的环境变量。最基本的形式是单一分类树。如名所示，分类树类似于自上而下的树。一棵树包含一个根节点和一些从根节点分支出来的内部节点，并以一组终节点或者叶子节点结束。我来逐步解释分类树。

该模型由三个步骤组成。第一个是生长树。首先将全部数据集作为一个组形成分类树的根节点。然后树通过迭代将数据分解成越来越同类的分支类群。每次划分都是基于环境变量将数据分成最优的 2 组，保证至少有一组数据非常均质。例如对干旱区生长的物种，可能会使用降雨量作为划分物种分布有与分布无的阈值。如果一组数据不同质，意味着它可能是分布有/无记录的混合数据，那它需要进一步的分解。例如，干旱地区的物种可能更喜欢沙土。该模型将继续分解，直到第二步，即停止树。这是基于一组规则停止分解的过程。例如，当所有的组都相对同质，则不可能进一步分解，这意味着它们具有相似的环境变量值，因此无需进一步改善模型。当每个终节点中的观测数据个数低于预设的最小值时，或者达到树中的最大分解次数时，分解也可以停止。最后一步是树剪枝。在这一步中，我们减少树的复杂性，来避免训练数据的过度拟合。这是通过保留最重要的分解来实现的。

分类树是一个非常有用的工具，因为它们相对容易理解和解释。然而，它们可能有些不稳定，因为树的根节点附近的小错误会极大改变输出结果。这个问题导致了如随机森林和增强回归树等新方法出现，它们建立在这种基本的分类树方法上但更加先进。我们先来看一下随机森林模型。

## 随机森林

随机森林模型使用随机的数据子集得到大量决策树。使用“装袋”的方法选择这些随机子集。每个随机子集包含约三分之二的数据。另外三分之一的数据不用来构建树，这部分数据被称为'袋子外'数据，之后用来评估模型。

在随机森林分析中，用环境变量的随机子集分解决策树。树木没有剪枝直至生长到最大尺寸，然后对所有树的预测值进行平均，以得到最强分类模型的预测变量。这个过程的优点是用了大量的树，其错误率远低于单个分类树的分析结果。

## 增强回归树

我在这里解释的最后一个基于树的模型是增强回归树模型。如名所示，这类模型是两种技术的结合：决策树模型和增强方法。像随机森林模型一样，增强回归树模型通过不断拟合多个决策树来提高模型的准确性。这两种方法的区别之一是选择构建树的数据的方法。这两种技术都是从全部数据中随机选择构建每个新树的数据。所有随机子集具有相同的数据量，都是从完整的数据集中选择。用过的数据会放回到完整数据集中，以便在随后的树中再次选择。如上所述，随机森林模型使用装袋的方法选择数据。使用这种方法，每个数据点被每个新的随机子集选中的概率相同。增强回归树使用增强方法，在后续建立的树中对输入数据进行加权。加权方法是，在之前的树中建模较差的数据有较高的概率被新树选中。所以，在这个过程中，拟合完第一棵树之后，模型在拟合下一棵树的时候就会考虑该树的预测残差等。通过考虑以前构建树的拟合结果，模型不断改进以增加准确性。这种顺序方法的增强具有唯一性。

增强回归树有两个重要的参数需要用户设定。第一个是树复杂度（tc），它控制着树的分割次数。一个 tc 值是 1 的树只有 1 个分割，这意味着该模型不考虑环境变量间的相互作用。tc 值是 2 的树会有两次分割等。另一个参数是学习率（lr），它决定每棵树对增长模型的贡献。lr 值小会导致构建许多树。这两个参数确定最佳预测所需的树个数。有一个规则，建议找到树复杂度和学习率值的组合，使得模型至少有 1000 棵树。

增强回归树是一种强大的算法，在大型数据集或者环境变量相对于观测数据较多的情况下，结果很好。对缺失值和异常值有很好的鲁棒性。

这就是三个基于树的模型的概述。总结一下，分类树算法构建单个分类树，随机森林构建基彼此独立的多个分类树 这些分类树由全部数据的随机子集构建，而增强回归树也建立多个分类树，但这些分类树相互依赖，每棵新树都考虑到了之前分类树的误分。一般来说，随机森林比单一分类树的模型结果更好，增强回归树优于随机森林。

人工神经网络

我在本单元中讲解的最后一个机器学习模型是人工神经网络。

术语“人工神经网络”是指一大批受到生物神经网络，特别是有大量连接的神经元来处理信息的大脑启发的模型。类似地，人工神经网络由大量节点和连接组成。它们通常按层次组织：将数据送入模型的输入层，多个隐藏层，以及模型结果的输出层。

我将重点介绍使用单个隐藏层的人工神经网络模型，它是一个前向反馈网络，这就是说信息只在一个方向流动，在网络中不存在循环或环。我也会解释经常用于训练该模型的反向传播的概念。

那么 我们从输入层开始吧。这是你的环境输入数据，每个输入节点代表一个环境变量。输入层中每个节点的信息都进入隐藏层。可以根据这些连接的重要性，给予它们特定的权重。通常在模型最初随机分配这些权重，但模型可以在后续运行中学习和优化权重，稍后我会解释。隐藏层中的节点由环境变量的不同组合组成，它们从输入层以输入乘以连接的权重值的方式得到信息。

所以在这个例子中，隐藏层中的第一个节点从每个输入节点接收信息，但是每个输入节点的贡献是不同的，这取决于连接的权重。输入乘以它们的权重并求和。连接权重越高，输入影响就越大。对隐藏层中的每个节点都进行此计算。

每个隐藏层的节点的加权和会传递到所谓的“激励函数”中，其将加权输入信号转换成可理解的输出信号。有很多不同形式的激励函数，但是最常用的是产生一个结果介于 0 和 1 之间的 S 型曲线的逻辑函数。然后将激活函数的结果传递到输出层，在物种分布模型中是指预测物种在给定位置是否有分布。类似于输入层和隐藏层之间的连接，隐藏层和输出层之间的连接也要做加权，因此输出也是隐藏节点加权和的结果。

作为模型训练的一部分，输出与期望的结果进行比较。在一个物种分布模型中，这些都是你已知物种是否有分布的地点的环境条件。预测模型结果和期望结果的差就是模型的误差，用来改进模型过程，称为反向传播。在反向传播过程中，每个连接的权重通过将旧权重乘以模型结果和期望结果的差值来重新计算。基于这些新权重的连接，隐藏层的节点可以计算出自己的误差，并用它来调整输入层连接的权重。当然，输入数据总是保持不变。毕竟权重已经调整过，模型重新用于前向反馈，从输入层并通过隐藏层的方式计算输出结果。

总之，输入是通过前向反馈来推进模型的，而误差是从输出到输入的反向传播。这个过程重复几次直到该模型达到预定的准确度或设定的运行次数。人工神经网络非常强大，可以处理大型数据集，但由于它们的复杂性，非常消耗时间，计算缓慢，并且它们需要用户很好的理解模型以便微调参数。

我希望在最近的三个单元中，你已经学到了很多具体算法，可用于你的物种分布模型。。在下一个单元中，我们将看一下如何评估你的模型结果。到时候见！