单元 5 分布数据模型

在这个物种分布模型在线开放课程之前的单元中，我们学到了很多关于建立模型所需的数据类型，也快速了解了一下可以用的不同算法。现在要详细解释这些算法。在这个单元中，我们看一下仅使用存在数据的模型。我会解释两个不同的地理模型，然后讲解最有名的框架模型 Bioclim。最后我会解释一个主流的机器学习模型 Maxent。

开始探索你感兴趣的物种分布的简单方法是通过将物种记录的位置绘制到地图上，并建立一个地理模型。这些地理模型只能使用存在数据，并简单地在其周围绘制一个形状。他们最有用的地方是可以快速了解物种出现的地区。

地理模型

让我们先从地理模型开始。正如我在单元 4 中提到的，地理模型是开始研究你感兴趣的物种分布的简单方法。这些模型只使用存在数据，并简单地绘制它们周围的区域。因为它们没有使用任何环境数据，主要用于快速了解物种的出现地区。

地理模型的两个例子是圆和凸包模型。圆形模型预测物种存在于物种观测出现地点的某个给定半径的圆周内。凸包模型在物种出现地点周围绘制一个空间凸包形状。这是将所有出现地点包含在内的最小多边形，且该多边形边界的所有角都向外凸出。连接任何两点的线必须完全落在凸包内。该模型预测一个物种可存在于凸包内的任何位置，且在凸包外部不存在该物种。

例如，国际自然保护联盟用凸包模型估计物种的出现区域。自然保护联盟认为凸包尺寸小于 100 km2 的物种极其濒危。在澳大利亚，澳洲毛鼻袋熊就是一种极度濒危的物种，其分布范围仅限于昆士兰埃平森林国家公园内的 3km^2 范围内。

这些地理模型易于理解和解释，但并不常用，因为他们没有考虑到控制或影响物种的环境条件。

简介型号：Bioclim

下一种模型是框架模型，这是最基本的真实物种分布模型。他们描述可以找到物种的环境条件。最著名的框架模型是 Bioclim，被认为是第一个物种分布模型。在十六世纪 60 年代中期，澳大利亚就开始开发这种模型了。

Bioclim 是一个相对简单易懂的模型。通常会采取 2 个以上的环境条件作为预测因子，我在这里用有 2 个环境变量和一组物种出现数据的二维图形展示 Bioclim 概念。Bioclim 用物种出现地点的环境变量的最大最小值构建了合适的环境空间。绿色区域代表适宜物种生存的环境条件。为了避免异常值的过度预测效应，结果范围可以根据用户指定的百分位数缩减，如内部 95%。如果你在模型中包含两个以上的环境变量，就会得到一个多维外接包围框。

为了预测任何给定地点的物种出现概率，Bioclim 比较了该位置的环境变量值与已知出现地点的环境变量值的百分位数分布。50th 的百分位数是指中位数，它恰好在数据的一半位置上。环境变量在某一未知地点的值越接近 50th 的值，该地就越适合物种生存，因此物种的出现概率越高。所以，出现在 50th 处的概率是 1，而在

10th 和 90th 处的概率相同。Bioclim 模型综合每个环境变量的得分得到总体的物种出现概率，所有环境变量在每个位置的权重是相等的。

虽然 Bioclim 已被广泛应用于物种分布模型中，但其有一些限制。该方法只能处理连续的环境变量，因此不能考虑土壤类型等分类变量。此外，它并没有考虑到变量之间的交互性。随着更复杂的算法的发展，Bioclim 算法在物种分布模型中的应用越来越少。

MAXENT

更复杂的算法之一是 Maxent，它是最大熵模型的代表算法。Maxent 是一种仅使用存在数据的算法，它比较了研究区域中物种出现的地点和所有已有的所有环境变量。它通过在整个研究区域抽取大量的点来定义这些可得的环境条件。这些点被称为背景点。因为背景点包括物种出现的已知位置，其与拟缺失点不同。背景点定义已有的环境数据。

Maxent 有两个主要组成部分。第一个是熵，这意味着校准模型来找到整个研究区域里最分散或最接近一致的物种分布。第二个是约束，这些规则基于观测到物种的位置的环境变量值。是在未知位置的每个环境的平均值必须接近已知出现地点的环境变量的平均值。

我用一个简单的例子说明这一点。这个栅格代表一个景观，在一些格网中，我们观测到一个物种的鸟存在。我们不知道该物种是否存在其他格网中。我们输入模型的环境变量之一是年平均气温，每个格网内都有一个该环境变量的值。Maxent 计算存在物种的所有格网的环境变量平均值，是 20 摄氏度。第一个约束是预测物种出现的格网的年平均气温必须为 20 摄氏度。第二约束是景观中所有格网的预测概率总和要等于 1。满足这两个约束的一个可能分布是设定每个物种观测数据所在格网的概率为 1/4，而其他格网的概率为 0。预测物种存在的格网的平均温度是 20 度，满足约束 1;概率也总计为 1，意味着也满足了约束 2。另一个选择是设定 12 个格网的概率是 1/12，总计也是 1，这 12 个格网的平均温度也是 20 度。平均温度为 40 度的 4 个格网的概率为 0。这种分布也符合上述的 2 个限制。第三个选择是给物种出现的格网赋予比其他单元格更多的权重。此选项也符合两个限制。这只是三个选项的一个说明，有更多的潜在分布满足预测物种出现的格网的年平均气温是 20 摄氏度的约束。Maxent 然后探究哪些潜在的分布最统一，在这一点上例子 2 符合。虽然例子中仅包含一个环境变量，当然，Maxent 能使用一组环境变量来对物种分布建模，这些变量被称为特征。

Maxent 考虑了六种类型的特征，并且每一种类型都允许不同形状的响应曲线，并对约束有不同的影响。第一个类型的特征是线性的连续变量。线性特征的约束如我在例子中的解释：每个环境变量在未知位置的平均值必须接近这些变量在已知出现位置的平均值。第二种是二次方特征，它是连续变量的平方和物种从其最佳条件变化的忍耐性。第三种类型是产品，它允许变量之间存在交互。然后是阈值特征类型，它将连续响应转化为二进制响应。铰链特征类型是线性和阈值类型的组合。最后一个类型是分类特征，它指的是具有不同类别响应的分类变量。默认情况下 Maxent 使用所有特征类型，但你可以选择使用其中几个来构建更简单的模型。

为了计算物种的潜在分布，Maxent 首先计算两个概率密度。对于所有物种存在点，概率密度描述了模型中的所有环境变量的相对相似性。所以，在这种情况下，对所有存在点而言，右边图中温度和降雨量的峰值数据是常见的。整个研究区域中基于背景点的图也是如此。因此，背景点的概率密度表征研究区域内的环境变量，而物种存在点的概率密度表征了物种出现的环境。然后 Maxent 计算这两个概率密度之间的比率，这给出了对于研究区域的每个点的该物种出现的相对环境适宜性。

Maxent 选择总体环境特征以及物种所在地的环境特征的相似性最大化的分布。这是 Maxent 的原始输出。为了更容易地解释结果，并估计给定位置的物种出现概率，Maxent 对原始输出进行逻辑变换。逻辑输出考虑了物种的流行情况，这是指占用区域的比例。然而，不能从只有存在数据的情况下得到确切的流行率。流行率的值可以由用户定义，Maxent 的默认值为 0.5。我们建议修改默认值，因为这不适用于稀有物种。

Maxent 的一个重要方面是正则化，这减少了模型的过度拟合。正则化有两种方式：首先，Maxent 放宽约束，不再要求对该模型中环境变量的确切平均值进行拟合，它考虑了平均值的置信区间。可以防止模型在输入数据附近的过近拟合。其次，该模型对复杂性进行限制，这意味着它排除了对模型没有显著改进的要素类型。

我希望这个单元能够让你更好地了解这些使用仅存在数据的模型如何工作。我期待着在下一个单元中见到你，我们会讨论统计回归模型。