This is a video transcript of Module 7 of the Online Open Course in Species Distribution Modelling. To access the full suite of videos click here.

Online Open Course - Species Distribution Modelling

Powered by the Biodiversity and Climate Change Virtual Laboratory

---

Module 7 Machine learning models

Welcome back! In the previous two modules of this online open course ins species distribution modelling we have looked at geographic, profile and statistical regression models to predict species distributions. In this module, we will look at another group of species distribution models: machine learning models.

Machine learning models use both presence and absence data. They typically use one part of the dataset to 'learn' and describe the dataset, and the other part to assess the accuracy of the model. There is a variety of different machine learning models, and in this module we will take a look at four of them. We first look at three algorithms that are based on decision trees, starting with the most basic classification trees, and then the more complex random forests and boosted regression trees. We then finish with artificial neural networks.

Classification Trees

Tree-based models partition the data into increasingly homogenous groups of presence or absence, based on their relationship to a set of predictor variables; these are the environmental variables that you put into the model. The most basic form is a single classification tree. As the name suggests, the classification tree resembles an upside-down tree. The tree consists of a root node from which some number of internal nodes branch out, ending in a set of terminal, or leaf, nodes. I will explain the functioning of a classification tree step by step.

The model consists of three steps. The first is growing the tree. Calibration of the tree starts with the complete dataset as one group, forming the root node. The tree is then grown by repeatedly splitting the data into increasingly homogeneous groups. Each split is based on the environmental variable that best divides the data into two groups, where at least one of the groups is very homogeneous. For example, presence versus absence of a species found in the arid zone might be split using a rainfall threshold. If a group is not homogeneous, which means that it might have a mix of presence and absence records, then it needs to be split further. For example, our arid species might prefer sandy soils. The model will continue to do this until the second step, which is called stopping the tree. This is where the splitting process is stopped

based on a set of rules. For example, when further splitting is impossible because all groups are relatively homogeneous, which means that they have similar values of the environmental variables, and thus no further improvements to the model can be made. Splitting can also be stopped when the number of observations in each terminal node would fall below a predefined minimum, or when some maximum number of splits in the tree is reached. The final step is pruning the tree. In this step, we reduce the complexity of the tree to avoid overfitting of the training data. This is achieved by keeping only the most important splits.

Classification trees are a very useful tool because they are relatively easy to understand and interpret. However, they can be somewhat unstable as small errors at the base of the tree, near the root node, can considerably change the outcome. This issue has led to the development of new methods such as Random Forests and Boosted Regression Trees, which are more advanced but built upon this basic Classification Tree method. Let's have a look at Random Forest models first.

Random Forests

The Random Forest model produces a large number of decision trees using random subsets of the data. These random subsets are selected in a procedure called 'bagging'. About two thirds of the data is included in each random subset. The other third of the data is not used to build the trees, and this part is called the 'out-of-the-bag' data. This part is later used to evaluate the model.

In a Random Forest analysis, each split within a decision tree is determined using a random subset of the environmental variables. The trees are grown to their maximum size without pruning, and then the predictions of all trees are averaged to find the set of predictor variables that produce the strongest classification model. The advantage of this process is that with a greater number of trees, the error rate is much lower than with a single classification tree analysis.

Boosted Regression Trees

The last tree-based model that I will explain here is the Boosted Regression Tree model. As the name suggests, these models are based on a combination of two techniques: decision tree models and boosting methods.

Like Random Forest models, Boosted Regression Trees repeatedly fit many decision trees to improve the accuracy of the model. One of the differences between these two methods is the way in which the data to build the trees is selected. Both techniques take a random subset of all data for each new tree that is built. All random subsets have the same number of data points, and are selected from the complete dataset. Used data is placed back in the full dataset and can be selected in subsequent trees. As mentioned before, Random Forest models use the bagging method for data selection. With this method, each data point has an equal probability of being selected for each new random subset. Boosted Regression Trees use the boosting method in which the input data are weighted in subsequent trees. The weights are applied in

such a way that data that were poorly modelled by previous trees have a higher probability of being selected in the new tree. So, in this process, after the first tree is fitted, the model will take into account the error in the prediction of that tree to fit the next tree, and so on. By taking into account the fit of previous trees that are built, the model continuously tries to improve its accuracy. This sequential approach is unique to boosting.

Boosted Regression Trees have two important parameters that need to be specified by the user.
The first one is the tree complexity (tc), this controls the number of splits in each tree. A tc value of 1 results in trees with only 1 split, and means that the model does not take into account interactions between environmental variables. A tc value of 2 results in two splits and so on.
The other parameter is the learning rate (lr), which determines the contribution of each tree to the growing model. As small value of lr results in many trees to be built. These two parameters together determine the number of trees that is required for optimal prediction. As a rule of thumb, it is advised to find the combination of tree complexity and learning rate values that result in a model with at least 1000 trees.

Boosted Regression Trees are a powerful algorithm and work very well with large datasets or when you have a large number of environmental variables compared to the number of observations, and they are very robust to missing values and outliers.

So, this was an overview of the three tree-based models. To summarize, classification tree algorithms build a single classification tree, random forests build multiple trees, independent of each other, on random subsets of all data, while boosted regression trees also build multiple trees, but these are dependent of each other, and each new tree takes into account the misclassifications of the previous tree. In general, random forests provide better models than single classification trees, but boosted regression trees perform better than random forests.

Artificial Neural Networks
The last machine learning model that I will explain in this module are the Artificial Neural Networks.

The term Artificial Neural Networks refers to a large group of models that are inspired by biological neural networks, in particular the brain, which consists of extremely large interconnected networks of neurons to process information. Similarly, Artificial Neural Networks consist of a large number of nodes and connections. These are typically organised in layers, with an input layer in which the data is fed into the model, a number of hidden layers, and the output layer which represents the result of the model.

I will focus here on the Artificial Neural Network model that uses a single hidden layer, and is a feed forward network, which means that the information only flows in one direction and there are no loops or cycles in the network. I will also explain the concept of back-propagation that is often used to train the model.

So, let's start with the input layer. This is your environmental input data, with each input node representing one environmental variable. The information from each node in the input layer is fed into the hidden layer. The connections can all be given a specific weight based on their importance. These weights are usually randomly assigned at the start of the model, but the model can learn and optimize the weights in subsequent runs, which I will explain later. The nodes in the hidden layer are thus comprised of different combinations of the environmental variables, and they receive the information from the input layer in a way in which the input is multiplied by the weight of the connection.

So in this example, the first node in the hidden layer receives information from each input node, but the contribution of each input node is different, depending on the weight of the connection. The inputs are multiplied by their weights and summed. The higher the weight of the connection, the more influence the input has. This calculation is done for each node in the hidden layer.

The weighted sums in each of the hidden layer nodes are passed into a so-called 'activation function', which transforms the weighted input signal into an understandable output signal. There are a lot of different forms of activation functions, but the logistic function that produces a sigmoid curve with an outcome between 0 and 1 is most often used. The outcome of the activation function is then passed on to the output layer, which in a species distribution model is the prediction whether a species will be present or absent in a given location. Similar to the connections between the input and the hidden layers, the connections between the hidden and output layer are weighted, and thus the output is also the result of the weighted sum of the hidden nodes.

As part of the training of the model, the output is compared to the desired output. In a species distribution model, these are the environmental conditions of the locations of which you know whether a species is present. The difference between the predicted outcome of the model and the desired outcome is the error of the model, and this is used to improve the model in a process called back-propagation. In the back-propagation process the weight of each connection is recalculated by multiplying the old weights by the difference between the output from the model and the desired output. Based on these new weighted connections the nodes in the hidden layer can calculate their own error, and use this to adjust the weights of the connections to the input layer. Of course, the input data always stay the same. After all the weights have been adjusted, the model recalculates the output in the feed forward way, so from the input layer through the hidden layer.

So to summarize, the input is fed forward through the model, while the errors are back-propagated from the output towards the input. This process is repeated several times until the model reaches a pre-defined accuracy, or a set number of runs. Artificial Neural Networks are very powerful and can handle large datasets, but due to their complexity they are very time-consuming and thus slow, and they require a good understanding by the user to fine tune the parameters.

I hope you have learnt a great deal about the specific algorithms that you can use for your species distribution model in these last three modules. In the next module, we will have a look at how to evaluate the outcome of your model. I hope to see you back in module 8!

Attribution

Please cite this video as follows:

Huijbers CM, Richmond SJ, Low-Choy SJ, Laffan SW, Hallgren W, Holewa H (2016) SDM Online Open Course, module 7: machine learning models. Biodiversity and Climate Change Virtual Laboratory, http://www.bccvl.org.au/training/. DDMMYY of access.

Acknowledgements

Lead partners: Griffith University, James Cook University

Thanks to: University of New South Wales, Macquarie University, University of Canberra

Funded by: National eResearch Tools and Resources Project (NeCTAR)

The BCCVL is supported by the National eResearch Tools and Resources Project (NeCTAR), an initiative of the Commonwealth being conducted as part of the Super Science Initiative and financed from the Education Investment Fund, Department of Education. The University of Melbourne is the lead agent for the delivery of the NeCTAR project and Griffith University is the sub-contractor.