Lab 8 – Classification and Regression Trees (CART, MRT & RF)

Classification And Regression Tree analysis (CART) and its extension to multiple simultaneous response variables, Multivariate Regression Tree analysis (MRT) can be viewed as an alternative approach to gradient analysis that we covered in the previous lab. It works with two datasets: one or more predictor variables, and one or more response variables, and I am interested in establishing associations between the two datasets that may be interpreted as causal relationships (e.g. a set of climate variables influencing species composition across various ecosystems).

The method can be executed in univariate and multivariate form for both predictor and response variables. As the simplest case, I can have one predictor and one response variable just like in a linear correlation. However the relationship can take any form (linear, curvilinear, discrete steps), and both the predictor and response variables can be categorical or continuous, or they can be a mix of both. Thus, the analytical approach is extremely flexible, has virtually no assumption, and is surprisingly powerful for both prediction and explanation.

CART and MRT work by splitting the dataset of response variables into two groups, so that the variance in response variables within the groups is minimized. In that, CART and MRT are doing the same as cluster analysis: finding groups of similar observations. What makes CART and MRT different is that external variables (the predictors) are imposed as a constraint to create the clusters. To give an example, I can’t just group a number of species inventory plots into a cluster based on their species composition. Rather I must find environmental thresholds that allow me to create clusters of inventory plots with similar species composition (say larger and smaller than 5°C mean annual temperature).

The method works empirically by investigating various thresholds in various predictor variables to find the first split in the response variable dataset that explains most of the variation between groups (actually minimizing variation within groups). After the first split, the process repeats for the two sub-groups, etc. until no significant amount of additional variance can be explained by further splits.

8.1. CART with one response variable

For this exercise, download the data set “AB_Tree_Plots.csv” from the website. This dataset contains species frequency data for three species of trees in different sample plots along with six climate variables measured at these sites. In addition, the frequency data is also converted to “Presence/Absence”, where presence is any frequency>0. The ecosystem and biome is also given for each sample plot.

To try out “Classification Tree” aspect of CART we use the class variable “Presence/Absence” for a species, and for the “Regression Tree” aspect of CART we can use the numerical species frequency as dependent data. We want to analyze with CART if climate variables can predict the frequency or presence or absence of tree species. You can also try to predict an ecosystem or biome class based on climate variables.

- At present, the latest version of R actually breaks some functionality of the mvpart package that we use for this, so run an older version R-2.15.3. This is quite common, and as a remedy you can install multiple versions of R and download the package under a previous release. To work with a previous release, you have to start R from the start menu and import the data set with the full path:
  ```
  dat1=read.csv("C:/Users/.../AB_Tree_Plots.csv")
  ```

- Alternatively you can try a forced install into the latest version of R with a developer toolkit. Download and install the appropriate version of RTools from here: http://tinyurl.com/z8w5lvs. Then, run this:
  ```
  install.packages("devtools")
  devtools::install_github("cran/mvpart")
  ```

This may work for any unsupported package, but you may encounter bugs (i.e. it’s unsupported).
To run classification a regression tree, you need to install the `mvpart` package. There are other packages available that can handle univariate CART and/or CART, (e.g. `tree` and `rpart`), but `mvpart`, which is primarily designed for multivariate regression trees, can handle this as well. Let’s start with CART, a univariate regression tree with a continuous response variable:

```r
library(mvpart)
out1 = mvpart(BETUPAP ~ MAT + MWMT + MCMT + MAP + MSP, dat1, xv="p", all.leaves=T)
summary(out1)
```

- The software will ask you what size of tree you want, and offers you a scree-plot for decision guidance. The green line is equivalent to a “variance explained by the split” statistic. So, each successive split will explain less variance. You should pick a tree size under the red line between the orange mark (well supported splits that explain significant variation) and the red mark (reasonably well supported, explaining some additional variation). This is due to the option `xv="p"` “pick”. If you leave it out, the software will choose the orange marker.

- There seems to be clear inflection point for four groups, so click on that in the graphic, and you’ll get the regression tree. The option `all.leaves=T` generates the number of observations and the average frequency at each node and leaf. So, the leftmost group has 64 plots with an average frequency of Betula papyrifera (BETUPAP) of 0.17%. The species is most frequent in the 4 samples in the rightmost group at 4.15%.

- The criteria indicated above the splits tell you how these frequency groups relate to predictor variables: The highest frequency of the species is in environments with cold winter temperatures. Intermediate frequencies at sites with still cool winter temperatures, but warm summers (i.e. high continentality). So, from left to right we are probably looking at southern, high elevation, sub-boreal, and boreal ecosystems.

- Looking at the readout of the summary command, we can see that the first split explained 36% of the variance (1.0000-0.6405), while the second and third split account together for 43% of the variation (0.6405-0.2143). This combined reporting of explained variance occurs sometimes when a subsequent split explains more variance than a previous split. If the “nsplit” does not skip anything, you can read off the variance explained directly in the CP column.
The other set of numbers that are important to look at are alternative predictor variables that may have explained as much or nearly as much of the variation. For the first split, MAT is second with 23% and the same goes for the second split with 19%, which is considerably less. So cold winter temperatures are clearly a driver for this species: it tolerates them well as compared to other species.

To run classification a classification tree (CART), you need to specify your response variable as a factor or class variable.

```r
dat1$ABIELAS_PA = as.factor(dat1$ABIELAS_PA)
out1 = mvpart(ABIELAS_PA ~ MAT + MWMT + MCMT + MAP + MSP + DRYNESS, dat1, xv = "p", all.leaves = T)
summary(out1)
```

As you can see, subalpine fir (ABIELAS) is present where it is moist and relatively warm, i.e. lower montane and subalpine ecosystem. It is excluded from the dry boral in the first split and the alpine in the second split.

The leaf plots are frequency histograms of your class variable. In our case we only have two levels (0, light blue; 1, dark blue). And you can use this vote for predictions. The prediction for the left and middle group is class "0" and for the right it is "1", indicated above the absence/presence ratio in the plot.
8.2. Multivariate regression tree analysis

Multivariate regression tree (MRT) analysis is a multivariate extension of CART, making it a true constrained gradient analysis with multiple predictors and multiple response variables. Normality in the response variable is perhaps somewhat desirable (to get a decent within-group variance estimate), but transformations of the predictor variable have no effect on the results whatsoever (makes sense – MRT/CART simply arrives at different cut values).

As an illustration how CART and MRT can handle interactions among predictor variables and unimodal distributions, imagine a simple regression tree: MAT>4°C =low, MAT≤4°C =intermediate, then split the intermediate group to MAT≤2°C =low, MAT>2°C =high. This would match a hump-shaped distribution with a maximum frequency at MAT around 3°C and low values below 2 and above 4°C. Similarly, I could have other variables (say high precipitation) having positive effects at MAT>4°C, but negative effects at MAT≤2°C in two further subsequent splits. This will cover interactions among predictors.

For this exercise, download the data set “AB_Tree_Plots_Normalized.csv” from the website. This dataset contains data of % forest cover for 5 species of trees in different sample plots along with 5 climate variables measured at these sites. In addition, the ecosystem and biome is given for each sample plot. We will use MRT to describe the relationship between the % cover data for multiple tree species (response variable) and climate variables (predictor variables) simultaneously.

- Start R from an empty workspace in a working directory, and import the data set AB_Tree_Plots_Normalized.csv. These are data for each tree species normalized to a mean of 0 and standard deviation of 1, which gives each species equal weight in the analysis and creates an output that can be more easily interpretable (as above or below average species frequency).
  
  ```
  dat2=read.csv("C:/Users/.../AB_Tree_Plots_Normalized.csv")
  head(dat2)
  attach(dat2)
  ```

- Mvpart requires the multivariate species response variable being defined as a matrix of numbers, so we do just that before we run the command. As you will see, a large number of groups are supported by this large dataset:
  
  ```
  spec2=as.matrix(dat2[,8:13])
  head(spec2)
  out2=mvpart(spec2~MAT+MWM+MCMT+MAP+MSP, dat2, all.leaves=T, xv="p")
  summary(out2)
  ```

- Interpretation: The first and most important split, explaining 21% of the total variance in the species frequency data, splits the forest communities by coldest mont temperature (MCMT), separating 56 plots with high frequency of boreal trees (to the right), and 80 plots rich in trees occurring in the foothills and montane trees (to the left). Can you continue with one or two other splits?

- Ultimately the leaves at the end represent species communities with similar species composition driven by different climate conditions. You can read out the group membership, which could be the basis for an ecosystem classification system. We will map your own ecosystem classification out for Alberta in the next exercise, so keep the dat2 object in your workspace (i.e. don’t close R)
  
  ```
  dat2$Group=out2$where
  table(out2$where)
  fix(dat2) # scroll to the very right to see your new Group variable
  ```

- The group numbering system includes nodes as well as leaves. Run the table command where you can see the leave numbers labeled from left to right with their corresponding number of observations that should match the multivariate regression tree plot.
8.3. Bootstrapped CART using the machine learning algorithm randomForest

CART can also be used to predict a response variable. Using the tree output from our first exercise, we could input habitat conditions of a new plot location, and follow the branches of the regression tree down to the ends, where we read off if the species should be at high, low or intermediate frequencies. Similarly, we could also predict class variables, for example which ecosystem do we expect, given a number of predictors (e.g. climate variables).

RandomForest is a bootstrapped version of CART, where many trees are built based on subsets of the data. From the 113 sample plots, the algorithm chooses a subset of perhaps 75 to build the first regression tree. In addition, not all predictor variables are used every time, but a random subset of perhaps 3 from the available 5 variables at each node. You can see that this gives me an infinite number of possible classification or regression trees. I can specify how many I want, e.g. 500. For the prediction of a new observation, the algorithm follows through all 500 trees to arrive at a prediction. For a continuous variable (e.g. species frequency) it calculates an average across the 500 trees, for a class variable (e.g. ecosystem type) it determines a majority vote (e.g. 363 votes for ecosystem A) for the final prediction.

For this exercise, we recycle the previously used databases “AB_Tree_Plots.csv”, “Climate_AB.csv”, “Climate_AB_2020s.csv”, etc. and generate some habitat projections under climate change, just as we did before with discriminant analysis.

- Install the two required packages and run the following code. We already knew which variables are important from a single regression tree above, but here we have more reliable summary statistics over many bootstrapped regression trees. The fewer trees you use, the faster the program, but the less reliable the prediction. Use the last plot to check if you have a sufficient number of trees specified.
  ```r
  library(randomForest)
  dat1=read.csv("AB_Tree_Plots.csv")
  head(dat1)
  out3=randomForest(Aبيلاس+MAT+MWMT+MCMT+MAP+MSP, dat1, ntree=100, importance=T)
  varImpPlot(out3)
  plot(out3)
  ```

- Now, we import spatial climate data and generate predictive habitat suitability maps. Subalpine fir frequencies are predicted as a function of multiple climate variables for all of Alberta based on the regression trees generated with the sample plot data:
  ```r
  library(lattice)
  climate1970=read.csv("AB_Climate.csv")
  attach(climate1970)
  PRED1970=predict(out3, climate1970)
  levelplot(PRED1970~X*Y, aspect="iso")
  ```

- Repeat for the 2020s, 2050s, and 2080s and you can see the species habitat move up-slope.
  ```r
  climate2020=read.csv("AB_Climate_2020s.csv")
  attach(climate2020)
  PRED2020=predict(out3, climate2020)
  levelplot(PRED2020~X*Y, aspect="iso")
  ```

- You can do the same for a class variable, in this case predicting ecosystem distributions rather than species distributions, as we did with discriminant analysis earlier. We can actually do that for our own ecosystem classification that came out of the multivariate regression tree analysis above. If you still have the object dat2 with the variable “Group” attached to the end, proceed. Otherwise re-run the
MRT with perhaps 8 to 12 splits. Let's map those inventory-plot based groups out, using climate as a predictor. Repeat for projected 2020s, 2050s and 2080s climate if you wish:

```r
head(dat2)
MYECO=as.factor(dat2$Group)
out4=randomForest(MYECO~MAT+MWMT+MCMT+MAP+MSP, dat2)
out4$confusion
```

```r
climatet970=read.csv("AB_Climate.csv")
MYECOMAP=predict(out4, climatet970)
ECOCOUNT=length(unique(MYECOMAP))
attach(climatet970)
plot(x,Y, pch=15, cex=0.3, col=rainbow(ECOCOUNT)[MYECOMAP])
```

### 8.4. Advanced RandomForest features (if you have time)

Let's use the built in dataset iris that we used before for discriminant analysis to classify species based on petal and sepal measurements. RandomForest can be used in the same way, except that we don't have to worry about any assumptions of normality or homogeneity of variances.

- Here we explore some more sophisticated features of the RandomForest package for classification:

  ```r
  library(randomForest)
  iris.rf=randomForest(Species~., data=iris, importance=TRUE, proximity=TRUE, ntree=500)
  ```

- Required number of trees gives errors for each species and the average for all species (black):

  ```r
  plot(iris.rf,lty=1)
  ```

- Misclassification error rates:

  ```r
  iris.rf$confusion
  ```

- Importance of individual predictor variables for classification (the further the value is on the right of the plot, the more important):

  ```r
  varImpPlot(iris.rf)
  ```

- The membership of a particular class as a function of a variable value can be displayed with this command:

  ```r
  partialPlot(iris.rf,iris,Petal.Width,"setosa")
  ```

- And we can predict unclassified observations. We make up some sample new observations from the original dataset to save some time importing (the first three rows are P. setosa, lets see if RandomForest gets that right:

  ```r
  newobs=iris[1:3,1:4]
  fix(newobs)
predict(iris.rf,newobs)
  ```

- This last plot conveys the confidence in your predictions for each individual sample. Colors represent species and points are samples. In this case, many samples can be predicted with great certainty (1) and only few classifications are questionable (approaching 0):

  ```r
  plot(margin(iris.rf))
  ```