
BINF*6210 Software Tools

Machine Learning & Random Forest



COTTENIELAB.org



Learning outcomes



Learning outcomes for this presentation

By the end of this presentation, combined with working through the associated R scripts, you should be able to:

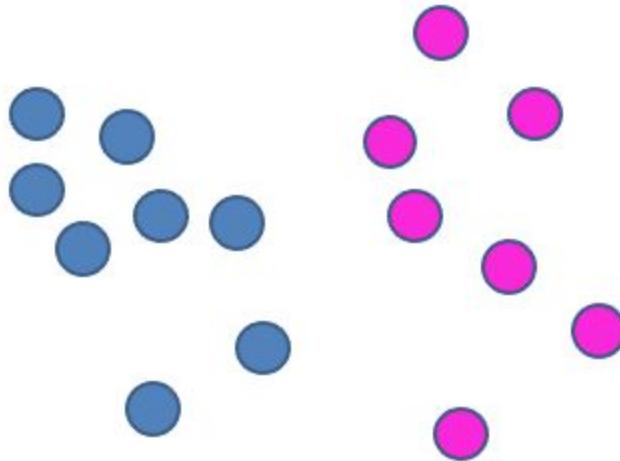
- Describe the difference between supervised and unsupervised machine learning and provide an example of when you would use each
- Describe at a high level the purpose and benefits of a random forest classifier
- Build your own classifier using functions from the R package randomForest

Supervised Machine Learning

Classification to Known Groups

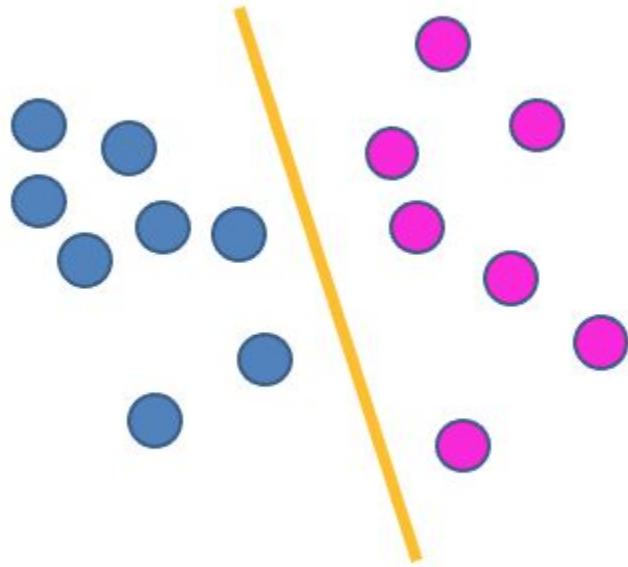
- Concerned with assigning observations to known groups
- We want to be able to make predictions about the group a new observation will belong to (response variable), based upon predictor variables
- We may also wish to determine which variables are the strongest predictors (i.e. for learning about mechanisms)
- Diverse applications, e.g.:
 - Predicting disease risk on the basis of clinical data or biomarkers
 - Learning about biological processes
 - Assigning sequences to taxonomic or biological categories

Step 1: we know the groups for the training data

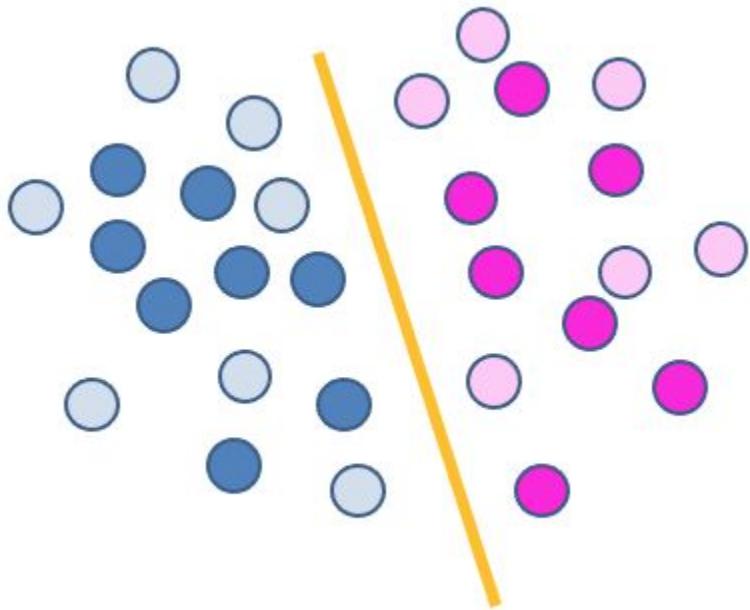


→ **labeled data**

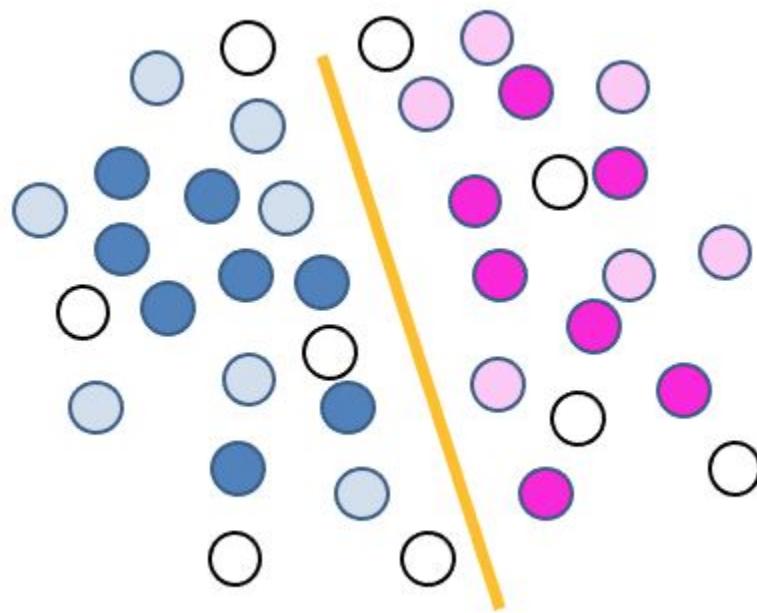
Step 2: find the rule(s) that separate the groups



Step 3: validate the rule(s) w/ separate, labeled data



Step 4: apply the rule(s) w/ unlabeled data

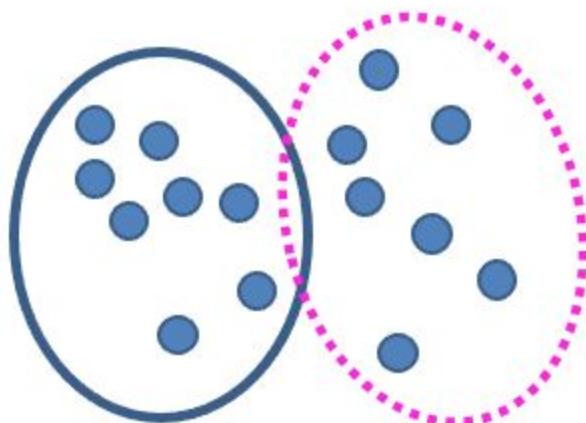


Script example

- We will look at one example of supervised machine learning (random forest algorithm implemented in randomForest R package)
- We will know the groups for the training data and will use simple sequence features as predictors
- We will look at 2 examples:
 - the groups are two different genes for classifier #1
 - the groups are two different taxonomic groups for classifier #2
- We will then see how well the predictive model can predict the groups for separate data

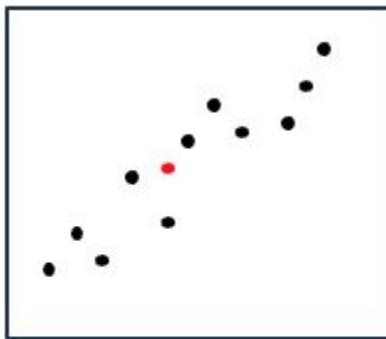
Unsupervised machine learning

- We do not know the groups
- Use algorithms to find patterns in the data/groups
- E.g. How many groups (BINs) are there in our data



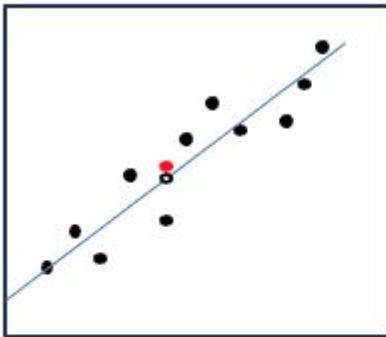
Random Forest

We have data and want to predict future data



- We know x , want to predict y value for white point.
- True value (currently unknown)
- Known data

Using a simple linear model

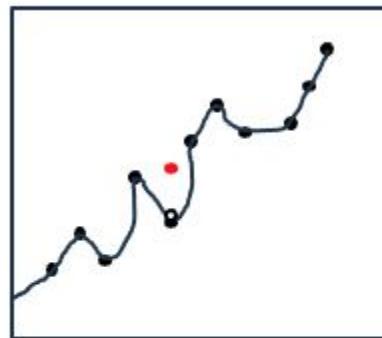
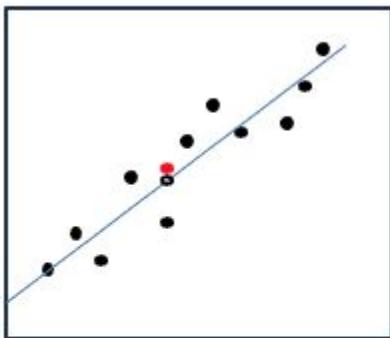


- We know x, want to predict y for this point.

- True value (currently unknown)

→ Using linear model, our predicted value for white point is quite similar to true value.

Using more complex prediction models



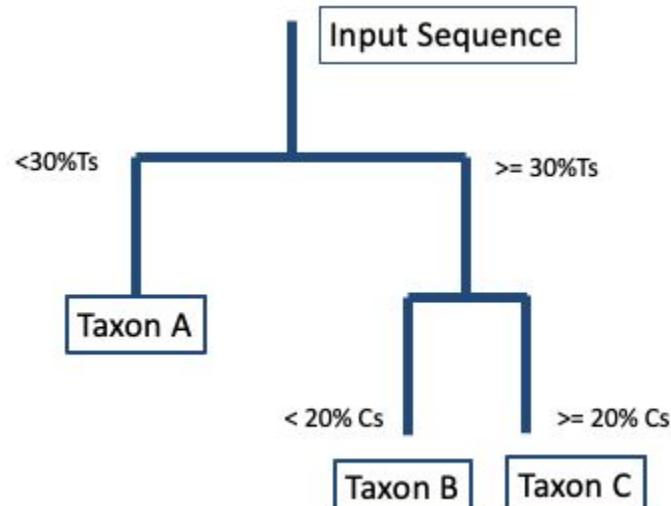
- We know x , want to predict y for this point.
- True value (currently unknown)

➡ Note that our prediction was more accurate with a simpler model here!

Supervised Machine Learning: Random Forest

- For each decision tree:
- “What feature will allow me to split the observations at hand in a way that the resulting groups are as different from each other as possible (and the members of each resulting subgroup are as similar to each other as possible)?”
- <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>

Example of a simple decision tree

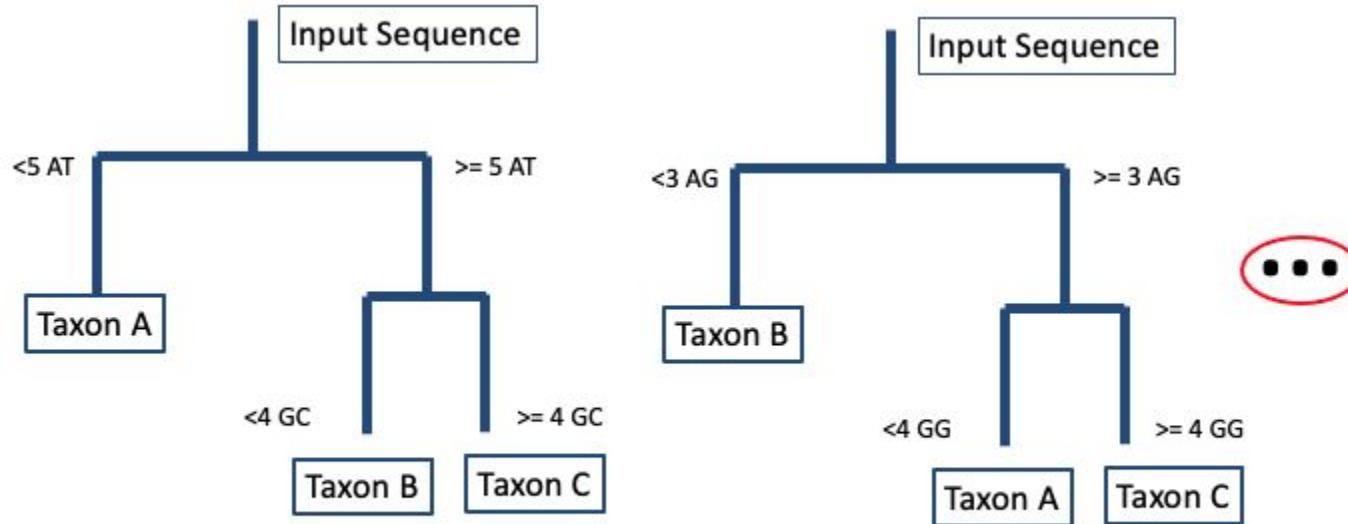


we want the resulting groups to be as homogenous as possible

Supervised Machine Learning: Random Forest

- Example of a classical machine learning algorithm
- We want to separate groups using their features
- What features allow us to separate the groups most accurately?
- Each individual decision tree is built by
 - randomly sampling the training data
 - randomly sampling the features for splitting the data into groups
 - -> forces decision trees to be more diverse.
- For a new case
 - Run through all prediction trees in random forest
 - Final group = majority of predicted group identities
- A few online tutorials
 - <https://victorzhou.com/blog/intro-to-random-forests/>
 - <https://www.blopig.com/blog/2017/04/a-very-basic-introduction-to-random-forests-using-r>

Random forest



Decision on final classification by voting

Example randomForest script

- Example script:
 - Data import & filtering
 - Sequence features
 - Training random forest model
 - Assessing role of feature number using cross-validation
 - Checking prediction accuracy
- Challenge: build your own classifier!
 - could be to separate genes, categories of genes (e.g. protein-coding vs. rRNA), taxonomic groups, etc.