Decision trees and Random Forests

AI for ecologists

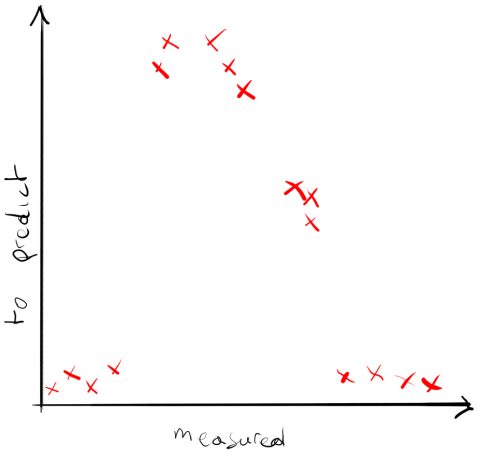
Paul Tresson

20/05/25

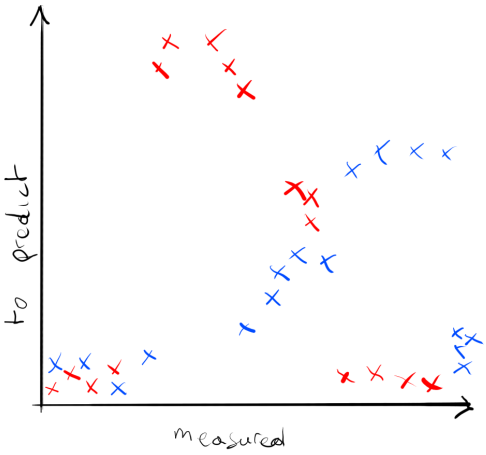
Institut de Recherche pour le Développement FRANCE
AMAPlab
FRB
CESAB

Introduction

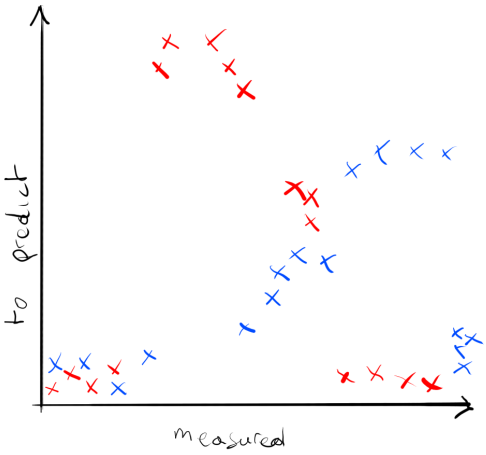
Motivation

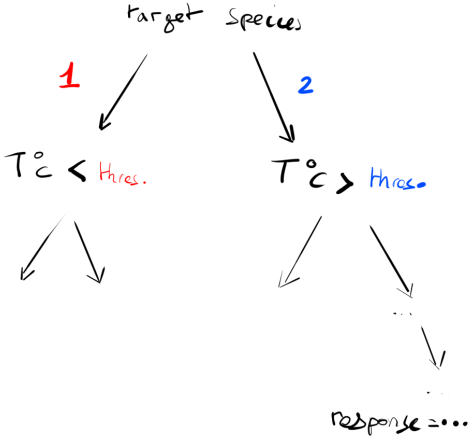


Motivation



Motivation


to predict
measured


target species
1
2
T°c < thres.
T°c > thres.
response = ...

Decision Trees

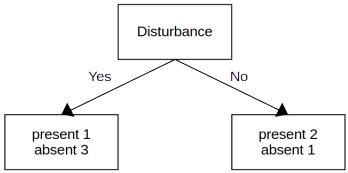
Simple example

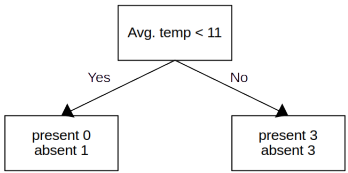
Disturbance	Habitat	Avg. temp.	Presence
Yes	Shrubs	10	0
Yes	Forests	12	0
No	Shrubs	18	1
No	Shrubs	25	1
Yes	Shrubs	28	1
Yes	Forests	30	0
No	Forests	33	0

adapted from StatQuest

Simple example

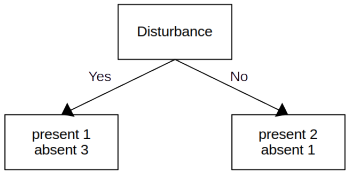
Disturbance	Habitat	Avg. temp.	Presence
Yes	Shrubs	10	0
Yes	Forests	12	0
No	Shrubs	18	1
No	Shrubs	25	1
Yes	Shrubs	28	1
Yes	Forests	30	0
No	Forests	33	0

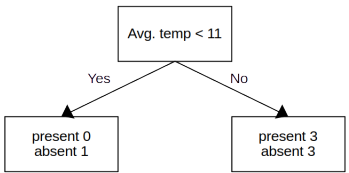




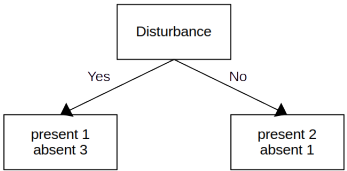
Gini impurity

$$\sum_{i=1}^{J}\left(p_i \sum_{k\neq i} p_k\right) = 1 - \sum_{i=1}^{J} p_i^2$$

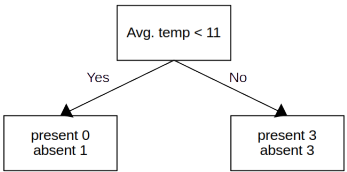




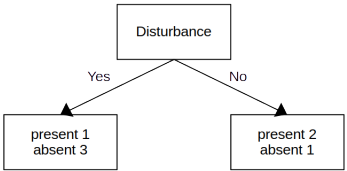
Gini impurity



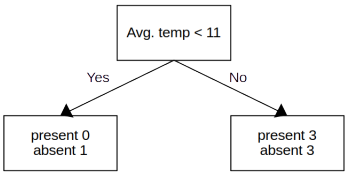
$$1 - (\frac{1}{1+3})^2 - (\frac{3}{1+3})^2 = 0.375$$



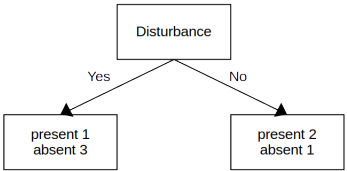
Gini impurity



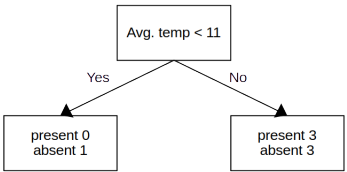
$$Leaf\ \ Gini = (\frac{4}{4+3})0.375$$



Gini impurity

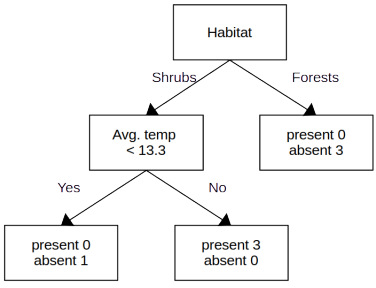


$$1 - (\frac{0}{0+1})^2 - (\frac{1}{0+1})^2 = 0$$

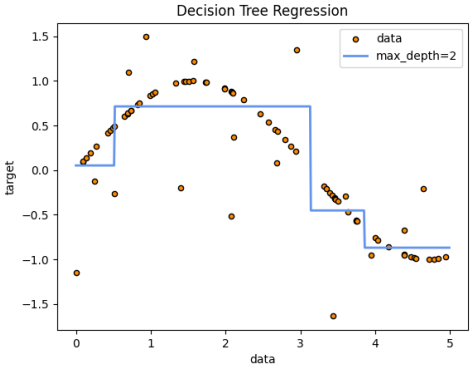


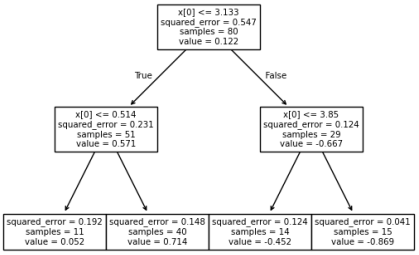
Building the tree

Disturbance	Habitat	Avg. temp.	Presence
Yes	Shrubs	10	0
Yes	Forests	12	0
No	Shrubs	18	1
No	Shrubs	25	1
Yes	Shrubs	28	1
Yes	Forests	30	0
No	Forests	33	0



Regression trees





adapted from sklearn documentation

Random Forests

RF advantages and drawbacks

Advantages

- different inputs

Drawbacks

RF advantages and drawbacks

Advantages

- different inputs
- different outputs

Drawbacks

RF advantages and drawbacks

Advantages

- different inputs
- different outputs
- $\approx$ explainable

Drawbacks

RF advantages and drawbacks

Advantages

- different inputs
- different outputs
- $\approx$ explainable
- pretty fast

Drawbacks

RF advantages and drawbacks

Advantages

- different inputs
- different outputs
- $\approx$ explainable
- pretty fast
- seasonned

Drawbacks

RF advantages and drawbacks

Advantages

- different inputs
- different outputs
- $\approx$ explainable
- pretty fast
- seasonned

Drawbacks

- need to test hyper-parameters

RF advantages and drawbacks

Advantages

- different inputs
- different outputs
- $\approx$ explainable
- pretty fast
- seasonned

Drawbacks

- need to test hyper-parameters
- **need for rich descriptors**

Decendants

- Gradient Boosting
- XGBoost

Usefull ressources

- `scikit-learn` docs !
- StatQuest

Thanks for you attention !

Let's practice !

References i