Decision Trees and Random Forests
Data science Certificate

Aymeric DIEULEVEUT

May 2022

- Supervised learning framework
- why
- how we can use a DT
- how to build it !
  * what do we need to choose
  * what needs to be learned, how ?
- how to implement a DT in practice (python)
- +/-

Outline

A Reminder on Supervised Learning

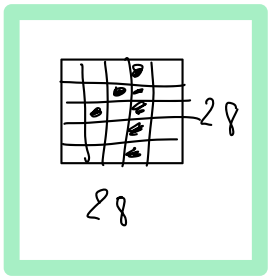
- Observations: $(X_i)_{i=1,\dots n}$. Each X_i has d components.
- Outputs: $(Y_i)_{i=1,\dots n}$ $\in \mathcal{Y}$

3 Examples:

- Iris Dataset: predict type of Iris from characteristics of the flower !
  ($\mathcal{Y}$: type of Iris ; d features)

classif, $\mathcal{Y}$ finite

- MNIST: handwritten digit predict.



classif

$Y = 1$

$\mathcal{Y} = \{0, \dots, 9\}$

- House price prediction

regression

$Y \to$ (price)

$X \to$ (size, location, etc)

Example: Iris Dataset

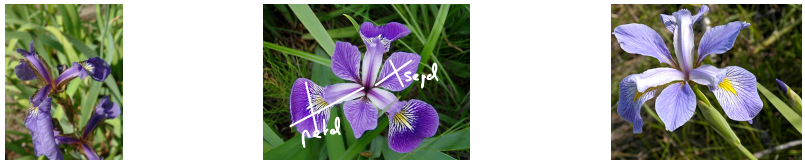


Figure: Iris: setosa, versiocolor, virginica

One of the most widely used toy dataset in classification:

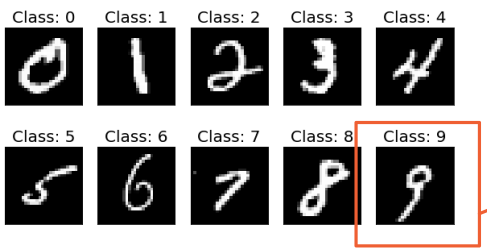
- 3 classes : *['setosa', 'versicolor', 'virginica']* = y.
- 50 points per class. → $n = 150$ number of training points
- 4 features: *['sepal length', 'sepal width', 'petal length', 'petal width']*

input

$d = 4$

4 continuous variables.

Example: MNIST Dataset



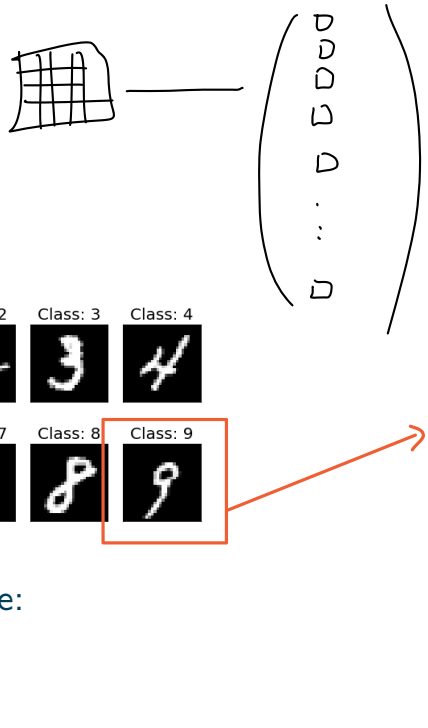


Figure:

Digit recognition from an image:

- 10 classes : $Y_i \in \{0, \dots, 9\}$
- 60k images, (50k train, 10k test), balanced → n
- 784 features. $= d$

↳ 28×28

Input $\in \mathbb{R}^{d=}$

House price Prediction

- Output : house value $Y \in \mathbb{R}$ → regression task!
- 1500 examples
- 81 features: Construction year, Neighborhood, Surface, Floor, Balcony: both quantitative and qualitative, some ordinal variables.

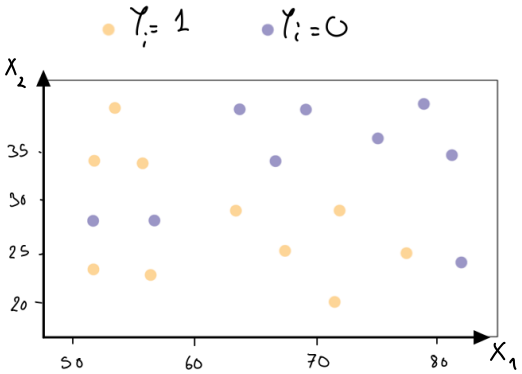
quantitative

Simplified version:

3 features: surface $\in \mathbb{R}$; neighborhood $\in \{0, \dots 20\}$; has garden $\in \{0, 1\}$

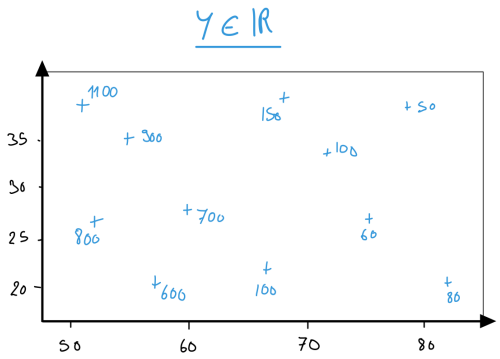
$Y = \begin{cases} 1 & \text{if price} > 1\text{M}. \\ 0 & \text{else.} \end{cases}$

Toy dataset



$\mathcal{Y} = \{0, 1\}$ classif ~~reg.~~

$d = 2$ $n = 19$



$\mathcal{Y} = \{0, 1\}$ reg.

$d = 2$ $n = 11$

Summary of those three datasets

	Iris	MNIST	House price prediction		Toy
Number of examples n	150	60k	1500		19
Number of the features d	4	$784 = 28^2$	81	3	2
Type of the features	quantit	quant	quant + qual.		quant
Space $\mathcal{Y}$	$\{0,1,2\}$	$\{0,\dots 9\}$	$\mathbb{R}_+$	$\{0,1\}$	$\{0,1\}$
Task	classif	classif	regressio	classif	classif

Pipeline: Training phase

train set $\left(X_i^{train}, Y_i^{train}\right)_{i=1..n}$

learn $\rightarrow \hat{f}_n$ such that

$$\hat{f}_n\left(X_i^{train}\right) \simeq Y_i^{train}$$

* type of fun°?
* how to train? [need hyperparam]

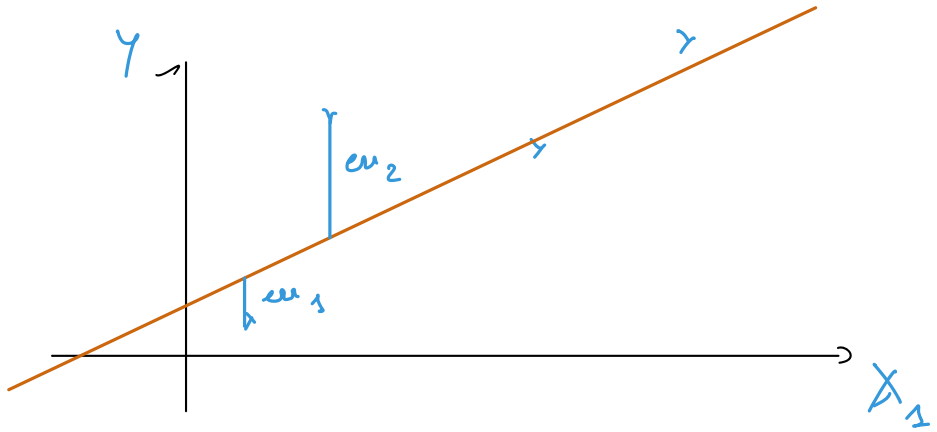
(Validat° phase: find good hyperparameters)

Test phase $X_i^{test} \longrightarrow \hat{f}_n\left(X_i^{test}\right)$ compare to Y_i^{test}

↳ error (test set)

Use it in practice ~~Y_i~~

Linear regression :



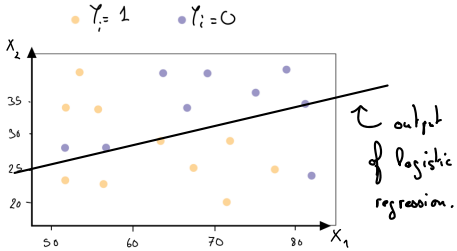
* type of pred° ? linear

* how to train ? Find the line that minimizes a criterion
= $\sum (err)^2$
↑ algo

Logistic regression

vector $\hat{\theta}$ /

$\sigma(\langle \hat{\theta}, X_i \rangle)$ = prob that $Y_i = 1$



↑ output of logistic regression.

predict with 1 if $\sigma(\langle \hat{\theta}, x \rangle) \geq \frac{1}{2}$ = { 1 if $\sigma(\langle \hat{\theta}, x \rangle) \geq \frac{1}{2}$; 0 otherwise

Limitations of linear models:

(1) * data is not always linearly separable!

(2) * interpretable is ok-ish but still not easy for a human.

Decision trees

* Regression: House-price dataset.

$$y \simeq 4000 \times \text{surf} + 16000 \times \text{floor} - 1000 \times \text{neighborhood} + 6000\ \text{surf}(2\text{mh}) + \cdots +$$

81 terms in the sum.

learned!

100 m² (→ 2mh) 110 m² overall

$6000 \times 100 + 400 \times 110$

Decision Tree

Goals of Decision Trees

1. Supervised Learning: solve the classification or regression task
2. Provide some understanding: what on this example allows me to classify it? What are the important features ?

↳ have an easy to - interpret method

3. Go beyond linear models

Decision Tree

Goals of Decision Trees

1. Supervised Learning: solve the classification or regression task
2. Provide some understanding: what on this example allows me to classify it? What are the important features ?

Two questions:

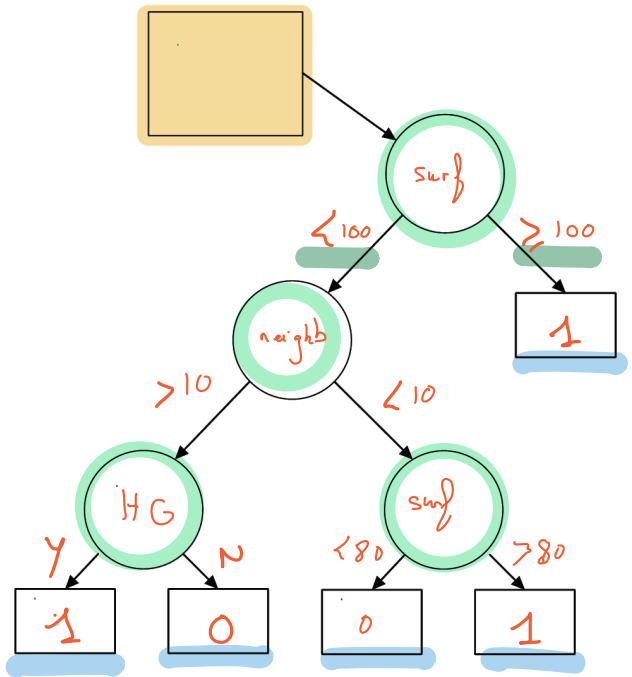
- How to train a decision tree? $\rightarrow$ **training phase (learning algorithm)** moderate
- How to predict with a given decision tree? $\rightarrow$ **inference phase** easy

Inference Phase

House Price - simplified

- Root = observation = input
- Nodes = tests — looking at one variable
- Branches = possible outcomes — only 2 always
- Leaves = final decision = output

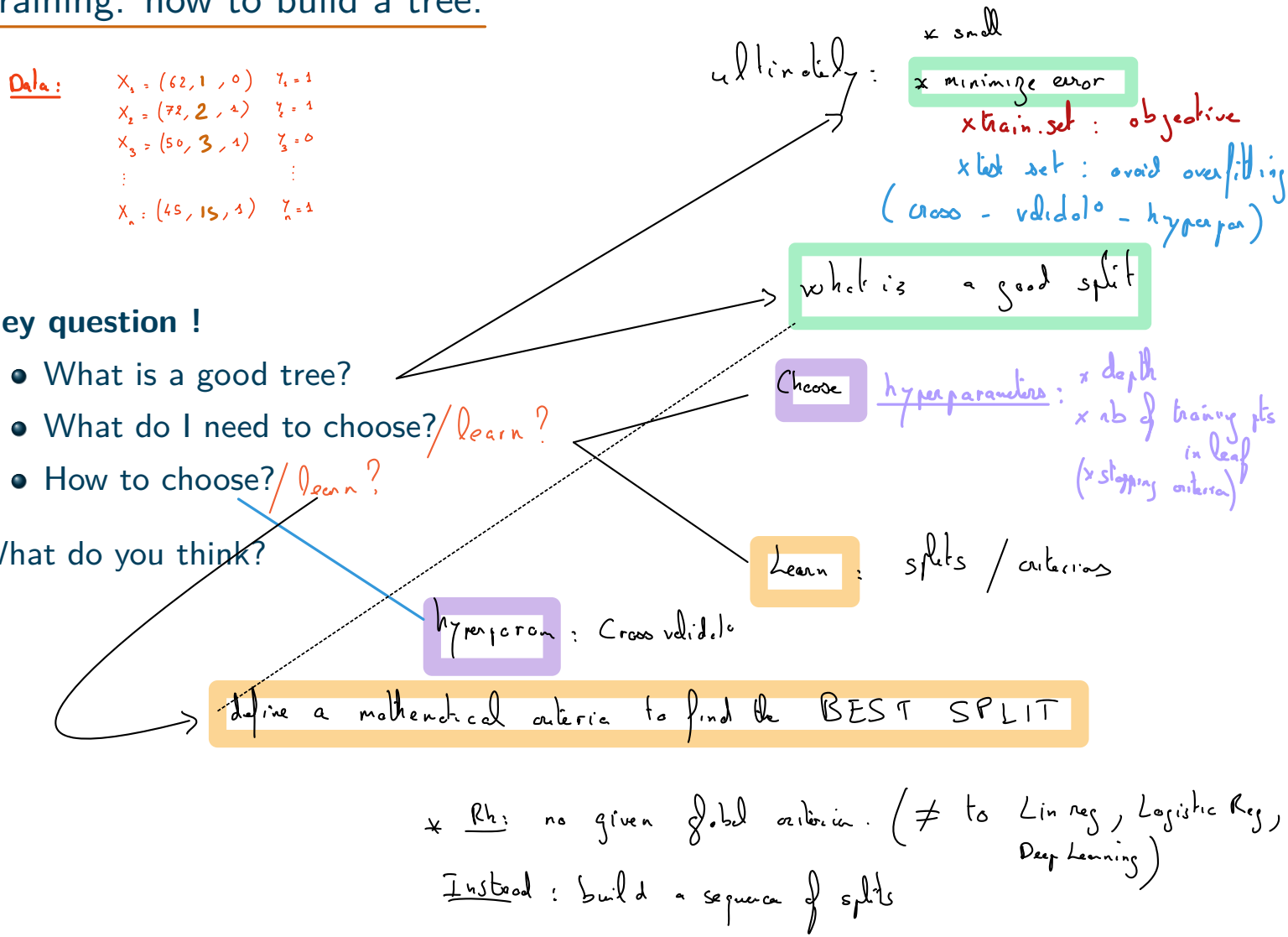
$X_{new} = (92, 6^*, 0)$



* Inference is very easy, very fast and easy to interpret! → turn into a policy! (ex: Covid and vaccinal° calendar)

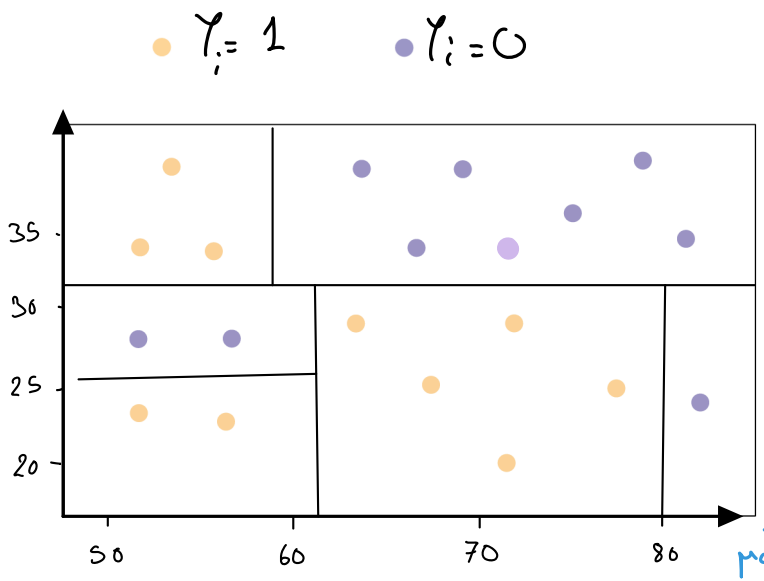
* Training: learn all nodes and splits!

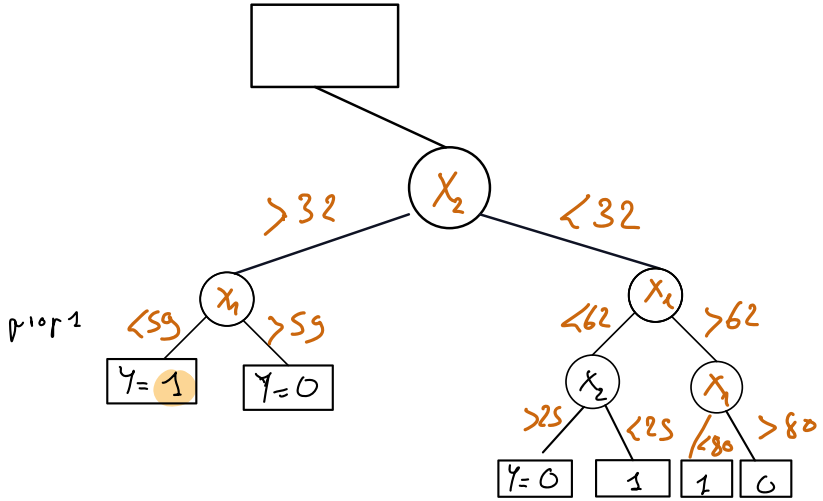
Training: how to build a tree.



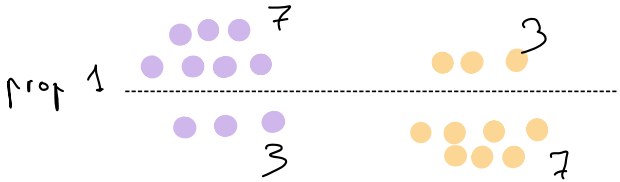
Building trees : CART

best split ?





prob of Y/0

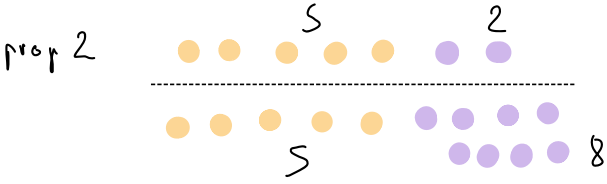


measure : misclassif

$0,3 \times 0,7$ } $0,21$ $= 42\%$
$0,7 \times 0,3$ } $+0,21$

$\oplus$

$42\% \rightarrow 84\%$



$2\left(\frac{2}{7} \times \frac{5}{7}\right) = \frac{20}{49} \simeq 40\%$ +

$2\left(\frac{5}{13} \times \frac{8}{13}\right) = \frac{80}{169} \simeq 47\%$ } 87%

Stopping criteria

-

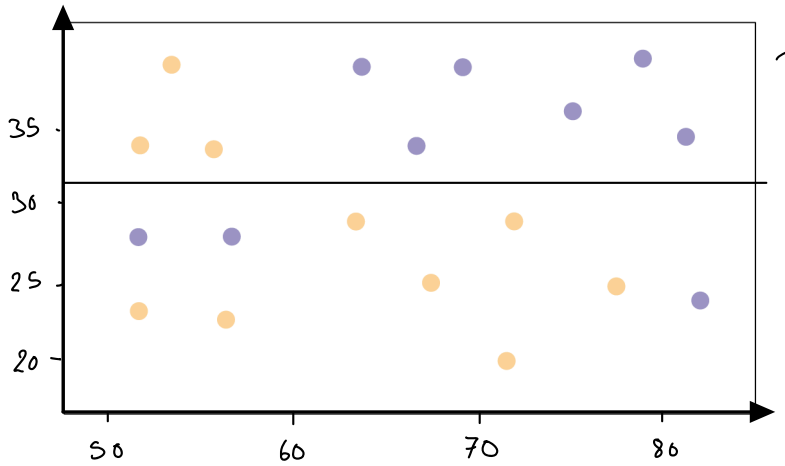
Complexity.

→ $d \times n$ impurity at each split.

splits $\in n$ worst case

1 : perfect case

Gini's impurity [classification : J classes]



$0,3 \times 0,7$
$+ 0,7 \times 0,3$

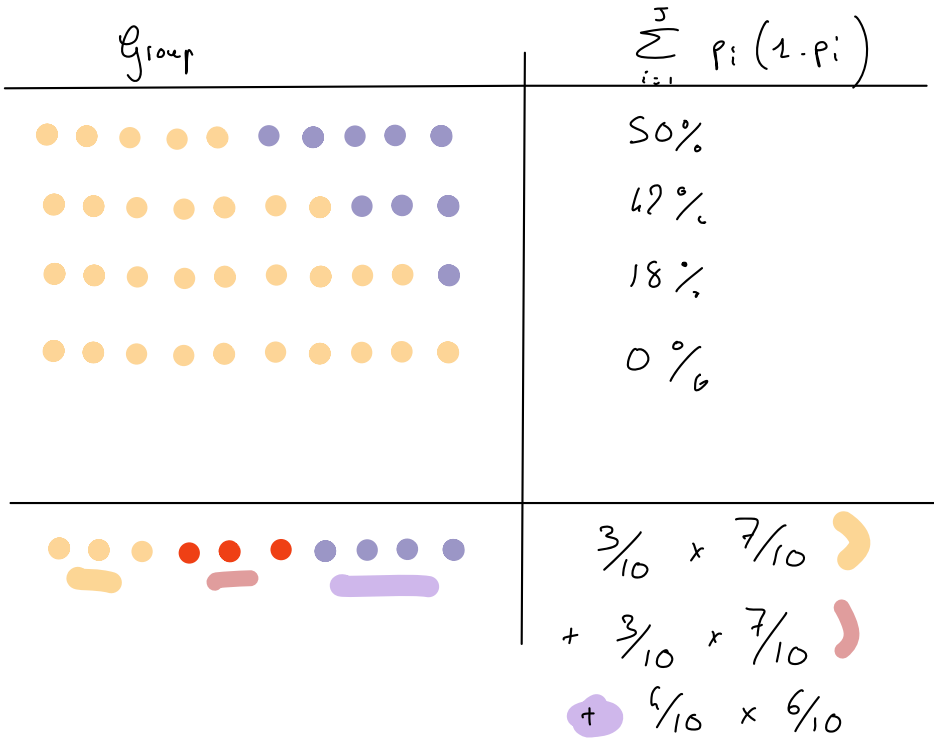
Algorithm: for all possible splits
compute $\sum_{\text{groups of the split}} I_{\text{group}}$

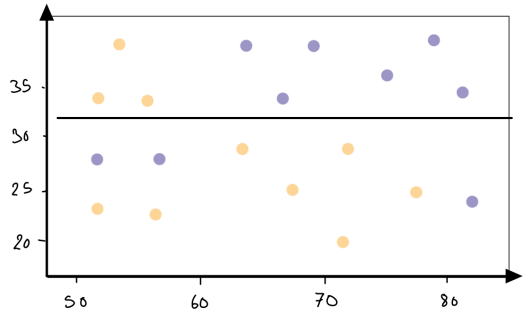
Gini impurity measures **how often a randomly chosen element from the set** would be **incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset**.

$$I_G(p) = \sum_{i=1}^{J} p_i(1-p_i) = 1 - \sum_{i=1}^{J} {p_i}^2$$

(J → nb of classes)

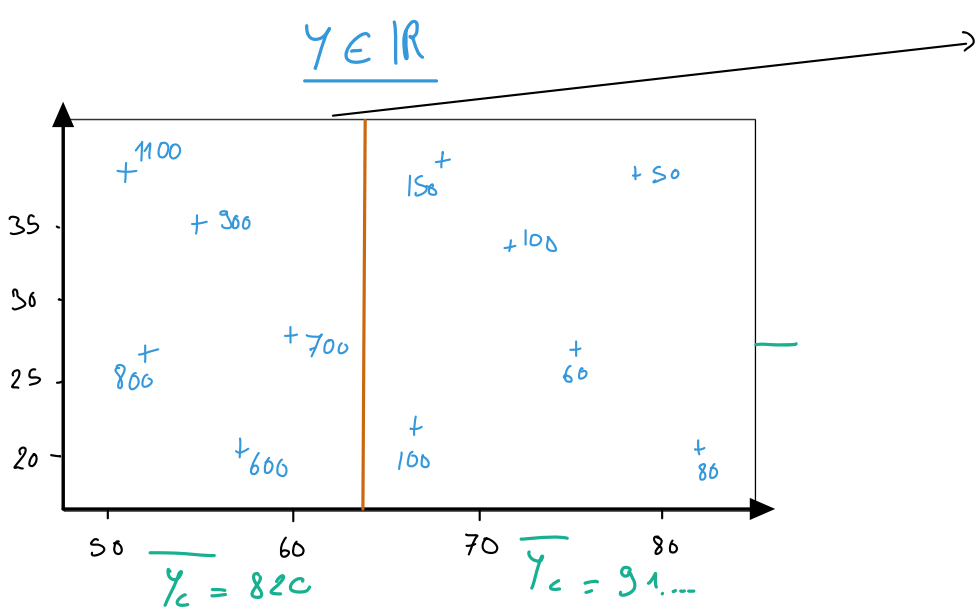
For two classes, related to the variance if classified as Bernoulli r.v..


Group
$\sum_{i=1}^{J} p_i(1-p_i)$
50%
42%
18%
0%
3/10 x 7/10
+ 3/10 x 7/10
+ 6/10 x 6/10


35
30
25
20
50
60
70
80

Building trees : CART

Regression case



$$MSE = \frac{1}{S}\left[(1100 - 820)^2 + (300 - 820)^2 + (800 - 820)^2 + (700 - 820)^2 + (600 - 820)^2 \right]$$

How to chose the best cut ?

- e.g., Given a possible cut, measure the reduction of impurity :
  - in regression: mean-squared-error $\sum_{i \in C} (Y_i - \bar{Y}_C)^2$
  - in classification: Gini's impurity.

$$\bar{Y}_C = \frac{\sum_{i \in C} Y_i}{\#C}$$

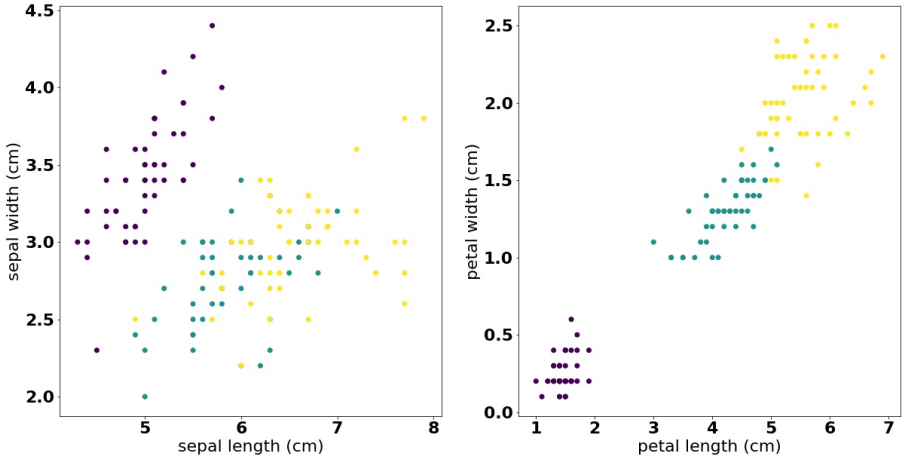
- Find dimension and threshold with optimal impurity reduction.
- Iterate until stopping criterion is met.

Example: Iris Dataset

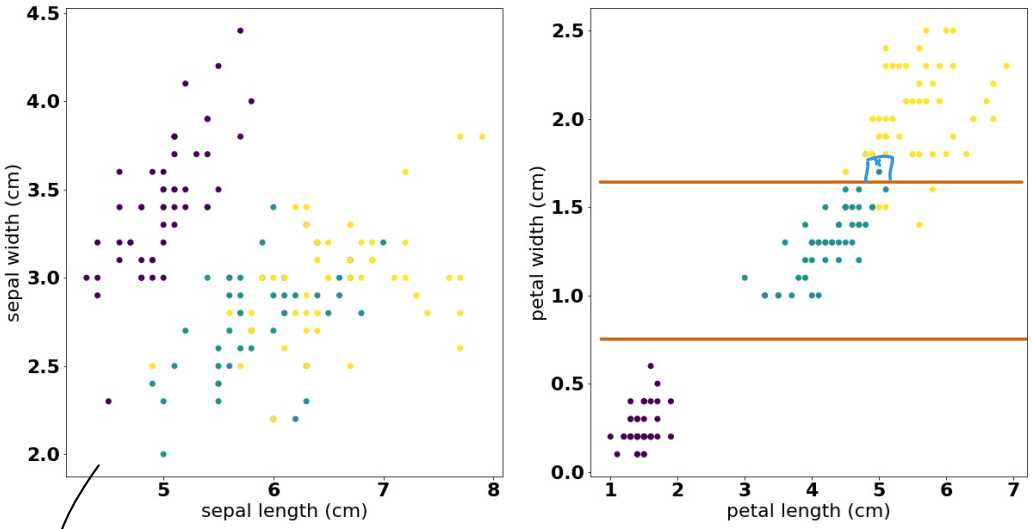
Figure: Iris: setosa, versiocolor, virginica

One of the most widely used toy dataset in classification:

- 3 classes : *['setosa', 'versicolor', 'virginica']*
- 50 points per class.
- 4 features: *['sepal length', 'sepal width', 'petal length', 'petal width']*



Exercise: Guess Classification tree for Iris Dataset



In pratice : Scikit Learn

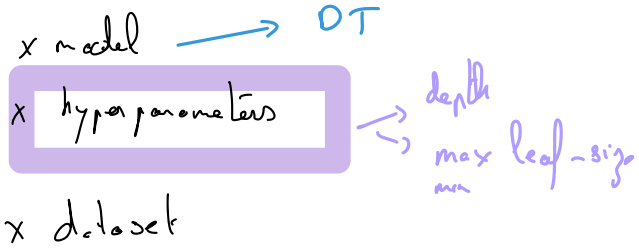
Universal Python library for Machine Learning:

- Very easy to use
- Very well documented
- Open Source

`https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html`

Example: Iris Dataset

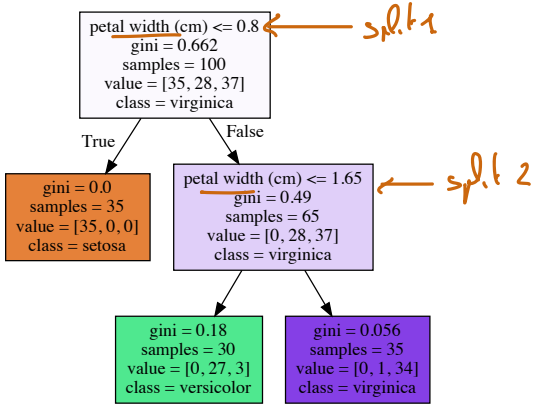
What do we need to specify ?



Output : learned model → DT .

```python
from sklearn.datasets import load_iris
from sklearn.model_selection import cross_val_score
from sklearn.tree import DecisionTreeClassifier, plot_tree
from sklearn.model_selection import train_test_split

# Choose the Learning algorithm
clf = DecisionTreeClassifier(max_depth = 2,
                             random_state=0)

#Load the dataset
iris = load_iris()
X_train, X_test, y_train, y_test = train_test_split(
    iris.data, iris.target, test_size=0.33, random_state=43)

#Check accuracy
cross_val_score(clf, iris.data, iris.target, cv=10)

# Fit the model
clf.fit(X_train,y_train)

# Plot tree
plot_tree(clf, feature_names = fn,
                class_names=cn,
                filled = True)
plt.show()
```

Output:



Example:

```python
# Check
X_test[0,:], iris.target_names[y_test[0]]
```

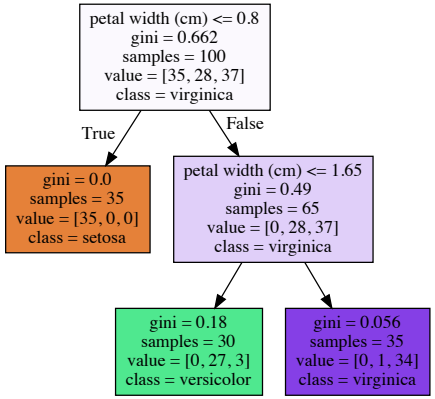
Returns:

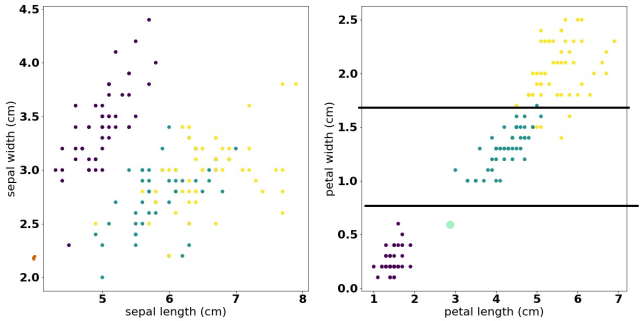
(array([4.8, 3.1, 1.6, 0.2]), 'setosa')

Code: https://colab.research.google.com/drive/1dXYIjacNAeASgl54vIWyWjMQlfgigSVA?usp=sharing#scrollTo=VeJHkJd7RsDw

Example: Iris Dataset

Output:





Example:

```
# Check
X_test[0,:], iris.target_names[y_test[0]]
```

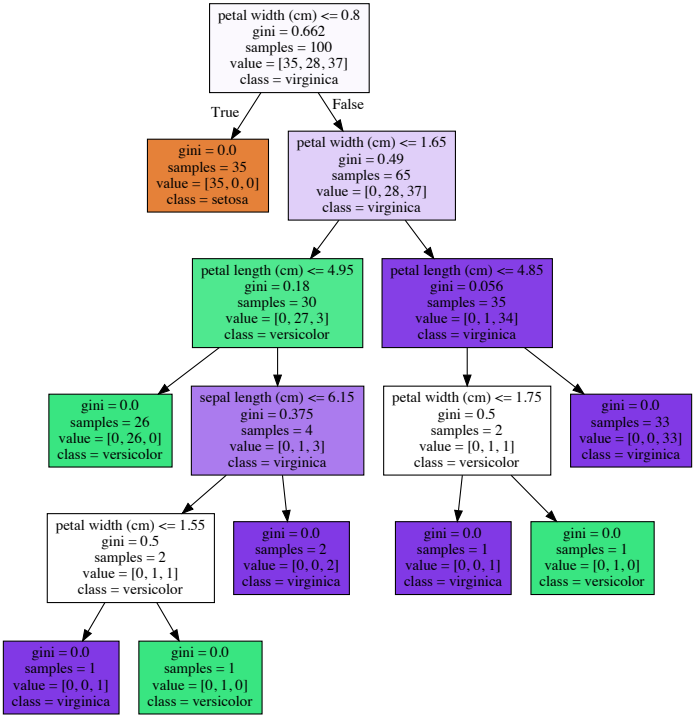
Returns:
(array([4.8, 3.1, 1.6, 0.2]), 'setosa')

Code: `https://colab.research.google.com/drive/1dXYIjacNAeASgl54vIWyWjMQlfgigSVA?usp=sharing#scrollTo=VeJHkJd7RsDw`

What if I use a deeper tree ? Or if I do not specify the depth in the model definition ?

Deeper trees

```
# Choose the Learning algorithm
clf = DecisionTreeClassifier(random_state=0)
```



- When does the algorithm stop then ?
- What do you think about it ?

overfitting

Stopping / Pruning

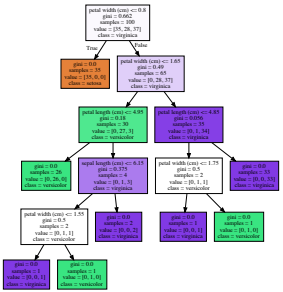
Trees that achieve perfect classification may suffer from over-fitting. (see example in the lab)

To avoid this problem, we can reduce the complexity of the trees:

1. Stopping rules
   - Set a minimum number of samples inside each leaf.
   - Set a maximum depth.
2. Pruning
   - Reduced error pruning.
   - Cost complexity pruning.

Pruning : example of Cost complexity pruning.

We create a sequence of Trees by changing one subtree at each step into a leaf, until we get only the root: the subtree is chosen to minimize the error made.[1] Then the test error is evaluated along the sequence, to choose the optimal pruning.



[1] *more details here*

Cart: pros and cons

Pros:

- Simple to understand, interprete, visualize
- Implicitly perform features/variable selection
- can handle both categorical and numerical data
- little effort for data preparation
- Non linear relationships do not affect performance
- Can work with large number of observations.

Cons:

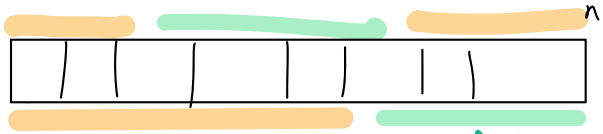
1. Can create over-complex trees: overfitting
2. Unstability: small variations of data can generate completely different trees
3. Cannot guarantee global optimal tree
4. Biased if some classes dominate

iterative algo — not a global criteria

Solution : Using Random Forests !

Overfitting

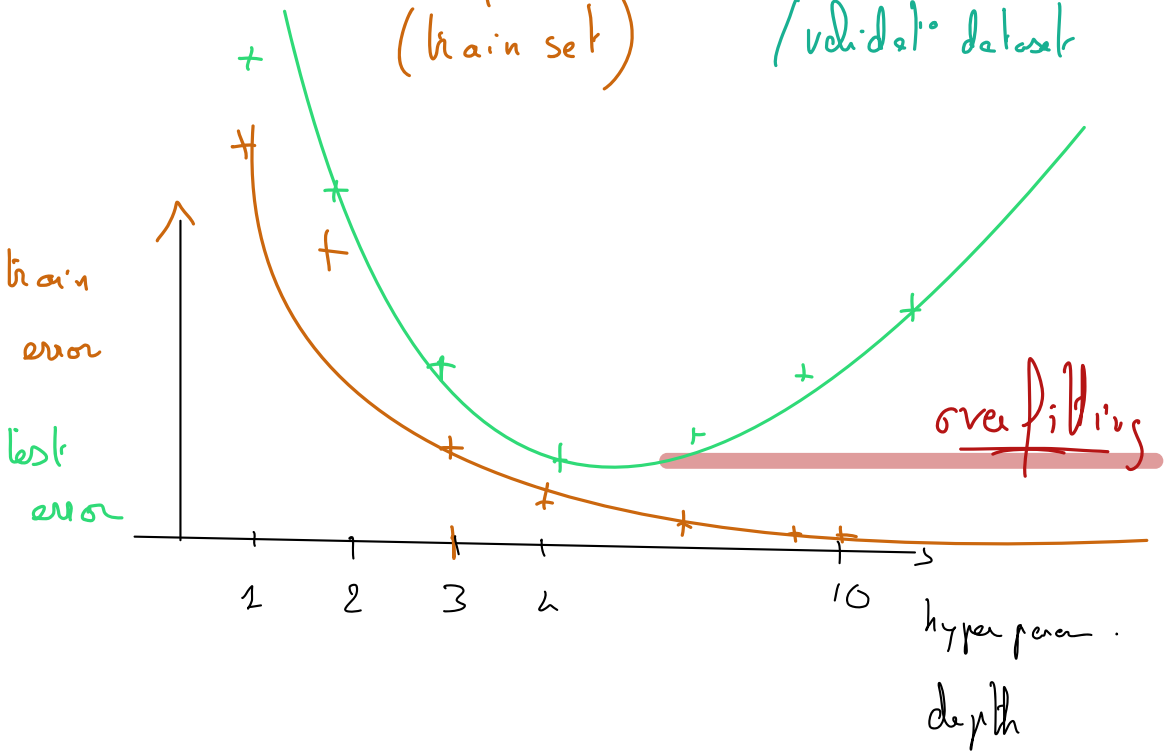
Train / validate data



Test data

Fit (train set)

evaluate / validate dataset



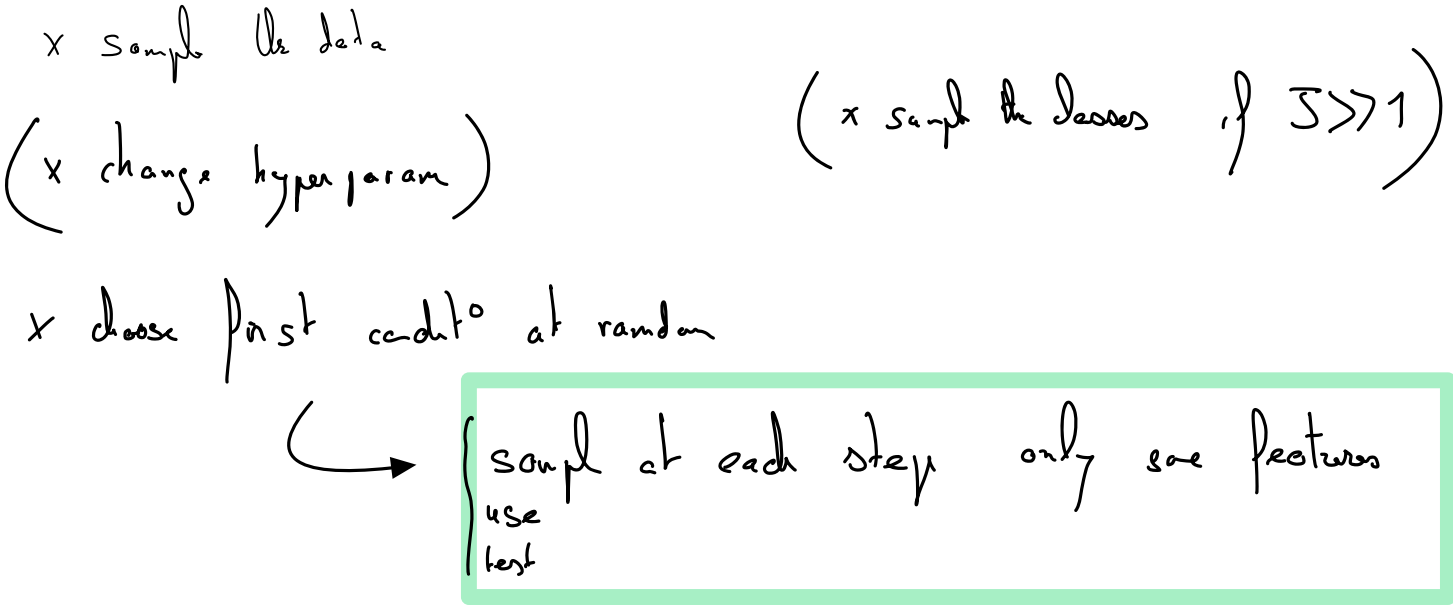
Outline

Random Forests

To put it in simple words: random forests builds multiple decision trees and merges them together to get a more accurate and stable predictions .

↳key idea: create variable trees + aggregate them

How to create variability ?



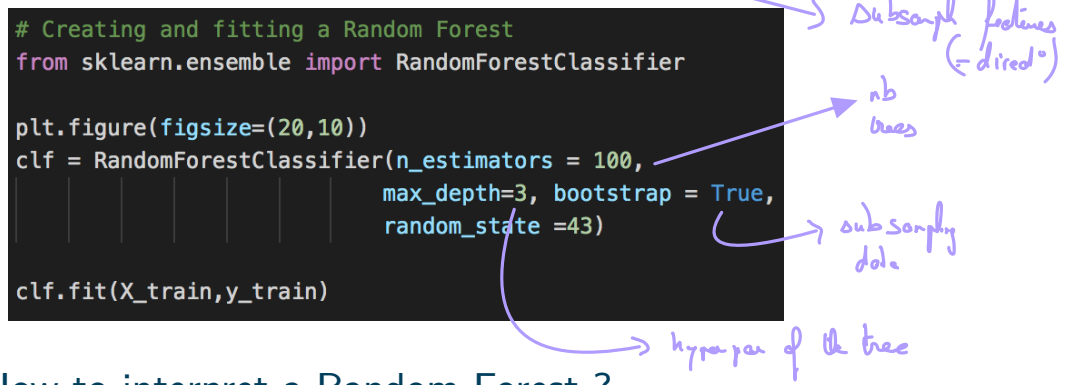
Random Forests

To put it in simple words: random forests builds multiple decision trees and merges them together to get a more accurate and stable predictions .
↳key idea: create variable trees + aggregate them

How to create variability ?

- Instead of the most important features, search the best feature among a random subset (ntry).
- Work on subsets of data (bootstrap=True).
- More `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html`

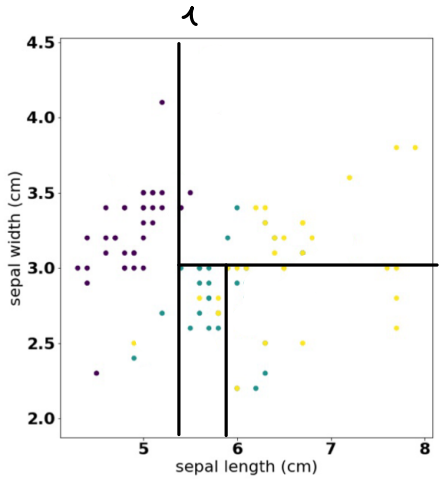
```
# Creating and fitting a Random Forest
from sklearn.ensemble import RandomForestClassifier

plt.figure(figsize=(20,10))
clf = RandomForestClassifier(n_estimators = 100,
                             max_depth=3, bootstrap = True,
                             random_state =43)

clf.fit(X_train,y_train)
```



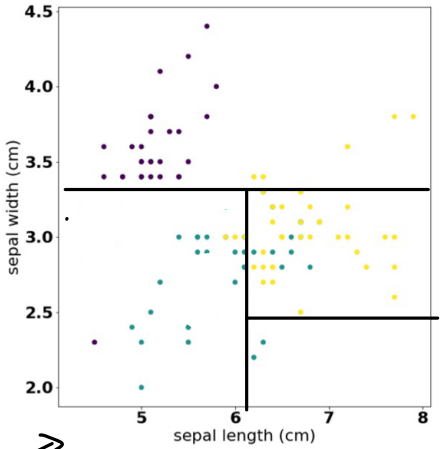
How to aggregate ? How to interpret a Random Forest ?

Bootstrap. = subsampling data

Tree 1



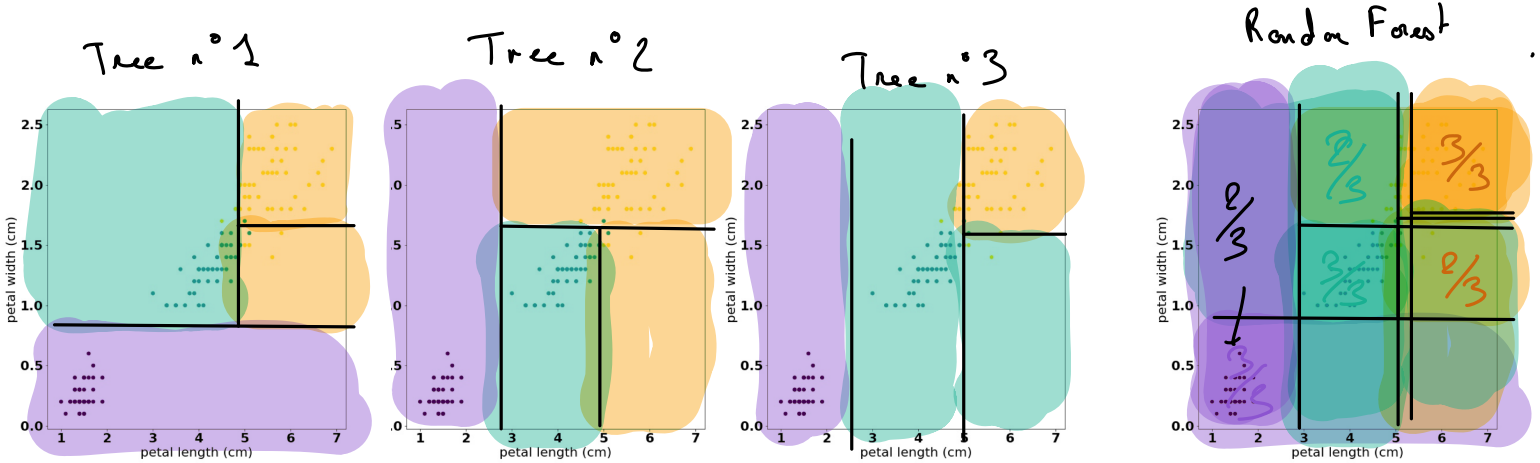
Tree n° 2



subsampled points

Aggregation

We using a voting or averaging process !



Aggregation

We using a voting or averaging process !

Feature importance

- how much tree nodes using a particular feature reduce impurity along the trees.
- suggests feature selection rule: drop out features with low importance to avoid overfitting.
- Attribute **feature_importance_** in RandomForestRegressor of Sklearn.

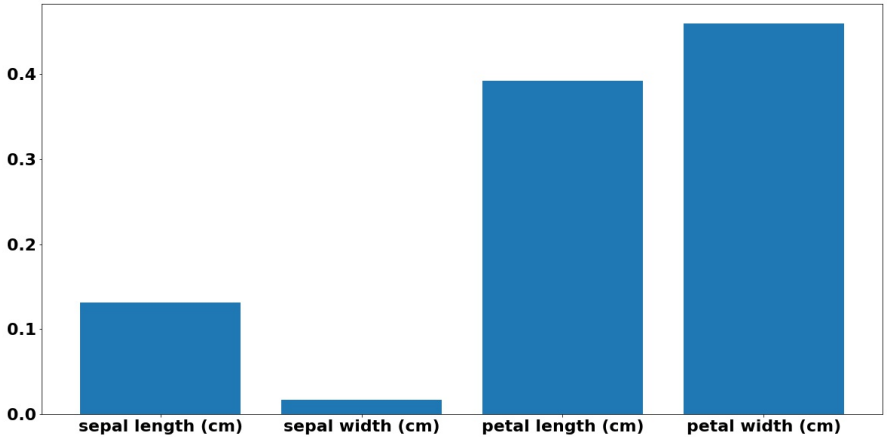


Figure: Feature importane output for a Random Forest Classifier in Python

Comparison with decision trees

Cons:

- Less interpretable
- Slower to run

Pros:

- Better in higher dimension
- More stable outputs
- Feature selection

Improving predictions - more arguments.

1. **n_estimators** : number of trees in the forest
   - improves prediction / stability
   - slows down the algorithm
2. **max_features** : maximum number of features considered to split a node
3. **min_sample_leaf** : minimum number of samples that should remain in a leaf node.

RF pros and cons

RF Pros:

1. works for classification and regression
2. default hyperparameters often produce reasonable prediction -> easy to use.
3. avoid overfitting if some good trees in the forest and easy feature selection -> high dimensional pb.
4. hard to beat in performance.
5. Bonus : supports multiple outputs!

RF Cons:

1. ~~Fast to train~~ but slow to provide new predictions -> ineffective for real-time predictions
2. Good for prediction but bad for description.

What about Regression ?

Questions?

Boosting

Another import technique (very powerful !)

Main idea:

- Give more importance to difficult point iteratively
- Incrementally building an ensemble by training each new instance to emphasize the training instances previously mis-modeled.

Example: AdaBoost

See: `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html`

Bonus: Multiple Outputs

- **Goal:** predict several outputs simultaneously
- **Solution:**
  - each leaf contains a value for each output
  - to chose the splits, we use a (weighted) average of the impurity of each output.

Bonus: complexity

Total complexity for one decision tree:

Worst case:

$$O(n^2 d)$$

Balanced case:

$$O(nd)$$

Tips

- Decision trees tend to overfit: limit the depth or use `min_samples_leaf` (5 is a good initial pick)
- Visualize the tree with a small depth first
- Balance your dataset: trees tend to be biased towards dominating class.

Conclusion

1. Trees are one of the simplest method (the most intuitive / human like)
2. Random Forests give excellent results in many applications.

Lab :

- Example on a synthetic dataset of time series
- Application to inflation prediction in Brazil.