Machine learning for classification (tree, bag, RF, Xtree)

M. Mougeot

ENSIIE, October 2024

Contents

Introduction

Goal of the practical sessions

- To understand classification machine learning methods, from a methodological and practical point of view.
- To apply models and to tune the appropriate parameters on several data sets using the 'Python' language.
- To interprete 'Python' outputs.

Warnings and Advices

- The goal of this practical session is not "just to program with Python"" but more specifically to understand the framework of Modeling, to learn how to developp appropriate models for answering to a given operationnal question on a given data set. The MAL course belongs to the **Data Science courses** . For each MAL practical session, you should **first understand** the mathematical and statistical backgrounds, **then write your own program with 'Python'** to practically answer to the questions.

Instructions for the firt exercices

- The firt part of this document introduces how to program Binary classification Machine learning models using the Python language. In order to understand more deeply the properties of the models from a modelling point of view given the values of the hyper parameters of each method, the firt exercices use 2D simulated data, so that the boundaries of the classification area may be visualized.

Instructions for the last exercice: the TP project.

- In the last exercice, the goal is to calibrate a machine learning decision model able to diagnose a binary medical output thanks to several covariables.
- This last practical work must be carried on with a 'group of two students'.
- This MAL project aims to developp a Jupyter notebook to solve a classification problem . Your names have to be written in the first lines of the program file with comments. Please note that without this information, no grade will be attributed to the missing name project.

For this practical session, the following librairies need bo be uploaded in the python environment.

```
import random as rd
import numpy as np
import matplotlib.pyplot as plt
from matplotlib.colors import ListedColormap
import math
```

The data

A labeled Data Set

With the help of function *gauss*() of the library `random`, or function *random*() of the library 'numpy, simulate a two dimensional sample of size $N = 200$ as illustrated in Figure 1 (left).

In order to visualise the MAP decision boundaries. A grid of $N = 15 * 15$ inputs is generated with regularly spaced points computed on the support on the previous training data set.

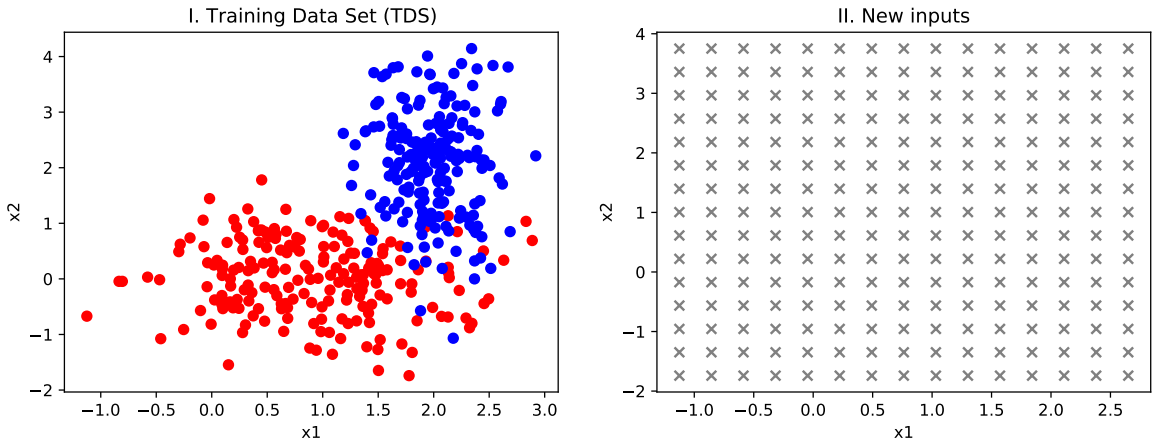

Figure 1: Data sets

Classification Decision Tree

The *DecisionTreeClassifier*() of the library 'tree' implements the decision tree for classification.

Model calibration

Following instructions calibrate a model on the training data set given inputs X and output Y.

```
# Decision Tree
from sklearn import tree
from sklearn.tree import DecisionTreeClassifier

tree = tree.DecisionTreeClassifier()
treefit = tree.fit(X, Y);
pY_train=treefit.predict_proba(X);

#Score and decision  on the training set
predxclass=np.argmax(pY_train,axis=1); #print(predclass)

#Accuracy
E_train=(G != predxclass).sum()/len(G)
#print("Error on the complete training set %5.2f->",E_train)
```

Model description

The tree model can be displayed thanks to the following instructions (see annex for printed tree)

```
from sklearn.tree import export_text
r = export_text(treefit); #print(r)
```

Score and decision boundaries

Compute the class prediction for all the inputs of the grid data set and visualize the decision boundaries.

```
## <matplotlib.contour.QuadContourSet object at 0x7fde82ffe1f0>
```

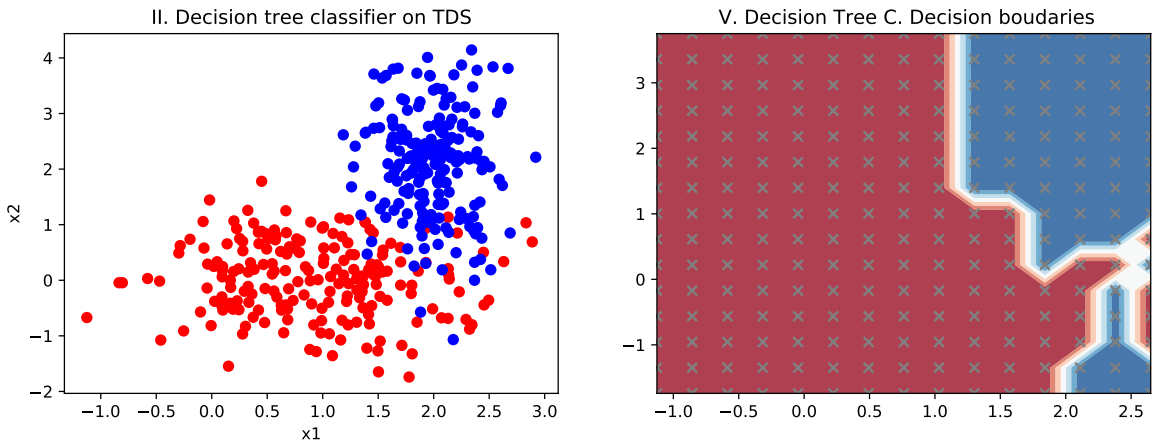

Figure 2: Data sets

Meta parameters: split, leaf and deviance

What are the meta parameters of the decision tree ? What are the default values of the meta parameters of the $DecisionTreeClassifier()$ function ?

Modify the Python instructions in order to set the minimum sample split parameter to 20 and the minimum sample leaf to 10. Update the model taking into account the new parameter values and visualize the new decision boundaries as for Figure 3.

Modify the Python instructions in order to modify the tuning of the deviance parameter.

```
## <matplotlib.contour.QuadContourSet object at 0x7fde830ab550>
```

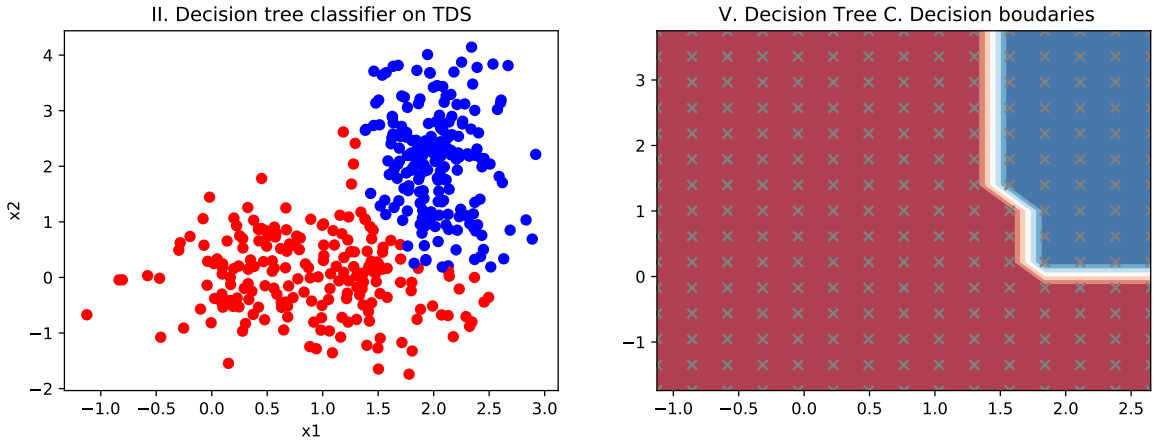

Figure 3: Data sets

How do you suggest to tune the meta parameters ?

Bagging

Model calibration

Following instructions calibrate a bagging model based on classification trees on the training data set given inputs X and output Y.

```
#Bagging
from sklearn import tree
from sklearn.tree import DecisionTreeClassifier
from sklearn.tree import export_text
from sklearn.ensemble import BaggingClassifier

treemod = tree.DecisionTreeClassifier()
bagmod=BaggingClassifier(base_estimator=treemod, n_estimators=10, random_state=0)
treemodfit=treemod.fit(X, y);
bagmodfit=bagmod.fit(X, y);

pY_train=bagmodfit.predict_proba(X);

#Score and decision  on the training set
predxclass=np.argmax(pY_train,axis=1); #print(predclass)
```

Score and decision boundaries

```
## <matplotlib.contour.QuadContourSet object at 0x7fde8360a610>
```

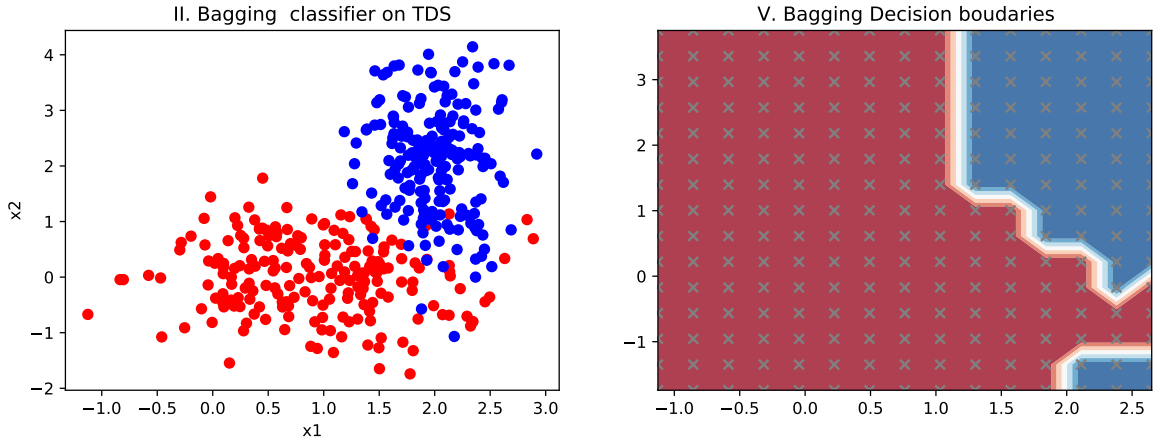

Figure 4: Data sets

Meta parameters

What are the meta parameters of the bagging function applied to decision tree classifiers? Modify the default values of the Bagging parameters and check the differences on the decision boundaries graph.

Random Forest

Model calibration

Following instructions calibrate a bagging model based on classification trees on the training data set given inputs X and output Y.

```
#Random forest
from sklearn.ensemble import RandomForestClassifier
from sklearn import tree
from sklearn.tree import DecisionTreeClassifier
from sklearn.tree import export_text

#tree = tree.DecisionTreeClassifier()
RF = RandomForestClassifier(max_depth=2, random_state=0)
RFfit = RF.fit(X, Y);
pY_train=RFfit.predict_proba(X);

#Score and decision computation on the training set
predxclass=np.argmax(pY_train,axis=1); #print(predclass)
```

Score and decision boundaries

```
## <matplotlib.contour.QuadContourSet object at 0x7fde836f4790>
```

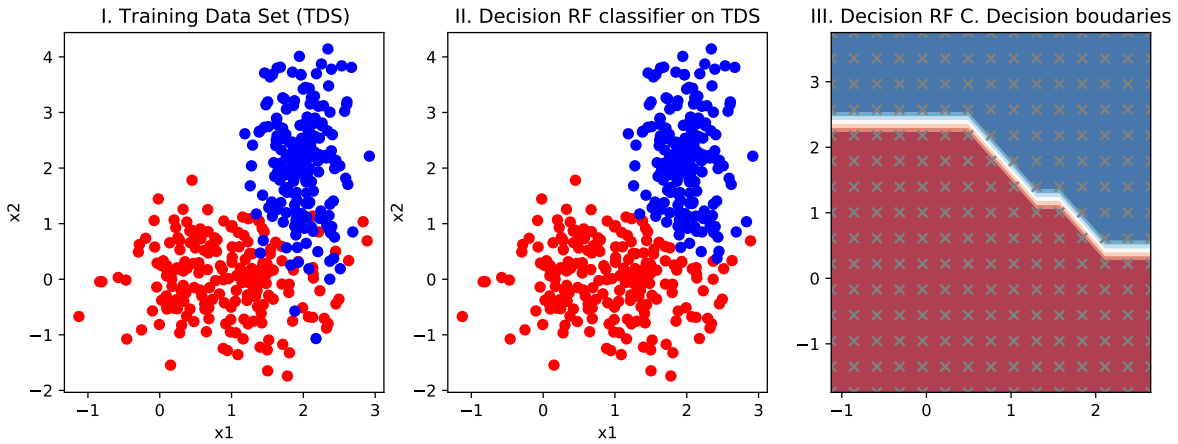

Figure 5: Data sets

Meta parameters

What are the meta parameters of the random forest function applied to decision tree classifiers? Modify the default values of the Bagging parameters and check the differences on the decision boundaries graph.

Random Forest

Model calibration

Following instructions calibrate a random forest model based on classification trees on the training data set given inputs X and output Y.

```
#Random forest
from sklearn.ensemble import RandomForestClassifier
from sklearn import tree
from sklearn.tree import DecisionTreeClassifier
from sklearn.tree import export_text

#tree = tree.DecisionTreeClassifier()
RF = RandomForestClassifier(max_depth=2, random_state=0, oob_score = True)
RFfit = RF.fit(X, Y);
```

Model criteria

Several criteria are commonly used to evaluate the RF model as the global score (accuracy), theOOB (Out Of Bag) and the importance variables:

```
score=RF.score;
OOB=RF.oob_score_
IF=RF.feature_importances_
```

Score and decision boundaries

```
## <matplotlib.contour.QuadContourSet object at 0x7fde837fb4c0>
```

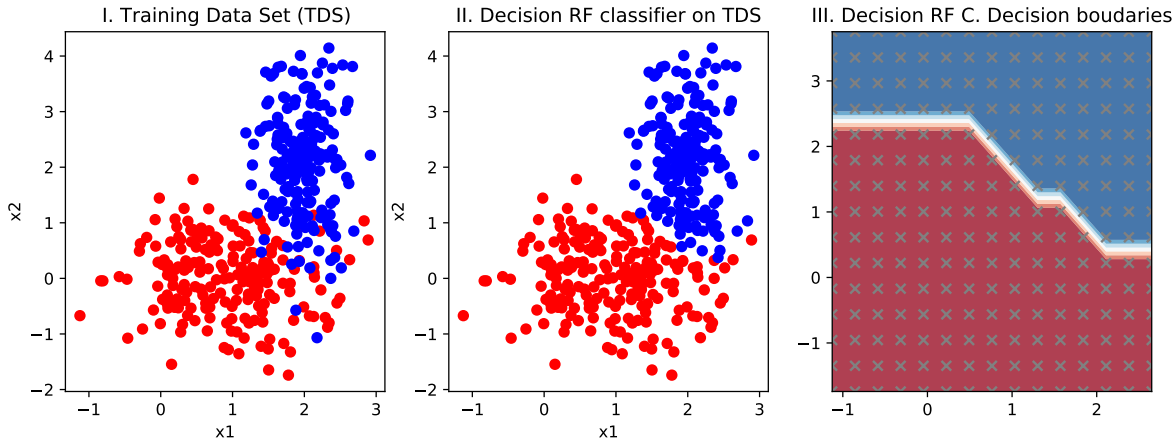

Figure 6: Random Forest

Extra Trees

Model calibration

Following instructions calibrate a random forest model based on classification trees on the training data set given inputs X and output Y.

```
#Random forest
from sklearn.ensemble import ExtraTreesClassifier

#tree = tree.DecisionTreeClassifier()
ExTC = ExtraTreesClassifier(max_depth=2, random_state=0)
ExTCfit = ExTC.fit(X, Y);
pY_train=ExTCfit.predict_proba(X);
```

Model criteria

Several criteria are commonly used to evaluate the ExT model as the global score (accuracy), the OOB (Out Of Bag) score (in the case of bootstap samples are generated) and the importance variables:

```
from sklearn.ensemble import ExtraTreesClassifier
ExTC = ExtraTreesClassifier(max_depth=2, random_state=0,bootstrap=True,oob_score=True)
ExTCfit = ExTC.fit(X, Y);
pY_train=ExTCfit.predict_proba(X);
score=ExTC.score;

OOB=ExTC.oob_score_
IF=ExTC.feature_importances_
```

Score and decision boundaries

```
## <matplotlib.contour.QuadContourSet object at 0x7fde840e0e80>
```

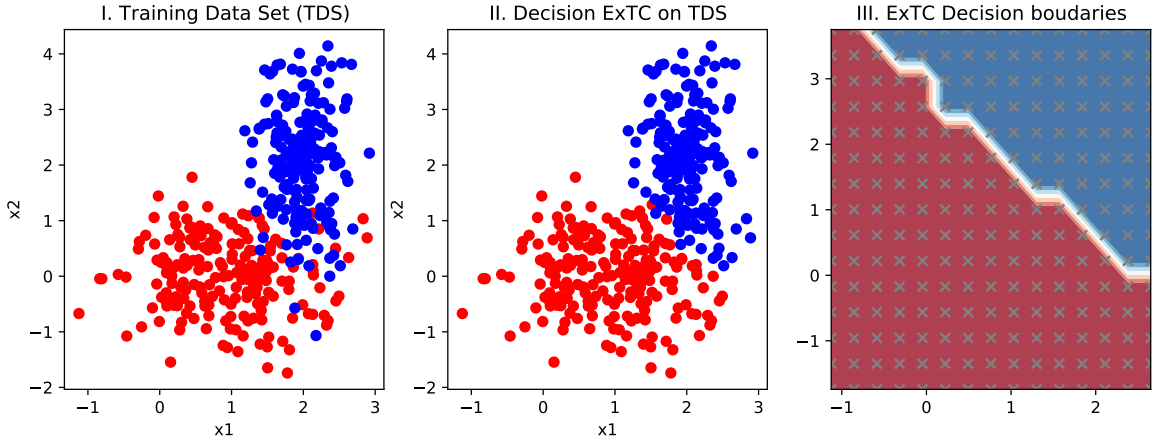

Figure 7: Extra Tree

EXERCICE TO DELIVER. Ice Classification in the Groenland

The aim of this study is to evaluate the ability of several learning machines for building a predictive classifier to evaluate, thanks to infrasound records, the (low or high) quantity of ice at a given spot in the Groenland.

This exercice should be conducted by a group of two students (no more). Each file should clearly mentionned the author names.

The requested work includes a report of 8 pages maximum presenting a synthesis of the studied models and the obtained results as well as the Python code file developed for this study.

Both files (report and code) will be submit on the ENSIIE Exam web site, directory **MAL2024TP** . The work should be download **before November, Tuesday 17th**.

Context

In this application, there are 4 potential target ouputs corresponding to infrasound signals (Y1, Y3, Y4, Y5) (file Targets.csv). The covariables are defined by 9 variables providing by the European Centre for Medium-Range Weather Forecasts (ECMWF) .

Description of the Output Variable

The displacement of large volumes of air leads to low-frequency acoustic waves in the atmosphere. This so-called infrasound is inaudible to humans as it has frequencies lower than 20 Hz. In the raw data, the 4 output variables are quantitative infra-sound records. For this classification study, **the quantitative valued target signals will be transformed into binary information using a approriate given threshold**.

Description of the Input Variables

- **climate information**: the European Weather Center provides information on 2 meter below sea temperature (t2m); Sea-surface temperature (SST); and wind speed (u10, v10)
- Sea Ice Concentration information (SIC)
- Groenland liquid water discharge simulated by Region Climate Models for 5 regions (r1_MAR, r2_MAR, r3_MAR, r4_MAR, r5_MAR)

More detailed information can be found in the Geophysical Research Letters, “Long-Term Infrasonic Monitoring of Land and Marine- Terminating Glaciers in Greenland”, Research letter, DOI 10.1029/2021GL097113, AGU advancing earth and space conference.

The data

The following instructions let to read the Input (data_features.csv) and output (data_Targets) data in Python.

```
import pandas as pd

#Load the data
tab = pd.read_csv('data_Features.csv')
tabY = pd.read_csv('data_Targets.csv')
```

Description of the work

- In this work, you have to study only one given target variable (Y1)
- Transform the target variable into a binary variable using an appropriate and motivated threshold.
- Train, test and compare different binary machine learning classifiers to predict the binary target variable.
- Conclusion. Draw conclusions about the performances and the models used.