Development project in Machine Learning

TAF MCE

Elsa Dupraz
elsa.dupraz@imt-atlantique.fr

November 9th, 2021

Objectives

Versioning with GIT

Project description

Deliverable

Objectives

Versioning with GIT

Project description

Deliverable

Objectives

- Develop good programming practices
- Use standard development tools
- Get used to collaborative work

- Work on Machine-Leaning datasets

Objectives

Versioning with GIT

Project description

Deliverable

Objectives

Versioning with GIT

Project description

Deliverable

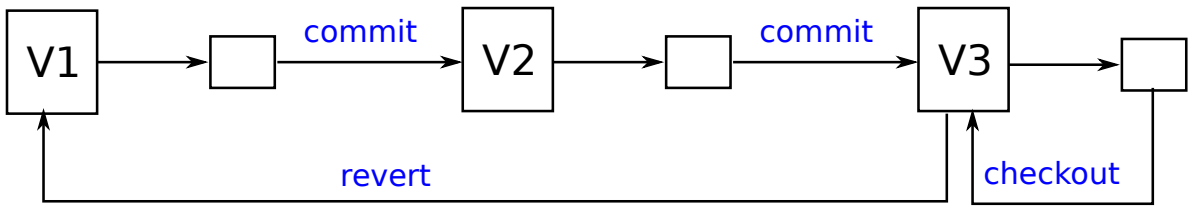
Versioning with GIT

- GIT is a versioning system
- It permits to keep track of the successive code versions
- It allows several persons to work on the same files, and can merge the various contributions
- It very efficiently deals with branches

Tutorial: `https://openclassrooms.com/fr/courses/1233741-gerez-vos-codes-source-avec-git`

How GIT works

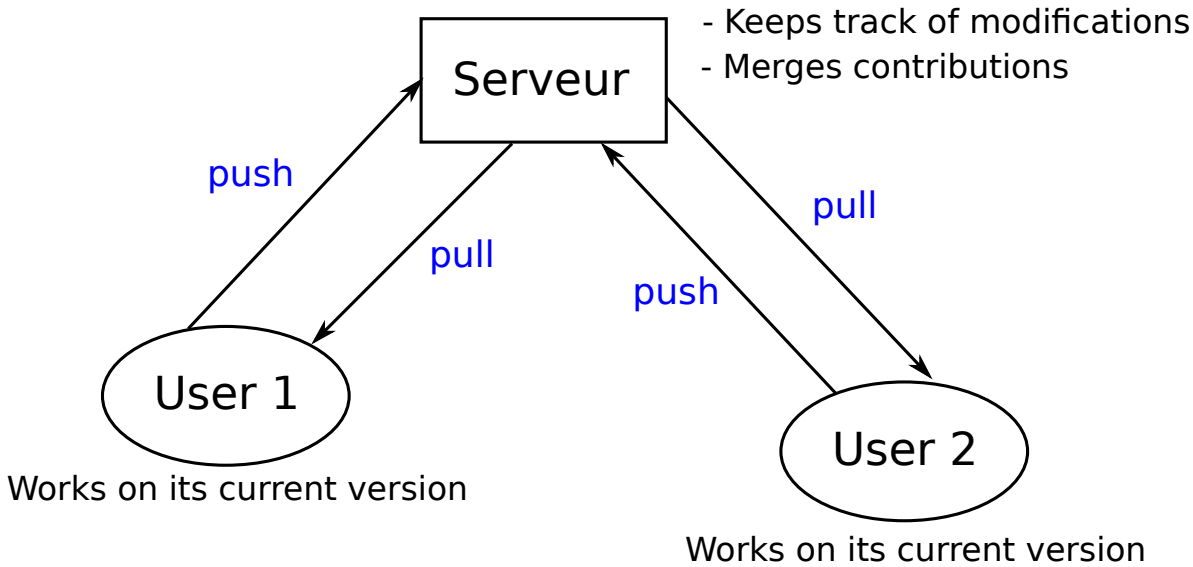
- Versioning:



Tutorial: `https://openclassrooms.com/fr/courses/1233741-gerez-vos-codes-source-avec-git`

How GIT works

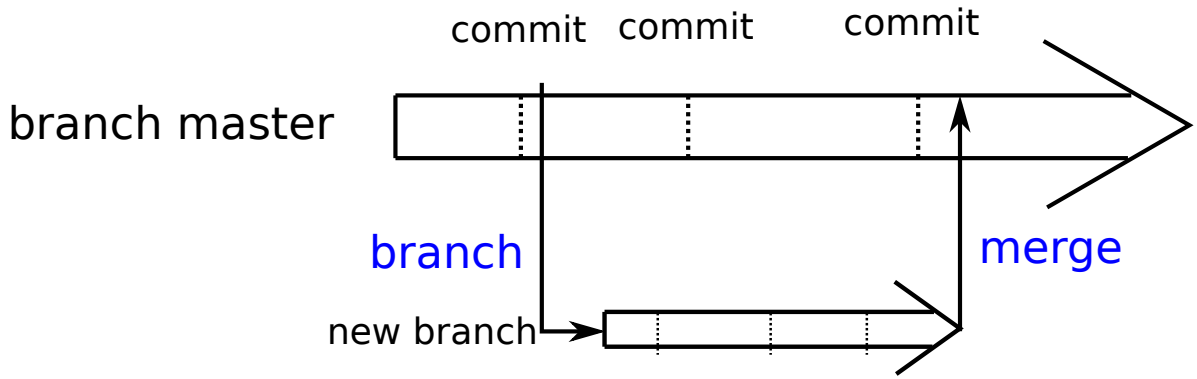
- Architecture:



Tutorial: `https://openclassrooms.com/fr/courses/1233741-gerez-vos-codes-source-avec-git`

How GIT works

- Branches:



Tutorial: `https://openclassrooms.com/fr/courses/1233741-gerez-vos-codes-source-avec-git`

Subjects

The objective of the project is to apply a Machine Learning model onto two different datasets:

- Option 1: Binary Classification
  - Banknote Authentication Dataset: `https://archive.ics.uci.edu/ml/datasets/banknote+authentication`
  - Chronic Kidney Disease: `https://www.kaggle.com/mansoordaku/ckdisease`
- Option 2: Linear Regression
  - Boston housing dataset: `https://www.kaggle.com/altavish/boston-housing-dataset`
  - Prostate cancer: `https://web.stanford.edu/~hastie/ElemStatLearn/data.html`

Constitute groups of 3 to 4 students and pick one of the two options.

Machine Learning Workflow

1. Import the dataset
2. Clean the data, perform pre-processing
   - Replace missing values by average or median values
   - Center and normalize the data
3. Split the dataset
   - Split between training set and test set
   - Split the training set for cross-validation
4. Train the model (including feature selection)
5. Validate the model

Objective: collaboratively implement this workflow and apply it to different ML problems/datasets

Indications (ML)

- You should first clean the dataset (handle missing values and categorical values)
- You may implement feature selection: bruteforce, by looking at correlations, from an ACP (for classification), by using Ridge regression (for linear regression), etc.
- Do not forget to save a part of your dataset as your test set. It will not be used for training, but only to assess the quality of your method.
- You may also use cross-validation to adjust the method (choice of the kernel, feature selection, etc.)
- You should automate your process as much as possible.

Indications (Development)

- Create a git repository for your group: `https://redmine-df.telecom-bretagne.eu/`
- Write the Python functions implementing the workflow in one single .py file.
- Apply the workflow onto the two datasets, using either a Python script or a notebook.
- Important: Your .py file containing the functions must be the same when applied to one or the other dataset
- Each student of the group should write at least one function. Indicate the writer of each function in comment.

Objectives

Versioning with GIT

Project description

Deliverable

Report

Report: 5 to 10 pages (without appendix). You may think of using GIT to write it :-)

- Present the project, the data, the methods you used, the main development steps.
- Show and comment your results
- Include one part describing what you think are good programming practices
- In appendix, provide the logs of your git repository and your code.

One report per group should be sent by e-mail to `elsa.dupraz@imt-atlantique.fr` before the 15th of December, 8PM.

Final comments

- Advice for good programming practices:
  `https://mikecroucher.github.io/reproducible_ML/`
- Register your groups before the 12th of November, at
  `https://semestriel.framapad.org/p/ml_project-9qqx`