ML 101: Practical exercises

Pierre-François Gimenez
CentraleSupélec/Inria

Hands-on Machine Learning for Security
September 24, 2021

Exercises structure

Exercises structure

- There are three parts that focus on different parts of the ML process
- We are not at school with linear, mandatory questions. Each exercise is an opportunity to experiment. Feel free to dig deeper or skip some parts.
- The link to the archive is in https://team.inria.fr/cidre/hands-on-machine-learning-for-security/
- We will use the Orange software. No programming required.

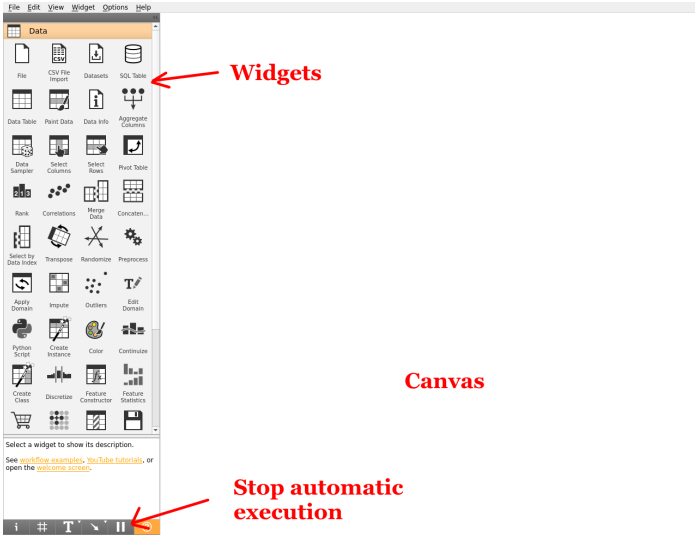
Yufei and I are available if you need some advice or help!

Orange overview

Orange

- Orange is a free and open source data mining software
- Download it at https://orangedatamining.com/download/
- It uses a graphical workflow to define the machine learning pipeline
- Add-ons for network analysis, image analysis, text mining, time series, etc.
- By default, operations are done on the fly, as the workflow is created. You may want to disable it on older hardware!

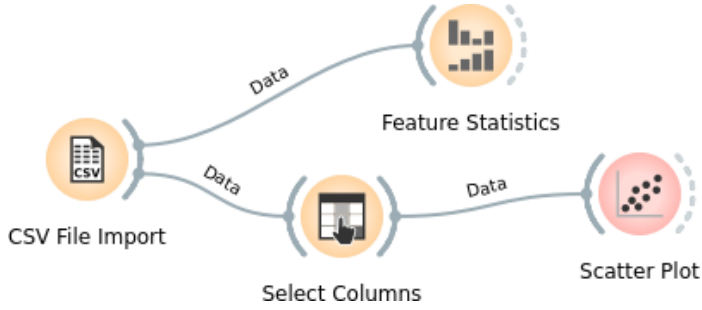
Orange overview



Orange overview

The workflow

- Nodes are computation units that can load file, learn a model, display it, etc.
- Click and drag a widget to add a node!
- Edges control the data flow. Double-clicking on them gives more options
- The input channels are on the left of the nodes, the output channels are on the right



Orange overview

Widget categories

- Data: to load, preprocess and save data
- Visualize: many plot possibilities
- Model: various ML models
- Evaluate: to produce ROC curve, confusion matrix, etc.
- Unsupervised: clustering and dimension reduction

Keep the documentation open, it will be useful:
https://orangedatamining.com/widget-catalog/

The dataset

KDD Cup 1999

- A classical dataset used for a competition within KDD-99
- The goal of this dataset is to evaluate anomaly detection in network data
- Features description: http://kdd.ics.uci.edu/databases/kddcup99/task.html
- Each line is a network connection that can be normal or bad (with four attack categories)
- We will only work with a random subset to speed up the computation. It is included in this archive.
- **This dataset is outdated and criticized, avoid it in research**

CentraleSupélec

The parts

Part 1: data mining and visualization

- Explore the data and try to make sense of them
- Select the features (unsupervised and supervised)

Part 2: model selection

- Experiments with multiple models at once
- Evaluate them with various metrics
- Tweak their parameters to get the best performances

Part 3: understanding the errors

- Visualize the errors of a model
- Try to make sense of them!

CentraleSupélec

Part 1: data mining and visualization

Data mining and preprocessing

Get to know the dataset

- Use the file `kddcup_unmerged.csv`. There are 19 attack types.
- Transfer the data from the CSV file to a Data Table
- The "Distribution" widget can also be useful
- Check the features and their meaning (cf. slide 7)
- Can you identify problems in the dataset?

Data preprocess

- The "Preprocess" widget allow some preprocessing steps
- You can transform features to continuous or categorical if necessary
- It includes some feature selection as well

CentraleSupélec

Feature selection

Unsupervised feature selection

- Check the correlations of the features. Highly correlated features can generally be removed as they are redundant
- For continuous features, you can experiment with the PCA (principal component analysis). PCA will create new, uncorrelated features that captures the information (called "variance" here). But new features may be difficult to intuitively grasp
- For discrete features, the equivalent is the Correspondence Analysis
- You can manually remove features with the "Select Columns" widget

Supervised feature selection

- With "Select Columns", inform that "label" is the target feature
- Check the information gain relatively to the label with "Rank"
- Compare your conclusions with the unsupervised feature selection

CentraleSupélec

Part 2: model selection

CentraleSupélec

Data preparation and models identification

Exercise

- Use the dataset `kddcup_merged.csv`. Attacks of similar families have been merged:
  - DOS: denial-of-service, e.g. syn flood;
  - R2L: unauthorized access from a remote machine, e.g. guessing password;
  - U2R: unauthorized access to local superuser (root) privileges, e.g., “buffer overflow” attacks;
  - probing: surveillance and other probing, e.g., port scanning.
- Identify the learning setting (regression, one-class learning, etc.)
- You can specify the target feature with "Select Columns"
- Find the models proposed by Orange compatible with this learning settings (cf. slide 7)
- Most of the time, Orange will automatically perform adapted preprocessing (continuization, normalization, etc.)

CentraleSupélec

Model training and evaluation

Exercise

- Experiments with various models, starting with the most simple
- Try ensemble techniques as well: tree boosting, AdaBoost and Stacking
- Check the slides of ML 101 for some reminders!
- Compare them with a ROC Analysis (target: "normal")
- Compare their Confusion Matrix
- Take this opportunity to look into explainable models with visualization: CN2 rules and decision tree

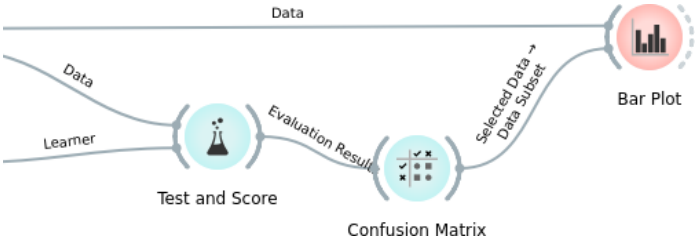
CentraleSupélec

Part 3: understanding the errors

Error extraction

How to extract errors

- Use the dataset `kddcup_merged.csv`.
- Choose a model easy to interpret, like a decision tree or rules induction
- Learn, test and score a model, and use the confusion matrix to identify its errors
- You can use the output of the confusion matrix to plot some errors



Error interpretation and how to correct it

Error analysis

- First find some predictions errors with the confusion matrix
- Check with the model why these are erroneous
- That's not an easy task!

Conclusion

What is your conclusion ? For example...

- Some important features are missing
- The dataset is too noisy (there is a difference between the measures and the reality)
- We should collect more data on this edge case
- We should experiment with a more complex model
- There errors are not relevant (they are outliers)