Machine Learning and Data Mining

scikit-learn Session Lab 3

Albert Bifet
albert.bifet@telecom-paristech.fr

October 11, 2016

scikit-learn

- scikit-learn is the leading machine learning software in Python
- scikit-learn is a project started in Paris, Telecom ParisTech and Inria
- scilkit-learn is easy to use and extend

scikit-learn Session Lab

- In this lab, we are going to create a classifier to use in scikit-learn
- Classifiers in scikit-learn has two main methods:
  - Build a model: `fit(self, X, Y)`
  - Make a prediction: `predict(self, X)`
- Classifiers are built using this template.

```
class NewClassifier:

    def __init__(self):
        # TODO

    def fit(self, X, Y):
        # TODO
        return self

    def predict(self, X):
        # TODO
        return Y
```

scikit-learn Session Lab Assignment

Write a jupyter notebook with the following tasks:

1. Write a majority class classifier: a classifier that predicts the class label that is more frequent in the dataset
2. Use the majority class classifier to evaluate one dataset, and justify why the evaluation results using the new classifier are correct

3. OPTIONAL: create another classifier with higher performance than the majority class classifier

4. NEXT LAB: Kaggle competition
   - Register at Kaggle and read:
     `https://www.kaggle.com/c/titanic`
   - Our competition:
     `https://inclass.kaggle.com/c/marketing-dataset`