Data science for lazy people, Automated Machine Learning

Big Data Congress Lithuania 2018

Diego Hueltes
@jdiegoh
https://www.sli.do
#bigdata2018
Data cleaning

STILL NOT CLEAN.
STILL NOT CLEAN!
# Data cleaning

<table>
<thead>
<tr>
<th>Sepal Length</th>
<th>Sepal Width</th>
<th>Petal Length</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.9</td>
<td>3.1</td>
<td>1.5</td>
<td>NA</td>
</tr>
<tr>
<td>5.4</td>
<td>3.7</td>
<td>1.5</td>
<td>NA</td>
</tr>
<tr>
<td>4.8</td>
<td>3.4</td>
<td>1.6</td>
<td>NA</td>
</tr>
<tr>
<td>4.8</td>
<td>3.0</td>
<td>1.4</td>
<td>0.1</td>
</tr>
<tr>
<td>5.7</td>
<td>3.0</td>
<td>1.1</td>
<td>0.1</td>
</tr>
<tr>
<td>5.8</td>
<td>4.0</td>
<td>1.2</td>
<td>0.2</td>
</tr>
<tr>
<td>5.7</td>
<td>4.4</td>
<td>1.5</td>
<td>0.4</td>
</tr>
<tr>
<td>5.4</td>
<td>3.9</td>
<td>1.3</td>
<td>0.4</td>
</tr>
<tr>
<td>5.1</td>
<td>3.5</td>
<td>1.4</td>
<td>0.3</td>
</tr>
<tr>
<td>5.7</td>
<td>3.8</td>
<td>1.7</td>
<td>0.3</td>
</tr>
<tr>
<td>5.1</td>
<td>3.8</td>
<td>1.5</td>
<td>0.3</td>
</tr>
</tbody>
</table>
Data cleaning
Raw data → Data Cleaning → Feature Preprocessing → Feature Selection → Model Selection → Parameter Optimization → Model Validation → Feature Construction
Feature selection
Feature selection
Raw data → Data Cleaning → Feature Preprocessing → Feature Selection → Model Selection → Model Validation

- Feature Construction → Feature Selection
Feature preprocessing

Alcohol and Malic Acid content of the wine dataset

- input scale
- Standardized $\mathcal{N}(\mu=0, \sigma=1)$
- min-max scaled $[\min=0, \max=1]$
Feature construction
Raw data -> Data Cleaning -> Feature Preprocessing -> Feature Selection

Feature Selection -> Model Selection

Model Selection -> Parameter Optimization

Parameter Optimization -> Model Validation
Model selection
Parameter optimization

RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                   max_depth=2, max_features='auto', max_leaf_nodes=None,
                   min_impurity_decrease=0.0, min_impurity_split=None,
                   min_samples_leaf=1, min_samples_split=2,
                   min_weight_fraction_leaf=0.0, n_estimators=10, n_jobs=1,
                   oob_score=False, random_state=0, verbose=0, warm_start=False)
Model validation
lazy

Oxford dictionary

Unwilling to work or use energy
lazy

Unwilling to work or use energy in repetitive tasks
Raw data → Data Cleaning → Feature Selection
Data Cleaning → Feature Preprocessing → Feature Construction
Feature Selection → Model Selection
Feature Preprocessing → Model Selection
Feature Construction → Parameter Optimization
Model Selection → Parameter Optimization
Parameter Optimization → Model Validation
Automated Machine Learning

TPOT  auto_ml
Auto-Sklearn  Google Cloud AutoML
Auto-Weka  AutoKeras
Machine-JS  H2O AutoML
DataRobot  ...

TPOT is a Python tool that automatically creates and optimizes machine learning pipelines using genetic programming.

https://github.com/EpistasisLab/tpot
auto-sklearn frees a machine learning user from algorithm selection and hyperparameter tuning. It leverages recent advantages in Bayesian optimization, meta-learning and ensemble construction.

https://github.com/automl/auto-sklearn
Genetic programming
Mutation

Crossover

Source: http://w3.onera.fr/smac/?q=tracker
Bayesian optimization
source: https://advancedoptimizationatharvard.wordpress.com/2014/04/28/bayesian-optimization-part-ii/
**TPOT** is a Python tool that automatically creates and optimizes *machine learning pipelines* using *genetic programming*.

https://github.com/EpistasisLab/tpot
from tpot import TPOTClassifier, TPOTRegressor

tpot = TPOTClassifier()
tpot.fit(X_train, y_train)
predictions = tpot.predict(X_test)

tpot = TPOTRegressor()
tpot.fit(X_train, y_train)
predictions = tpot.predict(X_test)
TPOT - configuration

```python
TPOTClassifier(config_dict = {
    'sklearn.ensemble.RandomForestClassifier' : {
        'n_estimators': [100],
        'criterion': ['gini', 'entropy'],
        'max_features': np.arange(0.05, 1.01, 0.05),
        'min_samples_split': range(2, 21),
        'min_samples_leaf': range(1, 21),
        'bootstrap': [True, False]
    },
    'sklearn.feature_selection.RFE' : {
        'step': np.arange(0.05, 1.01, 0.05),
        'estimator': {
            'sklearn.ensemble.ExtraTreesClassifier' : {
                'n_estimators': [100],
                'criterion': ['gini', 'entropy'],
                'max_features': np.arange(0.05, 1.01, 0.05)
            }
        }
    }
}),
```
**auto-sklearn**

**auto-sklearn** frees a machine learning user from algorithm selection and hyperparameter tuning. It leverages recent advantages in **Bayesian optimization**, **meta-learning** and **ensemble construction**.

https://github.com/automl/auto-sklearn
auto-sklearn

import autosklearn.classification
import autosklearn.regression

automl = autosklearn.classification.AutoSklearnClassifier()
automl.fit(X_train, y_train)
predictions = automl.predict(X_test)

automl = autosklearn.regression.AutoSklearnRegressor()
automl.fit(X_train, y_train)
predictions = automl.predict(X_test)
auto-sklearn custom config

include_estimators
exclude_estimators
include_preprocessors
exclude_preprocessors
Olive oil full dataset...

Test in Google Colab, clean dataset.

Tomorrow at 12:15

Hall 5
Olive oil full dataset...

Test in Google Colab, clean dataset.

**TPOT**: 55% Accuracy

**auto-sklearn**: 56% Accuracy

**H2O automl**: 51% Accuracy
Benchmarking Automatic Machine Learning Frameworks

Adithya Balaji, Alexander Allen

Automated Machine Learning—a Paradigm Shift That Accelerates Data Scientist Productivity @ Airbnb

https://medium.com/airbnb-engineering/automated-machine-learning-a-paradigm-shift-that-accelerates-data-scientist-productivity-airbnb-f1f8a10d61f8
Automated Machine Learning

- Exploratory analysis
- Selective discovering
- New ideas for your model
- Model optimization
Progress isn’t made by early risers. It’s made by lazy people trying to find easier ways to do something.

– Robert A. Heinlein
Thank you!

@jdiegoh
DiegoHueltes
Diego Hueltes
diego@hueltes.com

https://www.hueltes.com/automl