UNIVERSIDADE CATÓLICA PORTUGUESA

BRAGA

Machine Learning

Session 3 - PL

Data Scaling and Feature Selection

Ciência de Dados Aplicada

2023/2024

scikit-learn

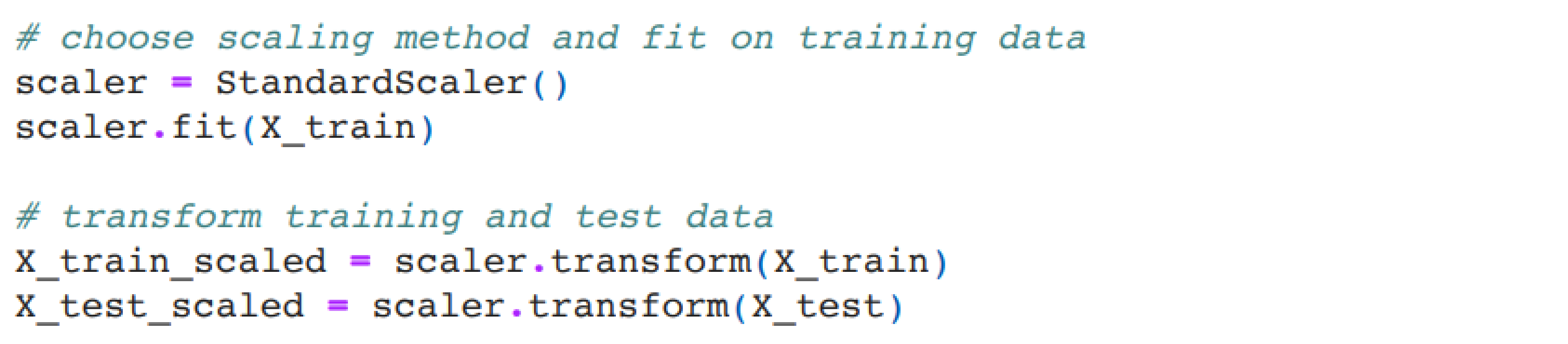
- **Scikit-learn** is a powerful library in Python for machine learning tasks, including data scaling and feature selection.

- Documentation: https://devdocs.io/scikit_learn/

- Tutorials: https://scikit-learn.org/stable/tutorial/index.html

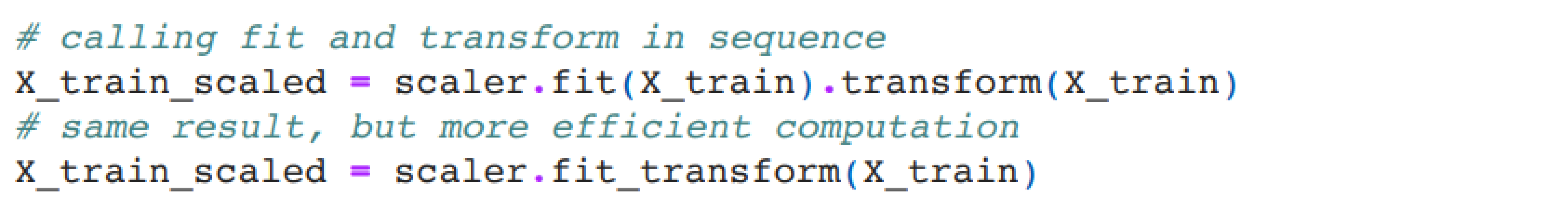
Data scaling in Python (scikit-learn)

- In scikit-learn all scalers follow the fit-transform methods:
  - "fit" prepares the scaler by learning from the data;
  - "transform" actually scales the data.



```
# choose scaling method and fit on training data
scaler = StandardScaler()
scaler.fit(X_train)

# transform training and test data
X_train_scaled = scaler.transform(X_train)
X_test_scaled = scaler.transform(X_test)
```



```
# calling fit and transform in sequence
X_train_scaled = scaler.fit(X_train).transform(X_train)
# same result, but more efficient computation
X_train_scaled = scaler.fit_transform(X_train)
```

Data scaling in Python (scikit-learn)

- StandardScaler:
  - https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html
- MinMaxScaler:
  - https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html
- Normalizer:
  - https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.Normalizer.html
- RobustScaler:
  - https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html
- Others:
  - https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html

Feature Selection in Python (scikit-learn)

- Feature selectors also follow the "fit" "transform" convention.

- Scikit-learn includes several feature selectors:
  - https://scikit-learn.org/stable/modules/feature_selection.html

Statistical tests in Python (scipy.stats)

- scipy.stats is a module within the SciPy library that provides a wide range of statistical functions and distributions for various statistical analyses.

- https://docs.scipy.org/doc/scipy/reference/stats.html

- Functions:
  - **T-test:** ttest_1samp, ttest_ind
  - **ANOVA**: f_oneway
  - **Non-parametric**: wilcoxon and kruskal
  - **Chi-square**: chisquare

Exercises:

- Notebooks on the github repository:
  - Notebook with examples:
    - notebooks/session3/examples.ipynb
  - Notebook with exercises:
    - notebooks/session3/exercises.ipynb