Transforms with scikit-learn

We have been using scikit-learn to this point, but we have not talked much about the package itself [1]. scikit-learn has a very well-designed API and a brief description here will help put the tools and vocabulary used in this course into perspective [2].

There are three interfaces meant to work together:

Transformer interface

- Used to convert data from one form to another

Estimator interface

- Used to build and fit models

Predictor interface

- Used to making predictions

NumPy arrays and SciPy sparse matrices are used as standardized input to these interfaces. The estimator interface provides a *.fit()* method for model training. All supervised and unsupervised algorithms use this interface. Feature extraction, feature selection and dimension reduction are also cases of the estimator interface. Some estimators also have a transform interface, like the StandardScaler.

```
from sklearn import preprocessing

scaler = preprocessing.StandardScaler().fit(X_train)
X_train =  scaler.transform(X_train)
X_test = scaler.transform(X_test)
```

We will refer to the different interfaces throughout the course with a focus in this unit on the transformer and estimator interfaces.

---

[1] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[2] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 108–122. 2013.