

Machine Learning and Data Mining

scikit-learn Session Lab

Albert Bifet
albert.bifet@telecom-paristech.fr



September 20, 2016

scikit-learn



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

scikit-learn Session Lab

- Install scikit-learn
 - Anaconda: <https://www.continuum.io/>
 - <http://scikit-learn.org/stable/install.html>
- Follow the scikit-learn Start Tutorial
 - <http://scikit-learn.org/stable/tutorial/basic/tutorial.html>

scikit-learn Session Lab

- Start the python Shell or jupyter
- Import classes

```
>>> import numpy as np
>>> from sklearn import datasets
```

- Load and parse the data file.

```
>>> iris = datasets.load_iris()
>>> iris_X = iris.data
>>> iris_y = iris.target
>>> np.unique(iris_y)
array([0, 1, 2])
```

- Split the data into training and test sets

```
>>> # Split iris data in train and test data
>>> # A random permutation, to split the data randomly
>>> np.random.seed(0)
>>> indices = np.random.permutation(len(iris_X))
>>> iris_X_train = iris_X[indices[:-10]]
>>> iris_y_train = iris_y[indices[:-10]]
>>> iris_X_test = iris_X[indices[-10:]]
>>> iris_y_test = iris_y[indices[-10:]]
```

scikit-learn Session Lab

- Train a k -nearest-neighbor model.

```
>>> # Create and fit a k-nearest-neighbor classifier
>>> from sklearn.neighbors import KNeighborsClassifier
>>> knn = KNeighborsClassifier()
>>> knn.fit(iris_X_train, iris_y_train)
KNeighborsClassifier(algorithm='auto', leaf_size=30,
                    metric='minkowski', metric_params=None,
                    n_jobs=1, n_neighbors=5, p=2,
                    weights='uniform')
```

- Evaluate model on test instances and compute test error

```
>>> from sklearn.metrics import accuracy_score
>>> knn.predict(iris_X_test)
array([1, 2, 1, 0, 0, 0, 2, 1, 2, 0])
>>> iris_y_test
array([1, 1, 1, 0, 0, 0, 2, 1, 2, 0])
>>> accuracy_score(iris_y_test, knn.predict(iris_X_test))
```

scikit-learn Session Lab Assignment

Write a jupyter notebook with the following tasks:

- 1 Write error of the classifier
- 2 What is the optimal parameter k of the k -nearest-neighbor classifier for this dataset?