ELEN0062 - Introduction to machine learning
Scikit-learn exercise

Jean-Michel Begon and Antonio Sutera

University of Liege

October 16, 2017

Statement

Run the following code:

```
from sklearn.datasets import load_digits
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

digits = load_digits()
X, y = digits.data, digits.target
```

1. Using numpy, compute the class proportions.
2. Divide the dataset into train/test sets. Use 70% of the samples as training set.
3. Instanciate a decision tree classifier.
4. Using the `fit` method, fit the decision tree on the training set.
5. Using the `predict` method, classify the test set.
6. Using the `accuracy_score` function, compute the accuracy.

Solution

```python
from sklearn.datasets import load_digits
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

digits = load_digits()
X, y = digits.data, digits.target

# Q1
labels = np.unique(y)
cls_freq = np.histogram(y, bins=len(labels))[0] / len(y)
print("Class \t Frequency")
for label, freq in zip(labels, cls_freq):
    print(label, "\t", "%.4f"%freq)
```

Solution

```
# Q2
X_ls, X_ts, y_ls, y_ts = train_test_split(X, y,
                                           train_size=.7)
# Q3—Q4
estimator = DecisionTreeClassifier().fit(X_ls, y_ls)

# Q5
y_pred = estimator.predict(X_ts)

# Q 6
print("Accuracy", accuracy_score(y_ts, y_pred))
```