CS342 Machine Learning: Lab #3
Quick Guide to scikit-learn for k-NN

Labs on Week 3 of Term 2

Week 17

Instructor:
Dr Theo Damoulas (T.Damoulas@warwick.ac.uk)

Tutors:
Helen McKay (H.McKay@warwick.ac.uk),
Joe Meagher (J.Meagher@warwick.ac.uk)
Karla Monterrubio-Gomez (K.Monterrubio-Gomez@warwick.ac.uk),
Jevgenij Gamper (J.Gamper@warwick.ac.uk)

KNeighborsClassifier

Before you start using *KNeighborsClassifier* function, you need to import it from *scikit-learn* package. *KNeighborsClassifier* is under *sklearn.neighbours.* You need to add the following line at the beginning of your python code:

```
from sklearn.neighbors import KNeighborsClassifier
```

In order to initiate the *KNeighborsClassifier*, you need to specify the k value (number of neighbors).

```
knn_model = KNeighborsClassifier(n_neighbors=3)
```

Fit the model using X as inputs and t as target values

```
knn_model = KNeighborsClassifier(n_neighbors=3)
knn_model.fit(X, t)
# OR just use one line
knn_model = KNeighborsClassifier(n_neighbors=3).fit(X, t)
```

Predict the class labels for the provided data X (training error)

```
p = knn_model.predict(X)
```

For more details on KNeighborsClassifier functions, see the Reference at `http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html#`.

Cross Validation: K-Fold

K-Fold divides all the N observations/samples in K equal-size groups called folds (for K = N, this is equivalent to Leave One Out Cross-validation). The K - 1 folds are used for training, and the left out fold is used for prediction/validation. Add the following line at the beginning of your python code:

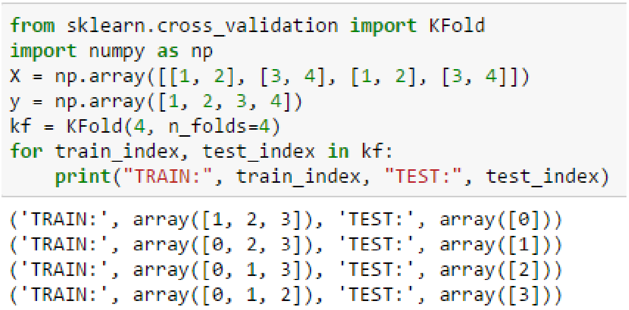
```
from sklearn.cross_validation import KFold
```

Provides train/test indices for every cross-validation iteration. N is the number of observations (or number of rows) and K is the number of folds.

```
folds = KFold(N, K=10)
```

This will return a list of train_index and test_index pairs (see Example below) for every iteration of cross-validation. You can use these to split your data each iteration of CV. For more details on KNeighborsClassifier functions, see the Reference at `http://scikit-learn.org/stable/modules/generated/sklearn.cross_validation.KFold.html#sklearn.cross_validation.KFold`.

Example



```
from sklearn.cross_validation import KFold
import numpy as np
X = np.array([[1, 2], [3, 4], [1, 2], [3, 4]])
y = np.array([1, 2, 3, 4])
kf = KFold(4, n_folds=4)
for train_index, test_index in kf:
    print("TRAIN:", train_index, "TEST:", test_index)
```

```
('TRAIN:', array([1, 2, 3]), 'TEST:', array([0]))
('TRAIN:', array([0, 2, 3]), 'TEST:', array([1]))
('TRAIN:', array([0, 1, 3]), 'TEST:', array([2]))
('TRAIN:', array([0, 1, 2]), 'TEST:', array([3]))
```