0112_use_pandas_for_training_test_split

December 22, 2018

1 Splitting data set into training and test sets using Pandas DataFrames methods

Note: this may also be performed using SciKit-Learn train_test_split method, but here we will use native Pandas methods.

1.1 Create a DataFrame

```
In [1]: # Create pandas data frame

        import pandas as pd

        name = ['Sam', 'Bill', 'Bob', 'Ian', 'Jo', 'Anne', 'Carl', 'Toni']
        age = [22, 34, 18, 34, 76, 54, 21, 8]
        gender = ['f', 'm', 'm', 'm', 'f', 'f', 'm', 'f']
        height = [1.64, 1.85, 1.70, 1.75, 1.63, 1.79, 1.70, 1.68]
        passed_physical = [0, 1, 1, 1, 0, 1, 1, 0]

        people = pd.DataFrame()
        people['name'] = name
        people['age'] = age
        people['gender'] = gender
        people['height'] = height
        people['passed'] = passed_physical

        print(people)
```

```
   name  age gender  height  passed
0   Sam   22      f    1.64       0
1  Bill   34      m    1.85       1
2   Bob   18      m    1.70       1
3   Ian   34      m    1.75       1
4    Jo   76      f    1.63       0
5  Anne   54      f    1.79       1
6  Carl   21      m    1.70       1
7  Toni    8      f    1.68       0
```

1.2 Split training and test sets

Here we take a random sample (25%) of rows and remove them from the original data by dropping index values.

```
In [2]: # Create a copy of the DataFrame to work from
        # Omit random state to have different random split each run

        people_copy = people.copy()
        train_set = people_copy.sample(frac=0.75, random_state=0)
        test_set = people_copy.drop(train_set.index)

        print ('Training set')
        print (train_set)
        print ('\nTest set')
        print (test_set)
        print ('\nOriginal DataFrame')
        print (people)
```

```
Training set
   name  age gender  height  passed
6  Carl   21      m    1.70       1
2   Bob   18      m    1.70       1
1  Bill   34      m    1.85       1
7  Toni    8      f    1.68       0
3   Ian   34      m    1.75       1
0   Sam   22      f    1.64       0

Test set
   name  age gender  height  passed
4    Jo   76      f    1.63       0
5  Anne   54      f    1.79       1

Original DataFrame
   name  age gender  height  passed
0   Sam   22      f    1.64       0
1  Bill   34      m    1.85       1
2   Bob   18      m    1.70       1
3   Ian   34      m    1.75       1
4    Jo   76      f    1.63       0
5  Anne   54      f    1.79       1
6  Carl   21      m    1.70       1
7  Toni    8      f    1.68       0
```