

Stack Overflow requires external JavaScript from another domain, which is blocked or failed to load.

Pandas dataset into an array for modelling in Scikit-Learn

Can we run scikit-learn models on Pandas DataFrames or do we need to convert DataFrames into NumPy arrays?

[python](#) [pandas](#) [scikit-learn](#)

edited Apr 6 '15 at 18:19



[Gyan Veda](#)

2,027 5 25 50

asked Mar 21 '14 at 14:31



[user40465](#)

146 4 12

migrated from stats.stackexchange.com Mar 21 '14 at 15:11

This question came from our site for people interested in statistics, machine learning, data analysis, data mining, and data visualization.

2 Answers

You can use `pandas.DataFrame` with `sklearn`, for example:

```
import pandas as pd
from sklearn.cluster import KMeans
```

Join Stack Overflow to learn, share knowledge, and build your career.

Email Sign Up

OR SIGN IN WITH

 Google

Facebook



Stack Overflow requires external JavaScript from another domain, which is blocked or failed to load.

```
(0.85, 33.3),  
(0.91, 31),  
(0.1, 15),  
(-0.23, 45)]
```

```
p_df = pd.DataFrame(data)  
kmeans = KMeans(init='k-means++', n_clusters=3, n_init=10)  
kmeans.fit(p_df)
```

Result:

```
>>> kmeans.labels_  
array([0, 0, 0, 2, 2, 2, 2, 0, 1], dtype=int32)
```

answered Mar 21 '14 at 14:44



Akavall

31.2k 20 107 150

1 It would be more informative to show how to deal with a DataFrame with named columns, since this is what causes more problems. A combination of this and the other answer would be great. – [lincolnfrías](#) Sep 25 '16 at 15:45

it works with named columns too – [rakesh](#) Apr 9 '17 at 21:57

Pandas DataFrames are very good at acting like Numpy arrays when they need to. If in doubt, you can always use the `values` attribute to get a Numpy representation (`df.values` will give you a Numpy array of the values in DataFrame `df`).

answered Mar 22 '14 at 22:00



Greg

1,876 12 17