Обзор библиотки Pandas

Щекалев Алексей Андреевич

ВМК МГУ

14.11.2016

Содержание

Раздел

Вывод данных

Когда мы сталкиваемся с задачами машинного обучения, нам нужно понять как устроены данные с которыми мы работаем.

Вывод в NumPy

```
print(data)
```

```
[[  0.00000000e+00   1.43019877e+09   7.00000000e+00 ...,   0.00000000e+00
    5.10000000e+01   0.00000000e+00]
 [  1.00000000e+00   1.43022034e+09   0.00000000e+00 ...,   0.00000000e+00
    6.30000000e+01   1.00000000e+00]
 [  2.00000000e+00   1.43022708e+09   7.00000000e+00 ...,   1.83000000e+03
    0.00000000e+00   6.30000000e+01]
 ...,
 [  1.14404000e+05   1.45029185e+09   1.00000000e+00 ...,   1.84600000e+03
    5.10000000e+01   6.30000000e+01]
 [  1.14405000e+05   1.45029299e+09   1.00000000e+00 ...,   2.04700000e+03
    6.30000000e+01   6.30000000e+01]
 [  1.14406000e+05   1.45031337e+09   7.00000000e+00 ...,   0.00000000e+00
    6.30000000e+01   0.00000000e+00]]
```

Рис.: Матрица объекты - признаки в NumPy

Вывод в Pandas

```
data.head()
```

	match_id	start_time	lobby_type	r1_hero	r1_level	r1_xp	r1_gold	r1_lh	r1_kills	r1_deaths	...
0	0	1430198770	7	11	5	2098	1489	20	0	0	...
1	1	1430220345	0	42	4	1188	1033	9	0	1	...
2	2	1430227081	7	33	4	1319	1270	22	0	0	...
3	3	1430263531	1	29	4	1779	1056	14	0	0	...
4	4	1430282290	7	13	4	1431	1090	8	1	0	...

5 rows × 147 columns

Рис.: Матрица объекты - признаки в Pandas

Раздел

Пример

Рассмотрим данные о моделях с обложки журнала Playboy с 1953 года по 2009

Пример

Рассмотрим данные о моделях с обложки журнала Playboy с 1953 года по 2009

```
import pandas as pd
girls = pd.read_excel('./data.xls')
```

Pandas умеет считывать фалы в форматах csv, xls, json и т.д.

Посмотрим как устроена выборка

```
girls.info()
```

```
RangeIndex: 604 entries, 0 to 603
Data columns (total 7 columns):
Месяц    604 non-null object
Год      604 non-null int64
S        604 non-null int64
T        604 non-null int64
B        604 non-null int64
L        604 non-null int64
W        604 non-null int64
```

Познакомимся с девушками поближе

```
girls.describe()
```

	Год	S	T	B	L	W
count	604.000000	604.000000	604.000000	604.000000	604.000000	604.000000
mean	1983.057947	89.293046	59.529801	87.942053	167.887417	52.168874
std	14.843740	3.994011	3.616909	3.479142	5.776711	4.040585
min	1953.000000	81.000000	46.000000	61.000000	150.000000	42.000000
25%	1970.000000	86.000000	58.000000	86.000000	165.000000	49.000000
50%	1983.000000	89.000000	61.000000	89.000000	168.000000	52.000000
75%	1996.000000	91.000000	61.000000	91.000000	173.000000	54.000000
max	2009.000000	104.000000	89.000000	99.000000	188.000000	68.000000

Проанализируем результат

В среднем

Средние показатели оказались вполне ожидаемыми: 89-60-88
Рост: 168
Вес: 52

Максимум/минимум

Заметим, что среди девушек присутствует модель с талией 89 см. Данный результат выглядит довольно подозрительно.

Проверим наши подозрения

Определим у какой модели талия 89 см.

```
girls[girls['T'] == 89]
```

	Месяц	Год	S	T	B	L	W
483	December	1998	86	89	86	173	52

Раздел

Плюсы библиотки Pandas

Плюсы

- Удобная загрузка данных при помощи функции `pd.read_`
- Удобный анализ обучающей выборки при помощи функции `pd.head()`
- Удобный доступ к информации о выборке при помощи фукции `pd.describe()`

Минусы

- Медленная скорость работы по сравнению с библиотекой NumPy