test_skll_reader

September 20, 2019

[1]:
```
import pandas as pd

from os.path import join
from skll import Reader, Writer
from skll import Reader2, Writer2
from sklearn.datasets import make_classification
```

[2]:
```
%load_ext memory_profiler
```

0.1 Code to generate data

```
def _to_delimited(x, y, path, sep):
    df = pd.DataFrame(x)
    df.columns = ['X{}'.format(i + 1) for i in df]
    df['y'] = y
    df.to_csv(path, sep=sep)

grid = [(1000, 20, 'test1'),
        (1000, 2000, 'test2'),
        (10000, 200, 'test3'),
        (100000, 200, 'test4')]

for n_samples, n_features, name in grid:
    X, y = make_classification(n_samples=n_samples,
                               n_features=n_features,
                               n_redundant=0,
                               n_repeated=0)
    _to_delimited(X, y, join('../test', name + '.csv'), sep=',')
    _to_delimited(X, y, join('../test', name + '.tsv'), sep='\t')
```

0.2 Test `Reader` objects

- `Reader`: New reader class
- `Reader2`: Old reader class

Data 1 - 1000 samples; 20 features

[3]:
```
%memit fs = Reader.for_path('../test/test1.csv').read()
%memit fs = Reader2.for_path('../test/test1.csv').read()
```

```
peak memory: 127.95 MiB, increment: 4.88 MiB
peak memory: 127.29 MiB, increment: -0.42 MiB
```

[4]:
```
%time fs = Reader.for_path('../test/test1.csv').read()
%time fs = Reader2.for_path('../test/test1.csv').read()
```

```
CPU times: user 50.6 ms, sys: 5.31 ms, total: 55.9 ms
Wall time: 53.7 ms
CPU times: user 187 ms, sys: 1.67 ms, total: 189 ms
Wall time: 189 ms
```

0.2.1 Data 2 - 1000 samples; 2000 features

[5]:
```
%memit fs = Reader.for_path('../test/test2.csv').read()
%memit fs = Reader2.for_path('../test/test2.csv').read()
```

```
peak memory: 392.92 MiB, increment: 262.68 MiB
peak memory: 289.30 MiB, increment: 110.73 MiB
```

[6]:
```
%time fs = Reader.for_path('../test/test2.csv').read()
%time fs = Reader2.for_path('../test/test2.csv').read()
```

```
CPU times: user 3.84 s, sys: 132 ms, total: 3.98 s
Wall time: 3.98 s
CPU times: user 16.8 s, sys: 97.8 ms, total: 16.9 s
Wall time: 16.9 s
```

0.2.2 Data 3 - 10000 samples; 200 features

[7]:
```
%memit fs = Reader.for_path('../test/test3.csv').read()
%memit fs = Reader2.for_path('../test/test3.csv').read()
```

```
peak memory: 407.56 MiB, increment: 202.13 MiB
peak memory: 274.65 MiB, increment: 52.79 MiB
```

[8]:
```
%time fs = Reader.for_path('../test/test3.csv').read()
%time fs = Reader2.for_path('../test/test3.csv').read()
```

```
CPU times: user 3.52 s, sys: 124 ms, total: 3.65 s
Wall time: 3.64 s
CPU times: user 16.7 s, sys: 104 ms, total: 16.8 s
Wall time: 16.8 s
```

0.2.3 Data 4 - 100000 samples; 200 features

[9]:
```
%memit fs = Reader.for_path('../test/test4.csv').read()
%memit fs = Reader2.for_path('../test/test4.csv').read()
```

```
peak memory: 2580.20 MiB, increment: 2322.33 MiB
peak memory: 1450.41 MiB, increment: 964.36 MiB
```

[10]:
```
%time fs = Reader.for_path('../test/test4.csv').read()
%time fs = Reader2.for_path('../test/test4.csv').read()
```

```
CPU times: user 35.7 s, sys: 1.29 s, total: 37 s
Wall time: 37.1 s
CPU times: user 2min 50s, sys: 1.44 s, total: 2min 51s
Wall time: 2min 52s
```

0.3 Test `Writer` objects

- `Writer`: New reader class
- `Writer2`: Old reader class

0.3.1 Data 1 - 1000 samples; 20 features

[12]:
```
fs = Reader.for_path('../test/test1.csv').read()
```

[13]:
```
%memit Writer.for_path('../test/_testout.csv', fs).write()
%memit Writer2.for_path('../test/_testout.csv', fs).write()
```

```
peak memory: 322.39 MiB, increment: 1.96 MiB
peak memory: 321.96 MiB, increment: 0.02 MiB
```

[14]:
```
%time Writer.for_path('../test/_testout.csv', fs).write()
%time Writer2.for_path('../test/_testout.csv', fs).write()
```

```
CPU times: user 39.5 ms, sys: 5.17 ms, total: 44.6 ms
Wall time: 43.1 ms
CPU times: user 614 ms, sys: 4 ms, total: 618 ms
Wall time: 618 ms
```

0.3.2 Data 2 - 1000 samples; 2000 features

[15]:
```
fs = Reader.for_path('../test/test2.csv').read()
```

[16]:
```
%memit Writer.for_path('../test/_testout.csv', fs).write()
%memit Writer2.for_path('../test/_testout.csv', fs).write()
```

```
peak memory: 319.60 MiB, increment: 51.96 MiB
peak memory: 301.10 MiB, increment: 0.10 MiB
```

[17]:
```
%time Writer.for_path('../test/_testout.csv', fs).write()
%time Writer2.for_path('../test/_testout.csv', fs).write()
```

```
CPU times: user 3.05 s, sys: 136 ms, total: 3.18 s
Wall time: 3.19 s
CPU times: user 39.9 s, sys: 62.1 ms, total: 40 s
Wall time: 40 s
```

0.3.3 Data 3 - 10000 samples; 200 features

[18]:
```
fs = Reader.for_path('../test/test3.csv').read()
```

[19]:
```
%memit Writer.for_path('../test/_testout.csv', fs).write()
%memit Writer2.for_path('../test/_testout.csv', fs).write()
```

```
peak memory: 346.43 MiB, increment: 45.32 MiB
peak memory: 334.27 MiB, increment: 0.00 MiB
```

[20]:
```
%time Writer.for_path('../test/_testout.csv', fs).write()
%time Writer2.for_path('../test/_testout.csv', fs).write()
```

```
CPU times: user 2.98 s, sys: 95.8 ms, total: 3.07 s
Wall time: 3.08 s
CPU times: user 41.3 s, sys: 216 ms, total: 41.6 s
Wall time: 41.6 s
```

0.3.4 Check TSVs

[21]:
```
%memit fs = Reader.for_path('../test/test2.tsv').read()
%memit fs = Reader2.for_path('../test/test2.tsv').read()
```

```
peak memory: 491.48 MiB, increment: 157.14 MiB
peak memory: 387.80 MiB, increment: 88.30 MiB
```

[22]:
```
%time fs = Reader.for_path('../test/test3.tsv').read()
%time fs = Reader2.for_path('../test/test3.tsv').read()
```

```
CPU times: user 3.5 s, sys: 112 ms, total: 3.62 s
Wall time: 3.61 s
CPU times: user 17.2 s, sys: 137 ms, total: 17.4 s
Wall time: 17.4 s
```

[23]:
```
%memit Writer.for_path('../test/_testout.csv', fs).write()
%memit Writer2.for_path('../test/_testout.csv', fs).write()
```

```
peak memory: 326.16 MiB, increment: 30.66 MiB
peak memory: 317.36 MiB, increment: 0.00 MiB
```

[24]:
```
%time Writer.for_path('../test/_testout.csv', fs).write()
%time Writer2.for_path('../test/_testout.csv', fs).write()
```

```
CPU times: user 3.03 s, sys: 97.3 ms, total: 3.12 s
Wall time: 3.13 s
CPU times: user 39.8 s, sys: 175 ms, total: 40 s
Wall time: 40 s
```