dask_example

January 31, 2020

[ ]:
```
import numpy as np
from sklearn import model_selection
import cudf as pd

max_depth = 8
n_trees = 100
n_trees_skl = 100
n_streams = 10

df = pd.read_csv('test.csv')
###df = pd.read_csv('test.csv').drop('Unnamed: 0', axis=1)
###X = df.drop(['C2'],1).astype(np.float32)
###y = df['C2'].astype(np.int32)

X_train, X_test, y_train, y_test = model_selection.train_test_split(X, y,
test_size=0.2)
```

[ ]:
```

```

[ ]:
```
from cuml.ensemble import RandomForestRegressor as cumlRFR

cuml_model = cumlRFR(max_depth=max_depth, n_estimators=n_trees,
                     ␣
 ↪n_streams=n_streams,n_bins=32,split_algo=0,split_criterion=2)

cuml_model.fit(X_train, y_train)
```

[ ]:
```
from sklearn.metrics import mean_squared_error
host_ytest = y_test.to_array()

cuml_y_pred_gpu = cuml_model.predict(X_test,predict_model='GPU')
print("cuML accuracy: ", mean_squared_error(host_ytest, cuml_y_pred_gpu))
```

[ ]:
```
from cuml.dask.ensemble import RandomForestRegressor as daskRFR
from cuml.dask.common import utils as dask_utils
from dask.distributed import Client, wait
from dask_cuda import LocalCUDACluster
```

```
import dask_cudf

n_partitions = 1

cluster = LocalCUDACluster(threads_per_worker=1, n_workers=n_partitions)
c = Client(cluster)
workers = c.has_what().keys()



# Shard the data across all workers
X_train_dask = dask_cudf.from_cudf(X_train, npartitions=n_partitions)
y_train_dask = dask_cudf.from_cudf(y_train, npartitions=n_partitions)
X_train_dask, y_train_dask = dask_utils.persist_across_workers(c,␣
 ↪[X_train_dask, y_train_dask], workers=workers)

# Build and train the model
dask_model = daskRFR(max_depth=max_depth, n_estimators=n_trees,
                     ␣
 ↪n_streams=n_streams,n_bins=16,split_algo=0,split_criterion=2)

dask_model.fit(X_train_dask, y_train_dask)
cuml_y_pred_dask = dask_model.predict(X_test.as_matrix()) #X_test is a cudf␣
 ↪dataframe
print("cuml dask accuracy:", mean_squared_error(host_ytest,cuml_y_pred_dask))
```

[ ]:

[ ]: