CatBoost Starter

CatBoost?

```
from catboost import CatBoostRegressor
# Initialize data
cat_features = [0,1,2]
train_data = [["a","b",1,4,5,6],["a","b",4,5,6,7],["c","d",30,40,50,60]]
test_data = [["a","b",2,4,6,8],["a","d",1,4,50,60]]
train_labels = [10,20,30]
# Initialize CatBoostRegressor
model = CatBoostRegressor(iterations=2, learning_rate=1, depth=2)
# Fit model
model.fit(train_data, train_labels, cat_features)
```

Xgboost 에서 카테고리변수를 사용하려면 One-hot Encoding등 인코딩을 거쳐야하는데
Catboost는 그러한 작업없이 변수의 위치만 설정해주면 문자열변수를 사용할 수 있다.
Lightgbm에도 비슷한 기능이 있지만 문자열 변수로는 받지 않는다.

https://towardsdatascience.com/catboost-vs-light-gbm-vs-xgboost-5f93620723db

Simple preprocessing

Train

	ip	app	device	os	channel	click_time	attributed_time	is_attributed	time_month	time_day	time_hr	time_min	time_sec
0	11846	12	1	13	259	2017-11-09 08:17:26	NaN	0	11	09	08	17	26
1	5147	19	0	0	347	2017-11-09 08:17:26	NaN	0	11	09	08	17	26
2	11782	9	1	8	127	2017-11-09 08:17:26	NaN	0	11	09	08	17	26
3	33867	35	1	19	21	2017-11-09 08:17:26	2017-11-09 09:05:37	1	11	09	08	17	26
4	110589	3	1	23	280	2017-11-09 08:17:26	NaN	0	11	09	08	17	26

Test

	click_id	ip	app	device	os	channel	click_time	time_month	time_day	time_hr	time_min	time_sec
0	0	5744	9	1	3	107	2017-11-10 04:00:00	11	10	04	00	00
1	1	119901	9	1	3	466	2017-11-10 04:00:00	11	10	04	00	00
2	2	72287	21	1	19	128	2017-11-10 04:00:00	11	10	04	00	00
3	3	78477	15	1	13	111	2017-11-10 04:00:00	11	10	04	00	00
4	4	123080	12	1	13	328	2017-11-10 04:00:00	11	10	04	00	00

Simple preprocessing

Train

	ip	app	device	os	channel	time_day	time_hr
0	11846	12	1	13	259	09	08
1	5147	19	0	0	347	09	08

Test

	ip	app	device	os	channel	time_day	time_hr
0	5744	9	1	3	107	10	04
1	119901	9	1	3	466	10	04

Format of Input Data

```
for col in train.columns:
    train[col] = train[col].astype('category')
for col in test.columns:
    test[col] =test[col].astype('category')
categorical_features_indices = np.where(train.dtypes != np.float32)[0]

train_pool = Pool(train,y , cat_features=categorical_features_indices)
test_pool = Pool(test, cat_features=categorical_features_indices)
```

입력시 CatBoost에 있는 Pool 함수를 이용해야한다.

Category 변수는 Pandas 에서 category로 설정해준 다음 파라미터를 이용해 위치를 설정해준다.

CatBoostRegressor vs CatBoostClassifier

CatBoostClassifier

```
model = CatBoostRegressor(learning_rate=0.037, depth=5, loss_function='Logloss')
model.fit(train_pool)
preds_class  = model.predict(test_pool)
sub['is_attributed'] = preds_class
sub.to_csv('catcat_sub0317.csv',index=False)
```

catcat_sub0317.csv 0.6295

11 days ago by SeoBeomSeok

add submission details

CatBoostRegressor

```
model = CatBoostRegressor(rsm=0.8, depth=5, learning_rate=0.037, eval_metric='MAE')
model.fit(train_pool)
cat_preds = model.predict(test_pool)
sub["is_attributed"] = cat_preds
sub.loc[sub['is_attributed'] < 0, 'is_attributed'] = 0
sub.loc[sub['is_attributed'] > 1, 'is_attributed'] = 1
```

catreg_sub0322.csv 0.8961

8 days ago by SeoBeomSeok

add submission details

변수=카테고리, CPU환경 ----->> 1번 학습시 3시간~4시간

CatBoostRegressor

변수=수치, CPU환경 ----->> 1번 학습시 1시간~1시간반

변수=수치, GPU환경 ----->> 1번 학습시 3분(but 파라미터 제한)