PGConf.СПб 2023

PostgresPro

ML in PostgeSQL

Александр Календарёв

PostgresPro

Apache MADlib

https://github.com/apache/madlib

madlib/dpage@PostgreSQL 12

Query Editor | Query History

```
1  -- Drop tables from previous runs
2  DROP TABLE IF EXISTS housing_linregr, housing_linregr_summary;
3
4  -- Train the model
5  SELECT madlib.linregr_train( 'housing',
6                               'housing_linregr',
7                               'medv',
8                               'ARRAY[crim, zn, indus, chas, nox, rm, age, dis, rad, tax, ptratio, b, lstat]'
9                             );
10 -- Predict the house prices, and compare to the actual data
11 SELECT housing.*,
12     predict,
13     medv - predict AS residual,
14     sqrt(avg(power(abs(medv - predict), 2)) OVER ()) AS rmse
15 FROM housing,
16     housing_linregr,
17     madlib.linregr_predict(coef,
18                            ARRAY[crim, zn, indus, chas, nox, rm, age, dis, rad, tax, ptratio, b, lstat]
19                           ) predict;
```

Data Output | Explain | Notifications | Messages

tax double precision	ptratio double precision	b double precision	lstat double precision	medv double precision	predict double precision	residual double precision	rmse double precision
296	15.3	396.9	4.98	24	29.098263530097405	-5.098263530097405	4.915902697381886
242	17.8	396.9	9.14	21.6	24.502275482309766	-2.902275482309765	4.915902697381886
242	17.8	392.83	4.03	34.7	31.227426410729453	3.4725735892705494	4.915902697381886
222	18.7	394.63	2.94	33.4	29.70710350127486	3.692896498725137	4.915902697381886
222	18.7	396.9	5.33	36.2	29.56479571582512	6.635204284174883	4.915902697381886
222	18.7	394.12	5.21	28.7	25.29376223617479	3.40623776382521	4.915902697381886
311	15.2	395.6	12.43	22.9	21.530411609932766	1.369588390067232	4.915902697381886
311	15.2	396.9	19.15	27.1	19.10333425509494	7.996665744905062	4.915902697381886

XGBoost TensorFlow Keras

под капотом
PL/Python C/C++

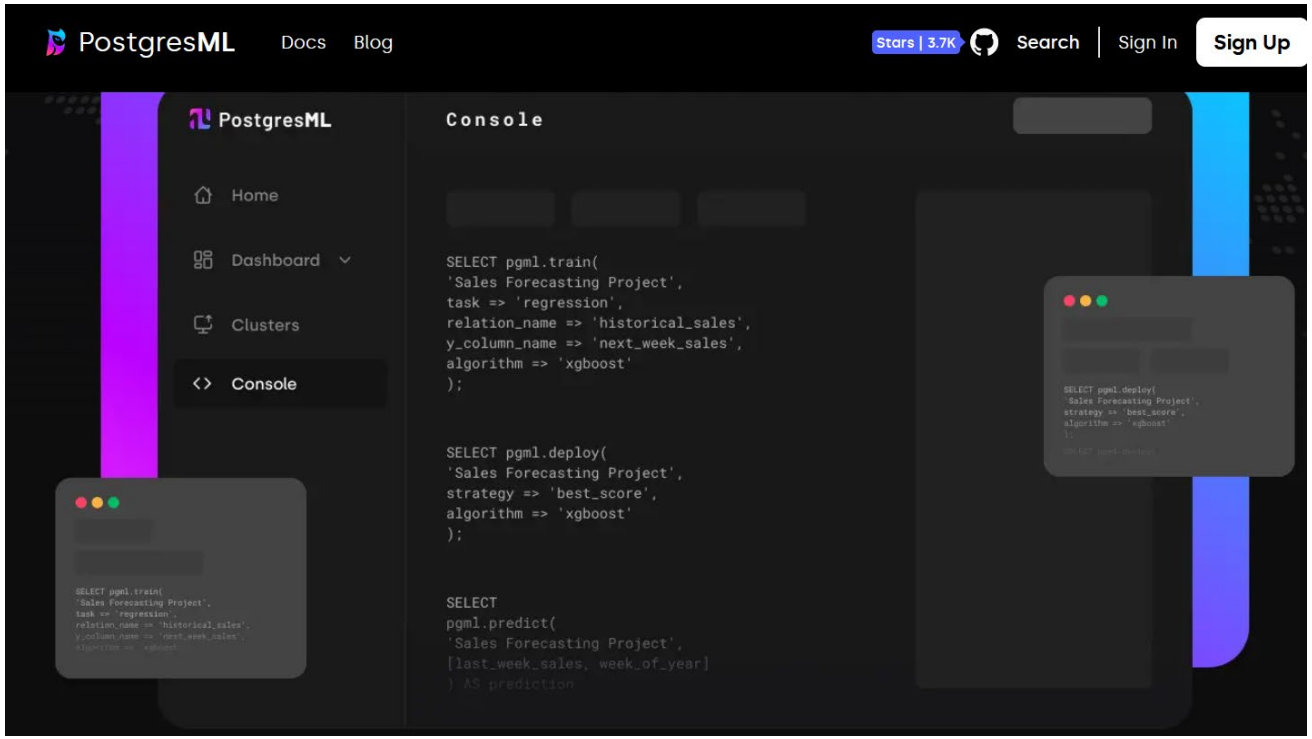
25 алгоритмов

коду 9-10 лет
В коммитов в 2023 г – около 10

PostgresPro

Postrgres ML

https://github.com/postgresml/postgresml



XGBoost

Yandex CatBoost

scikit learn machine learning in Python

PL/Rust под капотом

Порт в BERT GPT (NLP)

31 алгоритм

Проект с 2022г - живой

PostgresPro

https://github.com/akalend/pg_ml

Yandex CatBoost

Работает только с обученными моделями (libcatboostmodel)
Модели динамически загружаются
Получаем таблицу с которой можно работать

```
ml=# select ml_predict('/usr/local/pgsql/example/titanic.cbm', 'titanic2');
WARNING:  ALTER TABLE IF EXISTS titanic2_predict SET SCHEMA public; res=4
NOTICE:  processed=418
       ml_predict
-------------------------
 public.titanic2_predict
(1 row)

ml=# select * from titanic2_predict limit 5;
 row | passenger_id | pclass |                    name                    |  sex   | age  | sibsp | parch | ticket |  fare   | cabin | embarked | predict
-----+--------------+--------+--------------------------------------------+--------+------+-------+-------+--------+---------+-------+----------+----------
   1 |          892 |      3 | Kelly, Mr. James                           | male   | 34.5 |     0 |     0 | 330911 |  7.8292 | -999  | Q        | 0.142616
   2 |          893 |      3 | Wilkes, Mrs. James (Ellen Needs)           | female |   47 |     1 |     0 | 363272 |       7 | -999  | S        | 0.310916
   3 |          894 |      2 | Myles, Mr. Thomas Francis                  | male   |   62 |     0 |     0 | 240276 |  9.6875 | -999  | Q        | 0.083718
   4 |          895 |      3 | Wirz, Mr. Albert                           | male   |   27 |     0 |     0 | 315154 |  8.6625 | -999  | S        | 0.125321
   5 |          896 |      3 | Hirvonen, Mrs. Alexander (Helga E Lindqvist) | female |   22 |     1 |     1 | 3101298 | 12.2875 | -999  | S        |  0.39712
(5 rows)
```

PostgresPro

Типы входных данных:

Категориальные (текст)
catecorial feature

Числовые
feature

Временные ряды
time-series feature

PostgresPro

Типы входных данных

postgres@notebook-sasha: /usr/local/pgsql

```
adult=# \d titanic
                  Table "public.titanic"
    Column     |       Type       | Collation | Nullable | Default
---------------+------------------+-----------+----------+---------
 id            | integer          |           |          |
 passenger_id  | integer          |           |          |
 pclass        | integer          |           |          |
 name          | text             |           |          |
 sex           | text             |           |          |
 age           | double precision |           |          |
 sibsp         | integer          |           |          |
 parch         | integer          |           |          |
 ticket        | text             |           |          |
 fare          | double precision |           |          |
 cabin         | text             |           |          |
 embarked      | character(1)     |           |          |
 res           | boolean          |           |          |

adult=# select * from titanic limit 10;
```

Особенность libcatboostmodel:

```
ml=# SELECT ml_info('/usr/local/pgsql/example/titanic.cbm');
                        ml_info
-----------------------------------------------------
 dimension:1 numeric features:2 cagorial features:9
(1 row)
```

PostgresPro

Типы входных данных

Особенность libcatboostmodel:

Integer, bool – категориальные feature

```
ml=# SELECT ml_info('/usr/local/pgsql/example/titanic.cbm');
                        ml_info
-----------------------------------------------------
 dimension:1 numeric features:2 cagorial features:9
(1 row)
```

PostgreSQL

PostgresPro

Yandex CatBoost

Ограничения

Работает только с обученными моделями (libcatboostmodel)

Поля должны быть в том же порядке, как и в исходном датасете

catboost-for-titanic-top-7.ipynb ☆

Файл Изменить Вид Вставка Среда выполнения Инструменты Справка Последнее изменение: 7 сентября

+ Код + Текст

[ ]

res

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Sur
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	-999	Q	
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	-999	S	
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	-999	Q	
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	-999	S	
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	-999	S	
...	...	...	...	...	...	...	...	...	...	...	...	
413	1305	3	Spector, Mr. Woolf	male	-999.0	0	0	A.5. 3236	8.0500	-999	S	
414	1306	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758	108.9000	C105	C	
415	1307	3	Saether, Mr. Simon Sivertsen	male	38.5	0	0	SOTON/O.Q. 3101262	7.2500	-999	S	
416	1308	3	Ware, Mr. Frederick	male	-999.0	0	0	359309	8.0500	-999	S	
417	1309	3	Peter, Master. Michael J	male	-999.0	1	1	2668	22.3583	-999	C	

418 rows × 12 columns

```
postgres@notebook-sasha: /usr/local/pgsql
adult=# \d titanic
                    Table "public.titanic"
    Column    |       Type       | Collation | Nullable | Default
--------------+------------------+-----------+----------+---------
 id           | integer          |           |          |
 passenger_id | integer          |           |          |
 pclass       | integer          |           |          |
 name         | text             |           |          |
 sex          | text             |           |          |
 age          | double precision |           |          |
 sibsp        | integer          |           |          |
 parch        | integer          |           |          |
 ticket       | text             |           |          |
 fare         | double precision |           |          |
 cabin        | text             |           |          |
 embarked     | character(1)     |           |          |
 res          | boolean          |           |          |

adult=# select * from titanic limit 10;
```

Планы

Сделать полноценный SQL интерфейс

Оттестировать все виды моделей

Добавить параметры настройки, функции оценки

Заменить на XGBoost, возможность обучать модель

XGBoost

PostgresPro

Как это должно работать на самом деле?

CREATE MODEL <name> <options> <QUERY>

- Установить обработчик на QueryParser (Optimizer)

Синтактический анализ запроса:

Есть лексема MODEL?

- Нет: стандартная обработка
- Да:
  - Создание объекта модели
  - Сохранение метаданных (tbl pg_ml_model)
  - Выполнение запроса
  - Тренировка и сохранение модели Model.Fit(Dataset) в bin формате

Как это должно работать на самом деле?

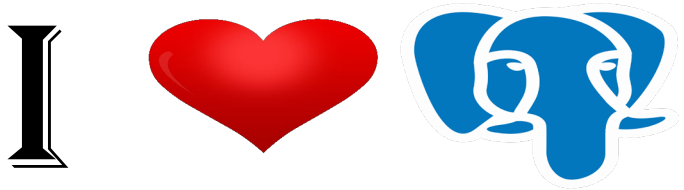
PREDICT <name> <QUERY>

- Установить обработчик на QueryParser (Optimizer)
  Проверить есть лексема PREDICT?
  - Нет: стандартная обработка
  - Да:
    - Поиск модели в метаданных (tbl pg_ml_model)
    - Загрузка объекта модели
    - Выполнение запроса
    - Применение Model.Predict(Row) к каждой строке
    - Формирование выходного набора данных

PostgresPro

Приглашаются желающие

к сотрудничеству



PostgresPro

Давайте делать PostgreSQL лучше