Vertica ML Python Workshop

Exercise 6: Duplicates

Ouali Badr

December 5, 2019

Executive Summary

> *"Science knows no country, because knowledge belongs to humanity, and is the torch which illuminates the world."*
>
> **Louis Pasteur**

VERTICA ML PYTHON allows the users to use Vertica advanced analytics and Machine Learning with a Python front-end Interface. In this exercise, you'll learn some basics to begin your fantastic Data Science Journey with the API. As a summary:

- Find duplicates
- Drop the duplicates

Contents

1 Presentation

Finding duplicates is one of the task in Data Preparation. They are adding bias in the data and many ML algorithms will consider duplicates as a regular patterns. It is important to delete all the bias added by these elements before applying Machine Learning.

During this exercise, you'll use the Iris dataset and find the duplicates.

2 Functions used during the Exercise

2.1 duplicated

Library: vertica_ml_python.vDataframe

```
vDataframe.duplicated(self, columns: list = [], count: bool = False)
```

Find all the vDataframe duplicates (the duplicates are defined according to specific columns of the vDataframe).

Parameters

- **columns:** *<list>*, optional
  List of the vDataframe columns.
- **count:** *<bool>*, optional
  If True, the function will return the number of duplicates.

Returns

The `tablesample` type containing the duplicates (the information will be stored in the `values` attribute). You can convert this object to pandas using the `to_pandas` method or to vDataframe using the `to_vdf` method.

Example

```
from vertica_ml_python.vdataframe import vdf_from_relation
relation = "((SELECT 1 AS x, 4 AS y) UNION ALL (SELECT 1 AS x, 4 AS y) UNION ALL (SELECT 1 AS x, 5 AS y)) z"
vdf = vdf_from_relation(relation, dsn = "VerticaDSN")

#Output
      x     y
0     1     4
1     1     4
2     1     5
Name: VDF, Number of rows: 3, Number of columns: 2

vdf.duplicated()

#Output
      x     y     occurrence
```

```
0     1     4              2
Name: Duplicated Rows, Number of rows: 1, Number of columns: 3
```

2.2 drop_duplicates

Library: vertica_ml_python.vDataframe

```
vDataframe.drop_duplicates(self, columns: list = [])
```

Drop the vDataframe duplicates (the duplicates are defined according to specific columns of the vDataframe).

Parameters

- **columns:** *<list>*, optional
  List of the vDataframe columns.

Returns

The vDataframe itself.

Example

```
from vertica_ml_python.vdataframe import vdf_from_relation
relation = "((SELECT 1 AS x, 4 AS y) UNION ALL (SELECT 1 AS x, 4 AS y) UNION
    ALL (SELECT 1 AS x, 5 AS y)) z"
vdf = vdf_from_relation(relation, dsn = "VerticaDSN")

#Output
      x     y
0     1     4
1     1     4
2     1     5
Name: VDF, Number of rows: 3, Number of columns: 2

vdf.drop_duplicates()

#Output
      x     y
0     1     4
1     1     5
Name: VDF, Number of rows: 2, Number of columns: 2
```

3 Questions

Turn on Jupyter with the 'jupyter notebook' command. Start the notebook exercise6.ipynb and answer to the following questions.

- **Question 1:** Find the data duplicates.
- **Question 2:** Drop the duplicates.
- **Question 3:** Why is it important to find and drop the duplicates ?