Spark MLLib

Everybody is doing ML now

Expert
Services

PySpark ML

PySpark ML este librăria de Apache Spark ML pentru Python. Este foarte similară cu populara librărie Sklearn, oferind o structură similară de apelare în două etape. Ea a fost concepută pentru a oferi oricărei persoane capabilitatea de a efectua aplica algoritmi de învățare automată pe date.

PySpark ML oferă:

- Un tip de coloană nou, pentru a lucra cu atributele datelor, numită Vector.
- O clasă pentru construirea coloanelor de tip Vector.
- Diverse clase pentru standardizarea datelor.
- Diverse clase pentru modele de învățare automată.

Codul PySpark ML este unul foarte simplu și foarte intuitiv, întrucât complexitatea este în ce facem cu datele.

Vector - Expresii de calcul

Pentru a eficientiza transferul datelor către modelele de învățare automată, PySpark folosește un tip special de date, numit Vector. Ea reprezintă o listă de valori (vector matematic în mai multe dimensiuni) reținută foarte eficient.

PySpark ML oferă clase speciale pentru a construi un astfel de vector din datele unui Data Frame , în modulul:

```
from pyspark.ml.feature import *
```

- Adăugarea unui coloane noi la un Data Frame de tip vector concatenând coloanele existente

```
vector_assembler = VectorAssembler(inputCols=['vechime', 'varsta'], outputCol='atribute')
```

```
vector_assembler = VectorAssembler().setInputCols(['vechime', 'varsta']).setOutputCol('atribute')
new_data_df = vector_assembler.transform(data_df)
```

- Precum metodele Data Frame-urilor, un nou obiect de tip Data Frame este returnat care are adăugată sau actualizată coloana atribute care va reține o listă (vector) formată cu valorile coloanelor vechime și vârstă.

Vector – Conversia valorilor

Un Vector poate avea ca elemente doar valori numerice. Așadar, el poate fi creat doar din coloane cu valori numerice. Alte tipuri de date necesită să fie convertite la valori numerice înainte de aplicarea algoritmilor standard.

PySpark ML oferă clase speciale pentru a construi un astfel de vector din datele unui Data Frame , în modulul:

```
from pyspark.ml.feature import *
```

- Adăugarea unui coloane noi la un Data Frame de tip vector concatenând coloanele existente

```
vector_assembler = VectorAssembler(inputCols=['vechime', 'varsta'], outputCol='atribute')
new_data_df = tx.transform(data_df)
```

  - Un nou obiect de tip Data Frame este returnat care are adăugată sau actualizată coloana atribute care va reține o listă (vector) formată cu valorile coloanelor vechime și vârstă.

- ✓ Un Vector poate fi creat doar din coloane cu valori numerice. Alte tipuri de date necesită conversia lor.