5_linear_regression_quiz

November 24, 2019

1 Linear Regression Quiz

Use this Jupyter notebook to find the answer to the quiz in the previous section. There is an answer key in the next part of the lesson.

In [1]:
```
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, concat, count, lit, udf, avg
from pyspark.sql.types import IntegerType, StringType
from pyspark.ml.feature import RegexTokenizer, VectorAssembler
from pyspark.ml.regression import LinearRegression
```

In [2]:
```
spark = SparkSession.builder \
    .master("local") \
    .appName("Creating Features") \
    .getOrCreate()
```

1.0.1 Read Dataset

In [3]:
```
stack_overflow_data = 'Train_onetag_small.json'
```

In [4]:
```
df = spark.read.json(stack_overflow_data)
df.persist()
```

Out[4]:
```
DataFrame[Body: string, Id: bigint, Tags: string, Title: string, oneTag: string]
```

1.0.2 Build Description Length Features

In [6]:
```
df = df.withColumn("Desc", concat(col("Title"), lit(' '), col("Body")))


regexTokenizer = RegexTokenizer(inputCol="Desc", outputCol="words", pattern="\\W")
df = regexTokenizer.transform(df)
body_length = udf(lambda x: len(x), IntegerType())
df = df.withColumn("DescLength", body_length(df.words))
```

In [7]:
```
assembler = VectorAssembler(inputCols=["DescLength"], outputCol="DescVec")
df = assembler.transform(df)
```

In [8]:
```
df.head()
```

Out[8]:
```
Row(Body="<p>I'd like to check if an uploaded file is an image file (e.g png, jpg, jpeg,
```

2 Question

Build a linear regression model using the length of the combined Title + Body fields. What is the value of r^2 when fitting a model with `maxIter=5, regParam=0.0, fitIntercept=False, solver="normal"`?

```
In [14]: number_of_tags = udf(lambda x: len(x.split(" ")), IntegerType())
         df = df.withColumn("NumTags", number_of_tags(df.Tags))

In [15]: df.groupby("NumTags").agg(avg(col("DescLength"))).orderBy("NumTags").show()
         data = df.select(col("NumTags").alias("label"), col("DescVec").alias("features"))
         data.head()
```

```
+-------+------------------+
|NumTags|   avg(DescLength)|
+-------+------------------+
|      1|143.68776158175783|
|      2| 162.1539186134137|
|      3|181.26021064340088|
|      4|201.46530249110322|
|      5|227.64375266524522|
+-------+------------------+
```

```
Out[15]: Row(label=5, features=DenseVector([96.0]))

In [16]: lr = LinearRegression(maxIter=5, regParam=0.0, fitIntercept=False, solver="normal")

In [20]: lrModel_q1 = lr.fit(data)

In [18]: lrModel_q1.summary

Out[18]: <pyspark.ml.regression.LinearRegressionTrainingSummary at 0x7f65dacbc7f0>

In [19]: lrModel_q1.summary.r2

Out[19]: 0.4455149596308462

In [ ]:
```