

ICP6 – Machine Learning Algorithms

Algorithm – Classification/Naive Bayes

Data Set - Immunotherapy dataset

Columns - sex, age, Time, Number_of_Warts, Type, Area, induration_diameter, (7 - features)

Result_of_Treatment(label)

This dataset contains information about wart treatment results of 90 patients using immunotherapy. This dataset has 7 different features like sex, age, time etc., and one label which has two categories (wart treatment curing the disease = 1, wart treatment not curing the disease = 0)

In the below code we have only considered three features sex, age and induration_diameter features to predict the Result_of_Treatment. And the cross-validation parameter is set to 80-20 which means 80% of the data = 72 samples are used to train the algorithm and 20% of the data = 18 samples are used to test the model. Random function is used to divide the samples to training and testing. Then the accuracy is calculated which is correct predictions in test set/total predictions in test set.

Code :

```
ICP6 NaiveBayes.py
NaiveBayes.py Immunotherapy.csv x
1 from pyspark.ml.classification import NaiveBayes
2 from pyspark.ml.evaluation import MulticlassClassificationEvaluator
3 import os
4 from pyspark.ml.feature import VectorAssembler
5 import numpy as np
6 # Load training data
7 from pyspark.python.pyspark.shell import spark
8 from pyspark.sql import SparkSession
9
10 os.environ["SPARK_HOME"] = "/Users/lalithalett/Downloads/spark-2.3.1-bin-hadoop2.7/"
11 os.environ["HADOOP_HOME"] = "/usr/local/cellar/hadoop3.1.0"
12 os.environ["PYSPARK_PYTHON"] = "/usr/local/cellar/python3.6.5.1/bin/python3.6"
13 os.environ["PYSPARK_DRIVER_PYTHON"] = "/usr/local/cellar/python3.6.5.1/bin/python3.6"
14
15 data = spark.read.format("csv").option("header", "true").load("/Users/lalithalett/PycharmProjects/ICP6/Immunotherapy.csv")
16
17 spark = SparkSession.builder.getOrCreate()
18
19 data = spark.read.load("Immunotherapy.csv", format="csv", header=True, delimiter=",")
20 data = data.withColumn("AGE_FACTOR", data["age"] - 0).withColumn("Area", data["Area"] - 0).withColumn("I_D", data["induration_diameter"] - 0).withColumn("label", data["sex"] - 0)
21 data.show(100)
22 assem = VectorAssembler(inputCols=["AGE_FACTOR", "Area", "I_D"], outputCol="features")
23 data = assem.transform(data)
24
25
26 # Split the data into train and test
27 splits = data.randomSplit([0.8, 0.2], 1234)
28 train = splits[0]
29 test = splits[1]
30
31 # create the trainer and set its parameters
32 nb = NaiveBayes(smoothing=1.0, modelType="multinomial")
33
34 # train the model
35 model = nb.fit(train)
36
37 # select example rows to display.
38 predictions = model.transform(test)
39 predictions.show(100)
40
41 # compute accuracy on the test set
42 evaluator = MulticlassClassificationEvaluator(labelCol="label", predictionCol="prediction",
43                                              metricName="accuracy")
44 accuracy = evaluator.evaluate(predictions)
45 print("Test set accuracy = " + str(accuracy))
```

Python Console Terminal Run TODO
Packages installed successfully: Installed packages: 'numpy' (today 7:38 PM) 29:18 LFS UTF-8 Event Log

Output:

```
ICP6 NaiveBayes.py
NaiveBayes.py Immunotherapy.csv
21 data = data.withColumn("AGE_FACTOR", data["age"] - 0).withColumn("Area", data["Area"] - 0).withColumn("I_D", data["induration_diameter"] - 0).withColumn("label", data["sex"] - 0)
22 data.show(100)
23 assem = VectorAssembler(inputCols=["AGE_FACTOR", "Area", "I_D"], outputCol="features")
24 data = assem.transform(data)
25
26
27 # Split the data into train and test
28 splits = data.randomSplit([0.8, 0.2], 1234)
29 train = splits[0]
30 test = splits[1]
```

Run: NaiveBayes

[sex]	[age]	[Time]	[Number_of_Warts]	[Type]	[Area]	[induration_diameter]	[Result_of_Treatment]	[AGE_FACTOR]	[I_D]	[label]
1	15	11	6	1	30.0	25	0	15.0	25.0	1.0
1	15	3	2	1	990.0	70	1	15.0	70.0	1.0
1	15	4	4	1	25.0	7	1	15.0	7.0	1.0
1	23	5.75	2	1	43.0	7	1	23.0	7.0	1.0
1	35	6.75	4	1	43.0	0	1	35.0	0.0	1.0
1	35	8.75	10	2	69.0	7	1	35.0	7.0	1.0
1	40	5.5	8	3	69.0	5	1	40.0	5.0	1.0
1	45	10	8	1	50.0	7	1	45.0	7.0	1.0
1	15	1.75	1	1	49.0	7	0	15.0	7.0	2.0
1	15	6.5	10	1	56.0	7	1	15.0	7.0	2.0
1	15	8	3	1	55.0	7	1	15.0	7.0	2.0
1	20	6.75	2	1	6.0	6	1	20.0	6.0	2.0
1	21	18.75	8	1	57.0	5	0	21.0	5.0	2.0
1	49	9	4	1	14.0	9	1	49.0	9.0	2.0
1	53	7.25	5	1	63.0	7	1	53.0	7.0	2.0

2018-07-23 22:50:30 WARN BLAS:61 - Failed to load implementation from: com.github.fommil.netlib.NativeSystemBLAS
2018-07-23 22:50:30 WARN BLAS:61 - Failed to load implementation from: com.github.fommil.netlib.NativeRefBLAS

[sex]	[age]	[Time]	[Number_of_Warts]	[Type]	[Area]	[induration_diameter]	[Result_of_Treatment]	[AGE_FACTOR]	[I_D]	[label]	features	rawPrediction	probability	[prediction]
1	15	11	6	1	30.0	25	0	15.0	25.0	1.0	[15.0, 30.0, 0.25, 0]	[-94.913720240870...	[0.00208947004149...	1.0
1	15	3	2	1	990.0	70	1	15.0	70.0	1.0	[15.0, 990.0, 70.0]	[-515.63434972363...	[1.0, 3.6312224202...	0.0
1	15	4	4	1	25.0	7	1	15.0	7.0	1.0	[15.0, 25.0, 7.0]	[-49.913631672934...	[0.16149173767334...	1.0
1	23	5.75	2	1	43.0	7	1	23.0	7.0	1.0	[23.0, 43.0, 7.0]	[-68.822827995989...	[0.25580115688555...	1.0
1	35	6.75	4	1	43.0	0	1	35.0	0.0	1.0	[35.0, 43.0, 0.0]	[-89.161802196624...	[0.07509705346745...	1.0
1	35	8.75	10	2	69.0	7	1	35.0	7.0	1.0	[35.0, 69.0, 7.0]	[-96.825181771488...	[0.43339448711152...	1.0
1	40	5.5	8	3	69.0	5	1	40.0	5.0	1.0	[40.0, 69.0, 5.0]	[-99.798900937367...	[0.46303485571853...	1.0
1	45	10	8	1	50.0	7	1	45.0	7.0	1.0	[45.0, 50.0, 7.0]	[-100.41941242186...	[0.1212080717400...	1.0
1	15	1.75	1	1	49.0	7	0	15.0	7.0	2.0	[15.0, 49.0, 7.0]	[-58.548208690900...	[0.50851555854235...	0.0
1	15	6.5	10	1	56.0	7	1	15.0	7.0	2.0	[15.0, 56.0, 7.0]	[-61.856293654473...	[0.62018213466712...	0.0
1	15	8	3	1	55.0	7	1	15.0	7.0	2.0	[15.0, 55.0, 7.0]	[-60.696652945391...	[0.61187801129679...	0.0
1	20	6.75	2	1	6.0	6	1	20.0	6.0	2.0	[20.0, 6.0, 6.0]	[-57.78698515372...	[0.0253826535780...	1.0
1	21	18.75	8	1	57.0	5	0	21.0	5.0	2.0	[21.0, 57.0, 5.0]	[-106.36926804301...	[0.17011146432648...	1.0
1	49	9	4	1	14.0	9	1	49.0	9.0	2.0	[49.0, 14.0, 9.0]	[-183.62357456010...	[0.00204606364782...	1.0
1	53	7.25	5	1	63.0	7	1	53.0	7.0	2.0	[53.0, 63.0, 7.0]	[-121.18027250119...	[0.10544977402969...	1.0

Test set accuracy = 0.4375
Process finished with exit code 0

Python Console Terminal Run TODO
Packages installed successfully: Installed packages: 'numpy' (today 7:38 PM)
136:1 LFS UTF-8

Test set accuracy = 0.4375