МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«Национальный исследовательский ядерный университет «МИФИ»
Обнинский институт атомной энергетики –
филиал федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский ядерный университет «МИФИ»
(ИАТЭ НИЯУ МИФИ)

Отделение интеллектуальных кибернетических систем

ЛАБОРАТОРНАЯ РАБОТА №4
«Машинное обучение »
по дисциплине
«Большие данные»

Выполнил студент 1 курса
группы ИВТ-М20
Лискунов Р. Г.

Проверил:
кандидат технических наук
Грицюк С. В.

Обнинск, 2020

Цель работы

Определить простую задачу машинного обучения и решить ее. Точность на испытательном наборе должна быть не менее 60%.

Краткая теория

Apache Spark MLlib используется для создания приложения машинного обучения. Приложение выполняет прогнозный анализ на открытом наборе данных. MLlib — это основная библиотека Spark, которая предоставляет множество служебных программ, полезных для задач машинного обучения, таких как:

1. Классификация;
2. Регрессия;
3. Кластеризация;
4. Моделирование сингулярного разложения и анализа по методу главных компонент;
5. Проверки гипотез и статистической выборки.

Общие сведения о классификации и логистической регрессии

Классификация — это распространенная задача машинного обучения, которая представляет собой процесс сортировки входных данных по категориям. Это задание алгоритма классификации, позволяющее определить, как назначить "метки" входным данным, которые мы предоставляем. Например, можно представить алгоритм машинного обучения, который принимает в качестве входных данных данные о акции. Затем делит биржевую акцию на две категории: акции, которые следует продавать и акции, которые следует хранить.

Логистическая регрессия — один из алгоритмов классификации. API Spark для логистической регрессии подходит для задач двоичной классификации или разделения входных данных на две группы. Дополнительные сведения о логистической регрессии можно посмотреть в документации.

В целом, процесс логистической регрессии создает логистическую функцию. Используем функцию для прогнозирования вероятности того, что входной вектор принадлежит одной группе или другой.

Ход работы

В процессе работы мы рассмотрим набор данных, состоящий из популярных детских имен. Сперва мы подключим контекст Spark, а также укажем в качестве dataframe, описанный выше набор данных. Первые 20 строк набора показаны на рисунке 1.

```
+-------------+------+--------------------------+------------------+-----+----+
|Year of Birth|Gender|Ethnicity                 |Child's First Name|Count|Rank|
+-------------+------+--------------------------+------------------+-----+----+
|2016         |FEMALE|ASIAN AND PACIFIC ISLANDER|Olivia            |172  |1   |
|2016         |FEMALE|ASIAN AND PACIFIC ISLANDER|Chloe             |112  |2   |
|2016         |FEMALE|ASIAN AND PACIFIC ISLANDER|Sophia            |104  |3   |
|2016         |FEMALE|ASIAN AND PACIFIC ISLANDER|Emma              |99   |4   |
|2016         |FEMALE|ASIAN AND PACIFIC ISLANDER|Emily             |99   |4   |
|2016         |FEMALE|ASIAN AND PACIFIC ISLANDER|Mia               |79   |5   |
|2016         |FEMALE|ASIAN AND PACIFIC ISLANDER|Charlotte         |59   |6   |
|2016         |FEMALE|ASIAN AND PACIFIC ISLANDER|Sarah             |57   |7   |
|2016         |FEMALE|ASIAN AND PACIFIC ISLANDER|Isabella          |56   |8   |
|2016         |FEMALE|ASIAN AND PACIFIC ISLANDER|Hannah            |56   |8   |
|2016         |FEMALE|ASIAN AND PACIFIC ISLANDER|Grace             |54   |9   |
|2016         |FEMALE|ASIAN AND PACIFIC ISLANDER|Angela            |54   |9   |
|2016         |FEMALE|ASIAN AND PACIFIC ISLANDER|Ava               |53   |10  |
|2016         |FEMALE|ASIAN AND PACIFIC ISLANDER|Joanna            |49   |11  |
|2016         |FEMALE|ASIAN AND PACIFIC ISLANDER|Amelia            |44   |12  |
|2016         |FEMALE|ASIAN AND PACIFIC ISLANDER|Evelyn            |42   |13  |
|2016         |FEMALE|ASIAN AND PACIFIC ISLANDER|Ella              |42   |13  |
|2016         |FEMALE|ASIAN AND PACIFIC ISLANDER|Arya              |42   |13  |
|2016         |FEMALE|ASIAN AND PACIFIC ISLANDER|Ariana            |40   |14  |
|2016         |FEMALE|ASIAN AND PACIFIC ISLANDER|Maya              |39   |15  |
+-------------+------+--------------------------+------------------+-----+----+
only showing top 20 rows
```

Рисунок 1. Топ 20 строк набора данных

Поставим задачу исследовать взаимосвязь числа имен, их ранга и их классификацию по этническому происхождению. В ходе исследования мы хотим попробовать предсказать, что большая или, наоборот, меньшая часть детей принадлежит конкретно взятому этносу.

Для этого мы создам дополнительные объекты языка программирования – dataframe, которые будут содержать дополнительную для нас информацию, которые касаются выбранных столбцов.

В процессе работы у нас появляются трудности с объединением данных, поэтому мы используем VectorAssembler – это преобразователь, который объединяет заданный список столбцов в один векторный столбец. Это полезно для объединения необработанных функций и функций, созданных различными преобразователями функций, в один вектор функций.

Работать напрямую с данными, хоть и в случае моего небольшого набора данных, не составляет труда, но для удобства обращения воспользуемся StringIndexer, который кодирует строковый столбец меток в столбец индексов меток. Также введем столбец features, который будет агрегировать значения числа детей, их ранга и этноса. Результаты представлены на рисунке 2.

Year of Birth	Gender	Ethnicity	Child's First Name	Count	Rank	Ethnicity	DistEthnicity	features	label
2016	FEMALE	ASIAN AND PACIFIC ISLANDER	olivia	172	1	1	ASIAN AND PACIFIC ISLANDER	[172.0,1.0]	3.0
2016	FEMALE	ASIAN AND PACIFIC ISLANDER	chloe	112	2	1	ASIAN AND PACIFIC ISLANDER	[112.0,2.0]	3.0
2016	FEMALE	ASIAN AND PACIFIC ISLANDER	sophia	104	3	1	ASIAN AND PACIFIC ISLANDER	[104.0,3.0]	3.0
2016	FEMALE	ASIAN AND PACIFIC ISLANDER	emma	99	4	1	ASIAN AND PACIFIC ISLANDER	[99.0,4.0]	3.0
2016	FEMALE	ASIAN AND PACIFIC ISLANDER	emily	99	4	1	ASIAN AND PACIFIC ISLANDER	[99.0,4.0]	3.0
2016	FEMALE	ASIAN AND PACIFIC ISLANDER	mia	79	5	1	ASIAN AND PACIFIC ISLANDER	[79.0,5.0]	3.0
2016	FEMALE	ASIAN AND PACIFIC ISLANDER	charlotte	59	6	1	ASIAN AND PACIFIC ISLANDER	[59.0,6.0]	3.0
2016	FEMALE	ASIAN AND PACIFIC ISLANDER	sarah	57	7	1	ASIAN AND PACIFIC ISLANDER	[57.0,7.0]	3.0
2016	FEMALE	ASIAN AND PACIFIC ISLANDER	isabella	56	8	1	ASIAN AND PACIFIC ISLANDER	[56.0,8.0]	3.0
2016	FEMALE	ASIAN AND PACIFIC ISLANDER	hannah	56	8	1	ASIAN AND PACIFIC ISLANDER	[56.0,8.0]	3.0
2016	FEMALE	ASIAN AND PACIFIC ISLANDER	grace	54	9	1	ASIAN AND PACIFIC ISLANDER	[54.0,9.0]	3.0
2016	FEMALE	ASIAN AND PACIFIC ISLANDER	angela	54	9	1	ASIAN AND PACIFIC ISLANDER	[54.0,9.0]	3.0
2016	FEMALE	ASIAN AND PACIFIC ISLANDER	ava	53	10	1	ASIAN AND PACIFIC ISLANDER	[53.0,10.0]	3.0
2016	FEMALE	ASIAN AND PACIFIC ISLANDER	joanna	49	11	1	ASIAN AND PACIFIC ISLANDER	[49.0,11.0]	3.0
2016	FEMALE	ASIAN AND PACIFIC ISLANDER	amelia	44	12	1	ASIAN AND PACIFIC ISLANDER	[44.0,12.0]	3.0
2016	FEMALE	ASIAN AND PACIFIC ISLANDER	evelyn	42	13	1	ASIAN AND PACIFIC ISLANDER	[42.0,13.0]	3.0
2016	FEMALE	ASIAN AND PACIFIC ISLANDER	ella	42	13	1	ASIAN AND PACIFIC ISLANDER	[42.0,13.0]	3.0
2016	FEMALE	ASIAN AND PACIFIC ISLANDER	arya	42	13	1	ASIAN AND PACIFIC ISLANDER	[42.0,13.0]	3.0
2016	FEMALE	ASIAN AND PACIFIC ISLANDER	ariana	40	14	1	ASIAN AND PACIFIC ISLANDER	[40.0,14.0]	3.0
2016	FEMALE	ASIAN AND PACIFIC ISLANDER	maya	39	15	1	ASIAN AND PACIFIC ISLANDER	[39.0,15.0]	3.0

only showing top 20 rows

Рисунок 2. Результаты агрегирования данных

Ввиду наличия большого объема данных, нам представляется возможным разбить их на более мелкие части. Таким образом, мы подготовим данные для логистической регрессии.

Зададим необходимые параметры для обучения и поддержания точности на уровне, указанном в цели работы. Полученные результаты отобразим на рисунке 3.

Year of Birth	Gender	Ethnicity	Child's First Name	Count	Rank	Ethnicity	DistEthnicity	features	label	rawPrediction	probability	prediction
2011	FEMALE	ASIAN AND PACIFIC...	ABIGAIL	24	24	1	ASIAN AND PACIFIC...	[24.0,24.0]	3.0	[-1.6939870962549...	[0.02283377010068...	3.0
2011	FEMALE	ASIAN AND PACIFIC...	ADA	13	35	1	ASIAN AND PACIFIC...	[13.0,35.0]	3.0	[-1.3669206701107...	[0.04070083391468...	3.0
2011	FEMALE	ASIAN AND PACIFIC...	AIZA	10	38	1	ASIAN AND PACIFIC...	[10.0,38.0]	3.0	[-1.2777207357078...	[0.04742760875143...	3.0
2011	FEMALE	ASIAN AND PACIFIC...	ALEENA	12	36	1	ASIAN AND PACIFIC...	[12.0,36.0]	3.0	[-1.3371873586431...	[0.04284096088815...	3.0
2011	FEMALE	ASIAN AND PACIFIC...	ALYSSA	26	22	1	ASIAN AND PACIFIC...	[26.0,22.0]	3.0	[-1.7534537191902...	[0.02050837490033...	3.0
2011	FEMALE	ASIAN AND PACIFIC...	ANGELINA	26	22	1	ASIAN AND PACIFIC...	[26.0,22.0]	3.0	[-1.7534537191902...	[0.02050837490033...	3.0
2011	FEMALE	ASIAN AND PACIFIC...	ARIA	12	36	1	ASIAN AND PACIFIC...	[12.0,36.0]	3.0	[-1.3371873586431...	[0.04284096088815...	3.0
2011	FEMALE	ASIAN AND PACIFIC...	ARIANA	15	33	1	ASIAN AND PACIFIC...	[15.0,33.0]	3.0	[-1.4263872930460...	[0.03670909900365...	3.0
2011	FEMALE	ASIAN AND PACIFIC...	ARIANA	15	33	1	ASIAN AND PACIFIC...	[15.0,33.0]	3.0	[-1.4263872930460...	[0.03670909900365...	3.0
2011	FEMALE	ASIAN AND PACIFIC...	ARIANNA	11	37	1	ASIAN AND PACIFIC...	[11.0,37.0]	3.0	[-1.3074540471754...	[0.04508201834151...	3.0

only showing top 10 rows

Рисунок 3. Результаты предсказаний

Несмотря на повторяющиеся имена в таблице результатов, мы можем увидеть, что теперь наша модель может предсказать по рангу и числу вероятный этнос ребенка. Крайний правый столбец, если исходить из наших предложений, совпадает с действительным значением, что является верным результатом исследования.

Листинг кода

```
package LabFour

import org.apache.log4j.Level.WARN
import org.apache.log4j.LogManager
import org.apache.spark.ml.classification.{LogisticRegression, LogisticRegressionModel}
import org.apache.spark.ml.feature.{StringIndexer, VectorAssembler}
import org.apache.spark.sql.functions.{col, lower}
import org.apache.spark.sql.{DataFrame, SparkSession}

object LabFour {
  val PATH: String = "src/main/data"
  val NODES: Int = 3

  def main(args: Array[String]): Unit = {
    val spark: SparkSession = SparkSession
      .builder()
      .appName("Lab4")
      .master(s"local[$NODES]")
      .getOrCreate
    LogManager.getRootLogger.setLevel(WARN)

    val dataframe: DataFrame = spark
      .read
      .format("csv")
      .option("header", "true")
      .option("delimiter", ",")
      .option("inferSchema", value = true)
      .load(s"$PATH/var.csv")

    dataframe.show(false)

    val seq: Seq[(Int, String)] = Seq(
      (0, "WHITE NON HISPANIC"),
      (1, "ASIAN AND PACIFIC ISLANDER"),
      (2, "HISPANIC"),
      (3, "BLACK NON HISPANIC")
    )

    import spark.implicits._
    val mapper: DataFrame = seq.toDF("Ethnicity", "DistEthnicity")

    val mapped: DataFrame = dataframe
      .join(
        mapper, dataframe("Ethnicity") === mapper("DistEthnicity"),
        "inner"
      )
    val columns: Array[String] = Array("Count", "Rank")

    val assembler: VectorAssembler = new VectorAssembler()
```

```
    .setInputCols(columns)
    .setOutputCol("features")
  val feature: DataFrame = assembler.transform(mapped)

  val indexer: StringIndexer = new StringIndexer()
    .setInputCol("DistEthnicity")
    .setOutputCol("label")

  val label: DataFrame = indexer
    .fit(feature)
    .transform(feature)
  label.show(false)

  val seed: Int = 5043
  val Array(training, test) = label
    .randomSplit(Array(0.7, 0.3), seed)

  val regression: LogisticRegression = new LogisticRegression()
    .setMaxIter(100)
    .setRegParam(0.02)
    .setElasticNetParam(0.8)
  val model: LogisticRegressionModel = regression
    .fit(training)

  val prediction: DataFrame = model
    .transform(test)
  prediction.show(10)
 }
}
```

Вывод

В ходе данной лабораторной работы я изучил особенности и возможности способов работы со Spark ML и предсказания данных на основе логистической регрессии. Для этого я использовал набор данных популярных детских имен. Полученная модель умеет по числу детей, которые родились в определенный год и по рангу имени, определять этнос ребенка. Данное исследование позволяет выявить наибольшее или наименьшее число детей, которые относятся к выбранному этносу. Сами значения могут быть использованы, например, при изготовке вакцин. Так, при росте числа новорожденных в Нью-Йорке следует изготовлять вакцину с большей вероятностью для определённой этнической группы.