CS 5433 - Big Data Management

Spring 2022

Group Project

Programming Assignment 3

This project will be done in groups of three using Spark ml.

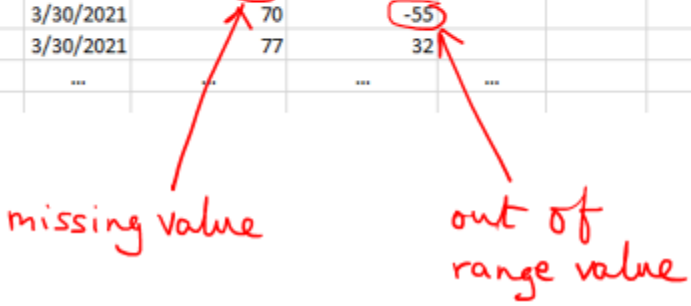
1 TASK 1: DATA CORRECTION

Missing data (cells with no values), and out of range values (for example, a temperature of 1000°G in a weather dataset) are listed below in section 9.

(a) Define "out of range" values only for the cells listed in section 9 For example, you may decide that any temperature t where $-30 > t > 120$ is out of range. Note: there are no strict rules on this, so decide on some range that looks reasonable

(b) Only for the cells listed in section 9, replace the missing data and out of range values with values from another row that is most similar to the row where data is missing or out of range. For example:

	A	B	C	D	E	F
1	City	Date	Highest Temp	Lowest Temp	Humidity	…
2			°F	°F		…
3	Stillwater	4/1/2021	77	45	..	
4	Oklahoma City	4/1/2021	79	47		
5	Tulsa	4/1/2021	81	46		
6	Dallas	4/1/2021	71	50		
7	Tulsa	3/31/2021	30	41		
8	New York	3/31/2021	55	35		
9	Stillwater	3/31/2021		30		
10	Tulsa	3/30/2021	70	-55		
11	Stillwater	3/30/2021	77	32		
12	…	…	…	…	…	



Row 9 column C (cell 9C) has a value missing. The closest value appears to row 11. Therefore cell 9C is given a value of 77

Row 10 column D has an out or range value. The closest row appears to be row 6, although the city is different. Therefore cell 10D is given a value of 50. The closest

row may also be row 5 as the city is the same, but the temperature difference is greater than for row 6. The similarity algorithm will identify the closest row.

There are a number of similarity algorithms such as Cosine similarity, Jacard similarity etc. You may choose any similarity algorithm.
Missing or out of range values will only be numerical (integers or float) values. You may have to use one-hot encoding for text (such as 'Stillwater' for example).

2 TASK 2: PREDICTION ALGORITHM

Each group will implement and predict using

(a) Linear regression
(b) Random Forest

Clearly identify the variable you are predicting. Predict on the same variable for both algorithms.

3 TASK 3: MEASURE ACCURACY

Measure the accuracy of both predictions. You may use RMSE (Root Mean Square Error) or some other metric. Define your metric if you are not using RMSE or R^2 which were covered in class

4 TASK 4:– CREATE A SPARK PIPELINE

Create a pipeline for tasks 1 – 3 above.

5 TASK 5: BONUS – IDENTIFY THE PRINCIPAL COMPONENTS (15% BONUS)

The principal components are the features which contribute most to the prediction. More details using Spark ml for principal component analysis (PCA) can be found at https://linuxtut.com/en/7714a1cc9be56588b72a/ (Principal component analysis with Spark ML).

There are many other sites related to PCA.

6 DELIVERABLES:

6.1 Group Deliverable I (one per group):

Deadline: April 29

Submit on canvas

1. All the source code you submit should be well commented [Penalty for not commenting adequately 25%]
2. Your source code should run on the Hadoop cluster.
3. Submissions

a. README File for each question [Group_No_README_x]. The README file should include:
   i. A brief description of the dataset (number of records, number of features or columns, the feature or label you are predicting, the other features etc.(½ page maximum)
   ii. Identify the variable you are predicting. What is the rationale?
   iii. A brief description of the work done for each task (½ page maximum per task)
   iv. The results from programs should be put in separate files [Group_No_Task_x_Output].
   v. Screenshots of results of all tasks [Group_No_Task_x_Screenshots].
   vi. Discussion of results of each task (½ page maximum per task). What do the results mean? Compare the results from the two prediction schemes.
   vii. Instructions to compile and run your program.
b. Commented source code for each task [Group_No_Task_x_Code].
   i. All the source files zipped as a single zip file [Group_No_PA2.zip]. Any other format will not be accepted

You may implement in the language of your choice. Use Spark on the department cluster.

6.2 Individual Deliverables (one per individual)

Deadline: April 29

Submit on canvas

To be submitted separately by each individual

1. Evaluation: The following table should be completed by each student
   50% of actual evaluation score will be deducted if you do not submit an evaluation of your team members.

Evaluation of team members (do not evaluate yourself)	**Name of team member 1** ……	**Marks for team member 1**	**Name of team member 2** ……	**Marks for team member 2**
Communication (out of 50):	Explain reason for marks	(out of 50)	Explain reason for marks	(out of 50)
Contribution (out of 100)	Explain reason for marks	(out of 100)	Explain reason for marks	(out of 100)

2. Summary of your contribution

Summary of your contribution:

7 BREAKDOWN OF MARKS

Task 1	10
Task 2	20
Task 3	15
Task 4	15
Evaluation	10
TOTAL	70
Task 5 (optional)	10

8 COLLABORATION POLICY:

You should complete this programming assignment individually. Any doubts/clarification about the questions should be directed to either the instructor/TA. Using code from web & other resources must be acknowledged in your report.