Assignment 2 Specification

(Individual Assignment)

FIT 5202 - Data processing for Big Data

Due: Friday October 18, 2019, 11:55 PM (Local Campus Time)
Worth: 20% of the final marks

Rain in Australia: Predict rain tomorrow in Australia

Predicting rain or weather is a common problem in machine learning. Different machine learning algorithms can be used to model and predict rainfall. In this assignment, we ask you to complete the analysis to predict whether there will be rain tomorrow or not. In particular, you are required to apply the tools of machine learning to visualize and predict the possibility of rainfall in Australia.

Required Datasets (available in Moodle):

- Rain in Australia (**weatherAUS.csv**)

The dataset is originally from Kaggle Dataset and can be found at this link. It has been modified to serve our purpose i.e. (to do the binary classification).

A. Creating Spark Session and Loading the Data

Step 01: Import Spark Session and initialize Spark

pyspark is the Spark Python API that exposes the Spark programming model to Python. You are already familiar with **sparkContext** from Assignment 1. **sparkContext** was used as a channel to access all spark functionality. In order to use APIs of SQL, HIVE, and Streaming, separate contexts need to be created. From SPARK 2.0.0 onwards **sparkSession** provides a single point of entry to interact with underlying Spark functionality and allows programming Spark with Dataframe and Dataset APIs. All the functionality available with **sparkContext** are also available in **sparkSession**.

Write the code to create a sparkSession object, with 4 local cores. To create a sparkSession with 4 core you have to use configure it as *local[4].* Give a name to your program using *appName*().

Step 02: Load the dataset and print the schema and total number of entries

In **sparkSession** you can use *spark_session.read.csv()* method to load data as CSV format. You can download the dataset from Moodle. After you load the csv file into a dataframe using spark session, write the code to print the total number of entries in the dataset.

B. Data Cleaning and Processing

Data cleaning and processing is an important aspect for any machine learning task. We have to carefully look into the data and based on the types, quality of the data, we have to plan our cleaning procedures.

Step 03: Delete columns from the dataset

During the data cleaning and processing phase, we delete unnecessary data from the dataset to improve the efficiency and accuracy of our model. You have to think which columns are not contributing to the rain prediction. To keep things simple, you are required to delete the following columns due to data quality and accuracy.

- Date
- Location
- Evaporation
- Sunshine
- Cloud9am
- Cloud3pm
- Temp9am
- Temp3pm

However, if you want to keep any of these columns, you can keep them if you process them in an intelligent way that improve the accuracy, that is fine, **however not mandatory**.

Step 04: Print the number of missing data in each column.

We already have an initial idea about the data structure from the schema. Even in plain eyes, we can observe that there are lots of NA (null) values in the given dataset. Your job in this step is to print the number of NA(null) values in each column.

Step 05: Fill the missing data with average value and maximum occurrence value.

In this step you have to fill in all the missing data with average value (for numeric column) or maximum frequency value (for non-numeric column).

Firstly, identify the columns which have numeric values (e.g., MinTemp, MaxTemp), calculate the average and fill the null value with the average.

Secondly, identify the columns with non-numeric values (e.g., WindGustDir, WindDir9am) and find the for frequent item (e.g., wind direction). Now fill the null values with that item for that particular column.

Step 06: Data transformation

In this step, you have to transform the data so that it will be useful to process by the machine learning algorithm. Before transforming your non-numerical data, do the type casting (to double) of the numerical value columns as they are defined as "String" (see, the schema of the dataset). For the non-numerical value column (i.e., WindGustDir, WindDir9am, WindDir3pm, RainTomorrow) use the StringIndexer method to convert them into numbers.

Step 07: Create the feature vector and divide the dataset

In this step, you have to create the feature vector from the given columns. When you create you feature vector, remember to exclude the column that you will be using for testing the accuracy of your model.

After creation of your feature vector, you have split your dataset into two (e.g., training and testing). In this assignment, you have to spit the dataset randomly and between 70 percent and 30 percent.

C. Apply Machine Learning Algorithms

Step 08: Apply machine learning classification algorithms on the dataset and compare their accuracy. Plot the accuracy as bar graph.

You have to use *DecisionTreeClassifier(), RandomForestClassifier(), and LogisticRegression(), GBTClassifier()* methods in spark to calculate the probability of the rain fall tomorrow based on the other related data points (e.g., temperature, wind, humidity). Finally, you have to draw the graph (e.g. bar chart) to demonstrate the comparison of their accuracy.

Step 09: Calculate the confusion matrix and find the precision, recall, and F1 score of each classification algorithm. Explain how the accuracy of the predication can be improved?

Finding the accuracy of the model does not always represent the quality of the model for a given dataset. Number of false positive and false negative identification also plays an important role when we decide about any particular classification model. The way we can calculate is called confusion matrix. You can use confusionMatrix() method to calculate the confusion matrix. From the confusion matrix show the precision, recall and f1 score of each classification model. Explain how you can improve the accuracy of the predication.

Complete all the steps mentioned above save the file as **Assignment 2.ipynb**

Assignment Marking

The rubric for the assignment is available in the Moodle for your reference. The marking of this assignment is based on quality of work that you have submitted rather than just quantity. The marking starts from zero and goes up based on the tasks you have successfully completed and it's quality for example how well the code submitted follows *programming standards, code documentation, presentation of the assignment, readability of the code, organisation of code and so on*. Please find the PEP 8 -- Style Guide for Python Code here for your reference.

Submission

You should submit your final version of the assignment solution online via Moodle; You must submit the following:

- A zip file of your Assignment 2 folder, named based on your authcate name (e.g. psan002). This should contain your **Assignment 2.ipynb.**This should be a ZIP file and *not any other kind of compressed folder (e.g. .rar, .7zip, .tar).*
- The assignment submission should be uploaded and finalised by Friday October 18th, 11:55 PM (Local Campus Time).
- Your assignment will be assessed based on the contents of the Assignment 2 folder you have submitted via Moodle. When marking your assignments, we will use the same ubuntu setup as provided to you in Week 01.

Other Information

Where to get help

You can ask questions about the assignment on the Assignment Discussion Forum on the unit's Moodle page. This is the preferred venue for assignment clarification-type questions. You should check this forum (and the News forum) regularly, as the responses of the teaching staff are "official" and can constitute amendments or additions to the assignment specification. Also, you can visit the consultation sessions if the problem and the confusions are still not solved.

Plagiarism and collusion

Plagiarism and collusion are serious academic offences at Monash University. Students must not share their work with any other students. Students should consult the policy linked below for more information.

https://www.monash.edu/students/academic/policies/academic-integrity

See also the video linked on the Moodle page under the Assignment block.

Students involved in collusion or plagiarism will be subject to disciplinary penalties, which can include:

- The work not being assessed

- A zero grade for the unit
- Suspension from the University
- Exclusion from the University

Late submissions

Submission must be made by the due date otherwise penalties will be enforced. You must negotiate any extensions formally with your campus unit lecturer via the in-semester special consideration process:

http://www.monash.edu.au/exams/special-consideration.html

There is a **5% penalty per day including weekends** for the late submission.