University of Sunshine Coast

ICT 707

Data Science and Practices

Big Data Assignment

Task 3

Submitted by: Gagandeep Kaur

Student number:1121869

Advancement of Data Science

Data is increasing on daily basis which provides an opportunity to improve effectiveness of production in various industry sectors, understating customer behavior and predicting success of a product etc. Data science empower business to take decision on real data driven facts and automate various business process. Its demand has escalated to 56% in 2018-2019. Advancement in Artificial intelligence (AI) and machine learning (ML) are used to give insights in given data sets can be visualized instantly.

Here are few advancement of Data Science in past two years:

- **Automation in data Science and AutoML Framework:** data scientist use to run manually data set against various algorithms to produce few models. However, with the help of AutoML machine learning platform productivity is boosted because it can develop various models simultaneous and doesn't require manual efforts, results are produced in mush less time. This framework automate model design and training.

  Automatic Data cleaning is a heavily researched area. IBM offers many automation and tooling for data cleaning. Google is investing in CloudAutoML. Majorly all data science pipeline process will become automated.

- **Graph analytics:** this will provide flexible and powerful tools that are capable of combining structural and non-structural data set. It analyses complicated data points and relationships using graphs. Graphs are easier to understand and provide maximum insight hence is best option to represent complex data points.

  Graph Analytics can be used to prevent financial crimes, analyzing power and water grids to find flaws, filtering out bots on social media to minimize fake news and information etc.

- **Data Fabric**: this technique collects data from various sources such as API, reusable data services, pipeline, semantic tiers and encapsulate organization data to provide transformable access to data. It provides scalability while data being agile which assist in keeping data in intelligible way for users and applications.

  It allows unparallel access to process, manage, store and data as needed. Business Intelligence and Data Science relies Data Fabrics as it gives clean access to large amount of data.

- **Augmented Analytics:** this technique focus on removing incorrect conclusion or bias for optimized decisions, this refers to better insight from data. Insight it offers are relatively better with reduced dependency on data scientists and machine learning experts. Increasing adoption of cloud computing and IoT and connected devices are major drives of augmented analytics. Many businesses prefer augmented analytics over traditional analytics to reduce human error and bias.

- **Natural Language processing and Conversational Analytics**: data science analysis was based on number, for a text to be process it need to be converted into numbers first. However, with advancement of natural language processing in deep learning large text can be integrated in data analysis. Neural network can extract and process quickly large text. It is able to classify text into two groups, first it determine the sentiment of the text and secondly it analysis similarity of data. At last all thee information is stored in single vector of numbers.

  This technology enable us to explore more complex data. NLP and conversational analytics will complement augmented analytics.

Comparison between Spark 3.0 and 2.4

Apace spark 3.0 was released on 10th June 2020. It has various enhanced features from its previous version 2.4. spark is open source and is popular for its use in big data processing, data science, data science and machine learning. There are many effective enhance done that benefit higher-level library, structured streaming and MLlib, high level API,SQL and DataFrames. Various optimization are added in this release.

- **Faster SQL queries:** In spark 3.0 SQL gain many improvements almost 46% of total improvements done in Spark 3.0 account for SQL engine innovation. In TPC-DS 30TB benchmark, Spark 3.0 is two times faster than Spark 2.4. (Spark Release 3.0.0 | Apache Spark, 2020)

- **Highlights in Spark 3.0**: adaptive query execution; dynamic partition pruning; ANSI SQL compliance; significant improvements in pandas APIs; new UI for structured streaming; 40x speedups for calling R user-defined functions; accelerator-aware scheduler; and SQL reference documentation.

- **Adaptive Query Execution:** Spark 2.2 added cost-based optimization to the existing rule based query optimizer. Spark 3.0 now has runtime adaptive query execution (AQE). With AQE, runtime statistics retrieved from completed stages of the query plan are used to re-optimize the execution plan of the remaining query stages. Databricks benchmarks yielded speed-ups ranging from 1.1x to 8x when using AQE. (McDonald et al., 2020)

- **Dynamic Partition pruning:** Spark 2.x static partition pruning improves performance by allowing Spark to read only a subset of the directories and files for queries that match partition filter criteria. Spark 3.0 brings this data pruning technique at runtime for queries that resemble data warehouse queries, which join a partitioned fact table with filtered values from dimension tables. Reducing the amount of data read and processed results in significant time savings. (McDonald et al., 2020)

Machine learning implementation

Data set

Data set used is a book recommendation data set goodbooks-10k. I downloaded this data set from Kaggle. Csv file used in ratings.csv. it has three columns book_id, user_id and rating all have integer values. Rating is oints given by a user to a book, this point can be between 1 to 5.

Link to the data source:

https://www.kaggle.com/zygmunt/goodbooks-10k

Collaborative filtering

Feature of the model

- Collaborative filtering is based on user behavior, it make suggestion depending on peoples liking and disliking. If people have disagreed in past then the will disagree in future as well. It look into past data of user behavior.
- If two users have exhibited similar preferences then assume that they are similar to each other in terms of taste.
- It is used for recommending movies, books, products that user might like by using known preferences of other users that exhibit similar behaviour.
- ALS algorithm is used to predict recommendations.

Key parameters

- numBlocks: the number of blocks the users and items will be partitioned into in order to parallelize computation (defaults to 10).
- rank: the number of latent factors in the model (defaults to 10).
- maxIter: the maximum number of iterations to run (defaults to 10)
- regParam: the regularization parameter in ALS (defaults to 1.0).
- implicitPrefs: whether to use the explicit feedback ALS variant or one adapted for implicit feedback data (defaults to false which means using explicit feedback).
- alpha: a parameter applicable to the implicit feedback variant of ALS that governs the baseline confidence in preference observations (defaults to 1.0).
- nonnegative: whether or not to use nonnegative constraints for least squares (defaults to false).

Configuration evaluation

Divided data into 60% training data and 40% testing data.

Used ASL model to train data and prediction

ASL algorithm results in Least square mean of 2.0.

Logistic Regression

Feature of the model

- Logistic regression is a popular method to predict a categorical response
- In spark.ml logistic regression can be used to predict a binary outcome by using binomial logistic regression
- is a probabilistic model
- its predictions are bounded between 0 and 1
- one of the most widely used linear classification models

Key parameters

- **weights** – Weights computed for every feature.
- **intercept** – Intercept computed for this model. (Only used in Binary Logistic Regression. In Multinomial Logistic Regression, the intercepts will not bea single value, so the intercepts will be part of the weights.)
- **numFeatures** – The dimension of the features.
- **numClasses** – The number of possible outcomes for k classes classification problem in Multinomial Logistic Regression. By default, it is binary logistic regression so numClasses will be set to 2.

Configuration evaluation

used Vector Assembler to transform data and added a label that have integer values. Vector assembler output a feature colume. Divided data into 80% training data and 20% testing data. Used Logistic Regression to train the data. Binary classification value resulted to 1.0.

Reference

- McDonald, C., McDonald, C., Evans, R., Lowe, J. and McDonald, C., 2020. *Optimizing And Improving Spark 3.0 Performance With Gpus* | *NVIDIA Developer Blog*. [online] NVIDIA Developer Blog. Available at: <https://developer.nvidia.com/blog/optimizing-and-improving-spark-3-0-performance-with-gpus/> [Accessed 19 October 2020].

- Spark.apache.org. 2020. *Spark Release 3.0.0* | *Apache Spark*. [online] Available at: <https://spark.apache.org/releases/spark-release-3-0-0.html#:~:text=In%20TPC%2DDS%2030TB%20benchmark,widely%20used%20language%20on%20Spark.&text=This%20release%20improves%20its%20functionalities,and%20more%20Pythonic%20error%20handling.> [Accessed 19 October 2020].

- DATAVERSITY. 2020. *Data Science Trends In 2020 - DATAVERSITY*. [online] Available at: <https://www.dataversity.net/data-science-trends-in-2020/#> [Accessed 19 October 2020].

- Medium. 2020. *Data Science Trends For 2020*. [online] Available at: <https://towardsdatascience.com/data-science-trends-for-2020-9b2ee27af499> [Accessed 19 October 2020].

- Seif, G., 2020. *The 4 Hottest Trends In Data Science For 2020 - Kdnuggets*. [online] KDnuggets. Available at: <https://www.kdnuggets.com/2019/12/4-hottest-trends-data-science-2020.html> [Accessed 19 October 2020].