Large scale data analysis – Spring 2024

Exam

The hand-in is a report submitted on LearnIT as a **pdf** file by

Friday, 31 May 2024, 14:00

The exam is composed of three sections, each with multiple questions. Each question can be answered within 0.5 - 1 page (250 - 500 words). You will get your grade based on your written answers to the questions. We encourage the use of code snippets and plots to elaborate on your answers. Code snippets do not count towards the total number of words per question.

Important: Your solutions and answers must be made by you and you only. This applies to program code, examples, tables, charts, and explanatory text that answer the exam questions. You are not allowed to create the exam solutions as group work, nor to consult with fellow students or any AI writing assistant tools, pose questions on internet fora, or the like. You are allowed to ask for clarification of possible mistakes, misprints, and so on, in the course LearnIT discussion forum. You should occasionally check the forum for news about mistakes and unclarity.

Your report **must** contain the following declaration:

I hereby declare that I have answered these exam questions myself without any outside help.

(Name) *(Date)*

A) Scalable data processing (35%)

The first assignment concerned a study that looked into the usage of the word "authentic" in Yelp reviews and how it might be signaling different characteristics from cuisine to cuisine.

- Explain how you found the businesses that have been reviewed by more than 5 influencer users using Spark's Dataframe API. Include screenshots from your query. (5%)
- Did you observe a difference in the amount of authenticity language used in the different areas? Explain the steps you took to arrive at the conclusion. Include a screenshot of the results (or sample of it). (10%)
- How did you test the hypothesis? Include your results. (10%)
- Which metric did you choose to evaluate your ML prediction model? Justify your choice and show your evaluation results. (10%)

B) ML lifecycle (40%)

In the second assignment one of the tasks was to implement a prediction pipeline for wind energy using weather data. You also designed a system that did model selection by experimental evaluation and packaged your final model in a format that ensures reproducibility across platforms.

- How did you align the data from the two data sources used in the modeling? Explain your choice and include screenshots from your code. (5%)
- Mention one feature transformation you implemented and argue why it is important. Include the code showing this transformation. (5%)
- Describe one example of data drift that can occur in the context of the assignment. (5%)
- Which models and/or hyperparameters did you experiment with using MLflow? Show the results of your experiments. (10%)
- Explain which steps you have taken to ensure that your wind prediction experiments are reproducible. Compare with different approaches. (15%)

C) Lecture material (25%)

- Briefly explain the four use cases discussed in the class where Wayang provides its biggest benefits? Could you think of any other use case? (5%)

- Assume a log analysis application where the input is logs from a web server and each line is of the form: <IP> -- <Date:time> <"URL_call">
  Example:

```
66.24.69.97 -- [17/May/2024:10:25:44 +0000] "GET http://www.google.com/bot.html"
66.24.69.97 -- [17/May/2024:10:26:44 +0000] "GET https://itu.dk/research.html"
66.24.69.97 -- [17/May/2024:10:28:44 +0000] "GET https://itu.dk/programmes/bds.html"
71.19.157.179 -- [17/May/2024:10:30:12 +0000] "GET http://www.google.com/faq.html"
66.24.69.97 -- [17/May/2024:31:10:44 +0000] "GET https://itu.dk/contact.html"
```

Write in pseudocode a Spark query using RDDs for outputting the total number of distinct IP addresses associated with calls to each domain page (e.g., http://www.google.com/faq.html and http://www.google.com/bot.html is the same domain page). Describe how Spark will work under the hood to answer this query and provide an illustration with 3 Spark workers. (10%)

- Generate a dataset of 500 3D points with 2 classes using make_blobs in sklearn; set the random state to be equal to your student number. Show a plot of this dataset. Use PCA, incremental PCA, and sparse PCA to reduce the dimensions to 2D; plot the data in each case. Briefly discuss when you would use incremental or sparse PCA. (5%)

- Identify two differences between distributed learning and federated learning in terms of datasets typically used. Explain how these differences were modeled in the experimental settings in the federated learning paper discussed in class. (5%)

Bonus question (extra 5%) - Describe your favourite concept, tool, technique, or algorithm from the class (that you haven't already discussed in the previous questions). Mention why you think it is important in the context of big data management.