Universität Heidelberg
Wintersemester 18/19

Group Parallel and Distributed Systems (PVS)
Artur Andrzejak, Diego Costa

Problem Set 5 for lecture Distributed Systems I (IVS1)

Due: 27.11.2018, 14:00 Uhr

Exercise 1 (9 Points)

In this exercise, we will continue the tutorial of Machine-Learning with PySpark[1] from Data Camp. In the Problem Set 3 - exercise 4, we have covered the content up to Data Exploration (inclusive). For this exercise we will be focusing on the following topics:

- Data Preprocessing
- Building A Machine Learning Model With Spark ML
- Evaluating the Model

At the end of the tutorial you should have a first-hand understanding on Spark MLlib. Aditionally, take a look at the data types documentation in MLlib [2] and answer the following questions:

a) What are the vector data type used for? What is the difference between the representation used on sparse and dense vector in Spark?

b) The tutorial introduced 3 new features not initially present in the dataset. Analyze the dataset and come up with two more features for your prediction task. Explain why such new features might be relevant for the prediction.

c) Run the code with your newly selected features. What features have most influence on the prediction? How can one measure the quality of a linear regression model prediction?

A big benefit of using ML Pipelines is to easily enable hyperparameter optimization. The current linear regression model still leaves room for improvement, which can be achieved by exploring how distinct set of hyperparameters afect the quality of your prediction.

d) First, modify your current implementation to use Spark pipelines[3]. Your pipeline should have only two stages: the scaler (StandardScaler) and the linear regression model.

e) Take a look at the example code on how to use CrossValidation[4] and incorporate it into your code. This hyper-parameter tuning algorithm allows you to easily explore different configurations for your linear regression model. Configure your CrossValidation to use the configuration grid specified in Table 1:

Hyperparameter	Values
regParam	[0.3, 0.1, 0.01]
fitIntercept	[False, True]
elasticNetParam	[0.0, 0.5, 0.8, 1.0]

Table 1: Hyperparameter to explore using the cross validation algorithm

[1] https://www.datacamp.com/community/tutorials/apache-spark-tutorial-machine-learning
[2] https://spark.apache.org/docs/2.3.0/mllib-data-types.html
[3] https://spark.apache.org/docs/latest/ml-pipeline.html#example-pipeline
[4] https://spark.apache.org/docs/latest/ml-tuning.html#cross-validation

f) Run your model with the crossvalidation algorithm. What is the best model configuration? Compare the coefficients and R2 metrics of your best model with the original version (from the tutorial). **Hint:** In PySpark, you can access the best model with *`crossvalModel.bestModel.stage[-1]`*.

Exercise 2 (3 Points)

Read the article *RESTful Web Services vs. Big Web Services: Making the Right Architectural Decision*[5] and answer the following questions:

a) What strengths and weaknesses do SOAP/WSDL-based web services (Big Web Services) have, according to the authors?

b) What is your own opinion with regards to strengths and weaknesses of RESTful Web Services?

c) Which decisions do software architects / developers have to make in the case of SOAP/WSDL-based web services and which ones in the case of RESTful web services?

Exercise 3 (2 Points)

There are different options to communicate across a distributed environment. For each of the following scenarios, pick one of these options: RPC (like Java RMI), web service (like RESTful WS) or a custom application layer protocol on top of TCP/IP. Explain your choices. Are there situations where your choice depends on what is on offer by third parties?

- A realtime online multiplayer game
- Incorporate current exchange rates for international currencies in a desktop calculator app
- A cross-platform distributed file system
- A gaming platform where human and AI contestants are supposed to compete in turn-based games (like chess) that are ruled and recorded by a central server
- A service that replies to messages including a bitmap image with a set of labels that apply for the content of the image
- A distributed computing platform that runs on a diverse grid of machines

Exercise 4 (2 Points)

In the presentation *P2P, DSM, and Other Products from the Complexity Factory*[6], researcher Willy Zwaenepoel criticises the academic research for its complexity biases. Read the slides and cites examples on technologies that have worned out due to its complexity. Why simple solutions tend to perdure in industry when complex approaches seem to dominate academia?

[5] http://www.jopera.org/files/www2008-restws-pautasso-zimmermann-leymann.pdf
[6] https://infoscience.epfl.ch/record/88080/files/Keynote_P2P%20DSM%20and%20Other%20Products%20from_v2_2.pdf