Advanced computational methods - Problem set 3

Hrvoje Stojic

February 15, 2016

0.1 Problem 1

We will do a logistic regression with PySpark on an AWS cluster of servers. You will use Bank marketing dataset from UCI repository. You will implement the regression through a chain of map-reduce operations to get some practice with such concepts, but also automatically, using the Spark's MLlib functions for linear regression, that hide all of these details.

This is not a big dataset and using a cluster for such a small dataset will result in poorer performance than working with it on a single computer. We could have worked with bigger datasets, but it would have involved a lot of waiting for some of the steps. It will serve well to illustrate the principles, after this you can easily start working with much bigger datasets.

- You will find the dataset in `Box/datasets/BankMarketing` folder. It is slightly modified version of the original dataset (you have a script there saying how exactly). I dropped categorical variables since these require special treatment and we will not cover it here.

- Log into AWS, inform yourself a bit about S3 storage solution, and create a bucket on S3 for this problem set. In the bucket create folders such as `data`, `bootstrap` and `log`.

- Upload the `bank.csv` text file to the data folder in your bucket. Upload the `Box/resources/install-jupyter-notebook-pySpark.sh` file to the boostrap folder.

- Install the AWS CLI (using python, not apt-get, see handout), configure them and then start the cluster with `aws emr create-cluster` command that I used to start the cluster during the class, you can find it below. Make sure that you are using the boostrap action option and that you have put correct paths to your bucket and your key-pair name. It is a small dataset so you do not need much resources. I recommend using at least 3 instances.
- Use the port forwarding method via SSH to access the Jupyter notebook port on the server and open your browser on appropriate port (see below). After accessing the notebook interface, upload the `PS3.ipynb` file that you can find in `Box/problemSets`.

- Follow the instructions in the notebook file to fill in the missing details.

- After you finish, download the final notebook, keeping it in the same `PS3.ipynb` format and put it in your github repository in folder `PS3`.

Command to launch EMR cluster on AWS via AWS CLI

```
aws emr create-cluster \
--name Spark-Jupyter-Cluster \
--ami-version 3.8.0 \
--instance-type m3.xlarge \
--instance-count 4 \
--applications Name=GANGLIA Name=SPARK,Args=[-g,-d,spark.executor.memory=10g] \
--bootstrap-actions Path=s3://your_bucket/bootstrap/install-jupyter-notebook-
pySpark.sh,Name=Install_Jupyter \
--region eu-west-1 \
--use-default-roles \
--ec2-attributes KeyName=myKey \
--enable-debugging \
--log-uri s3://your_bucket/Log/ \
--termination-protected
```

After the cluster is running, the notebook server runs on port 9999 (defined in boostrap script). You can connect to it by opening a tunnel from your local machine to your EMR master node.

```
ssh -i your_key_pair.pem -L 9000:localhost:9999 hadoop@master_public_DNS
```

You can find `master_public_DNS` in your EMR console on AWS, or via AWS CLI where instead of `j-someID` you should put the output of of `aws emr` command.

```
aws emr describe-cluster --region eu-west-1 --cluster-id j-someID
```

After you open the tunnel, open your browser and point to the following URL to access the notebook: `http://localhost:9000`.

Important details:

- Deadline is February 21, at 12h.
- Recall that you have to keep your code under version control in a single repository at Github, I will check your submission there. Place all the files for this problem set in a folder in this repo under the name "PS3".

- Please, pay attention to the names of the files I instruct you to use. Names facilitate examining the code greatly, there is 30 of you and it takes me far more time if everybody produces different set of files with different names.

- Create a simple text file with the name `Readme.md` in the problem set folder, if you want to leave me any instructions or messages regarding the problem set.

- I will give extra points for extra nice solutions!