Data Analytics Project (Project description)

1 Project Summary

1.1 Objective

Apply Spark to a machine learning exercise to understand the development process and common product outputs.

1.2 Method

Obtain a dataset and mining goal from any internet source. A contest site, like Kaggle.com, is helpful because it describes the data and a relatively simple problem (for simpler contests, such as less than $1000). You may solve expired contests and review the associated forums to gain insight into the process. (See the first week's notebook on data for further details on how to get data)

1.3 Approval

Send the instructor a link to the dataset along with a very brief summary of what you intend to complete to obtain approval. This is intended simply as a check to ensure that the dataset is of the appropriate size.

1.4 Deliver

To obtain credit, deliver a Python Spark notebook that is:

- fully inclusive (Run All is all that is required, including dataset download and processing)
- significantly uses the parallel processing capabilities of Spark to process the data
- easy to understand (i.e., provides a tutorial for the reader)

1.5 Advice

Simple is good. Focus on a simple prediction problem and solve it using a simple method (e.g., Logistic Regression). Only after completing a simple analysis go onto a more models.

1.6 Examples

There are a few project example on our web site, under project-examples folder. They include Logistic Regression, Random Forrest (decision trees), and sentiment analysis. Databricks provides a good example of graph analysis, with their On-Time Flight Performance with GraphFrames.

1.7 Project Ideas

The following sites provide examples of the kinds of problems that may be considered (in addition to the contest sites):

- http://1000projects.org/projects/data-mining-projects/
- http://www.galitshmueli.com/student-projects
- http://searchbusinessanalytics.techtarget.com/feature/Simple-data-mining-examples-and-datasets
- https://web.stanford.edu/class/cs345a/slides/06-projects.pdf

1.8 Challenge Projects

See the challenge_projects folder in our resources folder for some challenge projects.

2 Deliverables

2.1 Proposal

This proposal is simply to allow for a discussion between the teams and the instructor as a means to focus the projects so that they are successful.

Send an email to the instructor describing your group's proposed project. In the email, describe the following:

1. Team members
2. Name of projects
3. Sources of data
    a. Include an estimation of the number of rows/records you will be working with
4. Kind of mining task(s)
    a. E.g., prediction, clustering, classification, etc.
5. Key visualizations that you will include
6. Difficulties you anticipate

2.2 M1 Evaluation

Factor	Criteria	Comments	Points
Working			30
	Notebook entirely self-contained (no need to upload)		
	Notebook executes all cells without error		
Objective			5
	Project goal explained		
Data collected			5
	Data is downloaded w/i notebook		
	Data lineage explained		
Data explained			5
	Key fields are explained		
	Predictor field is explained		
	Data is summarized w/ stats & graphs		
Data cleaning			5
	Data is cleaned as necessary		
	Cleaning is explained		
Data transformation			5
	Data is transformed for modeling		
	Uncommon transformations are explained		
Data modeling			15

	Data is modeled (using Spark APIs)		
	Minimum of 1 model		
	Param grid		
	Cross validation		
	Uncommon models are explained		
Prediction (or discovery)			10
	Prediction or discovery from model		
	Result explained w/r to project goal		
Model eval			10
	Model is evaluated		
	Result explain		
Visualization			5
	Visualization of model eval		
	Visualization of prediction (as appropriate)		
Professionalism			5
	Headings		
	Use of mark down to explain		
	Inclusion of images as necessary		
	Professional looking (readable, fonts, typos, etc)		
	Good enough to send to prospective employer		
Summary			
Total			100

2.3 M2 Evaluation

Working			20	
	Notebook entirely self-contained (no need to upload)			
	Notebook executes all cells without error			
	Includes (by wget, addFile) a custom Python code from a file on your own GitHub (Dropbox, etc)			
Objective			5	

	Project goal explained			
Data collected			5	
	Data is downloaded w/i notebook			
	Data lineage explained			
Data explained			5	
	Key fields are explained			
	Predictor field is explained			
	Data is summarized w/ stats & graphs			
Data cleaning			5	
	Data is cleaned as necessary			
	Cleaning is explained			
Data transformation			5	
	Data is transformed for modeling			
	Uncommon transformations are explained			
Data modeling			10	
	Data is modeled (using Spark APIs)			
	Minimum of 3 models			
	Param grid			
	Cross validation			
	Uncommon models are explained			
Prediction (or discovery)			10	
	Prediction or discovery from model			
	Result explained w/r to project goal			
Model eval			20	
	Models are evaluated			
	Model evals are compared			
	Result explain			
	Recommendation is given			
Visualization			10	
	Visualization of model eval			
	Visualization of prediction (as appropriate)			

Professionalism			5	
	Headings			
	Use of mark down to explain			
	Inclusion of images as necessary			
	Professional looking (readable, fonts, typos, etc)			
	Good enough to send to prospective employer			
Summary				
Subtotal			100	100
Extra credit	One or more of the following, for maximum of 15 extra credit points			
	Extraordinary visualizations		5	
	Extraordinary modeling		5	
	Create ML app from PySpark		5	
	Place app in Docker stack		5	
	Introduce & explain (in detail) the use of a Spark API that was excluded from course. (You must use it in your analysis.)		5	
Grand Total			115	100

3 Presentation

At the end of the term, you will present your project to the class. The presentation is meant to demonstrate your project as well as your technical presentation skills.

3.1 Format

Deliver a brief presentation, lasting about 10 minutes, which explains your project. You may present your project in any format you deem appropriate; however, a typical presentation structure is as follows.

- Introduction of team, project, data, models, results, and conclusions using PowerPoint
- Quick overview of the sections in the notebook used in your computations
- Explanation of one part of the project notebook that is most interesting, novel, useful, etc.
- Concluding slide asking for questions

3.2 Evaluation

Criteria	Points
Timely	15
Explains project goals, methods, results	65
Conveys something interesting about the project	10
Professional is style and substance	10
total	100