Case Study 1 – Classification and Regression Algorithms

Instructions

- The goal of this assignment is to try out the various Spark algorithms in Scala an Python and to compare and contrast the various linear model algorithms using model evaluation criteria
- Develop algorithms in Python and Scala and submit the Scala code and Python code with instructions to run the code
- Write an executive summary report answering the asked questions and compare and contrast the various algorithms.
- The assignment is due Midnight: 6/21/2015.
- Zip the source file, instructions and executive summary and put in google drive and share it with analyticsneu@gmail.com

Problem statement

We will use the white vinho verde wine samples data, from the north of Portugal. The goal is to model wine quality based on physicochemical tests (see [Cortez et al., 2009], [Web Link]).Use Python and Apache Spark's MLLib capabilities and use Python and Scala to try out the following Classification and Regression Algorithms. We will use the Wine data set available at: https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv

The names are available at: https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality.names . We will first try out the algorithms in Python using SciPy's capabilities and tryout the following models

Type	Models	Regularization (L0,L1,L2)
Classification	sklearn.linear_model.SGDClassifier (SVM => Hinge loss)	
	sklearn.linear_model.SGDClassifier (Logarithmic regression => log loss)	
	sklearn.linear_model.LogisticRegression (LBFGS version)	
Regression	sklearn.linear_model.LinearRegression	
	sklearn.linear_model.SGDRegressor	
	sklearn.linear_model.Ridge	
	sklearn.linear_model.Lasso	

Then, use Apache Spark and try out the following models.

Type	Algorithms	PySpark Regularization (L0,L1,L2)	Scala Regularization (L0,L1,L2)
Classification	SVMWithSGD		
	LogisticRegressionWithLBFGS		
	LogisticRegressionWithSGD		
Regression	LinearRegressionWithSGD		
	RidgeRegressionWithSGD		
	LassoWithSGD		

Questions:

1. Summarize your implementation and provide instructions to run your code in

   a. in plain Python

   b. Pyspark

   c. Scala

Ensure code is well documented and self-explanatory. Comment on configuration, results and implementation ease. Comment on the various loss functions used and regularizers.

2. What performance metrics did you implement and use to evaluate Classification algorithms?

3. What performance metrics did you implement and use to evaluate Regression algorithms?

4. Compare and contrast using Just Scipy libraries vs using Apache Spark. How did the results vary? When would you use just Python libraries and when would you use Apache Spark?