GUIDE TO MAPREDUCE AND SPARK (UNITS 5 AND 6)

COURSE: MACHINE LEARNING I

MASTER IN BIG DATA ANALYTICS

Ricardo Aler Mur

About MAPREDUCE:

- First it is explained what is meant by large scale machine learning, and shown that there are several ways in which machine learning algorithms can be parallelized: task, data, and pipeline parallelism
- Some examples of task parallelism are commented (mainly, embarrasing parallelism or obvious parallelism).
- But the main kind of parallelism that is used nowadays is data parallelism.
- One of the main paradigms for data parallelism is MapReduce
- MapReduce is particularly useful when hundreds or thousands of computers connected via a network are available, and data can be partitioned into the different computers. The main idea of MapReduce is not to move the data, but to move the processes to where data is located.
- The MapReduce model is explained by explaining its main processes: map, sort and shuffle, and reduce. An example for counting words is explained. The combiner functions are explained to increase efficiency.

- Three algorithms are programmed in the MapReduce model:
    - KNN
    - Decision trees (by distributing the computation of the entropy function)
    - The clustering algorithm k-means
- Finally, it is explained that nowadays data parallelism is moving towards a new programming model called Spark, although many of the MapReduce ideas are valid for Spark.

About SPARK:

- The limitations of the MapReduce programming model are explained, and Spark is shown to solve them.
- Basic concepts are introduced, specially the RDD (Resilient Distributed Dataset) and the concept of transformation and action.
- Transformations transform a RDD into another RDD, but its execution is lazy. That means that noting happens when the transformation is applied.
- Only when an action is executed, all the transformations are actually applied and run.
- Some examples of transformations and actions are explained.
- A more complex RDD is introduced: pair RDDs, where every instance constains a key and a value.
- Some specific transformations for pair RDDs are explained: *reduceByKey and flatMap.*
- It is shown that the MapReduce programming model can be programmed in Spark with Map and ReduceByKey.
- Two of the main Spark libraries for Machine Learning are introduced: Mlib and ML, the latter being the most recent one. Mlib relies on RDDs and the LabelledPoint data type, while ML relies on a more complex data strcture called DataFrame.
- The non-supervised K-means algorithm is explained now within the Spark programming model.

Associated material

In addition to the lecture slides and some exercises (check the course guide), a new assignment with pySpark is proposed. This assignment has two parts. The first part asks the student to complete a programming exercise with spark, taking advantage of its data

paralellism. The second assignment addresses a Machine Learning problem, but this time, the resources available in Spark will be used.