DSC 102
Systems for Scalable Analytics

Arun Kumar

Exercise 4

Time tip: Roughly 45sec to 1min per 1pt
Tips on MapReduce Problems

General approach/advice on how to cast a data analytics computation onto the MapReduce API:

Step 1) Identify the exact data access pattern of the computation over the dataset. Draw it out to see it visually if you like.

Step 2) Identify how to decompose the bulk of the whole computation into a bunch of independent chunk computations on sub-elements (rows/columns/tiles). Typically, scalability along rows is the most preferable because most modern large-scale datasets have large numbers of rows.

Step 3) Identify how to aggregate those decomposed parts to get the final result as if it was computed in a single-threaded in-RAM manner. This aggregation step may not always be needed though.

Step 4) Align the sharding with Step 2. Put the independent chunk computations in the Mapper. Identify what the Mapper’s intermediate output (emit) data structure should be. Put the aggregation in Step 3 and any post processing in the Reducer.
Q1) [6pts] Write pseudocode for (or just describe precisely in prose) a MapReduce job to compute the Frobenius norm (aka L2 norm) of a given large matrix. It should be scalable along the number of rows. Make sure to explain your assumption on how the dataset is stored/sharded to begin with.
Exercise

Q2) [8pts] Write pseudocode for (or just describe precisely in prose) a single MapReduce job to compute the Person correlation matrix of a given data matrix. It should be scalable along the number of rows of the dataset.

Make sure to explain your assumption on how the dataset is stored/sharded to begin with.

Hint: The Pearson correlation coefficient between two variables $x$ and $y$ can be rewritten as follows (given $n$ rows):

$$ r_{x,y} = \frac{n \sum_{i=1}^{n} (x_i y_i) - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n \left( \sum_{i=1}^{n} x_i^2 \right) - \left( \sum_{i=1}^{n} x_i \right)^2} \sqrt{n \left( \sum_{i=1}^{n} y_i^2 \right) - \left( \sum_{i=1}^{n} y_i \right)^2} } $$
Q3) [8pts] Write pseudocode for (or just describe precisely in prose) a MapReduce job to compute the Gramian of a given large matrix. It should be scalable along the number of rows.

Make sure to explain your assumption on how the dataset is stored/sharded to begin with.
Q4) Suppose you are given a large dataset with 50 numeric and 9 categorical features (domain size of 50 each). The HDFS file size is 3 TB.

A. [10pts] Write pseudocode for (or just describe precisely in prose) MapReduce job(s) to compute this dataset’s correlation matrix. Hint: You would need could do two separate MapReduce jobs.

B. [4pts] What is the rough total disk I/O cost of the above in TB? Include both reads and writes of intermediate data and output.

C. [4pts] Briefly explain how you would scale this computation on an on-premise cluster.

D. [4pts] Briefly explain how you would scale this computation on AWS.
Q5) [3 x 3pts] Suppose you are given a large dataset file for ML training that is of size 120 GB. What is the lowest possible I/O cost (in GB) of each of the following feature engineering operations? Ignore final output write costs and any potential gains due to caching.

A. Quadratic (order 2) feature interactions
B. Binning a numeric feature into 10 given intervals
C. Whitening a numeric feature
D. One-hot encoding of a categorical feature (assume feature’s domain has only 5000 unique values)
Q6) [4 x 8pts] Write pseudocode (or just describe precisely) using MapReduce/Spark operations to perform the following data science operations at scale:

A. Quadratic (order 2) feature interactions
B. Binning a numeric feature with given bins
C. Whitening a numeric feature
D. One-hot encoding of a categorical feature (assume feature’s domain has only 5000 unique values and is given)
Q7) [3pts] Which of the following hyperparameter tuning approaches is the most popular in practice as per surveys?

A. Grid search
B. Random search
C. Hyperband
D. All of A, B, C
E. None of the above
Q8) [4pts] Suppose you are performing model selection for a RandomForest model. For hyper-parameter tuning, you try 3 values of number of trees and 4 values of maximum tree height. To aid your interpretability, you also explore 5 different manually created subsets of features apart from the full feature set.

What is the total number of models built in this model selection workload?
Exercise

(Advanced/Extra Credit; Optional)

Q9) [10pts] You are given a large training dataset of \((Y,X_1,X_2)\) examples on HDFS for binary classification (i.e., \(Y = 0\) or \(1\)) with two categorical features \(X_1\) and \(X_2\). The domains of the features are known beforehand as \(D_{X1}\) and \(D_{X2}\) and have only tens of unique values.

Write pseudocode for a single MapReduce job to train a Naive Bayes model. It should be scalable along the number of rows.

Make sure to explain your assumption on how the dataset is stored/sharded to begin with.

Hint: Naive Bayes training only needs to estimate the distribution \(P(Y)\) and all class-conditional probability distributions \(P(X_i|Y)\) using frequency counts.