CSE 234
Data Systems for Machine Learning

Arun Kumar

Exercise 2
General approach/advice on how to cast a statistical/data analysis/ML computation onto the MapReduce / UDA 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 4a) For MapReduce: 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.

Step 4b) For UDA: Identify what the agg. state is based on Step 2. Put the independent chunk computations in Transition. Usually, a chunk is a tuple at a time. Put the aggregation in Step 3 in Merge. Put any postprocessing in Finalize.
Exercise

Q1) [7pts] (Quiz 1) Write pseudocode for 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.
Q2) [7pts] (From Quiz 1) Assume you are given a large matrix stored as a table with rows as tuples and columns as attributes.

Write pseudocode for a UDA to compute the Gramian of the matrix. It should be scalable along the number of rows.

Make sure to explain your aggregation state.
Exercise

Q3) [7pts] Write pseudocode for 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

Q4) [7pts] Assume you are given a large matrix stored as a table with rows as tuples and columns as attributes.

Write pseudocode for a UDA to compute the Frobenius norm (aka L2 norm) of the matrix. It should be scalable along the number of rows.

Make sure to explain your aggregation state.
Exercise

Q5) [10pts] You are given a large training dataset of (Y,X1,X2) examples on HDFS for binary classification (i.e., Y = 0 or 1) with two categorical features X1 and X2. The domains of the features are known beforehand as DX1 and DX2 and have only tens of unique values.

Write pseudocode for a 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(Xi|Y) using frequency counts.
Exercise

Q6) [10pts] You are given a large training dataset of (Y,X1,X2) examples as a table for binary classification (i.e., Y = 0 or 1) with two categorical features X1 and X2. Each of Y and Xi are stored as a separate attribute. The domains of the features are known beforehand as DX1 and DX2 and have only tens of unique values.

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

Make sure to explain your aggregation state.

Hint: Naive Bayes training only needs to estimate the distribution P(Y) and all class-conditional probability distributions P(Xi|Y) using frequency counts.
Q7) [10pts] Write pseudocode for 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):

\[ r_{x,y} = \frac{n \sum_{i=1}^{n} (x_i y_i) - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2} \sqrt{n(\sum_{i=1}^{n} y_i^2) - (\sum_{i=1}^{n} y_i)^2}} \]
Q8) [10pts] Assume you are given a large data matrix stored as a table with rows as tuples and columns as attributes.

Write pseudocode for a single UDA to compute the Person correlation matrix. It should be scalable along the number of rows of the data matrix.

Make sure to explain your aggregation state.

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

$$r_{x,y} = \frac{n \sum_{i=1}^{n} (x_i y_i) - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n \left( \sum_{i=1}^{n} x_i^2 \right) - (\sum_{i=1}^{n} x_i)^2} \sqrt{n \left( \sum_{i=1}^{n} y_i^2 \right) - (\sum_{i=1}^{n} y_i)^2}}$$