Parallel and Distributed Computing
Lab7
SparkML

In this lab you will work with Spark's Machine Learning package, SparkML.

The Dataset

The data we will use is on heart disease and comes from the UCI (https://archive.ics.uci.edu/ml/datasets/Heart+Disease). This data has been modified for the lab. There are two files, heartTraining.csv and heartTesting.csv. Both files contain a person id, age, sex, and cholesterol level. They both also have a field for a prediction of heart disease or not. In the training set these are actual values, but in the testing set they are all set to a value of no.

Building the Pipeline

First you will build the ML pipeline on the training data. This pipeline ends with a simple logistical regression on the data to determine if the person has heart disease or not.

Age should be broken into categories: below 40, 40-49, 50-59, 60-69, 70 and above.

Cholesterol should be left as a number.

The Sex and Prediction string fields will obviously need to be turned into numbers.

Train (fit) the pipeline on the training data.

Testing Data

Once you have the pipeline model built, test it on the testing data. As a final result you should display the id, probability and prediction data (just the first 20 lines are fine).

You should also test it on the original training data to see how well the regression predictions match the human predictions. With only these limited factors (age, sex, cholesterol) and no tuning of the regression, the predictions will not be highly accurate. But for this lab, it will be fine. Please include this measurement with your results.

Submission

You are upload 2 documents as your submission. The first is your Python source code (PY). Make sure to have your name in a comment at the top of your source code. Second is a document that has a cut-paste of your results from running your code.