**Mining-Assignment#1**

**Raja Mummidi**

**M08802112**

Procedure:

Taken the Letter Recognition dataset from UCI Library. Removed to the first column from the dataset and split the dataset into 4 equal parts of the length 5000 records each. Given each file to mapper function and calculated co-variance matrix, mean and number of records for each file. The output from the mapper function given to reducer function and merged covariance of two files at a time. Finally merged all outputs from Reducer function to calculate final co-variance matrix.

Small g =

function [covMat,meanMat,n]=Mapperfn(dataset)

covMat=cov(dataset);

meanMat = mean(dataset)

n=size(dataset,1);

end

G is

function [mergeCovMat,meanMat,N]=

Reducerfn(covMat1,covMat2,meanMat1,meanMat2,n1,n2)

N= n1+n2;

meanMat = (n1\*meanMat1 + n2\*meanMat2) / N;

mergeCovMat = (n1\*covMat1 + n2\*covMat2

+ n1\*(meanMat1-meanMat)'\*(meanMat1-meanMat)

+ n2\*(meanMat2-meanMat)'\*(meanMat2-meanMat))/ N;

end

**Results:**

Covariance using map reduce paradigm and covariance using inbult cov function.

cov(orignalMat)-finalCovMat

Columns 1 through 11

-0.0005 -0.0007 -0.0005 -0.0004 -0.0004 0.0000 -0.0000 -0.0000 -0.0000 -0.0001 -0.0000

-0.0007 -0.0016 -0.0007 -0.0009 -0.0006 -0.0000 0.0000 0.0000 -0.0001 -0.0002 0.0001

-0.0005 -0.0007 -0.0006 -0.0005 -0.0005 -0.0000 -0.0000 0.0001 -0.0000 -0.0001 -0.0000

-0.0004 -0.0009 -0.0005 -0.0008 -0.0005 -0.0000 0.0000 -0.0001 -0.0000 -0.0000 0.0000

-0.0004 -0.0006 -0.0005 -0.0005 -0.0007 -0.0001 0.0000 0.0000 0.0001 0.0001 0.0001

0.0000 -0.0000 -0.0000 -0.0000 -0.0001 -0.0006 0.0003 0.0000 0.0001 -0.0001 0.0003

-0.0000 0.0000 -0.0000 0.0000 0.0000 0.0003 -0.0008 0.0001 0.0000 -0.0002 -0.0006

-0.0000 0.0000 0.0001 -0.0001 0.0000 0.0000 0.0001 -0.0011 0.0002 0.0003 -0.0000

-0.0000 -0.0001 -0.0000 -0.0000 0.0001 0.0001 0.0000 0.0002 -0.0009 -0.0001 0.0001

-0.0001 -0.0002 -0.0001 -0.0000 0.0001 -0.0001 -0.0002 0.0003 -0.0001 -0.0009 -0.0001

-0.0000 0.0001 -0.0000 0.0000 0.0001 0.0003 -0.0006 -0.0000 0.0001 -0.0001 -0.0010

0.0000 0.0000 0.0000 -0.0000 0.0000 0.0000 0.0002 -0.0001 -0.0001 0.0001 -0.0001

-0.0003 -0.0003 -0.0004 -0.0002 -0.0005 -0.0001 0.0000 -0.0001 0.0003 0.0002 -0.0000

-0.0000 0.0000 -0.0000 -0.0000 -0.0000 0.0001 -0.0003 0.0001 0.0000 -0.0000 -0.0003

-0.0002 -0.0003 -0.0002 -0.0003 -0.0004 -0.0001 0.0001 -0.0000 -0.0003 0.0001 0.0002

0.0000 0.0000 0.0001 0.0000 0.0000 -0.0001 0.0001 -0.0001 0.0000 0.0001 0.0002

Columns 12 through 16

0.0000 -0.0003 -0.0000 -0.0002 0.0000

0.0000 -0.0003 0.0000 -0.0003 0.0000

0.0000 -0.0004 -0.0000 -0.0002 0.0001

-0.0000 -0.0002 -0.0000 -0.0003 0.0000

0.0000 -0.0005 -0.0000 -0.0004 0.0000

0.0000 -0.0001 0.0001 -0.0001 -0.0001

0.0002 0.0000 -0.0003 0.0001 0.0001

-0.0001 -0.0001 0.0001 -0.0000 -0.0001

-0.0001 0.0003 0.0000 -0.0003 0.0000

0.0001 0.0002 -0.0000 0.0001 0.0001

-0.0001 -0.0000 -0.0003 0.0002 0.0002

-0.0006 0.0000 0.0001 -0.0000 -0.0001

0.0000 -0.0008 -0.0000 -0.0001 0.0000

0.0001 -0.0000 -0.0004 0.0000 0.0001

-0.0000 -0.0001 0.0000 -0.0010 -0.0001

-0.0001 0.0000 0.0001 -0.0001 -0.0004

**Observations:**

There is no much difference in original covariance and its counterpart using map reduce paradigm. So, calculating covariance can be achieved by Map Reduce.

**Code:**

**Main function :**

%Reading the dataset into table

tblCharRecDS= readtable('original-letter-recognition.csv','ReadVariableNames',false);

C=table2cell(tblCharRecDS);

C=C(:,2:17);

orignalMat=cell2mat(C);

%Partitioning the dataset

[file1,file2,file3,file4]=DatasetPartition(orignalMat);

[covMat1,meanMat1,n1]=Mapperfn(file1);

[covMat2,meanMat2,n2]=Mapperfn(file2);

[covMat3,meanMat3,n3]=Mapperfn(file3);

[covMat4,meanMat4,n4]=Mapperfn(file4);

[mergeCovMat1,mergeMeanMat1,N1]=Reducerfn(covMat1,covMat2,meanMat1,meanMat2,n1,n2);

[mergeCovMat2,mergeMeanMat2,N2]=Reducerfn(covMat3,covMat4,meanMat3,meanMat4,n3,n4);

[finalCovMat,finalMeanMat,N]=Reducerfn(mergeCovMat1,mergeCovMat2,mergeMeanMat1,mergeMeanMat2,N1,N2);

cov(orignalMat)-finalCovMat

**DatasetPartition:**

function [file1,file2,file3,file4]=DatasetPartition(orignalMat)

rows=size(orignalMat,1);

randIdx=randperm(rows);

file1Idx=randIdx(1,1:5000);

file2Idx=randIdx(1,5001:10000);

file3Idx=randIdx(1,10001:15000);

file4Idx=randIdx(1,15001:20000);

file1=orignalMat(file1Idx,:);

file2=orignalMat(file2Idx,:);

file3=orignalMat(file3Idx,:);

file4=orignalMat(file4Idx,:);

end

**Mapperfn**

function [covMat,meanMat,n]=Mapperfn(dataset)

covMat=cov(dataset);

meanMat = mean(dataset)

n=size(dataset,1);

end

**Reducerfn**

function [mergeCovMat,meanMat,N]=Reducerfn(covMat1,covMat2,meanMat1,meanMat2,n1,n2)

N= n1+n2;

meanMat = (n1\*meanMat1 + n2\*meanMat2) / N;

mergeCovMat = (n1\*covMat1 + n2\*covMat2 + n1\*(meanMat1-meanMat)'\*(meanMat1-meanMat) + n2\*(meanMat2-meanMat)'\*(meanMat2-meanMat))/ N;

end