Manual of Frovedis Python API

Contents

Chapter 1

Introduction

This manual contains Python API documantation. If you are new to Frovedis, please read the tutorial_python first.

Currently we only provide part of the API documentation. We are still updating the contents.

- Matrix
  - FrovedisDvector
  - FrovedisCRSMatrix
  - FrovedisBlockcyclicMatrix
  - pblas_wrapper
  - scalapack_wrapper
  - arpack_wrapper
  - getrf_result
  - gesvd_result
- Machine Learning
  - Linear Regression
  - Lasso Regression
  - Ridge Regression
  - Logistic Regression
  - Linear SVM
  - Matrix Factorization using ALS
  - kmeans

Chapter 2

FrovedisDvector

2.1 NAME

FrovedisDvector - A data structure used in modeling the in-memory dvector data of frovedis server side at client python side.

2.2 SYNOPSIS

class frovedis.matrix.dvector.FrovedisDvector(vec=None)

2.2.1 Public Member Functions

load (vec)
load_numpy_array (vec)
debug_print()
release()

2.3 DESCRIPTION

FrovedisDvector is a pseudo data structure at client python side which aims to model the frovedis server side `dvector<double>` (see manual of frovedis dvector for details).

Note that the actual vector data is created at frovedis server side only. Python side FrovedisDvector contains a proxy handle of the in-memory vector data created at frovedis server, along with its size.

2.3.1 Constructor Documentation

2.3.1.1 FrovedisDvector (vec=None)

Parameters
vec: It can be any python array-like object or None. In case of None (Default), it does not make any request to server.

Purpose

This constructor can be used to construct a FrovedisDvector instance, as follows:

```
v1 = FrovedisDvector()          # empty dvector, no server request is made
v2 = FrovedisDvector([1,2,3,4]) # will load data from the given list
```

Return Type

It simply returns "self" reference.

2.3.2 Pubic Member Function Documentation

2.3.2.1 load (vec)

Parameters
vec: It can be any python array-like object (but not None).

Purpose

This function works similar to the constructor. It can be used to load a FrovedisDvector instance, as follows:

```
v = FrovedisDvector().load([1,2,3,4]) # will load data from the given list
```

Return Type

It simply returns "self" reference.

2.3.2.2 load_numpy_array (vec)

Parameters
vec: Any numpy array with values to be loaded in.

Purpose
This function can be used to load a python side numpy array data into frovedis server side dvector. It accepts a python numpy array object and converts it into the frovedis server side dvector whose proxy along size information are stored in the target FrovedisDvector object.

Return Type
It simply returns "self" reference.

2.3.2.3 size()

Purpose
It returns the size of the dvector

Return Type
An integer value containing size of the target dvector.

2.3.2.4 debug_print()

Purpose
It prints the contents of the server side distributed vector data on the server side user terminal. It is mainly useful for debugging purpose.

Return Type
It returns nothing.

2.3.2.5 release()

Purpose
This function can be used to release the existing in-memory data at frovedis server side.

Return Type
It returns nothing.

2.3.2.6 FrovedisDvector.asDvec(vec)

Parameters
vec: A numpy array or python array like object or an instance of FrovedisDvector.

Purpose
This static function is used in order to convert a given array to a dvector. If the input is already an instance of FrovedisDvector, then the same will be returned.

Return Type

An instance of FrovedisDvector.

Chapter 3

FrovedisCRSMatrix

3.1 NAME

FrovedisCRSMatrix - A data structure used in modeling the in-memory crs matrix data of frovedis server side at client python side.

3.2 SYNOPSIS

class frovedis.matrix.sparse.FrovedisCRSMatrix(mat=None)

3.2.1 Public Member Functions

load (mat)
load_scipy_matrix (mat)
load_text (filename)
load_binary (dirname)
save_text (filename)
save_binary (dirname)
debug_print()
release()

3.3 DESCRIPTION

FrovedisCRSMatrix is a pseudo matrix structure at client python side which aims to model the frovedis server side `crs_matrix<double>` (see manual of frovedis crs_matrix for details).

Note that the actual matrix data is created at frovedis server side only. Python side FrovedisCRSMatrix contains a proxy handle of the in-memory matrix data created at frovedis server, along with number of rows and number of columns information.

3.3.1 Constructor Documentation

3.3.1.1 FrovedisCRSMatrix (mat=None)

Parameters
mat: It can be a string containing filename having text data to be loaded, or any scipy sparse matrix or any python array-like object or None. In case of None (Default), it does not make any request to server.

Purpose

This constructor can be used to construct a FrovedisCRSMatrix instance, as follows:

```
mat1 = FrovedisCRSMatrix() # empty matrix, no server request is made
mat2 = FrovedisCRSMatrix("./data") # will load data from given text file
mat3 = FrovedisCRSMatrix([1,2,3,4]) # will load data from the given list
```

Return Type

It simply returns "self" reference.

3.3.2 Pubic Member Function Documentation

3.3.2.1 load (mat)

Parameters
mat: It can be a string containing filename having text data to be loaded, or any scipy sparse matrix or any python array-like object (but it can not be None).

Purpose

This works similar to the constructor.
It can be used to load a FrovedisCRSMatrix instance, as follows:

```
mat1 = FrovedisCRSMatrix().load("./data")  # will load data from given text file
mat2 = FrovedisCRSMatrix().load([1,2,3,4]) # will load data from the given list
```

Return Type

It simply returns "self" reference.

3.3.2.2 load_scipy_matrix (mat)

Parameters
mat: Any scipy matrix with values to be loaded in.

Purpose
This function can be used to load a python side scipy sparse data matrix into frovedis server side crs matrix. It accepts a scipy sparse matrix object and converts it into the frovedis server side crs matrix whose proxy along with number of rows and number of columns information are stored in the target FrovedisCRSMatrix object.

Return Type
It simply returns "self" reference.

3.3.2.3 load_text (filename)

Parameters
filename: A string object containing the text file name to be loaded.

Purpose
This function can be used to load the data from a text file into the target matrix. Note that the file must be placed at server side at the given path and it should have contents stored in libSVM format, i.e., "column_index:value" at each row (see frovedis manual of make_crs_matrix_load() for more details).

Return Type
It simply returns "self" reference.

3.3.2.4 load_binary (dirname)

Parameters
dirname: A string object containing the directory name having the binary data to be loaded.

Purpose
This function can be used to load the data from the specified directory with binary data file into the target matrix. Note that the file must be placed at server side at the given path.

Return Type
It simply returns "self" reference.

3.3.2.5 save_text (filename)

Parameters
filename: A string object containing the text file name in which the data is to be saveed.

Purpose

This function is used to save the target matrix as text file with the filename at the given path. Note that the file will be saved at server side at the given path.

Return Type
It returns nothing.

3.3.2.6 save_binary (dirname)

Parameters
dirname: A string object containing the directory name in which the data is to be saveed as little-endian binary form.

Purpose

This function is used to save the target matrix as little-endian binary file with the filename at the given path. Note that the file will be saved at server side at the given path.

Return Type
It returns nothing.

3.3.2.7 numRows()

Purpose
It returns the number of rows in the matrix

Return Type
An integer value containing rows count in the target matrix.

3.3.2.8 numCols()

Purpose
It returns the number of columns in the matrix

Return Type
An integer value containing columns count in the target matrix.

3.3.2.9 debug_print()

Purpose
It prints the contents of the server side distributed matrix data on the server side user terminal. It is mainly useful for debugging purpose.

Return Type
It returns nothing.

3.3.2.10 release()

Purpose
This function can be used to release the existing in-memory data at frovedis server side.

Return Type
It returns nothing.

3.3.2.11 FrovedisCRSMatrix.asCRS(mat)

Parameters
mat: A scipy matrix, an instance of FrovedisCRSMatrix or any python array-like data.

Purpose
This static function is used in order to convert a given matrix to a crs matrix. If the input is already an instance of FrovedisCRSMatrix, then the same will be returned.

Return Type

An instance of FrovedisCRSMatrix.

Chapter 4

FrovedisBlockcyclicMatrix

4.1 NAME

FrovedisBlockcyclicMatrix - A data structure used in modeling the in-memory blockcyclic matrix data of frovedis server side at client python side.

4.2 SYNOPSIS

class frovedis.matrix.dense.FrovedisBlockcyclicMatrix(mat=None)

4.2.1 Overloaded Operators

operator= (mat)
operator+ (mat)
operator- (mat)
operator* (mat)
operator~ (mat)

4.2.2 Public Member Functions

load (mat)
load_numpy_matrix (mat)
load_text (filename)
load_binary (dirname)
save_text (filename)
save_binary (dirname)
transpose()
to_numpy_matrix ()
debug_print()
release()

4.3 DESCRIPTION

FrovedisBlockcyclicMatrix is a pseudo matrix structure at client python side which aims to model the frovedis server side `blockcyclic_matrix<double>` (see manual of frovedis blockcyclic_matrix for details).

Note that the actual matrix data is created at frovedis server side only. Python side FrovedisBlockcyclicMatrix contains a proxy handle of the in-memory matrix data created at frovedis server, along with number of rows and number of columns information.

4.3.1 Constructor Documentation

4.3.1.1 FrovedisBlockcyclicMatrix (mat=None)

Parameters
mat: It can be a string containing filename having text data to be loaded, or another FrovedisBlockcyclicMatrix instance for copy or any python array-like object or None. In case of None (Default), it does not make any request to server.

Purpose

This constructor can be used to construct a FrovedisBlockcyclicMatrix instance, as follows:

```
mat1 = FrovedisBlockcyclicMatrix() # empty matrix, no server request is made
mat2 = FrovedisBlockcyclicMatrix("./data") # will load data from given text file
mat3 = FrovedisBlockcyclicMatrix(mat2) # copy constructor
mat4 = FrovedisBlockcyclicMatrix([1,2,3,4]) # will load data from the given list
```

Return Type

It simply returns “self” reference.

4.3.2 Overloaded Operators Documentation

4.3.2.1 operator= (mat)

Parameters
mat: An existing FrovedisBlockcyclicMatrix instance to be copied.

Purpose
It can be used to copy the input matrix in the target matrix. It returns a self reference to support operator chaining.

For example,

```
m1 = FrovedisBlockcyclicMatrix([1,2,3,4])
m2 = m1 (copy operatror)
m3 = m2 = m1
```

Return Type
It returns “self” reference.

4.3.2.2 operator+ (mat)

Parameters
mat: An instance of FrovedisBlockcyclicMatrix or an array-like structure.

Purpose
It can be used to perform addition between two blockcyclic matrices. If the input data is not a FrovedisBlockcyclicMatrix instance, internally it will get converted into a FrovedisBlockcyclicMatrix instance first and then that will be added with the source matrix.

For example,

```
m1 = FrovedisBlockcyclicMatrix([1,2,3,4])
m2 = FrovedisBlockcyclicMatrix([1,2,3,4])
m3 = m2 + m1
```

Return Type
It returns the resultant matrix of the type FrovedisBlockcyclicMatrix.

4.3.2.3 operator- (mat)

Parameters
mat: An instance of FrovedisBlockcyclicMatrix or an array-like structure.

Purpose
It can be used to perform subtraction between two blockcyclic matrices. If the input data is not a FrovedisBlockcyclicMatrix instance, internally it will get converted into a FrovedisBlockcyclicMatrix instance first and then that will be subtracted from the source matrix.

For example,

```
m1 = FrovedisBlockcyclicMatrix([1,2,3,4])
m2 = FrovedisBlockcyclicMatrix([1,2,3,4])
m3 = m2 - m1
```

Return Type
It returns the resultant matrix of the type FrovedisBlockcyclicMatrix.

4.3.2.4 operator* (mat)

Parameters
mat: An instance of FrovedisBlockcyclicMatrix or an array-like structure.

Purpose
It can be used to perform multiplication between two blockcyclic matrices. If the input data is not a FrovedisBlockcyclicMatrix instance, internally it will get converted into a FrovedisBlockcyclicMatrix instance first and then that will be multiplied with the source matrix.

For example,

```
m1 = FrovedisBlockcyclicMatrix([1,2,3,4])
m2 = FrovedisBlockcyclicMatrix([1,2,3,4])
m3 = m2 * m1
```

Return Type
It returns the resultant matrix of the type FrovedisBlockcyclicMatrix.

4.3.2.5 operator~ ()

Purpose
It can be used to obtain transpose of the target matrix. If the input data is not a FrovedisBlockcyclicMatrix instance, internally it will get converted into a FrovedisBlockcyclicMatrix instance first and then the transpose will get computed.

For example,

```
m1 = FrovedisBlockcyclicMatrix([1,2,3,4])
m2 = ~m1
```

Return Type
It returns the resultant matrix of the type FrovedisBlockcyclicMatrix.

4.3.3 Pubic Member Function Documentation

4.3.3.1 load (mat)

Parameters
mat: It can be a string containing filename having text data to be loaded, or another FrovedisBlockcyclicMatrix instance for copy or any python array-like object (but it can not be None).

Purpose

This function works similar to the constructor. It can be used to load a FrovedisBlockcyclicMatrix instance, as follows:

```
mat1 = FrovedisBlockcyclicMatrix().load("./data")  # will load data from given text file
mat2 = FrovedisBlockcyclicMatrix().load(mat1)      # copy operation
mat3 = FrovedisBlockcyclicMatrix().load([1,2,3,4]) # will load data from the given list
```

Return Type

It simply returns "self" reference.

4.3.3.2 load_numpy_matrix (mat)

Parameters
mat: A numpy matrix with values to be loaded in.

Purpose
This function can be used to load a python side dense data matrix into a frovedis server side blockcyclic matrix. It accepts a numpy matrix object and converts it into the frovedis server side blockcyclic matrix whose proxy along with number of rows and number of columns information are stored in the target FrovedisBlockcyclicMatrix object.

Return Type
It simply returns "self" reference.

4.3.3.3 load_text (filename)

Parameters
filename: A string object containing the text file name to be loaded.

Purpose
This function can be used to load the data from a text file into the target matrix. Note that the file must be placed at server side at the given path.

Return Type
It simply returns "self" reference.

4.3.3.4 load_binary (dirname)

Parameters
dirname: A string object containing the directory name having the binary data to be loaded.

Purpose
This function can be used to load the data from the specified directory with binary data file into the target matrix. Note that the file must be placed at server side at the given path.

Return Type
It simply returns "self" reference.

4.3.3.5 save_text (filename)

Parameters
filename: A string object containing the text file name in which the data is to be saved.

Purpose

This function is used to save the target matrix as text file with the filename at the given path. Note that the file will be saved at server side at the given path.

Return Type
It returns nothing.

4.3.3.6 save_binary (dirname)

Parameters
dirname: A string object containing the directory name in which the data is to be saved as little-endian binary form.

Purpose

This function is used to save the target matrix as little-endian binary file with the filename at the given path. Note that the file will be saved at server side at the given path.

Return Type
It returns nothing.

4.3.3.7 transpose ()

Purpose

This function will compute the transpose of the given matrix.

Return Type
It returns the transposed blockcyclic matrix of the type FrovedisBlockcyclicMatrix.

4.3.3.8 to_numpy_matrix ()

Purpose
This function is used to convert the target blockcyclic matrix into numpy matrix.
Note that this function will request frovedis server to gather the distributed data, and send back that data in the rowmajor array form and the python client will then convert the received numpy array from frovedis server to python numpy matrix.

Return Type
It returns a two-dimensional dense numpy matrix

4.3.3.9 numRows()

Purpose
It returns the number of rows in the matrix

Return Type
An integer value containing rows count in the target matrix.

4.3.3.10 numCols()

Purpose
It returns the number of columns in the matrix

Return Type
An integer value containing columns count in the target matrix.

4.3.3.11 debug_print()

Purpose
It prints the contents of the server side distributed matrix data on the server side user terminal. It is mainly useful for debugging purpose.

Return Type
It returns nothing.

4.3.3.12 release()

Purpose
This function can be used to release the existing in-memory data at frovedis server side.

Return Type
It returns nothing.

4.3.3.13 FrovedisBlockcyclicMatrix.asBCM(mat)

Parameters
mat: An instance of FrovedisBlockcyclicMatrix or any python array-like structure.

Purpose
This static function is used in order to convert a given matrix to a blockcyclic matrix. If the input is already an instance of FrovedisBlockcyclicMatrix, then the same will be returned.

Return Type

An instance of FrovedisBlockcyclicMatrix.

Chapter 5

pblas_wrapper

5.1 NAME

pblas_wrapper - a frovedis module provides user-friendly interfaces for commonly used pblas routines in scientific applications like machine learning algorithms.

5.2 SYNOPSIS

import frovedis.matrix.wrapper.PBLAS

5.2.1 Public Member Functions

PBLAS.swap (v1, v2)
PBLAS.copy (v1, v2)
PBLAS.scal (v, al)
PBLAS.axpy (v1, v2, al=1.0)
PBLAS.dot (v1, v2)
PBLAS.nrm2 (v)
PBLAS.gemv (m, v1, v2, trans=False, al=1.0, b2=0.0)
PBLAS.ger (v1, v2, m, al=1.0)
PBLAS.gemm (m1, m2, m3, trans_m1=False, trans_m2=False, al=1.0, be=0.0)
PBLAS.geadd (m1, m2, trans=False, al=1.0, be=1.0)

5.3 DESCRIPTION

PBLAS is a high-performance scientific library written in Fortran language. It provides rich set of functionalities on vectors and matrices. The computation loads of these functionalities are parallelized over the available processes in a system and the user interfaces of this library is very detailed and complex in nature. It requires a strong understanding on each of the input parameters, along with some distribution concepts.

Frovedis provides a wrapper module for some commonly used PBLAS subroutines in scientific applications like machine learning algorithms. These wrapper interfaces are very simple and user needs not to consider all the detailed distribution parameters. Only specifying the target vectors or matrices with some other parameters (depending upon need) are fine. At the same time, all the use cases of a PBLAS routine can also be performed using Frovedis PBLAS wrapper of that routine.

This python module implements a client-server application, where the python client can send the python matrix data to frovedis server side in order to create blockcyclic matrix at frovedis server and then python client can request frovedis server for any of the supported PBLAS operation on that matrix. When required, python client can request frovedis server to send back the resultant matrix and it can then create equivalent python data.

The individual detailed descriptions can be found in the subsequent sections. Please note that the term "inout", used in the below section indicates a function argument as both "input" and "output".

5.3.1 Detailed Description

5.3.1.1 swap (v1, v2)

Parameters
v1: A FrovedisBlockcyclicMatrix with single column (inout)
v2: A FrovedisBlockcyclicMatrix with single column (inout)

Purpose
It will swap the contents of v1 and v2, if they are semantically valid and are of same length.

Return Value
On success, it returns nothing. If any error occurs, it throws an exception.

5.3.1.2 copy (v1, v2)

Parameters
v1: A FrovedisBlockcyclicMatrix with single column (input)
v2: A FrovedisBlockcyclicMatrix with single column (output)

Purpose
It will copy the contents of v1 in v2 (v2 = v1), if they are semantically valid and are of same length.

Return Value
On success, it returns nothing. If any error occurs, it throws an exception.

5.3.1.3 scal (v, al)

Parameters
v: A FrovedisBlockcyclicMatrix with single column (inout)
al: A double parameter to specify the value to which the input vector needs to be scaled. (input)

Purpose
It will scale the input vector with the provided "al" value, if it is semantically valid. On success, input vector "v" would be updated (in-place scaling).

Return Value
On success, it returns nothing. If any error occurs, it throws an exception.

5.3.1.4 axpy (v1, v2, al=1.0)

Parameters
v1: A FrovedisBlockcyclicMatrix with single column (input)
v2: A FrovedisBlockcyclicMatrix with single column (inout)
al: A double parameter to specify the value to which "v1" needs to be scaled (not in-place scaling) [Default: 1.0] (input/optional)

Purpose
It will solve the expression v2 = al*v1 + v2, if the input vectors are semantically valid and are of same length. On success, "v2" will be updated with desired result, but "v1" would remain unchanged.

Return Value
On success, it returns nothing. If any error occurs, it throws an exception.

5.3.1.5 dot (v1, v2)

Parameters
v1: A FrovedisBlockcyclicMatrix with single column (input)
v2: A FrovedisBlockcyclicMatrix with single column (input)

Purpose
It will perform dot product of the input vectors, if they are semantically valid and are of same length. Input vectors would not get modified during the operation.

Return Value
On success, it returns the dot product result of the type double. If any error occurs, it throws an exception.

5.3.1.6 nrm2 (v)

Parameters
v: A FrovedisBlockcyclicMatrix with single column (input)

Purpose
It will calculate the norm of the input vector, if it is semantically valid. Input vector would not get modified during the operation.

Return Value
On success, it returns the norm value of the type double. If any error occurs, it throws an exception.

5.3.1.7 gemv (m, v1, v2, trans=False, al=1.0, be=0.0)

Parameters
m: A FrovedisBlockcyclicMatrix (input)
v1: A FrovedisBlockcyclicMatrix with single column (input)
v2: A FrovedisBlockcyclicMatrix with single column (inout)
trans: A boolean value to specify whether to transpose "m" or not [Default: False] (input/optional)
al: A double type value [Default: 1.0] (input/optional)
be: A double type value [Default: 0.0] (input/optional)

Purpose
The primary aim of this routine is to perform simple matrix-vector multiplication.
But it can also be used to perform any of the below operations:

```
(1) v2 = al*m*v1 + be*v2
(2) v2 = al*transpose(m)*v1 + be*v2
```

If trans=False, then expression (1) is solved. In that case, the size of "v1" must be at least the number of columns in "m" and the size of "v2" must be at least the number of rows in "m".
If trans=True, then expression (2) is solved. In that case, the size of "v1" must be at least the number of rows in "m" and the size of "v2" must be at least the number of columns in "m".

Since "v2" is used as input-output both, memory must be allocated for this vector before calling this routine, even if simple matrix-vector multiplication is required. Otherwise, this routine will throw an exception.

For simple matrix-vector multiplication, no need to specify values for the input parameters “trans”, “al” and “be” (leave them at their default values).

On success, “v2” will be overwritten with the desired output. But “m” and “v1” would remain unchanged.

Return Value
On success, it returns nothing. If any error occurs, it throws an exception.

5.3.1.8 ger (v1, v2, m, al=1.0)

Parameters
v1: A FrovedisBlockcyclicMatrix with single column (input)
v2: A FrovedisBlockcyclicMatrix with single column (input)
m: A FrovedisBlockcyclicMatrix (inout)
al: A double type value [Default: 1.0] (input/optional)

Purpose
The primary aim of this routine is to perform simple vector-vector multiplication of the sizes “a” and “b” respectively to form an axb matrix. But it can also be used to perform the below operations:

```
m = al*v1*v2' + m
```

This operation can only be performed if the inputs are semantically valid and the size of “v1” is at least the number of rows in matrix “m” and the size of “v2” is at least the number of columns in matrix “m”.

Since “m” is used as input-output both, memory must be allocated for this matrix before calling this routine, even if simple vector-vector multiplication is required. Otherwise it will throw an exception.

For simple vector-vector multiplication, no need to specify the value for the input parameter “al” (leave it at its default value).

On success, “m” will be overwritten with the desired output. But “v1” and “v2” will remain unchanged.

Return Value
On success, it returns nothing. If any error occurs, it throws an exception.

5.3.1.9 gemm (m1, m2, m3, trans_m1=False, trans_m2=False, al=1.0, be=0.0)

Parameters
m1: A FrovedisBlockcyclicMatrix (input)
m2: A FrovedisBlockcyclicMatrix (input)
m3: A FrovedisBlockcyclicMatrix (inout)
trans_m1: A boolean value to specify whether to transpose “m1” or not [Default: False] (input/optional)
trans_m2: A boolean value to specify whether to transpose “m2” or not [Default: False] (input/optional)
al: A double type value [Default: 1.0] (input/optional)
be: A double type value [Default: 0.0] (input/optional)

Purpose
The primary aim of this routine is to perform simple matrix-matrix multiplication.
But it can also be used to perform any of the below operations:

```
(1) m3 = al*m1*m2 + be*m3
(2) m3 = al*transpose(m1)*m2 + be*m3
(3) m3 = al*m1*transpose(m2) + be*m3
(4) m3 = al*transpose(m1)*transpose(m2) + be*m3
```

(1) will be performed, if both “trans_m1” and “trans_m2” are False.

(2) will be performed, if trans_m1=True and trans_m2 = False.

(3) will be performed, if trans_m1=False and trans_m2 = True.

(4) will be performed, if both “trans_m1” and “trans_m2” are True.

If we have four variables nrowa, nrowb, ncola, ncolb defined as follows:

```
if(trans_m1) {
  nrowa = number of columns in m1
  ncola = number of rows in m1
}
else {
  nrowa = number of rows in m1
  ncola = number of columns in m1
}

if(trans_m2) {
  nrowb = number of columns in m2
  ncolb = number of rows in m2
}
else {
  nrowb = number of rows in m2
  ncolb = number of columns in m2
}
```

Then this function can be executed successfully, if the below conditions are all true:

```
(a) "ncola" is equal to "nrowb"
(b) number of rows in "m3" is equal to or greater than "nrowa"
(b) number of columns in "m3" is equal to or greater than "ncolb"
```

Since “m3” is used as input-output both, memory must be allocated for this matrix before calling this routine, even if simple matrix-matrix multiplication is required. Otherwise it will throw an exception.

For simple matrix-matrix multiplication, no need to specify the value for the input parameters “trans_m1”, “trans_m2”, “al”, “be” (leave them at their default values).

On success, “m3” will be overwritten with the desired output. But “m1” and “m2” will remain unchanged.

Return Value
On success, it returns nothing. If any error occurs, it throws an exception.

5.3.1.10 geadd (m1, m2, trans=False, al=1.0, be=1.0)

Parameters
m1: A FrovedisBlockcyclicMatrix (input)
m2: A FrovedisBlockcyclicMatrix (inout)
trans: A boolean value to specify whether to transpose “m1” or not [Default: False] (input/optional)
al: A double type value [Default: 1.0] (input/optional)
be: A double type value [Default: 1.0] (input/optional)

Purpose
The primary aim of this routine is to perform simple matrix-matrix addition. But it can also be used to perform any of the below operations:

```
(1) m2 = al*m1 + be*m2
(2) m2 = al*transpose(m1) + be*m2
```

If trans=False, then expression (1) is solved. In that case, the number of rows and the number of columns in "m1" should be equal to the number of rows and the number of columns in "m2" respectively.
If trans=True, then expression (2) is solved. In that case, the number of columns and the number of rows in "m1" should be equal to the number of rows and the number of columns in "m2" respectively.

If it is needed to scale the input matrices before the addition, corresponding "al" and "be" values can be provided. But for simple matrix-matrix addition, no need to specify values for the input parameters "trans", "al" and "be" (leave them at their default values).

On success, "m2" will be overwritten with the desired output. But "m1" would remain unchanged.

Return Value
On success, it returns nothing. If any error occurs, it throws an exception.

5.4 SEE ALSO

scalapack_wrapper, blockcyclic_matrix

Chapter 6

scalapack_wrapper

6.1 NAME

scalapack_wrapper - a frovedis module provides user-friendly interfaces for commonly used scalapack routines in scientific applications like machine learning algorithms.

6.2 SYNOPSIS

import frovedis.matrix.wrapper.SCALAPACK

6.3 WRAPPER FUNCTIONS

SCALAPACK.getrf (m)
SCALAPACK.getri (m, ipivPtr)
SCALAPACK.getrs (m1, m2, ipivPtr, trans=False)
SCALAPACK.gesv (m1, m2)
SCALAPACK.gels (m1, m2, trans=False)
SCALAPACK.gesvd (m, wantU=False, wantV=False)

6.4 DESCRIPTION

ScaLAPACK is a high-performance scientific library written in Fortran language. It provides rich set of linear algebra functionalities whose computation loads are parallelized over the available processes in a system and the user interfaces of this library is very detailed and complex in nature. It requires a strong understanding on each of the input parameters, along with some distribution concepts.

Frovedis provides a wrapper module for some commonly used ScaLAPACK subroutines in scientific applications like machine learning algorithms. These wrapper interfaces are very simple and user needs not to consider all the detailed distribution parameters. Only specifying the target vectors or matrices with some other parameters (depending upon need) are fine. At the same time, all the use cases of a ScaLAPACK routine can also be performed using Frovedis ScaLAPACK wrapper of that routine.

This python module implements a client-server application, where the python client can send the python matrix data to frovedis server side in order to create blockcyclic matrix at frovedis server and then python

client can request frovedis server for any of the supported ScaLAPACK operation on that matrix. When required, python client can request frovedis server to send back the resultant matrix and it can then create equivalent python data (see manuals for FrovedisBlockcyclicMatrix to python data conversion).

The individual detailed descriptions can be found in the subsequent sections. Please note that the term "inout", used in the below section indicates a function argument as both "input" and "output".

6.4.1 Detailed Description

6.4.1.1 getrf (m)

Parameters
m: A FrovedisBlockcyclicMatrix (inout)

Purpose
It computes an LU factorization of a general M-by-N distributed matrix, "m" using partial pivoting with row interchanges.

On successful factorization, matrix "m" is overwritten with the computed L and U factors. Along with the return status of native scalapack routine, it also returns the proxy address of the node local vector "ipiv" containing the pivoting information associated with input matrix "m" in the form of GetrfResult. The "ipiv" information will be useful in computation of some other routines (like getri, getrs etc.)

Return Value
On success, it returns the object of the type GetrfResult as explained above. If any error occurs, it throws an exception explaining cause of the error.

6.4.1.2 getri (m, ipivPtr)

Parameters
m: A FrovedisBlockcyclicMatrix (inout)
ipiv: A long object containing the proxy of the ipiv vector (from GetrfResult) (input)

Purpose
It computes the inverse of a distributed square matrix using the LU factorization computed by getrf(). So in order to compute inverse of a matrix, first compute it's LU factor (and ipiv information) using getrf() and then pass the factored matrix, "m" along with the "ipiv" information to this function.

On success, factored matrix "m" is overwritten with the inverse (of the matrix which was passed to getrf()) matrix. "ipiv" will be internally used by this function and will remain unchanged.

For example,

```
res = SCALAPACK.getrf(m)      // getting LU factorization of "m"
SCALAPACK.getri(m,res.ipiv()) // "m" will have inverse of the initial value
```

Return Value
On success, it returns the exit status of the scalapack routine itself. If any error occurs, it throws an exception explaining cause of the error.

6.4.1.3 getrs (m1, m2, ipiv, trans=False)

Parameters
m1: A FrovedisBlockcyclicMatrix (input)
m2: A FrovedisBlockcyclicMatrix (inout)

ipiv: A long object containing the proxy of the ipiv vector (from GetrfResult) (input)
trans: A boolean value to specify whether to transpose "m1" [Default: False] (input/optional)

Purpose
It solves a real system of distributed linear equations, AX=B with a general distributed square matrix (A) using the LU factorization computed by getrf(). Thus before calling this function, it is required to obtain the factored matrix "m1" (along with "ipiv" information) by calling getrf().

For example,

```
res = SCALAPACK.getrf(m1) // getting LU factorization of "m1"
SCALAPACK.getrs(m1,m2,res.ipiv())
```

If trans=False, the linear equation AX=B is solved.
If trans=True, the linear equation transpose(A)X=B (A'X=B) is solved.

The matrix "m2" should have number of rows >= the number of rows in "m1" and at least 1 column in it.

On entry, "m2" contains the distributed right-hand-side (B) of the equation and on successful exit it is overwritten with the distributed solution matrix (X).

Return Value
On success, it returns the exit status of the scalapack routine itself. If any error occurs, it throws an exception explaining cause of the error.

6.4.1.4 gesv (m1, m2)

Parameters
m1: A FrovedisBlockcyclicMatrix (inout)
m2: A FrovedisBlockcyclicMatrix (inout)

Purpose
It solves a real system of distributed linear equations, AX=B with a general distributed square matrix, "m1" by computing it's LU factors internally. This function internally computes the LU factors and ipiv information using getrf() and then solves the equation using getrs().

The matrix "m2" should have number of rows >= the number of rows in "m1" and at least 1 column in it.

On entry, "m1" contains the distributed left-hand-side square matrix (A), "m2" contains the distributed right-hand-side matrix (B) and on successful exit "m1" is overwritten with it's LU factors, "m2" is overwritten with the distributed solution matrix (X).

Return Value
On success, it returns the exit status of the scalapack routine itself. If any error occurs, it throws an exception explaining cause of the error.

6.4.1.5 gels (m1, m2, trans=False)

Parameters
m1: A FrovedisBlockcyclicMatrix (input)
m2: A FrovedisBlockcyclicMatrix (inout)
trans: A boolean value to specify whether to transpose "m1" [Default: False] (input/optional)

Purpose
It solves overdetermined or underdetermined real linear systems involving an M-by-N distributed matrix (A) or its transpose, using a QR or LQ factorization of (A). It is assumed that distributed matrix (A) has full rank.

If trans=False and M >= N: it finds the least squares solution of an overdetermined system.
If trans=False and M < N: it finds the minimum norm solution of an underdetermined system.
If trans=True and M >= N: it finds the minimum norm solution of an underdetermined system.
If trans=True and M < N: it finds the least squares solution of an overdetermined system.

The matrix "m2" should have number of rows >= max(M,N) and at least 1 column.

On entry, "m1" contains the distributed left-hand-side matrix (A) and "m2" contains the distributed right-hand-side matrix (B). On successful exit, "m1" is overwritten with the QR or LQ factors and "m2" is overwritten with the distributed solution matrix (X).

Return Value
On success, it returns the exit status of the scalapack routine itself. If any error occurs, it throws an exception explaining cause of the error.

6.4.1.6 gesvd (m, wantU=False, wantV=False)

Parameters
m: A FrovedisBlockcyclicMatrix (inout)
wantU: A boolean value to specify whether to compute U matrix [Default: False] (input)
wantV: A boolean value to specify whether to compute V matrix [Default: False] (input)

Purpose
It computes the singular value decomposition (SVD) of an M-by-N distributed matrix.

On entry "m" contains the distributed matrix whose singular values are to be computed.

If wantU = wantV = False, then it computes only the singular values in sorted oder, so that sval(i) >= sval(i+1). Otherwise it also computes U and/or V (left and right singular vectors respectively) matrices.

On successful exit, the contents of "m" is destroyed (internally used as workspace).

Return Value
On success, it returns the object of the type GesvdResult containing the singular values and U and V components (based on the requirement) along with the exit status of the native scalapack routine. If any error occurs, it throws an exception explaining cause of the error.

6.5 SEE ALSO

blockcyclic_matrix, pblas_wrapper, arpack_wrapper, getrf_result, gesvd_result

Chapter 7

arpack_wrapper

7.1 NAME

arpack_wrapper - a frovedis module supports singular value decomposition on sparse data using arpack routines.

7.2 SYNOPSIS

import frovedis.matrix.wrapper.ARPACK

7.2.1 Public Member Functions

ARPACK.computeSVD (data, k)

7.3 DESCRIPTION

This module provides interface to compute singular value decomposition on sparse data using arpack native routines at frovedis server side.

7.3.1 Detailed Description

7.3.1.1 computeSVD (data, k)

Parameters
data: Any scipy sparse matrix or python array-like structure or an instance of FrovedisCRSMatrix.
k: An integer value to specify the number of singular values to compute.

Purpose
It computes the singular value decomposition on the sparse data at frovedis side. Once done, it returns a GesvdResult object containing the proxy of the results at frovedis server.

When required, the spark client can convert back the frovedis server side SVD result to numpy data by calling to_numpy_results() function on GesvdResult structure.

For example,

```
res = ARPACK.computeSVD(data,2) // compute 2 singular values for the given data
p_res = res.to_numpy_results()
```

Return Value
On success, it returns an object of GesvdResult type containing the proxy of SVD results at frovedis server side. If any error occurs, it throws an exception.

Chapter 8

getrf_result

8.1 NAME

getrf_result - a structure to model the output of frovedis wrapper of scalapack getrf routine.

8.2 SYNOPSIS

import frovedis.matrix.results.GetrfResult

8.2.1 Public Member Functions

release()
ipiv()
stat()

8.3 DESCRIPTION

GetrfResult is a client python side pseudo result structure containing the proxy of the in-memory scalapack getrf result (node local ipiv vector) created at frovedis server side.

8.3.1 Public Member Function Documentation

8.3.1.1 release()

Purpose
This function can be used to release the in-memory result component (ipiv vector) at frovedis server.

Return Type
It returns nothing.

8.3.1.2 ipiv()

Purpose
This function returns the proxy of the node_local "ipiv" vector computed during getrf calculation. This value will be required in other scalapack routine calculation, like getri, getrs etc.

Return Type
A long value containing the proxy of ipiv vector.

8.3.1.3 stat()

Purpose
This function returns the exit status of the scalapack native getrf routine on calling of which the target result object was obtained.

Return Type
It returns an integer value.

Chapter 9

gesvd_result

9.1 NAME

gesvd_result - a structure to model the output of frovedis singular value decomposition methods.

9.2 SYNOPSIS

import frovedis.matrix.results.GesvdResult

9.2.1 Public Member Functions

to_numpy_results()
save(svec, umat=None, vmat=None)
save_binary(svec, umat=None, vmat=None)
load (svec, umat=None, vmat=None, mtype='B')
load_binary (svec, umat=None, vmat=None, mtype='B')
debug_print()
release()
stat()
getK()

9.3 DESCRIPTION

GesvdResult is a python side pseudo result structure containing the proxies of the in-memory SVD results created at frovedis server side. It can be used to convert the frovedis side SVD result to python equivalent data structures.

9.3.1 Public Member Function Documentation

9.3.1.1 to_numpy_results()

Purpose
This function can be used to convert the frovedis side SVD results to python numpy result structures.

If U and V both are computed, it returns: (numpy matrix, numpy array, numpy matrix)
indicating (umatrix, singular vector, vmatrix).

When U is calculated, but not V, it returns: (numpy matrix, numpy array, None)
When V is calculated, but not U, it returns: (None, numpy array, numpy matrix)
When neither U nor V is calculated, it returns: (None, numpy array, None)

Return Type
It returns a tuple as explained above.

9.3.1.2 save(svec, umat=None, vmat=None)

Parameters
svec: A string object containing name of the file to save singular vectors as text data. (mandatory)
umat: A string object containing name of the file to save umatrix as text data. (optional)
vmat: A string object containing name of the file to save vmatrix as text data. (ptional)

Purpose
This function can be used to save the result values in different text files at server side. If saving of U and V components are not required, "umat" and "vmat" can be None, but "svec" should have a valid filename.

Return Type
It returns nothing.

9.3.1.3 save_binary (svec, umat=None, vmat=None)

Parameters
svec: A string object containing name of the file to save singular vectors as binary data. (mandatory)
umat: A string object containing name of the file to save umatrix as binary data. (optional)
vmat: A string object containing name of the file to save vmatrix as binary data. (optional)

Purpose
This function can be used to save the result values in different files as little-endian binary data at server side. If saving of U and V components are not required, "umat" and "vmat" can be None, but "svec" should have a valid filename.

Return Type
It returns nothing.

9.3.1.4 load(svec, umat=None, vmat=None, mtype='B')

Parameters
svec: A string object containing name of the file from which to load singular vectors as text data for the target result. (mandatory)
umat: A string object containing name of the file from which to load umatrix as text data for the target result. (optional)
vmat: A string object containing name of the file from which to load vmatrix as text data for the target result. (optional)
mtype: A character value, can be either 'B' or 'C'. (optional)

Purpose
This function can be used to load the result values in different text files at server side. If loading of U and V components are not required, "umat" and "vmat" can be None, but "svec" should have a valid filename.

If mtype = 'B' and umat/vmat is to be loaded, then they will be loaded as blockcyclic matrices at server side.

If mtype = 'C' and umat/vmat is to be loaded, then they will be loaded as colmajor matrices at server side.

Return Type
It returns nothing.

9.3.1.5 load_binary(svec, umat=None, vmat=None, mtype='B')

Parameters
svec: A string object containing name of the file from which to load singular vectors as binary data for the target result. (mandatory)
umat: A string object containing name of the file from which to load umatrix as binary data for the target result. (optional)
vmat: A string object containing name of the file from which to load vmatrix as binary data for the target result. (optional)
mtype: A character value, can be either 'B' or 'C'. (optional)

Purpose
This function can be used to load the result values in different little-endian binary files at server side. If loading of U and V components are not required, "umat" and "vmat" can be None, but "svec" should have a valid filename.

If mtype = 'B' and umat/vmat is to be loaded, then they will be loaded as blockcyclic matrices at server side.

If mtype = 'C' and umat/vmat is to be loaded, then they will be loaded as colmajor matrices at server side.

Return Type
It returns nothing.

9.3.1.6 debug_print()

Purpose
This function can be used to print the result components at server side user terminal. This is useful in debugging purpose.

Return Type
It returns nothing.

9.3.1.7 release()

Purpose
This function can be used to release the in-memory result components at frovedis server.

Return Type
It returns nothing.

9.3.1.8 stat()

Purpose
This function returns the exit status of the scalapack native gesvd routine on calling of which the target result object was obtained.

Return Type
An integer value.

9.3.1.9 getK()

Purpose
This function returns the number of singular values computed.

Return Type
An integer value.

Chapter 10

Linear Regression

10.1 NAME

Linear Regression - A regression algorithm to predict the continuous output without any regularization.

10.2 SYNOPSIS

class frovedis.mllib.linear_model.LinearRegression (fit_intercept=True, normalize=False,
copy_X=True, n_jobs=1, solver='sag', verbose=0)

10.2.1 Public Member Functions

fit(X, y, sample_weight=None)
predict(X)
save(filename)
load(filename)
debug_print()
release()

10.3 DESCRIPTION

Linear least squares is the most common formulation for regression problems. It is a linear method with the loss function given by the **squared loss**:

```
L(w;x,y) := 1/2(wTx-y)^2
```

Where the vectors x are the training data examples and y are their corresponding labels which we want to predict. w is the linear model (also known as weight) which uses a single weighted sum of features to make a prediction. The method is called linear since it can be expressed as a function of wTx and y. Linear regression does not use any regularizer.

The gradient of the squared loss is: (wTx-y).x

Frovedis provides implementation of linear regression with two different optimizers: (1) stochastic gradient descent with minibatch and (2) LBFGS optimizer.

The simplest method to solve optimization problems of the form **min f(w)** is gradient descent. Such first-order optimization methods well-suited for large-scale and distributed computation. Whereas, L-BFGS is an optimization algorithm in the family of quasi-Newton methods to solve the optimization problems of the similar form.

Like the original BFGS, L-BFGS (Limited Memory BFGS) uses an estimation to the inverse Hessian matrix to steer its search through feature space, but where BFGS stores a dense nxn approximation to the inverse Hessian (n being the number of features in the problem), L-BFGS stores only a few vectors that represent the approximation implicitly. L-BFGS often achieves rapider convergence compared with other first-order optimization.

This module provides a client-server implementation, where the client application is a normal python scikit-learn program. Scikit-learn has its own linear_model providing the Linear Regression support. But that algorithm is non-distributed in nature. Hence it is slower when comparing with the equivalent Frovedis algorithm (see frovedis manual for ml/linear_regression) with big dataset. Thus in this implementation, a scikit-learn client can interact with a frovedis server sending the required python data for training at frovedis side. Python data is converted into frovedis compatible data internally and the scikit-learn ML call is linked with the respective frovedis ML call to get the job done at frovedis server.

Scikit-learn side call for Linear Regression quickly returns, right after submitting the training request to the frovedis server with a unique model ID for the submitted training request.

When operations like prediction will be required on the trained model, scikit-learn client sends the same request to frovedis server on the same model (containing the unique ID) and the request is served at frovedis server and output is sent back to the scikit-learn client.

10.3.1 Detailed Description

10.3.1.1 LinearRegression()

Parameters
fit_intercept: A boolean parameter specifying whether a constant (intercept) should be added to the decision function. (Default: True)
normalize: A boolean parameter. (unused)
copy_X: A boolean parameter. (unsed)
n_jobs: An integer parameter. (unused)
solver: A string parameter specifying the solver to use. (Default: 'sag')
verbose: A integer parameter specifying the log level to use. (Default: 0)

Purpose
It initialized a LinearRegression object with the given parameters.

The parameters: "normalize", "copy_X" and "n_jobs" are not yet supported at frovedis side. Thus they don't have any significance when calling the frovedis linear regression algorithm. They are simply provided for the compatibility with scikit-learn application.

"solver" can be either 'sag' for frovedis side stochastic gradient descent or 'lbfgs' for frovedis side LBFGS optimizer when optimizing the linear regression model.

"verbose" value is set at 0 by default. But it can be set to 1 (for DEBUG mode) or 2 (for TRACE mode) for getting training time logs from frovedis server.

Return Value
It simply returns "self" reference.

10.3.1.2 fit(X, y, sample_weight=None)

Parameters
X: A scipy sparse matrix or any python array-like object or an instance of FrovedisCRSMatrix.
y: Any python array-like object or an instance of FrovedisDvector.
sample_weight: Python array-like optional parameter. (unused)

Purpose
It accepts the training feature matrix (X) and corresponding output labels (y) as inputs from the user and trains a linear regression model with those data at frovedis server.

It doesn't support any initial weight to be passed as input at this moment. Thus the "sample_weight" parameter will simply be ignored. It starts with an initial guess of zeros for the model vector and keeps updating the model to minimize the cost function until convergence is achieved (default convergence tolerance value is 0.001) or maximum iteration count is reached (default 1000, is not configurable at this moment).

For example,

```
# loading sample CRS data file
mat = FrovedisCRSMatrix().load("./sample")
lbl = FrovedisDvector([1.1,0.2,1.3,1.4,1.5,0.6,1.7,1.8])

# fitting input matrix and label on linear regression object
lr = LinearRegression(solver='sgd', verbose=2).fit(mat,lbl)
```

Return Value
It simply returns "self" reference.
Note that the call will return quickly, right after submitting the fit request at frovedis server side with a unique model ID for the fit request. It may be possible that the training is not completed at the frovedis server side even though the client scikit-learn side fit() returns.

10.3.1.3 predict(X)

Parameters
X: A scipy sparse matrix or any python array-like object or an instance of FrovedisCRSMatrix.

Purpose
It accepts the test feature matrix (X) in order to make prediction on the trained model at frovedis server.

Return Value
It returns a numpy array of double (float64) type containing the predicted outputs.

10.3.1.4 save(filename)

Parameters
filename: A string object containing the name of the file on which the target model is to be saved.

Purpose
On success, it writes the model information (weight values etc.) in the specified file as little-endian binary data. Otherwise, it throws an exception.

Return Value
It returns nothing.

10.3.1.5 load(filename)

Parameters
filename: A string object containing the name of the file having model information to be loaded.

Purpose
It loads the model from the specified file (having little-endian binary data).

Return Value
It simply returns "self" instance.

10.3.1.6 debug_print()

Purpose
It shows the target model information (weight values etc.) on the server side user terminal. It is mainly used for debugging purpose.

Return Value
It returns nothing.

10.3.1.7 release()

Purpose
It can be used to release the in-memory model at frovedis server.

Return Value
It returns nothing.

10.4 SEE ALSO

lasso_regression, ridge_regression, dvector, crs_matrix

Chapter 11

Lasso Regression

11.1 NAME

Lasso Regression - A regression algorithm to predict the continuous output with L1 regularization.

11.2 SYNOPSIS

class frovedis.mllib.linear_model.Lasso (alpha=0.01, fit_intercept=True, normalize=False,
precompute=False, copy_X=True, max_iter=1000,
tol=1e-4, warm_start=False, positive=False,
random_state=None, selection='cyclic',
verbose=0, solver='sag')

11.2.1 Public Member Functions

fit(X, y, sample_weight=None)
predict(X)
save(filename)
load(filename)
debug_print()
release()

11.3 DESCRIPTION

Linear least squares is the most common formulation for regression problems. It is a linear method with the loss function given by the **squared loss**:

```
L(w;x,y) := 1/2(wTx-y)^2
```

Where the vectors x are the training data examples and y are their corresponding labels which we want to predict. w is the linear model (also known as weight) which uses a single weighted sum of features to make a prediction. The method is called linear since it can be expressed as a function of wTx and y. Lasso regression uses L1 regularization to address the overfit problem.

The gradient of the squared loss is: (wTx-y).x
The gradient of the regularizer is: sign(w)

Frovedis provides implementation of lasso regression with two different optimizers: (1) stochastic gradient descent with minibatch and (2) LBFGS optimizer.

The simplest method to solve optimization problems of the form **min f(w)** is gradient descent. Such first-order optimization methods well-suited for large-scale and distributed computation. Whereas, L-BFGS is an optimization algorithm in the family of quasi-Newton methods to solve the optimization problems of the similar form.

Like the original BFGS, L-BFGS (Limited Memory BFGS) uses an estimation to the inverse Hessian matrix to steer its search through feature space, but where BFGS stores a dense nxn approximation to the inverse Hessian (n being the number of features in the problem), L-BFGS stores only a few vectors that represent the approximation implicitly. L-BFGS often achieves rapider convergence compared with other first-order optimization.

This module provides a client-server implementation, where the client application is a normal python scikit-learn program. Scikit-learn has its own linear_model providing the Lasso Regression support. But that algorithm is non-distributed in nature. Hence it is slower when comparing with the equivalent Frovedis algorithm (see frovedis manual for ml/lasso_regression) with big dataset. Thus in this implementation, a scikit-learn client can interact with a frovedis server sending the required python data for training at frovedis side. Python data is converted into frovedis compatible data internally and the scikit-learn ML call is linked with the respective frovedis ML call to get the job done at frovedis server.

Scikit-learn side call for Lasso Regression quickly returns, right after submitting the training request to the frovedis server with a unique model ID for the submitted training request.

When operations like prediction will be required on the trained model, scikit-learn client sends the same request to frovedis server on the same model (containing the unique ID) and the request is served at frovedis server and output is sent back to the scikit-learn client.

11.3.1 Detailed Description

11.3.1.1 Lasso()

Parameters
alpha: A double parameter containing the learning rate. (Default: 0.01)
fit_intercept: A boolean parameter specifying whether a constant (intercept) should be added to the decision function. (Default: True)
normalize: A boolean parameter (unused)
precompute: A boolean parameter (unused)
copy_X: A boolean parameter (unsed)
max_iter: An integer parameter specifying maximum iteration count. (Default: 1000)
tol: A double parameter specifying the convergence tolerance value, (Default: 1e-4)
warm_start: A boolean parameter (unused)
positive: A boolean parameter (unused)
random_state: An integer, None or RandomState instance. (unused)
selection: A string object. (unused)
verbose: An integer object specifying the log level to use. (Default: 0)
solver: A string object specifying the solver to use. (Default: 'sag')

Purpose
It initialized a Lasso object with the given parameters.

The parameters: "normalize", "precompute", "copy_X", "warm_start", "positive", "random_state" and "selection" are not yet supported at frovedis side. Thus they don't have any significance in this call. They are simply provided for the compatibility with scikit-learn application.

"solver" can be either 'sag' for frovedis side stochastic gradient descent or 'lbfgs' for frovedis side LBFGS optimizer when optimizing the linear regression model.

"verbose" value is set at 0 by default. But it can be set to 1 (for DEBUG mode) or 2 (for TRACE mode) for getting training time logs from frovedis server.

Return Value
It simply returns "self" reference.

11.3.1.2 fit(X, y, sample_weight=None)

Parameters
X: A scipy sparse matrix or any python array-like object or an instance of FrovedisCRSMatrix.
y: Any python array-like object or an instance of FrovedisDvector.
sample_weight: Python array-like optional parameter. (unused)

Purpose
It accepts the training feature matrix (X) and corresponding output labels (y) as inputs from the user and trains a linear regression model with L1 regularization with those data at frovedis server.

It doesn't support any initial weight to be passed as input at this moment. Thus the "sample_weight" parameter will simply be ignored. It starts with an initial guess of zeros for the model vector and keeps updating the model to minimize the cost function until convergence is achieved or maximum iteration count is reached.

For example,

```
# loading sample CRS data file
mat = FrovedisCRSMatrix().load("./sample")
lbl = FrovedisDvector([1.1,0.2,1.3,1.4,1.5,0.6,1.7,1.8])

# fitting input matrix and label on lasso object
lr = Lasso(solver='sgd', verbose=2).fit(mat,lbl)
```

Return Value
It simply returns "self" reference.
Note that the call will return quickly, right after submitting the fit request at frovedis server side with a unique model ID for the fit request. It may be possible that the training is not completed at the frovedis server side even though the client scikit-learn side fit() returns.

11.3.1.3 predict(X)

Parameters
X: A scipy sparse matrix or any python array-like object or an instance of FrovedisCRSMatrix.

Purpose
It accepts the test feature matrix (X) in order to make prediction on the trained model at frovedis server.

Return Value
It returns a numpy array of double (float64) type containing the predicted outputs.

11.3.1.4 save(filename)

Parameters
filename: A string object containing the name of the file on which the target model is to be saved.

Purpose
On success, it writes the model information (weight values etc.) in the specified file as little-endian binary data. Otherwise, it throws an exception.

Return Value
It returns nothing.

11.3.1.5 load(filename)

Parameters
filename: A string object containing the name of the file having model information to be loaded.

Purpose
It loads the model from the specified file (having little-endian binary data).

Return Value
It simply returns "self" instance.

11.3.1.6 debug_print()

Purpose
It shows the target model information (weight values etc.) on the server side user terminal. It is mainly used for debugging purpose.

Return Value
It returns nothing.

11.3.1.7 release()

Purpose
It can be used to release the in-memory model at frovedis server.

Return Value
It returns nothing.

11.4 SEE ALSO

linear_regression, ridge_regression, dvector, crs_matrix

Chapter 12

Ridge Regression

12.1 NAME

Ridge Regression - A regression algorithm to predict the continuous output with L2 regularization.

12.2 SYNOPSIS

class frovedis.mllib.linear_model.Ridge (alpha=0.01, fit_intercept=True, normalize=False,
copy_X=True, max_iter=1000,
tol=1e-3, solver='sag',
random_state=None, verbose=0)

12.2.1 Public Member Functions

fit(X, y, sample_weight=None)
predict(X)
save(filename)
load(filename)
debug_print()
release()

12.3 DESCRIPTION

Linear least squares is the most common formulation for regression problems. It is a linear method with the loss function given by the **squared loss**:

```
L(w;x,y) := 1/2(wTx-y)^2
```

Where the vectors x are the training data examples and y are their corresponding labels which we want to predict. w is the linear model (also known as weight) which uses a single weighted sum of features to make a prediction. The method is called linear since it can be expressed as a function of wTx and y. Ridge regression uses L2 regularization to address the overfit problem.

The gradient of the squared loss is: (wTx-y).x
The gradient of the regularizer is: w

Frovedis provides implementation of ridge regression with two different optimizers: (1) stochastic gradient descent with minibatch and (2) LBFGS optimizer.

The simplest method to solve optimization problems of the form **min f(w)** is gradient descent. Such first-order optimization methods well-suited for large-scale and distributed computation. Whereas, L-BFGS is an optimization algorithm in the family of quasi-Newton methods to solve the optimization problems of the similar form.

Like the original BFGS, L-BFGS (Limited Memory BFGS) uses an estimation to the inverse Hessian matrix to steer its search through feature space, but where BFGS stores a dense nxn approximation to the inverse Hessian (n being the number of features in the problem), L-BFGS stores only a few vectors that represent the approximation implicitly. L-BFGS often achieves rapider convergence compared with other first-order optimization.

This module provides a client-server implementation, where the client application is a normal python scikit-learn program. Scikit-learn has its own linear_model providing the Ridge Regression support. But that algorithm is non-distributed in nature. Hence it is slower when comparing with the equivalent Frovedis algorithm (see frovedis manual for ml/ridge_regression) with big dataset. Thus in this implementation, a scikit-learn client can interact with a frovedis server sending the required python data for training at frovedis side. Python data is converted into frovedis compatible data internally and the scikit-learn ML call is linked with the respective frovedis ML call to get the job done at frovedis server.

Scikit-learn side call for Ridge Regression quickly returns, right after submitting the training request to the frovedis server with a unique model ID for the submitted training request.

When operations like prediction will be required on the trained model, scikit-learn client sends the same request to frovedis server on the same model (containing the unique ID) and the request is served at frovedis server and output is sent back to the scikit-learn client.

12.3.1 Detailed Description

12.3.1.1 Ridge()

Parameters
alpha: A double parameter containing the learning rate. (Default: 0.01)
fit_intercept: A boolean parameter specifying whether a constant (intercept) should be added to the decision function. (Default: True)
normalize: A boolean parameter (unused)
copy_X: A boolean parameter (unsed)
max_iter: An integer parameter specifying maximum iteration count. (Default: 1000)
tol: A double parameter specifying the convergence tolerance value, (Default: 1e-3)
solver: A string object specifying the solver to use. (Default: 'sag')
random_state: An integer, None or RandomState instance. (unused)
verbose: An integer object specifying the log level to use. (Default: 0)

Purpose
It initialized a Ridge object with the given parameters.

The parameters: "normalize", "copy_X" and "random_state" are not yet supported at frovedis side. Thus they don't have any significance in this call. They are simply provided for the compatibility with scikit-learn application.

"solver" can be either 'sag' for frovedis side stochastic gradient descent or 'lbfgs' for frovedis side LBFGS optimizer when optimizing the linear regression model.

"verbose" value is set at 0 by default. But it can be set to 1 (for DEBUG mode) or 2 (for TRACE mode) for getting training time logs from frovedis server.

Return Value
It simply returns "self" reference.

12.3.1.2 fit(X, y, sample_weight=None)

Parameters
X: A scipy sparse matrix or any python array-like object or an instance of FrovedisCRSMatrix.
y: Any python array-like object or an instance of FrovedisDvector.
sample_weight: Python array-like optional parameter. (unused)

Purpose
It accepts the training feature matrix (X) and corresponding output labels (y) as inputs from the user and trains a linear regression model with L2 regularization with those data at frovedis server.

It doesn't support any initial weight to be passed as input at this moment. Thus the "sample_weight" parameter will simply be ignored. It starts with an initial guess of zeros for the model vector and keeps updating the model to minimize the cost function until convergence is achieved or maximum iteration count is reached.

For example,

```
# loading sample CRS data file
mat = FrovedisCRSMatrix().load("./sample")
lbl = FrovedisDvector([1.1,0.2,1.3,1.4,1.5,0.6,1.7,1.8])

# fitting input matrix and label on ridge regression object
lr = Ridge(solver='sgd', verbose=2).fit(mat,lbl)
```

Return Value
It simply returns "self" reference.
Note that the call will return quickly, right after submitting the fit request at frovedis server side with a unique model ID for the fit request. It may be possible that the training is not completed at the frovedis server side even though the client scikit-learn side fit() returns.

12.3.1.3 predict(X)

Parameters
X: A scipy sparse matrix or any python array-like object or an instance of FrovedisCRSMatrix.

Purpose
It accepts the test feature matrix (X) in order to make prediction on the trained model at frovedis server.

Return Value
It returns a numpy array of double (float64) type containing the predicted outputs.

12.3.1.4 save(filename)

Parameters
filename: A string object containing the name of the file on which the target model is to be saved.

Purpose
On success, it writes the model information (weight values etc.) in the specified file as little-endian binary data. Otherwise, it throws an exception.

Return Value
It returns nothing.

12.3.1.5 load(filename)

Parameters
filename: A string object containing the name of the file having model information to be loaded.

Purpose
It loads the model from the specified file (having little-endian binary data).

Return Value
It simply returns "self" instance.

12.3.1.6 debug_print()

Purpose
It shows the target model information (weight values etc.) on the server side user terminal. It is mainly used for debugging purpose.

Return Value
It returns nothing.

12.3.1.7 release()

Purpose
It can be used to release the in-memory model at frovedis server.

Return Value
It returns nothing.

12.4 SEE ALSO

linear_regression, lasso_regression, dvector, crs_matrix

Chapter 13

Logistic Regression

13.1 NAME

Logistic Regression - A classification algorithm to predict the binary output with logistic loss.

13.2 SYNOPSIS

class frovedis.mllib.linear_model.LogisticRegression (penalty='l2', dual=False,
tol=1e-4, C=0.01, fit_intercept=True, intercept_scaling=1,
class_weight=None, random_state=None, solver='sag',
max_iter=1000, multi_class='ovr', verbose=0, warm_start=False,
n_jobs=1)

13.2.1 Public Member Functions

fit(X, y, sample_weight=None)
predict(X)
predict_proba (X) save(filename)
load(filename)
debug_print()
release()

13.3 DESCRIPTION

Classification aims to divide the items into categories. The most common classification type is binary classification, where there are two categories, usually named positive and negative. Frovedis supports binary classification algorithm only.

Logistic regression is widely used to predict a binary response. It is a linear method with the loss function given by the **logistic loss**:

```
L(w;x,y) := log(1 + exp(-ywTx))
```

Where the vectors x are the training data examples and y are their corresponding labels (Frovedis considers negative response as -1 and positive response as 1, but when calling from scikit-learn interface, user should pass 0 for negative response and 1 for positive response according to the scikit-learn requirement) which we want to predict. w is the linear model (also called as weight) which uses a single weighted sum of features to make a prediction. Frovedis Logistic Regression supports ZERO, L1 and L2 regularization to address the overfit problem.

The gradient of the logistic loss is: -y(1 - 1 / (1 + exp(-ywTx))).x
The gradient of the L1 regularizer is: sign(w)
And The gradient of the L2 regularizer is: w

For binary classification problems, the algorithm outputs a binary logistic regression model. Given a new data point, denoted by x, the model makes predictions by applying the logistic function:

```
f(z) := 1 / 1 + exp(-z)
```

Where z = wTx. By default (threshold=0.5), if f(wTx) > 0.5, the response is positive (1), else the response is negative (0).

Frovedis provides implementation of logistic regression with two different optimizers: (1) stochastic gradient descent with minibatch and (2) LBFGS optimizer.

The simplest method to solve optimization problems of the form **min f(w)** is gradient descent. Such first-order optimization methods well-suited for large-scale and distributed computation. Whereas, L-BFGS is an optimization algorithm in the family of quasi-Newton methods to solve the optimization problems of the similar form.

Like the original BFGS, L-BFGS (Limited Memory BFGS) uses an estimation to the inverse Hessian matrix to steer its search through feature space, but where BFGS stores a dense nxn approximation to the inverse Hessian (n being the number of features in the problem), L-BFGS stores only a few vectors that represent the approximation implicitly. L-BFGS often achieves rapider convergence compared with other first-order optimization.

This module provides a client-server implementation, where the client application is a normal python scikit-learn program. Scikit-learn has its own linear_model providing the Logistic Regression support. But that algorithm is non-distributed in nature. Hence it is slower when comparing with the equivalent Frovedis algorithm (see frovedis manual for ml/logistic_regression) with big dataset. Thus in this implementation, a scikit-learn client can interact with a frovedis server sending the required python data for training at frovedis side. Python data is converted into frovedis compatible data internally and the scikit-learn ML call is linked with the respective frovedis ML call to get the job done at frovedis server.

Scikit-learn side call for Logistic Regression quickly returns, right after submitting the training request to the frovedis server with a unique model ID for the submitted training request.

When operations like prediction will be required on the trained model, scikit-learn client sends the same request to frovedis server on the same model (containing the unique ID) and the request is served at frovedis server and output is sent back to the scikit-learn client.

13.3.1 Detailed Description

13.3.1.1 LogisticRegression()

Parameters *penalty*: A string object containing the regularizer type to use. (Default: 'l2')
dual: A boolean parameter (unused)
tol: A double parameter specifying the convergence tolerance value, (Default: 1e-4)
C: A double parameter containing the learning rate. (Default: 0.01)
fit_intercept: A boolean parameter specifying whether a constant (intercept) should be added to the decision

function. (Default: True)
intercept_scaling: An integer parameter. (unused)
*class*weight_: A python dictionary or a string object. (unused)
random_state: An integer, None or RandomState instance. (unused)
solver: A string object specifying the solver to use. (Default: 'sag')
max_iter: An integer parameter specifying maximum iteration count. (Default: 1000)
multi_class: A string object specifying type of classification. (Default: 'ovr')
verbose: An integer object specifying the log level to use. (Default: 0)
warm_start: A boolean parameter. (unused)
n_jobs: An integer parameter. (unused)

Purpose
It initialized a Lasso object with the given parameters.

The parameters: "dual", "intercept_scaling", "class_weight", "warm_start", "random_state" and "n_jobs" are not yet supported at frovedis side. Thus they don't have any significance in this call. They are simply provided for the compatibility with scikit-learn application.

"penalty" can be either 'l1' or 'l2' (Default: 'l2').

"solver" can be either 'sag' for frovedis side stochastic gradient descent or 'lbfgs' for frovedis side LBFGS optimizer when optimizing the linear regression model.

"multi_class" can only be 'ovr' as frovedis suppots binary classification algorithms only at this moment.

"verbose" value is set at 0 by default. But it can be set to 1 (for DEBUG mode) or 2 (for TRACE mode) for getting training time logs from frovedis server.

Return Value
It simply returns "self" reference.

13.3.1.2 fit(X, y, sample_weight=None)

Parameters
X: A scipy sparse matrix or any python array-like object or an instance of FrovedisCRSMatrix.
y: Any python array-like object or an instance of FrovedisDvector.
sample_weight: Python array-like optional parameter. (unused)

Purpose
It accepts the training feature matrix (X) and corresponding output labels (y) as inputs from the user and trains a linear regression model with specifed regularization with those data at frovedis server.

It doesn't support any initial weight to be passed as input at this moment. Thus the "sample_weight" parameter will simply be ignored. It starts with an initial guess of zeros for the model vector and keeps updating the model to minimize the cost function until convergence is achieved or maximum iteration count is reached.

For example,

```
# loading sample CRS data file
mat = FrovedisCRSMatrix().load("./sample")
lbl = FrovedisDvector([1,0,1,1,1,0,1,1])

# fitting input matrix and label on logistic regression object
lr = LogisticRegression(solver='sgd', verbose=2).fit(mat,lbl)
```

Return Value
It simply returns "self" reference.

Note that the call will return quickly, right after submitting the fit request at frovedis server side with a unique model ID for the fit request. It may be possible that the training is not completed at the frovedis server side even though the client scikit-learn side fit() returns.

13.3.1.3 predict(X)

Parameters
X: A scipy sparse matrix or any python array-like object or an instance of FrovedisCRSMatrix.

Purpose
It accepts the test feature matrix (X) in order to make prediction on the trained model at frovedis server.

Return Value
It returns a numpy array of double (float64) type containing the predicted outputs.

13.3.1.4 predict_proba(X)

Parameters
X: A scipy sparse matrix or any python array-like object or an instance of FrovedisCRSMatrix.

Purpose
It accepts the test feature matrix (X) in order to make prediction on the trained model at frovedis server. But unlike predict(), it returns the probability values against each input sample to be positive.

Return Value
It returns a numpy array of double (float64) type containing the prediction probability values.

13.3.1.5 save(filename)

Parameters
filename: A string object containing the name of the file on which the target model is to be saved.

Purpose
On success, it writes the model information (weight values etc.) in the specified file as little-endian binary data. Otherwise, it throws an exception.

Return Value
It returns nothing.

13.3.1.6 load(filename)

Parameters
filename: A string object containing the name of the file having model information to be loaded.

Purpose
It loads the model from the specified file (having little-endian binary data).

Return Value
It simply returns “self” instance.

13.3.1.7 debug_print()

Purpose
It shows the target model information (weight values etc.) on the server side user terminal. It is mainly used for debugging purpose.

Return Value
It returns nothing.

13.3.1.8 release()

Purpose
It can be used to release the in-memory model at frovedis server.

Return Value
It returns nothing.

13.4 SEE ALSO

linear_svm, dvector, crs_matrix

Chapter 14

Linear SVM

14.1 NAME

Linear SVM (Support Vector Machines) - A classification algorithm to predict the binary output with hinge loss.

14.2 SYNOPSIS

class frovedis.mllib.svm.LinearSVC (penalty='l2', loss='hinge', dual=True, tol=1e-4,
C=0.01, multi_class='ovr', fit_intercept=True,
intercept_scaling=1, class_weight=None, verbose=0,
random_state=None, max_iter=1000, solver='sag')

14.2.1 Public Member Functions

fit(X, y, sample_weight=None)
predict(X)
predict_proba (X) save(filename)
load(filename)
debug_print()
release()

14.3 DESCRIPTION

Classification aims to divide items into categories. The most common classification type is binary classification, where there are two categories, usually named positive and negative. Frovedis supports binary classification algorithms only.

The Linear SVM is a standard method for large-scale classification tasks. It is a linear method with the loss function given by the **hinge loss**:

```
L(w;x,y) := max{0, 1-ywTx}
```

Where the vectors x are the training data examples and y are their corresponding labels (Frovedis considers negative response as -1 and positive response as 1, but when calling from scikit-learn interface, user should pass 0 for negative response and 1 for positive response according to the scikit-learn requirement) which we want to predict. w is the linear model (also known as weight) which uses a single weighted sum of features to make a prediction. Linear SVM supports ZERO, L1 and L2 regularization to address the overfit problem.

The gradient of the hinge loss is: -y.x, if ywTx $<$ 1, 0 otherwise.
The gradient of the L1 regularizer is: sign(w)
And The gradient of the L2 regularizer is: w

For binary classification problems, the algorithm outputs a binary svm model. Given a new data point, denoted by x, the model makes predictions based on the value of wTx.

By default (threshold=0), if wTx >= 0, then the response is positive (1), else the response is negative (0).

Frovedis provides implementation of linear SVM with two different optimizers: (1) stochastic gradient descent with minibatch and (2) LBFGS optimizer.

The simplest method to solve optimization problems of the form **min f(w)** is gradient descent. Such first-order optimization methods well-suited for large-scale and distributed computation. Whereas, L-BFGS is an optimization algorithm in the family of quasi-Newton methods to solve the optimization problems of the similar form.

Like the original BFGS, L-BFGS (Limited Memory BFGS) uses an estimation to the inverse Hessian matrix to steer its search through feature space, but where BFGS stores a dense nxn approximation to the inverse Hessian (n being the number of features in the problem), L-BFGS stores only a few vectors that represent the approximation implicitly. L-BFGS often achieves rapider convergence compared with other first-order optimization.

This module provides a client-server implementation, where the client application is a normal python scikit-learn program. Scikit-learn has its own svm module providing the LinearSVC (Support Vector Classification) support. But that algorithm is non-distributed in nature. Hence it is slower when comparing with the equivalent Frovedis algorithm (see frovedis manual for ml/linear_svm) with big dataset. Thus in this implementation, a scikit-learn client can interact with a frovedis server sending the required python data for training at frovedis side. Python data is converted into frovedis compatible data internally and the scikit-learn ML call is linked with the respective frovedis ML call to get the job done at frovedis server.

Scikit-learn side call for Linear SVC quickly returns, right after submitting the training request to the frovedis server with a unique model ID for the submitted training request.

When operations like prediction will be required on the trained model, scikit-learn client sends the same request to frovedis server on the same model (containing the unique ID) and the request is served at frovedis server and output is sent back to the scikit-learn client.

14.3.1 Detailed Description

14.3.1.1 LinearSVC()

Parameters
penalty: A string object containing the regularizer type to use. (Default: 'l2')
loss: A string object containing the loss function type to use. (Default: 'hinge')
dual: A boolean parameter (unused)
tol: A double parameter specifying the convergence tolerance value, (Default: 1e-4)
C: A double parameter containing the learning rate. (Default: 0.01)
multi_class: A string object specifying type of classification. (Default: 'ovr')
fit_intercept: A boolean parameter specifying whether a constant (intercept) should be added to the decision function. (Default: True)
intercept_scaling: An integer parameter. (unused)

class_weight: A python dictionary or a string object. (unused)
verbose: An integer object specifying the log level to use. (Default: 0)
random_state: An integer, None or RandomState instance. (unused)
max_iter: An integer parameter specifying maximum iteration count. (Default: 1000)
solver: A string object specifying the solver to use. (Default: 'sag')

Purpose
It initialized a Lasso object with the given parameters.

The parameters: "dual", "intercept_scaling", "class_weight", and "random_state" are not yet supported at frovedis side. Thus they don't have any significance in this call. They are simply provided for the compatibility with scikit-learn application.

"penalty" can be either 'l1' or 'l2' (Default: 'l2').
"loss" value can only be 'hinge'.

"solver" can be either 'sag' for frovedis side stochastic gradient descent or 'lbfgs' for frovedis side LBFGS optimizer when optimizing the linear regression model.

"multi_class" can only be 'ovr' as frovedis suppots binary classification algorithms only at this moment.

"verbose" value is set at 0 by default. But it can be set to 1 (for DEBUG mode) or 2 (for TRACE mode) for getting training time logs from frovedis server.

Return Value
It simply returns "self" reference.

14.3.1.2 fit(X, y, sample_weight=None)

Parameters
X: A scipy sparse matrix or any python array-like object or an instance of FrovedisCRSMatrix.
y: Any python array-like object or an instance of FrovedisDvector.
sample_weight: Python array-like optional parameter. (unused)

Purpose
It accepts the training feature matrix (X) and corresponding output labels (y) as inputs from the user and trains a linear regression model with specifed regularization with those data at frovedis server.

It doesn't support any initial weight to be passed as input at this moment. Thus the "sample_weight" parameter will simply be ignored. It starts with an initial guess of zeros for the model vector and keeps updating the model to minimize the cost function until convergence is achieved or maximum iteration count is reached.

For example,

```
# loading sample CRS data file
mat = FrovedisCRSMatrix().load("./sample")
lbl = FrovedisDvector([1,0,1,1,1,0,1,1])

# fitting input matrix and label on linear SVC object
lr = LinearSVC(solver='sgd', verbose=2).fit(mat,lbl)
```

Return Value
It simply returns "self" reference.
Note that the call will return quickly, right after submitting the fit request at frovedis server side with a unique model ID for the fit request. It may be possible that the training is not completed at the frovedis server side even though the client scikit-learn side fit() returns.

14.3.1.3 predict(X)

Parameters
X: A scipy sparse matrix or any python array-like object or an instance of FrovedisCRSMatrix.

Purpose
It accepts the test feature matrix (X) in order to make prediction on the trained model at frovedis server.

Return Value
It returns a numpy array of double (float64) type containing the predicted outputs.

14.3.1.4 predict_proba(X)

Parameters
X: A scipy sparse matrix or any python array-like object or an instance of FrovedisCRSMatrix.

Purpose
It accepts the test feature matrix (X) in order to make prediction on the trained model at frovedis server. But unlike predict(), it returns the probability values against each input sample to be positive.

Return Value
It returns a numpy array of double (float64) type containing the prediction probability values.

14.3.1.5 save(filename)

Parameters
filename: A string object containing the name of the file on which the target model is to be saved.

Purpose
On success, it writes the model information (weight values etc.) in the specified file as little-endian binary data. Otherwise, it throws an exception.

Return Value
It returns nothing.

14.3.1.6 load(filename)

Parameters
filename: A string object containing the name of the file having model information to be loaded.

Purpose
It loads the model from the specified file (having little-endian binary data).

Return Value
It simply returns "self" instance.

14.3.1.7 debug_print()

Purpose
It shows the target model information (weight values etc.) on the server side user terminal. It is mainly used for debugging purpose.

Return Value
It returns nothing.

14.3.1.8 release()

Purpose
It can be used to release the in-memory model at frovedis server.

Return Value
It returns nothing.

14.4 SEE ALSO

logistic_regression, dvector, crs_matrix

Chapter 15

Matrix Factorization using ALS

15.1 NAME

Matrix Factorization using ALS - A matrix factorization algorithm commonly used for recommender systems.

15.2 SYNOPSIS

class frovedis.mllib.recommendation.ALS (max_iter=100, alpha=0.01, regParam=0.01,
seed=0, verbose=0)

15.2.1 Public Member Functions

fit(X, rank)
predict(id)
recommend_users (pid, k)
recommend_products (uid, k)
save(filename)
load(filename)
debug_print()
release()

15.3 DESCRIPTION

Collaborative filtering is commonly used for recommender systems. These techniques aim to fill in the missing entries of a user-item association matrix. Frovedis currently supports model-based collaborative filtering, in which users and products are described by a small set of latent factors that can be used to predict missing entries.

Frovedis uses the alternating least squares (ALS) algorithm to learn these latent factors. The algorithm is based on a paper "Collaborative Filtering for Implicit Feedback Datasets" by Hu, et al.

This module provides a client-server implementation, where the client application is a normal python scikit-learn program. Scikit-learn does not have any collaborative filtering recommender algorithms like ALS. In this implementation, scikit-learn side recommender interfaces are provided, where a scikit-learn client can interact with a frovedis server sending the required python data for training at frovedis side. Python data is

converted into frovedis compatible data internally and the scikit-learn ALS call is linked with the frovedis ALS call to get the job done at frovedis server.

Scikit-learn side call for ALS quickly returns, right after submitting the training request to the frovedis server with a unique model ID for the submitted training request.

When operations like recommendation will be required on the trained model, scikit-learn client sends the same request to frovedis server on the same model (containing the unique ID) and the request is served at frovedis server and output is sent back to the scikit-learn client.

15.3.1 Detailed Description

15.3.1.1 ALS ()

Parameters
max_iter: An integer parameter specifying maximum iteration count. (Default: 100)
alpha: A double parameter containing the learning rate (Default: 0.01)
regParam: A double parameter containing the regularization parameter (Default: 0.01)
seed: A long parameter containing the seed value to initialize the model structures with random values. (Default: 0)
verbose: An integer parameter specifying the log level to use. (Default: 0)

Purpose

It initialized an ALS object with the given parameters.

"verbose" value is set at 0 by default. But it can be set to 1 (for DEBUG mode) or 2 (for TRACE mode) for getting training time logs from frovedis server.

Return Value
It simply returns "self" reference.

15.3.1.2 fit(X, rank)

Parameters
X: A scipy sparse matrix or any python array-like object or an instance of FrovedisCRSMatrix.
rank: An integer parameter containing the user given rank for the input matrix.

Purpose
It accepts the training matrix (X) and trains a matrix factorization model on that at frovedis server.

It starts with initializing the model structures of the size MxF and NxF (where M is the number of users and N is the products in the given rating matrix and F is the given rank) with random values and keeps updating them until maximum iteration count is reached.

For example,

```
# loading sample CRS data file
mat = FrovedisCRSMatrix().load("./sample")

# fitting input matrix on ALS object
als = ALS().fit(mat,rank=4)
```

Return Value
It simply returns "self" reference.
Note that the call will return quickly, right after submitting the fit request at frovedis server side with a unique model ID for the fit request. It may be possible that the training is not completed at the frovedis server side even though the client scikit-learn side fit() returns.

15.3.1.3 predict(ids)

Parameters
ids: A python tuple or list object containing the pairs of user id and product id to predict.

Purpose
It accepts a list of pair of user ids and product ids (0-based ID) in order to make prediction for their ratings from the trained model at frovedis server.

For example,

```
# this will print the predicted ratings for the given list of id pairs
print als.predict([(1,1), (0,1), (2,3), (3,1)])
```

Return Value
It returns a numpy array of double (float64) type containing the predicted ratings.

15.3.1.4 recommend_users(pid, k)

Parameters
pid: An integer parameter specifying the product ID (0-based) for which to recommend users.
k: An integer parameter specifying the number of users to be recommended.

Purpose
It recommends the best "k" users with highest rating confidence in sorted order for the given product.

If k > number of rows (number of users in the given matrix when training the model), then it resets the k as "number of rows in the given matrix" in order to recommend all the users with rating confidence values in sorted order.

Return Value
It returns a python list containing the pairs of recommended users and their corresponding rating confidence values in sorted order.

15.3.1.5 recommend_products(uid, k)

Parameters
uid: An integer parameter specifying the user ID (0-based) for which to recommend products.
k: An integer parameter specifying the number of products to be recommended.

Purpose
It recommends the best "k" products with highest rating confidence in sorted order for the given user.

If k > number of columns (number of products in the given matrix when training the model), then it resets the k as "number of columns in the given matrix" in order to recommend all the products with rating confidence values in sorted order.

Return Value
It returns a python list containing the pairs of recommended products and their corresponding rating confidence values in sorted order.

15.3.1.6 save(filename)

Parameters
filename: A string object containing the name of the file on which the target model is to be saved.

Purpose
On success, it writes the model information (user-product features etc.) in the specified file as little-endian binary data. Otherwise, it throws an exception.

Return Value
It returns nothing.

15.3.1.7 load(filename)

Parameters
filename: A string object containing the name of the file having model information to be loaded.

Purpose
It loads the model from the specified file (having little-endian binary data).

Return Value
It simply returns "self" instance.

15.3.1.8 debug_print()

Purpose
It shows the target model information on the server side user terminal. It is mainly used for debugging purpose.

Return Value
It returns nothing.

15.3.1.9 release()

Purpose
It can be used to release the in-memory model at frovedis server.

Return Value
It returns nothing.

Chapter 16

kmeans

16.1 NAME

kmeans - A clustering algorithm commonly used in EDA (exploratory data analysis).

16.2 SYNOPSIS

class frovedis.mllib.cluster.KMeans (n_clusters=8, init='k-means++',
n_init=10, max_iter=300, tol=1e-4, precompute_distances='auto',
verbose=0, random_state=None, copy_x=True,
n_jobs=1, algorithm='auto')

16.2.1 Public Member Functions

fit(X, y=None)
predict(X)
save(filename)
load(filename)
debug_print()
release()

16.3 DESCRIPTION

Clustering is an unsupervised learning problem whereby we aim to group subsets of entities with one another based on some notion of similarity. K-means is one of the most commonly used clustering algorithms that clusters the data points into a predefined number of clusters (K).

This module provides a client-server implementation, where the client application is a normal python scikit-learn program. Scikit-learn has its own cluster module providing kmeans support. But that algorithm is non-distributed in nature. Hence it is slower when comparing with the equivalent Frovedis algorithm (see frovedis manual for ml/kmeans) with big dataset. Thus in this implementation, a scikit-learn client can interact with a frovedis server sending the required python data for training at frovedis side. Python data is converted into frovedis compatible data internally and the scikit-learn ML call is linked with the respective frovedis ML call to get the job done at frovedis server.

Scikit-learn side call for kmeans quickly returns, right after submitting the training request to the frovedis server with a unique model ID for the submitted training request.

When operations like prediction will be required on the trained model, scikit-learn client sends the same request to frovedis server on the same model (containing the unique ID) and the request is served at frovedis server and output is sent back to the scikit-learn client.

16.3.1 Detailed Description

16.3.1.1 KMeans()

Parameters *n_clusters*: An integer parameter specifying the number of clusters. (Default: 8)
init: A string object. (unused)
n_init: An integer parameter. (unused)
max_iter: An integer parameter specifying the maximum iteration count. (Default: 300)
tol: A double parameter specifying the convergence tolerance. (Default: 1e-4)
precompute_distances: A string object. (unused)
verbose: An integer object specifying the log level to use. (Default: 0)
random_state: An integer, None or RandomState instance. (unused)
copy_X: A boolean parameter. (unused)
n_jobs: An integer parameter. (unused)
algorithm: A string object. (unused)

Purpose
It initialized a KMeans object with the given parameters.

The parameters: "init", "n_init", "precompute_distances", "random_state", "copy_X", "n_jobs" and "algorithms" are not yet supported at frovedis side. Thus they don't have any significance in this call. They are simply provided for the compatibility with scikit-learn application.

"verbose" value is set at 0 by default. But it can be set to 1 (for DEBUG mode) or 2 (for TRACE mode) for getting training time logs from frovedis server.

Return Value
It simply returns "self" reference.

16.3.1.2 fit(X, y=None)

Parameters
X: A scipy sparse matrix or any python array-like object or an instance of FrovedisCRSMatrix.
y: None (simply ignored in scikit-learn as well).

Purpose

It clusters the given data points (X) into a predefined number (k) of clusters.

For example,

```
# loading sample CRS data file
mat = FrovedisCRSMatrix().load("./sample")

# fitting input matrix on kmeans object
kmeans = KMeans(n_clusters=2, verbose=2).fit(mat)
```

Return Value
It simply returns "self" reference.

Note that the call will return quickly, right after submitting the fit request at frovedis server side with a unique model ID for the fit request. It may be possible that the training is not completed at the frovedis server side even though the client scikit-learn side fit() returns.

16.3.1.3 predict(X)

Parameters
X: A scipy sparse matrix or any python array-like object or an instance of FrovedisCRSMatrix.

Purpose
It accepts the test data points (X) and returns the centroid information.

Return Value
It returns a numpy array of integer (int32) type containing the centroid values.

16.3.1.4 save(filename)

Parameters
filename: A string object containing the name of the file on which the target model is to be saved.

Purpose
On success, it writes the model information in the specified file as little-endian binary data. Otherwise, it throws an exception.

Return Value
It returns nothing.

16.3.1.5 load(filename)

Parameters
filename: A string object containing the name of the file having model information to be loaded.

Purpose
It loads the model from the specified file (having little-endian binary data).

Return Value
It simply returns "self" instance.

16.3.1.6 debug_print()

Purpose
It shows the target model information on the server side user terminal. It is mainly used for debugging purpose.

Return Value
It returns nothing.

16.3.1.7 release()

Purpose
It can be used to release the in-memory model at frovedis server.

Return Value
It returns nothing.