

ECL DOCUMENTATION



[Go Up](#)

Name	KMeans
Version	1.0.1
Description	KMeans Bundle for Clustering algorithm
License	http://www.apache.org/licenses/LICENSE-2.0
Copyright	Copyright (C) 2019 HPCC Systems
Authors	HPCCSystems
DependsOn	ML_Core 3.2.2
Platform	6.4.0

OVERVIEW

KMeans

ECL bundle for KMeans calculation

Table of Contents

KMeans.ecl
Classic KMeans Clustering
Types.ecl
Type definition module for KMeans
Test

KMeans

[Go Up](#)

IMPORTS

ML_Core | ML_Core.Types | ML_Core.ModelOps2 | PBblas.Types | Types.KMeans_Model |
Types.KMeans_Model.Ind1 |

DESCRIPTIONS

KMEANS

/ EXPORT	KMeans
(INTEGER max_iter = 100 , REAL t = 0.00001)	

Classic KMeans Clustering.

Clustering Algorithms are a branch of unsupervised machine learning algorithms. They automatically categorize observations(points) into groups without pre-defined labels. KMeans[1] is one of the most well-known clustering algorithms. Given the data points for clustering and the K initial centroids of each cluster, the KMeans algorithm can automatically group each sample into one cluster.

KMeans is a popular clustering method for cluster analysis in data mining. It iteratively update the cluster centroids until it reaches the tolerance. KMeans module is both highly data scalable and model scalable on HPC Systems Platform.

Reference. [1] Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), 100-108.

PARAMETER **max_iter** ||| INTEGER8 — The maximum number of iterations to run KMeans. It's an integer scalar value. The default value is 100.

PARAMETER **t** ||| REAL8 — The convergence tolerance. It's a real value scalar. KMeans will stop iterating when center movement of each cluster is smaller than t between two consecutive iterations. The default value is 0.00001.

Children

1. **Fit** : Train and return a KMeans model
 2. **Centers** : Extract the final coordinates of the centers of each cluster from the trained model
 3. **Predict** : Compute the cluster center for each new sample
 4. **Labels** : Function Labels() computes the closest center of each training sample from the trained Model
 5. **Iterations** : Extract the number of iterations that each work item took to converge, from the provided model
-

FIT

KMeans /

Fit
(DATASET(NumericField) sampleset, DATASET(NumericField) initCentroids)

Train and return a KMeans model.

Fit function takes the samples and initial centroids as inputs and returns a trained KMeans model.

PARAMETER **sampleset** ||| TABLE (NumericField) — The samples to be clustered in DATASET(NumericField) format. Each observation (e.g. record) is identified by 'id', and each feature is identified by field number (i.e. 'number').

PARAMETER **initCentroids** ||| TABLE (NumericField) — The initial K centroids for clustering in DATASET(NumericField) format. Each observation (e.g. record) is identified by 'id', and each feature is identified by field number.

RETURN **TABLE ({ UNSIGNED2 wi , REAL8 value , SET (UNSIGNED4) indexes })** — KMeans Model in the format of ML_Core.Types.Layout_Model2.

SEE ML_Core.Types.Layout_Model2

SEE ML_Core.Types.NumericField

SEE `Types.KMeans_Model`

CENTERS

[KMeans](#) /

Centers
<code>(DATASET(Types.Layout_Model2) mod)</code>

Extract the final coordinates of the centers of each cluster from the trained model.

PARAMETER `mod` ||| `TABLE (Layout_Model2)` — The fitted/trained KMeans model.

RETURN `TABLE ({ UNSIGNED2 wi , UNSIGNED8 id , UNSIGNED4 number , REAL8 value })` — centers
The Final coordinates of the center of each cluster in NumericField format.

SEE `ML_Core.Types.NumericField`

PREDICT

[KMeans](#) /

<code>DATASET(KTypes.Labels)</code>	Predict
<code>(DATASET(Types.Layout_Model2) mod, DATASET(Types.NumericField) newSamples)</code>	

Compute the cluster center for each new sample.

PARAMETER `mod` ||| `TABLE (Layout_Model2)` — The fitted/trained KMeans model.

PARAMETER `newSamples` ||| `TABLE (NumericField)` — The new samples to be clustered.

RETURN `TABLE ({ UNSIGNED2 wi , UNSIGNED8 id , UNSIGNED8 label })` — The index of the closest
center for each new sample.

SEE `Types.KMeans_Model.Labels`

SEE `ML_Core.Types.NumericField`

LABELS

[KMeans](#) /

<code>DATASET(KTypes.Labels)</code>	Labels
<code>(DATASET(Types.Layout_Model2) mod)</code>	

Function `Labels()` computes the closest center of each training sample from the trained Model.

PARAMETER `mod` ||| `TABLE (Layout_Model2)` — The fitted/trained KMeans model.

RETURN `TABLE ({ UNSIGNED2 wi , UNSIGNED8 id , UNSIGNED8 label })` — The closest center index for each training sample.

SEE `Types.KMeans_Model.Labels`

ITERATIONS

[KMeans](#) /

<code>DATASET(KTypes.n_Iters)</code>	Iterations
<code>(DATASET(Types.Layout_Model2) mod)</code>	

Extract the number of iterations that each work item took to converge, from the provided model.

PARAMETER `mod` ||| `TABLE (Layout_Model2)` — The fitted/trained KMeans model.

RETURN `TABLE ({ UNSIGNED2 wi , UNSIGNED8 iters })` — iterations The total number of iterations for each wi.

SEE `Types.KMeans_Model.n_Iters`

Types

[Go Up](#)

IMPORTS

ML_Core.Types |

DESCRIPTIONS

TYPES

Types

Type definition module for KMeans.

Children

1. [KMeans_Model](#) : Definition of the meaning of the indexes of the KMeans Model variables

KMEANS_MODEL

[Types](#) /

KMeans_Model

Definition of the meaning of the indexes of the KMeans Model variables.

Ind1 enumerates the first index, which is used to determine which type of data is stored:

- Centers stores the list of centers of clusters. The second index is the centerID. The third index is the number field of the center.
- samples stores the set of sample indexes (i.e. ids) associated with each centerId. The value is the Id of its closest center.
- Iterations stores the iterations associated with each wi. It represents how many iteration runs of each wi before it stops iterating. It does not have following index.

Children

1. [Ind1](#) : Index 1 represents the category of data within the model
2. [Centers_Indexes](#) : Centers_Indexes enumerates the second and third indexes of each center which is the parent index
3. [Samples_Indexes](#) : Samples_Indexes enumerates the indexes of each sample which is the parent index
4. [Labels](#) : Labels format defines the distance space where each cluster defined by a center and its closest samples
5. [n_iters](#) : The number of iterations for which each work item was trained

IND1

[Types](#) / [KMeans_Model](#) /

Ind1

Index 1 represents the category of data within the model.

VALUE reserved = 1. Reserved for future use.

VALUE centers = 2. The set of tree nodes within the model.

VALUE samples = 3. The particular record ids that are included in tree's sample .

VALUE iterations = 4. The iteration runs of each wi.

Children

1. [reserved](#) : No Documentation Found
 2. [centers](#) : No Documentation Found
 3. [samples](#) : No Documentation Found
 4. [iterations](#) : No Documentation Found
-

RESERVED

[Types](#) / [KMeans_Model](#) / [Ind1](#) /

CTypes.t_index	reserved
--------------------------------	--------------------------

No Documentation Found

RETURN **UNSIGNED4** —

CENTERS

[Types](#) / [KMeans_Model](#) / [Ind1](#) /

CTypes.t_index	centers
--------------------------------	-------------------------

No Documentation Found

RETURN **UNSIGNED4** —

SAMPLES

[Types](#) / [KMeans_Model](#) / [Ind1](#) /

CTypes.t_index	samples
----------------	----------------

No Documentation Found

RETURN **UNSIGNED4** —

ITERATIONS

[Types](#) / [KMeans_Model](#) / [Ind1](#) /

CTypes.t_index	iterations
----------------	-------------------

No Documentation Found

RETURN **UNSIGNED4** —

CENTERS_INDEXES

[Types](#) / [KMeans_Model](#) /

Centers_Indexes

Centers_Indexes enumerates the second and third indexes of each center which is the parent index. The parent index value is 2. It is used to store the id and the field value of each center.

RETURN **UNSIGNED2** —

VALUE id = 2. The center identifier.

VALUE number = 3. The field identifier.

SAMPLES_INDEXES

[Types](#) / [KMeans_Model](#) /

Samples_Indexes

Samples_Indexes enumerates the indexes of each sample which is the parent index. The parent index value is 3. It is used to store the sampleID. The value is the Id of its closest center.

RETURN **UNSIGNED2** —

VALUE id = 2. The sample identifier.

LABELS

[Types](#) / [KMeans_Model](#) /

Labels

Labels format defines the distance space where each cluster defined by a center and its closest samples.

FIELD **wi** ||| UNSIGNED2 — The model identifier.

FIELD **id** ||| UNSIGNED8 — The sample identifier.

FIELD **label** ||| UNSIGNED8 — The identifier of the closest center to the sample.

N_ITERS

[Types](#) / [KMeans_Model](#) /

n_iters

The number of iterations for which each work item was trained.

FIELD wi ||| UNSIGNED2 — The work item id.

FIELD iters ||| UNSIGNED8 — The number of iterations.

Test

[Go Up](#)

Table of Contents

Datasets

Datasets

[Go Up](#)

Table of Contents

[DSIris.ecl](#)

The file provide the information of the testing dataset: Public Dataset Iris

Test/ Datasets/ DSIris

[Go Up](#)

IMPORTS

ML_Core.Types |

DESCRIPTIONS

DSIRIS

DSIris

The file provide the information of the testing dataset: Public Dataset Iris. Reference [1] Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [[http:*archive.ics.uci.edu/ml](http://archive.ics.uci.edu/ml)]. Irvine, CA: University of California, School of Information and Computer Science.

Children

1. [Layout](#) : No Documentation Found
 2. [ds](#) : No Documentation Found
 3. [sklearn_rst](#) : No Documentation Found
 4. [sklearn_converge](#) : No Documentation Found
 5. [sklearn_alleg](#) : No Documentation Found
-

LAYOUT

[DSIris /](#)

Layout

No Documentation Found

FIELD sepal_length ||| REAL8 — No Doc

FIELD sepal_width ||| REAL8 — No Doc

FIELD petal_length ||| REAL8 — No Doc

FIELD petal_width ||| REAL8 — No Doc

FIELD class ||| REAL8 — No Doc

DS

[DSIris /](#)

ds

No Documentation Found

RETURN TABLE (**Layout**) —

SKLEARN_RST

[DSIris /](#)

sklearn_rst

No Documentation Found

RETURN TABLE ({ INTEGER8 id , REAL8 sepal_length , REAL8 sepal_width , REAL8 petal_length ,
REAL8 petal_width }) —

SKLEARN_CONVERGE

[DSIris](#) /

<code>sklearn_converge</code>

No Documentation Found

RETURN INTEGER8 —

SKLEARN_ALLEG

[DSIris](#) /

<code>sklearn_alleg</code>

No Documentation Found

RETURN TABLE ({ INTEGER8 id , INTEGER8 y }) —
