root

Go Up

Name	KMeans
Version	1.0.0
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

Table of Contents

Cluster_Types

Go Up

IMPORTS

ML_Core.Types |

DESCRIPTIONS

MODULE Cluster_Types

Cluster_Types

Type definition module for KMeans.

Children

1. KMeans_Model : Definition of the meaning of the indexes of the KMeans Model variables

MODULE KMeans_Model

Cluster_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

---

MODULE Ind1

Cluster_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

ATTRIBUTE reserved

Cluster_Types \ KMeans_Model \ Ind1 \

`Types.t_index`	**reserved**

No Documentation Found

RETURN **UNSIGNED4** —

ATTRIBUTE centers

Cluster_Types \ KMeans_Model \ Ind1 \

`Types.t_index`	**centers**

No Documentation Found

RETURN **UNSIGNED4** —

ATTRIBUTE samples

Cluster_Types \ KMeans_Model \ Ind1 \

`Types.t_index`	**samples**

No Documentation Found

RETURN **UNSIGNED4** —

ATTRIBUTE iterations

Cluster_Types \ KMeans_Model \ Ind1 \

Types.t_index	**iterations**

No Documentation Found

RETURN **UNSIGNED4** —

ATTRIBUTE Centers_Indexes

Cluster_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.

ATTRIBUTE Samples_Indexes

Cluster_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.

RECORD Labels

Cluster_Types \ KMeans_Model \

	Labels

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

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

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

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

RECORD n_iters

Cluster_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.

KMeans

Go Up

IMPORTS

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

DESCRIPTIONS

MODULE KMeans

	KMeans
(INTEGER max_iter = 10 , REAL t = 0.0)	

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 HPCC 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 maxinum number of iterations to run KMeans. It's an integer scalar value. The default value is 10.

PARAMETER **t** ||| REAL8 — The convergence tolerance. It's a real value scalar. The default value is 0.0.

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

---

FUNCTION Fit

KMeans \

Fit

```
(DATASET(Types.NumericField) d1, DATASET(Types.NumericField) d2)
```

Train and return a KMeans model.

Fit function takes the samples d1 and initial centroids d2 and returns a trained KMeans model.

PARAMETER **d1** ||| 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 **d2** ||| 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 Cluster_Types.KMeans_Model

FUNCTION Centers

KMeans \

	Centers
`(DATASET(Types.Layout_Model2) mod)`	

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

FUNCTION Predict

KMeans \

`DATASET(KTypes.Labels)`	**Predict**

`(DATASET(Types.Layout_Model2) mod,`
`DATASET(Types.NumericField) newSamples)`	

Compute the cluster center for each new sample.

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

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

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

SEE Cluster_Types.KMeans_Model.Labels

SEE ML_Core.Types.NumericField

FUNCTION Labels

KMeans \

DATASET(KTypes.Labels)	**Labels**
(DATASET(Types.Layout_Model2) mod)	

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 Cluster_Types.KMeans_Model.Labels

FUNCTION Iterations

KMeans \

DATASET(KTypes.n_Iters)	**Iterations**
(DATASET(Types.Layout_Model2) mod)	

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 Cluster_Types.KMeans_Model.n_Iters