BigML Cluster Cheat Sheet

Cluster configuration

Cluster configuration options

Option	Description	Default	API Name
K-means: number of clusters	Specifies the number of clusters to be found. Choose from 2 up to 300 clusters.	8	k
G-means: critical value	Determines how strictly the data distribution in the neighborhood of a candidate cluster should fit a Gaussian distribution. The strictest fit corresponds to a value of 1. Higher critical values will tend to find fewer clusters. You can set a maximum value of 20.	5	critical_value
Default numeric value	Replaces missing numeric values in the dataset with the field maximum, mean, median, minimum, or zero. If you don't enable this option, your instances with missing numeric values will be ignored. However, if all the instances contain at least a missing value, BigML automatically replaces them with the median.	Null	default_numeric_value
Scale fields	Assigns scales to particular fields.	False	field_scales
Auto-scaled	Scales all numeric fields so their standard deviations are 1. This makes each field have roughly equivalent influence.	True	balance_fields
Weights	Assigns different weights to the instances in your dataset. Any numeric field with no negative or missing values is valid as a weight assignable field. Each instance will be weighted individually according to the weight field's value.	False	weight_field
Summary fields	Sets summary fields, which aren't used to compute clusters, but their values are included when you create a dataset from a cluster. Non-preferred fields aren't eligible as summary fields. If you want to include a non-preferred field as a summary field, you will first need set that field as preferred.	[ ]	summary_fields

Sampling

Option	Description	Default	API Name
Rate	Allows you to set the proportion of the dataset you want to consider between 0% and 100%.	100%	sample_rate
Range	Specifies a linear subset of the dataset instances that you want to be considered for the sample (example: from instance 5 to instance 1,000). The rate you set will be computed over the range configured.	(1, max. rows in dataset)	range

Sampling (Cont.)

Option	Description	Default	API Name
Sampling	Chooses between a random sampling or a deterministic sampling. When using deterministic sampling the random-number generator will always use the same seed, producing repeatable results.	Random	seed
Replacement	Enables sampling with replacement, which allows a single instance to be selected multiple times. Sampling without replacement ensures that each instance cannot be selected more than once.	False	replacement
Out of bag	Enables out-of-bag sampling. If an instance is not selected as part of a sampling, it is considered out of bag. Enabling this will select only the out-of-bag instances for the currently defined sample. It can only be chosen when a sample is deterministic and the sample rate is less than 100%.	False	out_of_bag

Batch Centroid Configuration

Default Numeric Values

Option	Description	Default	API Name
Default numeric value	Replaces missing numeric values in your dataset by the field's maximum, mean, median, minimum, or zero. If you do not activate this option, your instances with missing numeric values will be ignored and you will not get a prediction for them.	Null	default_numeric_value

Output File Options

Option	Description	Default	API Name
Fields separator	Allows you to choose the best separator for your fields.	Comma	separator
Show/hide fields	Allows you to show or hide the rest of the fields in your output file.	True	output_fields
Headers	Allows you to show or hide the names of your columns in the output file.	True	header
Centroid column name	Allows you to set the name for the centroid column in your output file.	Centroid	centroid_name
Distance	Allows you to include the distance for each centroid per instance.	False	distance
Distance column name	Allows you to set the name for the distance column in your output file.	Centroid	distance_name
New line	Sets the character to use as the line break in the generated csv file: "LF", "CRLF".	LF	newline

Output Dataset

Option	Description	Default	API Name
Output dataset	Defines whether a dataset with the results should be automatically created or not.	True	output_dataset