BigML PCA Cheat Sheet

PCA Configuration

PCA Configuration Options

Option	Description	Default	API Name
Standardize	Allows you to automatically scale numeric fields to a 0-1 range. Standardizing implies assigning equal importance to all the fields when these are not measured on the same scale; otherwise it is often the case that each principal component is dominated by a single field.	True	standardize
Default numeric value	Replaces missing numeric values in the dataset with the field maximum, mean, median, minimum, or zero.	Null	default_numeric_value

Sampling

Option	Description	Default	API Name
Rate	Sets the proportion of the dataset you want to consider between 0% and 100%.	100%	sample_rate
Range	Specifies a subset of instances from which to sample, e.g., 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	Allows you to choose 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	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	Selects only the out-of-bag instances for the currently defined sample. If an instance is not selected as part of a sample, it is considered out-of-bag. It is only selectable when a sample is deterministic and the sample rate is less than 100%.	False	out_of_bag

Batch Projections

Limit number of components

Option	Description	Default	API Name
Cumulative variance	Allows you to limit the total components to be returned by setting a threshold between 0% and 100%. The prediction uses the minimum number of components such that the cumulative explained variance is greater than the given threshold.	Null	variance_threshold
Maximum components	Allows you to limit the number of components to be returned by setting an integer greater than 0.	Null	max_components

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
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