SplitDatasetTrainTest Documentation

Description: Splits a dataset (and cls file) into a number of train and test subsets

Author: Stefano Monti (Broad Institute), Marc-Danie Nazaire (Broad Institute), gp-help@broad.mit.edu

Summary: Partitions a dataset into one or more train/test sets. The partitioning is done using either a percentage split or cross-validation approach. The percentage split option splits a dataset according to the specified percentage into a train and a test file. The cross-validation method partitions the samples into k-folds partitions. Each of the k-folds partitions are used as the test dataset and the remaining k-1 folds are used as the train dataset. If the option to stratify the splits is selected the class template is used in order to split the different classes evenly between a train/test set. The output prefix parameter is used to name the generated train/test datasets. For example, if the prefix is all_aml the train/test datasets will be named: aml_all.train.0, aml_all.test.0, aml_all.train.1, aml_all.test.1, etc.

Parameters:

Name	Description
input.dataset.file	input dataset - .gct, .res
cls.file	class template - .cls
split.method	whether to split the data using a percentage split or cross validation approach
stratified	whether to create splits stratified with respect to the class template
folds	number of train/test folds to generate
percentage.split.proportion	proportion of data to be allocated to train file when split method is percentage split (ignored when split method is cross-validation)
seed	random number generator seed
output.prefix	output saved to output.prefix.{trn,tst}.n.{gct|res,cls}, where n is the split index

Platform dependencies:

Module type: Preprocess & Utilities
CPU type: any
OS: any
Language: R