MDITRE User Manual

December 21, 2021

1 Supported options for data configuration file file

All of the following options are optional, unless stated as required.

1.1 description

Options allowed in section 'description' are:

tag (Required) A short string used to name the serialized pickle object containing the processed dataset. The file is saved in the current directory with the filename in the format `[tag]_dataset_object.pickle`.

1.2 data

Options allowed in section 'data' are:

data_type '16s' (the default) or 'metaphlan'.

abundance_data (Required) Filename or path to a table of OTU abundance data

sample_metadata (Required) Filename or path to a table giving a subject and timepoint for each sample in the abundance table.

subject_data (Required) Filename or path to a table of data about each subject

jplace_file (Required unless data_type is 'metaphlan') Filename or path to pplacer results in `.jplace` format

sequence_key Filename or path to a FASTA file to be used to rename the OTUs listed in the abudance table, in the case where the name of each OTU is simply a sequence, as in some DADA2.

outcome_variable (Required) String, specifying which of the columns in the subject data table encodes the outcome for prediction.

outcome_positive_value (Required) The value of the outcome variable that corresponds to a positive outcome.

taxonomy_source String specifying how taxonomic annotations for each variable (i.e., subtree of the overall phylogeny) should be generated. Valid options are 'table', 'pplacer', and 'hybrid' (in short, 'table' labels OTUs from a table and higher clades based on the OTUs they contain, 'pplacer' labels all variables based on the pplacer results, and 'hybrid' labels variables based on pplacer results except where a species-level placement is given in a table.) If this option is not present, no such annotation will be done. If given, note that the table of annotations will be written to `[tag]_variable_annotations.txt` at the conclusion of the preprocessing step, unless phylogenetic aggregation is not performed.

placement_table Filename or path to a table of taxonomic placements for OTUs

pplacer_taxa_table Filename or path to a table from the pplacer reference package. Required if 'taxonomy_source' is 'pplacer' or 'hybrid'; otherwise ignored.

pplacer_seq_info Filename or path to a table from the pplacer reference package. Required if 'taxonomy_source' is 'pplacer' or 'hybrid'; otherwise ignored.

1.3 preprocessing

Options allowed in section 'preprocessing' are:

min_overall_abundance If this is specified, all OTUs with lower total abundance data than this value, summed across all samples, are dropped. (Abundance data is assumed to be measured in 16S read counts at this stage, but this is not enforced.)

min_sample_reads If this is specified, all samples where the total abundance data across all (remaining) OTUs sum to less than this value are dropped. (Abundance data is assumed to be measured in 16S read counts at this stage, but this is not enforced.)

trim_start The time to consider as the beginning of the study. (By default, this will be the timepoint of the earliest (remaining) sample.) Samples before this time will be dropped. If this is given, 'trim_stop' must be given also.

trim_stop The time to consider as the end of the study. (By default, this will be the timepoint of the latest (remaining) sample.) Samples after this time will be dropped.

density_filter_n_samples If this is given, 'density_filter_n_intervals' and 'density_filter_n_consecutive' must be given also. The duration of the study will be divided into the specified number of time intervals, and subjects from whom at least the specified number of samples are available in *every* time window formed from the specified number of consecutive intervals are retained; samples from all other subjects are dropped.

density_filter_n_intervals See above.

density_filter_n_consecutive See above.

take_relative_abundance If True, abundance data will be converted to relative abundances by normalizing so that the sum of the data for each sample is 1.0.

temporal_abundance_threshold If this option is given, tthe following two options must also be given: 'temporal_abundance_consecutive_samples' and 'temporal_abundance_n_subjects'. Variables which exceed the abundance threshold in at least the specified number of consecutive samples in the data from each of at least the specified number of subjects will be kept; others will be dropped.

temporal_abundance_consecutive_samples See above.

temporal_abundance_n_subjects See above.

pickle_dataset If true, the datastructure corresponding to the processed and filtered dataset will be serialized to `[tag]_dataset_object.pickle` at the conclusion of the preprocessing step.

2 General options

- **data**: path to the pickle file containing the preprocessed dataset
- **data_name**: Name of the dataset, used to create a directory to store the model output
- **save_as_csv**: Save all the model output as CSV files
- **verbose**: Print training output logs to the console

3 Training options

- **workers**: Number of cpu threads to use for pytorch data loading
- **epochs**: Number of training iterations to run the model
- **batch-size**: Number of training samples in a single batch. For full batch training it is set to the dataset size

- **deterministic**: Deterministic training for reproducibility
- **seed**: Random seed for reproducibility
- **cv_type**: Cross-Validation procedure to be used. Options are ”loo” (leave-one-out), ”kfold” and ”None”
- **kfolds**: Number of folds to use if cv_type chosen is ”kfold”
- **distributed**: Use multiprocessing for training
- **local_rank**: Rank of the current process if using distributed training

4 Optimization options

- **lr_fc**: Learning rate for the regression coefficients. Value of 0.001 works well in practice.
- **lr_bias**: Learning rate for the regression bias term. Value of 0.001 works well in practice.
- **lr_alpha**: Learning rate on detector selector parameters. Value of 0.001 works well in practice. Using a very high value may result in model not training.
- **lr_beta**: Learning rate on rule selector parameters. Value of 0.001 works well in practice. Using a very high value may result in model not training.
- **lr_thresh**: Learning rate on abundance threshold parameters. This value is dependant on the scale of the abundances. We found that value of 0.001 works well for all our datasets.
- **lr_slope**: Learning rate on slope threshold parameters. This value is dependent on the scale of rate of change of abundances. We found that a value of 0.0001 works well for all our datasets.
- **lr_time**: Learning rate on time window length parameters. Value of 0.01 works well in practice since the time windows are generally on a higher scale (days, months etc.)
- **lr_mu**: Learning rate on time window center parameters. Value of 0.01 works well in practice since the time window centers are generally on a higher scale (day, month etc.)
- **lr_kappa**: Learning rate on phylogenetic radius parameters. Value of 0.001 works well in practice.
- **lr_eta**: Learning rate on phylogenetic embedding parameters. Value of 0.001 works well in practice.

5 Model options

- **min_k_bc**: Initial temperature (sharpening factor) before annealing on the rule and detector selectors.
- **max_k_bc**: Final temperature (sharpening factor) after annealing on the rule and detector selectors.
- **min_k_thresh**: Initial temperature (sharpening factor) before annealing on the threshold detector response.
- **max_k_thresh**: Final temperature (sharpening factor) after annealing on the threshold detector response.
- **min_k_slope**: Initial temperature (sharpening factor) before annealing on the slope detector response.
- **max_k_slope**: Final temperature (sharpening factor) after annealing on the slope detector response.
- **min_k_time**: Initial temperature (sharpening factor) before annealing on the temporal focus response.
- **max_k_time**: Final temperature (sharpening factor) after annealing on the temporal focus response.
- **min_k_otu**: Initial temperature (sharpening factor) before annealing on the phylogenetic focus response.
- **max_k_otu**: Final temperature (sharpening factor) after annealing on the phylogenetic focus response.
- **z_mean**: Mean of the Negative Binomial prior on the detector selectors.
- **z_var**: Variance of the Negative Binomial prior on the detector selectors.
- **z_r_mean**: Mean of the Negative Binomial prior on the rule selectors.
- **z_r_var**: Variance of the Negative Binomial prior on the rule selectors.
- **w_var**: Variance of the Normal prior on the regression coefficients.