VARITY FRAMWORK USER GUIDE

1. INTRODUCTION 1

1.1 STEPS TO APPLY VARITY FRAMEWORK 1
1.2 TO BUILD VARITY MODELS 1
1.3 SYSTEM REQUIREMENT 2
1.4 TECHNICAL SUPPORT 2

2. VARITY DATA PREPARATION 2

3. VARITY SESSION CONFIGURATION 2

3.1 DATA OBJECT CONFIGURATION 2
3.2 ESTIMATOR OBJECT CONFIGURATION 3
3.3 PREDICTOR OBJECT CONFIGURATION 4
3.4 QUALITY INFORMATIVE PROPERTY CONFIGURATION 5
3.5 HYPERPARAMETER CONFIGURATION 6

4. VARITY COMMANDS 6

5. TABLE 1: REQUIRED PYTHON PACKAGES 10

1. Introduction

VARITY is a supervised machine learning approach to build specialized predictive models using training examples with optimized differential weights. Training examples are assembled into different training sets: 1) one core set of training examples which are known to have high quality. 2) one or more add-on sets of training examples with uncertain quality. For each training set, the weights of examples are determined using one or more logistic functions each takes one quality-informative property as input. The parameters of each logistic function are treated as hyper-parameters and optimized for performance on the core set of examples using cross-validation.

1.1 Steps to apply VARITY framework

1. Assemble training examples (high-quality core set and add-on sets with uncertain quality).
2. Identify quality-informative properties for each add-on set using domain knowledge (Optional: verify using moving window analysis).
3. Config hyper-parameters based on quality-informative properties (parameters of corresponding logistic functions).
4. Optional: Run nested cross-validation to evaluate performance.
5. Run hyper-parameter optimization to determine weight of each training example.
6. Train the final VARITY model with core set and weighted add-on sets.

1.2 System Requirement

Python version 3.7.2 and a few python packages listed in Table 1.

1.3 Technical Support

Please contact joe.wu.ca@gmail.com for technical support.

2. VARITY project folder

A project folder that has sub-folders with pre-defined names is needed to start a VARITY project, please refer to http://varity.varianteffect.org/downloads/project.tar.gz.

3. VARITY data preparation

All core and add-on sets of training examples need to be assembled into one single CSV file as an input, and data columns include:

- Feature columns (currently only supports Real or Integer type features)
- A "label" column: the dependent variable (0 or 1 for binary classification problem)
- A "extra_data" column: 0 for core examples, and 1 for add-on examples.
- A "set_name" column: name for different core and add-on sets.
- Candidate quality-informative property columns.

VARITY data files location: /project_path/data
The data used in manuscript "***Improved pathogenicity prediction for rare human missense variants***" is located in http://varity.varianteffect.org/downloads/VARITY_training.tar.gz

4. VARITY session configuration

VARITY session uses a configuration file (*/project_path/config/YOUR_SESSION_ID.vsc)* to setup everything needed for training a VARITY model. The configuration file contains settings for four different type of objects used in VARITY framework: 1) Data 2) Estimator 3) Predictor 4) Quality informative property 5) Hyperparameter. Each object has a list of attributes, and their names can be found in a "object definition" line (starts with "*" symbol) delimited by "|" sign. A new object instance can be created below the "object definition" line, starting with the object instance name and then values of each attribute. A valid session config file is needed before running any VARITY commands. The config file used in manuscript "***Improved pathogenicity prediction for rare human missense variants***" is located in http://varity.varianteffect.org/downloads/varity.vsc.

Data object configuration

The data object takes the input training data (consists of core and add-on training examples) and splits it in nested cross validation fashion. The core training examples were first splitted into *k1* outer-loop folds. For each outer-loop, the core training examples are splitted into an outer-loop training set and an outer-loop test set, and subsequently the outer-loop training examples are further splitted into *k2* inner-loop folds. The purpose of outer-loop cross-validation is to fairly evaluate model performance on core set of examples, and inner-loop cross-validation is to optimize hyperparameter for each outer-loop. The attributes for data object:

- **test_split_method**

The way of splitting for outer-loop cross-validation.

0: random split

1: stratified split (keep same prior for the training and test in each outer-loop fold)

- **test_split_folds**

Number of folds *(k1)* for outer-loop cross-validation

- **test_split_ratio**

Only used when the “test_split_folds” attribute is set to 1. When there is only one outer-loop fold, this value determines the fraction of the core set of examples as test set. If the value is set to 0, the corresponding inner-loop cross-validation will be based on the whole core set of examples, which is the appropriate configuration for the data object used for building a final VARITY model.

- **cv_split_method**

The way of splitting for inner-loop cross-validation.

0: random split

1: stratified split (keep same prior for the training and validation in each inner-loop fold)

- **cv_split_folds**

Number of folds for inner-loop cross-validation

- **cv_split_ratio**

Only used when the “cv_split_folds” attribute is set to 1. When there is only one inner-loop fold, this value determines the fraction of the outer-loop training examples as inner-loop validation set.

- **data_file**

The absolute path of the input training data.

4.1 Estimator object configuration

VARITY framework is designed to support a list of different machine learning algorithms. However, currently it only support the gradient boosted trees algorithm. The attributes for data object:

- **algo_name**

The name of the learning algorithm. The name for gradient boosted trees for binary classification is “xgb_c”

- **round_digits**

The significant digits of the output metrics (e.g., AUROC)

4.2 Predictor object configuration

The predictor object is associated with one data object instance and one estimator instance, and has following attributes:

- **type**

0: VARITY model predictor

1: Other existing predictor (predictions can be found as one column in the data file)

- **ml_type**

VARITY framework is designed to support different type of machine learning tasks. However, currently only the following type is supported:

“classification_binary”: binary classification

- **data**

the name of the associated data object instance

- **estimator**

the name of the associated estimator object instance

- **tune_obj**

VARITY framework is designed to support a list of metrics as the objective function for hyperparameter optimization and performance evaluation using cross-validation. Currently only the following metrics are supported:

“macro_cv_aubprc”: The Area Under Balanced Precision Recall Curve via cross-validation

“macro_cv_auroc”: The Area Under ROC curve via cross-validation

- **hyperopt_trials**

The number of hyperparameter tuning trials (HyperOpt trials)

- **trials_mv_size**

The number of trials in one moving window (used for select the best trial, see Section 4 save_best_hp command).

- **features**

A list of features used for the predictor. For other predictors with existing predictions (type = 1), just use the corresponding column name in the data file as a single feature. All features need to be put in square brackets ([]) and separated by comma.

4.3 Quality informative property (qip) configuration

- **predictors**

A list of predictors (see 3.3) associated with the current quality informative property.

- **weight_function**

The weight function that takes the current quality informative property as input. The output of the weight function is used for weighting training examples. Currently only 'logistic' is supported.

- **hyperparameters**

The name of the parameters for the weight function that are treated as hyperparameters.

- **set_list**

A list of training sets. The current weight function (quality informative property) applies on the training examples in the all the sets listed here.

- **set_type**

The type of the listed sets, either 'core' or 'addon'

- **qip_col**

The column name of quality informative property in the data files.

- **direction**

  0: order the quality-informative property from low to high in moving windows analysis.

  1: order the quality-informative property from high to low in moving windows analysis.

- **mv_size_precent**

  For moving window analysis, this parameter determines the number of examples in each moving window, which equals to product of [number of examples in an add-on set] and [mv_size_percent] (rounded).

- **mv_data_points**

  Number of moving windows for moving window analysis

- **enable**

  1: The quality informative property is enabled for hyper-parameter tuning

  0: The quality informative property is disabled for hyper-parameter tuning

4.4 Hyperparameter configuration

There are two types of hyperparameters for a VARITY predictor; 1) algorithm level parameters with respect to the estimator. 2) weighting parameters for training sets. Each hyperparameter has following attributes:

- **qip**

  The quality informative property defined in 3.4

- **hp_type**

1: logistic parameter (growth rate *k* and maximum value *L*)

2: logistic parameter (mid-point x_0)

3: algorithm parameter

- **from**
  The lower bound of the hyperparameter value.

- **to**

  The higher bound of the hyperparameter value. For filtering parameters, the higher bound is determined automatically based on number of examples in the associated add-on set.

- **step**

  The difference between each hyperparameter value from the lower bound to higher bound. For filtering parameters, this parameter indicates the number of examples in a "filtering block". The add-on set is filtered out in blocks instead of individual example.

- **default**

  The default value of the hyperparameter.

- **data_type**

  The data type of the hyperparameter (int or real)

- **data_interval**

  number of possible mid-points.

- **significant_digits**

  number of significant digits for the value of parameters, use 'None' for no restriction.

5. VARITY commands

To run VARITY commands, please take the following steps first:

1) Download code in this git repository (https://github.com/joewuca/varity/tree/master/python) to a local folder as your VARITY script folder.
2) Make sure you have installed python 3, and all associated packages needed for VARITY framework (See Table 1)
3) Make sure your python PATH has included the VARITY script folder.

VARITY framework currently supports the following commands:

- **init_session**

  *python3 varity_run.py **actions**=init_session **session_id**=YOUR_SESSION_ID **project_path**=PATH_OF_YOUR_PROJECT_FOLDER*

  This command initiates a session (creating all necessary VARITY framework objects) based on the current configuration file. Unless you reinitiate the session again, changes to the configuration file will not affect the initialized VARITY objects. You can reinitiate the session by adding argument ***reinitiate****=1* to the command line, but you might need to generate all your results again after session re-initiation.

  output:

  - */project_path/output/npy/[**session_id**]_[EACH DATA_INSTANCE_NAME]_savedata.npy*
  - */project_path/output/npy/[**session_id**]_[EACH PREDICTOR_INSTANCE_NAME]_hp_config_dict.npy*

- **mv_analysis**

  *python3 varity_run.py **actions**=mv_analysis **session_id**=YOUR_SESSION_ID **predictor**=PREDICTOR_INSTANCE_NAME **mv_qip**=QIP_INSTANCE_NAME **project_path**=PATH_OF_YOUR_PROJECT_FOLDER*

  Moving analysis first order the add-on sets ("set_list" attribute of ***mv_qip***) examples by the informative property ("qip_col" attribute of mv_qip) in ascending or descending order ("direction" attribute of ***mv_qip***), then create ["mv_data_points" attribute of ***mv_qip***] number of moving windows (size of each window equals to [number of examples in add-on sets] * ["mv_size_precent" attribute of ***mv_qip***]). The predictive utility of each window is estimated using 10-fold cross validation on the core training set, where the training examples in each fold were supplemented by all of the add-on examples in that moving window.

  output:

  - */project_path/output/csv/[**session_id**]_mv_analysis_[**predictor**]_[**mv_qip**].csv*

- **plot_mv_result**

*python3 varity_run.py **action**=plot_mv_result **session_id**=YOUR_SESSION_ID **predictor**= PREDICTOR_INSTANCE_NAME **mv_qip** =QIP_INSTANCE_NAME **project_path**=PATH_OF_YOUR_PROJECT_FOLDER*

Plot the moving analysis result (the predictive utility of each moving window)

output:

- */project_path/output/img/[**session_id**]_plot_mv_result_[**predictor**]_[**mv_qip**].png*

- **hp_tuning**

*python3 varity_run.py **action**=hp_tuning **session_id**=YOUR SESSION ID **predictor**=PREDICTOR_INSTANCE_NAME **cur_test_fold**= THE OUTER-LOOP FOLD **project_path**=PATH_OF_YOUR_PROJECT_FOLDER*

Hyperparameter tuning for the input ***predictor*** on the specified outer-loop fold. For the predictors for nested cross validation, the possible outer-loop fold value is from 0 to "test_split_folds" attribute of the corresponding data object minus one. For the final VARITY model predictor (compare to predictors used for nested cross-validation), there is only one dummy outer loop therefore the outer-loop fold should be set to 0.

output:

- */project_path/output/npy/[**session_id**]_[**predictor**]_[**filteing_hp**]_tf[**cur_test_fold**]_trials.pkl*
- */project_path/output/csv/[**session_id**]_[**predictor**]_[**filtering_hp**]_tf[**cur_test_fold**]_ trial_results.txt*

- **save_best_hp**

*python3 varity_run.py **action**=save_best_hp **session_id**=YOUR_SESSION_ID **predictor**=PREDICTOR_INSTANCE_NAME **cur_test_fold**=THE OUTER-LOOP FOLD **project_path**=PATH_OF_YOUR_PROJECT_FOLDER*

Select and save the optimum hyperparameter setting from all hyperparameter optimization (HyperOpt) trials using the following procedure: 1) Re-order all trials by mean metric ("tune_obj" attribute of ***predictor)*** on training sets (averaged over 10 training sets) from low to high; 2) calculate a moving window (we used window size 100) average of mean metric on validation sets; 3) define an "early stopping" point at the first moving window (the "fittest" region) for which mean metric on validation sets begins to descend; 4) Select as final the hyperparameters from the trial within this "fittest" region that achieved the highest mean metric on validation sets.

output:

- */project_path/output/npy/[**session_id**]_[**predictor**]_tf[**cur_test_fold**]_hp_dict.npy*
- */project_path/output/csv/[**session_id**]_[**predictor**]_tf[**cur_test_fold**]_best_hps.csv*

  - */project_path/output/img/[**session_id**]_[**predictor**]_tf[**cur_test_fold**]_hp_selection.png*

- **plot_hp_weight**

  *python3 varity_run.py **action**=plot_hp_weight **session_id**=YOUR_SESSION_ID **predictor**=PREDICTOR_INSTANCE_NAME **cur_test_fold**=THE OUTER-LOOP FOLD **filtering_hp**=HYPERPARAMETER_INSTANCE_NAME*

  This command plots the weight of each example in an add-on set or combined add-on sets ("source" attribute of ***filtering_hp***) ordered by the informative property ("orderby" of ***filtering_hp***). The filtering threshold and weight are based on the optimized hyperparameters.
  output:

  - */project_path/output/img/[**session_id**]_[**predictor**]_[**filtering_hp**].png*

- **test_cv_prediction**

  *python3 varity_run.py **action**=test_cv_prediction **session_id**=YOUR_SESSION_ID **predictor**=PREDICTOR_INSTANCE_NAME*

  This command runs nested cross-validation. For each outer-loop, It makes predictions the test set using model trained with the optimized hyper-parameters via inner-loop cross-validation.
  output:

  - */project_path/output/npy/[**session_id**]_[**predictor**]_test_cv_results.npy*
  - */project_path/output/csv/[**session_id**]_[**predictor**]_hp_test_cv_results.csv*

- **plot_test_result**

  *"python3 varity_run.py **action**=plot_test_result **session_id**=YOUR_SESSION_ID **predictor**= PREDICTOR_INSTANCE_NAME **compare_predictors**=[PREDICTORS_FOR_COMPARISON]*

  Plot the balanced precision recall curve and ROC curve using the results from nested cross-validation. The statistical test is carried out between each predictor specified in ***compare_predictors*** (usually non-VARITY predictors) and the predictor specified in ***predictor*** (usually VARITY predictor). For each outer-loop fold, the test set is filtered if there is a missing prediction from any of the predictor specified in ***compare_predictors***.
  output:

  - */project_path/output/npy/[**session_id**]_[**predictor**]_filter_1_auroc_interp.png*
  - */project_path/output/csv/[**session_id**]_[**predictor**]_ filter_1_aubprc_interp.png*

- **targe_prediction**

  *python3 varity_run.py **action**=target_prediction **session_id**=YOUR_SESSION_ID **predictor**= PREDICTOR_INSTANCE_NAME **cur_test_fold**=THE OUTER-LOOP FOLD **target_file**=TARGET_FILE_NAME **loo**= [0 OR 1]*

Predict the examples in ***target_file***. If the ***loo*** is set to 1, then only the target examples that have been used in training will be predicted using leave-one-example-out strategy.
output:

- */project_path/output/npy/[**session_id**]_[**predictor**_tf[**cur_test_fold**]_target_predicted.csv*
- */project_path/output/npy/[**session_id**]_[**predictor**_tf[**cur_test_fold**]_target_loo_predicted.csv*

6. Table 1: Required python packages

Package Name	Version	Description (Link to document)
Cython	0.29.14	C extension for python (https://cython.org/)
graphviz	0.13.2	Open source graph visualization software (https://graphviz.org/)
hyperopt	0.2.2	Bayesian hyperparameter optimization (https://github.com/hyperopt/hyperopt)
matplotlib	3.1.0	Python visualization (https://matplotlib.org/)
numpy	1.16.0	Scientific computing with Python (https://numpy.org/)
pandas	0.24.0	Data analysis and manipulation tool (https://pandas.pydata.org)
scikit-learn	0.20.2	Machine learning in Python (https://scikit-learn.org/stable/)
scipy	1.2.0	Python-based ecosystem of open-source software for mathematics, science, and engineering (https://www.scipy.org/)
seaborn	0.9.0	Statistical data visualization (https://seaborn.pydata.org/)
shap	0.34.0	A game theoretic approach to explain the output of any machine learning model (https://github.com/slundberg/shap)
xgboost	0.90	An optimized distributed gradient boosting library (https://xgboost.readthedocs.io/en/latest/)