MACHINE LEARNING DOCUMENTATION

Go Up

Name	HPCC_Causality
Version	1.0
Description	HPCC Causality Bundle
License	SeeLICENSE.TXT
Copyright	Copyright (C) 2022 HPCC SystemsÂő
Authors	HPCCSystems
DependsOn	ML_Core
Platform	8.4.0

OVERVIEW

HPCC_Causality

Causality Bundle for HPCC Systems Platform

This bundle supports research into Causal Analysis of data.

It provides three main modules: - Synth – Allows the generation of complex synthetic datasets with known causal relationships, using structural equation models. These can be used with the other layers of this toolkit or with other HPCC Machine Learning bundles. - Probability – Provides a rich probability analysis system, supporting conditional probabilities, conditionalizing, independence testing, and predictive capabilities. - Causality – Provides a range of Causal Analysis algorithms. This layer requires a Causal Model, which combined with a dataset, allows questions to be asked that are beyond the realm of Statistics. Specifically, this module allows: Model Hypothesis Testing, Causal Analysis, Causal Metrics, Counterfactual Analysis (future), and limited Causal Discovery.

The above methods are computationally intense, and are fully parallelized when running on an HPCC Systems Cluster.

It is built on the underlying capabilities of the Python-based “Because” Causal Analytic Library.

Installation

This bundle requires python3, PIP, and “Because” on each HPCC Node

Installing Because:

Clone the repository https://github.com/RogerDev/Because.git

Run: sudo pip3 install <path to Because>

Example: sudo pip3 install /source/Because

This must be done on each HPCC Cluster node. It is important to use sudo so that the bundle is installed for all users. Since HPCC nodes run as special user hpcc, installing as the current user would not allow the module to be found by hpcc.

Installing HPCC_Causality

On your client system, where HPCC Clienttools was installed, run ecl bundle install https://github.com/RogerDev/HPCC_Causality.git

Using HPCC_Causality

Each of the main modules provides documentation and examples of use. Documentation is provided within the module and examples are in the test folder.

Table of Contents

Causality

Go Up

IMPORTS

_versions.HPCC_Causality.V1_0.HPCC_Causality.Types |

_versions.ML_Core.V3_2_2.ML_Core.Types |

_versions.HPCC_Causality.V1_0.HPCC_Causality.internal.cModel |

DESCRIPTIONS

CAUSALITY

/ EXPORT	**Causality**
(DATASET(cModelTyp) mod, DATASET(NumericField) dat)	

Causal Model Module. Causal level methods require a combination of a causal model, and a dataset. Methods include: - ValidateModel – Analyze the data against the provided causal model and evaluate the degree of correspondence between the two. - Intervene – Simulate the effect on a target variable of a causal intervention on one or more variable - Metrics – Evaluate various causal metrics on desgnated pairs [source, destination] of variables.

PARAMETER **mod** ||| TABLE (cModelTyp) — A causal model in DATASET(cModel) format. The dataset should contain only a single record, defining the model.

PARAMETER **dat** ||| TABLE (NumericField) — The data in NumericField format. The field number should correspond to the order of variables specified in the model.

SEE Types.cModel

SEE ML_Core.Types.NumericField

Children

1. ValidateModel : Validate the causal model relative to the data
2. Intervene : Calculate the results of a Causal Intervention
3. Metrics : Calculate a set of causal metrics from a designated source variable to a designated destination variable
4. DiscoverModel : Analyze the data to estimate the causal relationships between variables

---

VALIDATEMODEL

Causality /

ValidationReport	**ValidateModel**
(UNSIGNED order=3, UNSIGNED strength=1)	

Validate the causal model relative to the data.

PARAMETER **order** ||| UNSIGNED8 — The largest number of variables to consider at a time (default 3). Higher values lead to exponentially increasing run times, and diminishing evaluation accuracy. Very large datasets are required in order to evaluate higher order evaluations (default=3, recommended).

PARAMETER **strength** ||| UNSIGNED8 — The thoroughness to be used in conditionalizing on variables. Allows a tradeoff between run-time and certainty of discrimination. Strength = 1 is sufficient to distinguish linear relationships, where higher numbers are needed to distinguish subtle non-linear relationships. Range [1,100]. For practical purposes, strength > 5 should not be needed. Default = 1.

RETURN **ROW ({ REAL8 Confidence , UNSIGNED8 NumTotalTests , SET (UNSIGNED8) NumTestsPerType , SET (UNSIGNED8) NumErrsPerType , SET (UNSIGNED8) NumWarnsPerType , SET (STRING) Errors , SET (STRING) Warnings })** — A detailed validation report in Types.ValidationReport format

SEE Types.ValidationReport

INTERVENE

Causality /

DATASET(Distr)	**Intervene**
(DATASET(ProbQuery) queries, UNSIGNED pwr=1)	

Calculate the results of a Causal Intervention. Interventions simulate the effect of setting a variable or variables to fixed values, while breaking the links from those variables' parents. The distribution of a target variable given the interventions is returned for each query. This is roughly equivalent to performing a randomized study. Interventions are of the form: - Distribution = (Var | List of interventions)

PARAMETER **queries** ||| TABLE (ProbQuery) — A list of queries. Exactly 1 target per query must be specified, and the target must be unbound (i.e. with zero arguments). One or more interventions can be provided for each variable. Interventions must be of an exact value (e.g. do(var = value)). This is indicated by a single arg in the intervention ProbSpec.

PARAMETER **pwr** ||| UNSIGNED8 — No Doc

RETURN **TABLE (PDist)** — A set of Types.Distr records, describing each of the queried distributions.

METRICS

Causality /

DATASET(cMetrics)	**Metrics**
(DATASET(ProbQuery) queries, UNSIGNED pwr=1)	

Calculate a set of causal metrics from a designated source variable to a designated destination variable. The following metrics are produce for each source / destination pair: - Average Causal Effect (ACE) – The average effect on the destination variable of a unit intervention on the source variable. - Controlled Direct Effect (CDE) – The direct effect on the destination variable of a unit intervention on the source variable. - Indirect Effect (IE) – The indirect effect (i.e. via intermediate variables) on the destination variable of a unit intervention on the source variable.

PARAMETER **queries** ||| TABLE (ProbQuery) — A list of queries, each with two targets [source, destination], and no conditions or interventions. Targets should be unbound (i.e. no args).

PARAMETER **pwr** ||| UNSIGNED8 — No Doc

RETURN **TABLE (cMetrics)** — Dataset of cMetrics records, one per query, with id corresponding to the id of the original query.

DISCOVERMODEL

Causality /

DATASET(DiscoveryReport)	**DiscoverModel**
(UNSIGNED pwr=1)	

Analyze the data to estimate the causal relationships between variables. Produces information that is useful for understanding the variables' relationships, and attempts to build a full causal model. Discovery is done hierarchically, first determining "clusters" that share a common set of exogenous variables. Then each cluster is analyzed for topology, and finally, the inter-cluster relationships are estimated. Note that this function does not use the model information supplied to the module except for a list of variable names. It, rather, produces an estimated of the model that generated the data.

PARAMETER **pwr** ||| UNSIGNED8 — The power to use for statisitical queries. Range [1, 100]. The higher power, the more accuracy, but longer runtime. Power=1 suffices for liner relationships. Power > 10 is not recommended due to very long runtimes.

RETURN **TABLE ({ SET (STRING) Exos , SET (STRING) Clusters , TABLE (SetMembers) ClustMembers , TABLE (SetMembers) ClustGraph , TABLE (SetMembers) VarGraph })** — A DATASET(DiscoveryReport) with a single record representing the results of the discovery.

SEE Types.DiscoveryReport

Probability

Go Up

IMPORTS

_versions.HPCC_Causality.V1_0.HPCC_Causality.Types |

_versions.ML_Core.V3_2_2.ML_Core.Types | python3 |

_versions.HPCC_Causality.V1_0.HPCC_Causality.internal.ProbSpace |

DESCRIPTIONS

PROBABILITY

/ EXPORT	**Probability**
(DATASET(NumericField) ds, SET OF STRING varNames)	

Probability Module Contains a set of probability functions to execute against a multivariate dataset. The dataset consists of a set of variable names, and a set of observations for each variable. The observations are in NumericField format, with the field number corresponding to the order of the variable names. Probability functions include: - P(…) – Unconditional, Conditional and Joint Numerical Probabilities. - E(…) – Unconditional and Conditional Expectations - Distr(…) – Unconditional and Conditional Distributions - Dependence(…) – Test of Dependence and Conditional Dependence Between Variables - isIndependent(…) – Boolean Independence and Conditional Independence Test. - Predict(…) – Machine Learning style regression without training required. - Classify(…) – Machine Learning style classification without training required.

PARAMETER **ds** ||| TABLE (NumericField) — – Set of multivariate observations in NumericField format. Each observation shares an id (1 - numObservations), and field numbers correspond to the order of variable names in the varNames parameter.

PARAMETER **varNames** ||| SET (STRING) — – An ordered list of variable name strings.

Children

1. P : Calculate a series of numerical probabilities

2. E : Calculate a series of numerical expected values
3. Distr : Calculate a series of Distributions
4. Dependence : Perform a series of dependency tests
5. isIndependent : Perform a series of dependency tests and evaluate the results as a Boolean
6. Predict : Perform a set of regression style predictions on a continuous variable
7. Classify : Perform a set of classification predictions on a discrete target variable

---

P

Probability /

DATASET(NumericField)	**P**
(DATASET(ProbQuery) queries)	

Calculate a series of numerical probabilities. Queries are of the form: - Exact Query – P(Var = Val | List of Conditions) - Range Query – P(Val1 <= Var <= Val2 | List of Conditions) - Joint Probability – P([Exact or Range Query 1, . . .] | List of Conditions)

PARAMETER **queries** ||| TABLE (ProbQuery) — A list of queries. One or more target may be specified for each query, and the targets must be bound (i.e. with 1 or 2 arguments).

RETURN **TABLE (NumericField)** — A set of NumericField records, with value being the probability of the query as field-number 1.

E

Probability /

DATASET(NumericField)	**E**
(DATASET(ProbQuery) queries)	

Calculate a series of numerical expected values. Expectations are of the form: - E(Var | List of Conditions)

PARAMETER **queries** ||| TABLE (ProbQuery) — A list of queries. Exactly 1 target per query must be specified, and the target must be unbound (i.e. with zero arguments).

RETURN **TABLE (NumericField)** — A set of NumericField records, with value being the Expected Value of each query.

DISTR

Probability /

DATASET(Distr)	**Distr**
(DATASET(ProbQuery) queries)	

Calculate a series of Distributions. Distributions are of the form: - Distr(Var | List of Conditions)

PARAMETER **queries** ||| TABLE (ProbQuery) — A list of queries. Exactly 1 target per query must be specified, and the target must be unbound (i.e. with zero arguments).

RETURN **TABLE (PDist)** — A set of Types.Distr records, describing each of the queried distributions.

DEPENDENCE

Probability /

DATASET(NumericField)	**Dependence**
(DATASET(ProbQuery) queries)	

Perform a series of dependency tests. Form: - Dependency(target1, target2 | List of conditions)

PARAMETER **queries** ||| TABLE (ProbQuery) — A list of queries. Exactly 2 targets per query must be specified.

RETURN **TABLE (NumericField)** — a list of p-values with .5 confidence, in NumericField format. Values less than .5 indicate probable independence. Values greater than .5 indicate probable dependence

---

ISINDEPENDENT

Probability /

DATASET(NumericField)	**isIndependent**
(DATASET(ProbQuery) queries)	

Perform a series of dependency tests and evaluate the results as a Boolean. Form: - isIndependent(target1, target2 | List of conditions)

PARAMETER **queries** ||| TABLE (ProbQuery) — A list of queries. Exactly 2 targets per query must be specified.

RETURN **TABLE ({ UNSIGNED2 wi , UNSIGNED8 id , UNSIGNED4 number , REAL8 value })** — A list of results as NumericField. Result of 1 indicates that the two targets are most likely independent. 0 indicates probable dependence.

PREDICT

Probability /

DATASET(NumericField)	**Predict**
(STRING target, SET OF STRING varNames, DATASET(NumericField) varDat)	

Perform a set of regression style predictions on a continuous variable. Form: - E(target | conditions)

PARAMETER **target** ||| STRING — The dependent variable (i.e. prediction target). The target should be a continuous variable.

PARAMETER **varNames** ||| SET (STRING) — The names of the independent variables to be used for prediction

PARAMETER **varDat** ||| TABLE (NumericField) — The values of the conditional variables in NumericField format. The field numbers correspond to the order of the varNames list.

RETURN **TABLE (NumericField)** — A DATASET(NumericField) with the prediction values in field number 1.

CLASSIFY

Probability /

DATASET(NumericField)	**Classify**
(STRING target, SET OF STRING varNames, DATASET(NumericField) varDat)	

Perform a set of classification predictions on a discrete target variable. Form: - E(target | conditions)

PARAMETER **target** ||| STRING — The dependent variable (i.e. prediction target). The target should be a discrete variable.

PARAMETER **varNames** ||| SET (STRING) — The names of the independent variables to be used for prediction

PARAMETER **varDat** ||| TABLE (NumericField) — The values of the conditional variables in NumericField format. The field numbers correspond to the order of the varNames list.

RETURN **TABLE (NumericField)** — A DATASET(NumericField) with the prediction values in field number 1.

Synth

Go Up

IMPORTS

std.system.Thorlib | _versions.ML_Core.V3_2_2.ML_Core.Types | _versions.HPCC_Causality.V1_0.HPCC_Causality.Types | python3 |

DESCRIPTIONS

SYNTH

/ EXPORT	**Synth**
(DATASET(SEM) semDef)	

Module to produce a synthetic, multivariate dataset from a Structural Equation Model (SEM). This allows creation of datasets with known distributional and or causal characteristics.

PARAMETER **semDef** ||| TABLE (SEM) — A Structural Equation Model in Types.SEM format.

SEE Types.SEM

SEE Test/Synth/synthTest.ecl for an example

Children

1. Generate : Generate the data

GENERATE

Synth /

DATASET(NumericField)	**Generate**
(UNSIGNED numRecs)	

Generate the data. Data generation is fully parallelized, each node generates numRecs / nNodes samples from the same multivariate distribution. *

PARAMETER **numRecs** ||| UNSIGNED8 — The number of samples (multivariate observations) to generate.

RETURN **TABLE (NumericField)** — The generated samples in NumericField format. The field numbers correspond to the order of variables specified in the SEM.

SEE ML_Core.Types.NumericField

Types

Go Up

DESCRIPTIONS

TYPES

Types

Module provides all common record types for the Causality Bundle

Children

1. ProbSpec : Record layout for Probability Query parameters Three forms are supported: - Variable Name alone (i.e
2. ProbQuery : Record Layout for Probability Queries
3. HistEntry : Histogram Entry
4. Distribution : Record to represent the Distribution of a single random variable Values are discretized
5. DatTypeEnum : Enumeration for Random Variable Data Type (see RV below)
6. RV : Random Variable Record type for causal model representation
7. cModel : Causal Model Definition Record
8. SEM : Record to represent a Structural Equation Model (SEM) See Synth/synthTest.ecl for details on use of fields
9. ValidationReport : Model Validation Report Shows result of a model validation test
10. cMetrics : Record type for the results of a metrics query
11. SetMembers : Represents a named set along with its members
12. DiscoveryReport : Results of the DiscoverModel function

PROBSPEC

Types /

ProbSpec

Record layout for Probability Query parameters Three forms are supported: - Variable Name alone (i.e. Unbound variable) - Variable Name and One Argument (i.e. Var = arg1) - Variable Name and Two Arguments (i.e. arg1 <= Var <= arg2)

FIELD **VarName** ||| STRING — – The variable name

FIELD **Args** ||| SET (REAL8) — – The arguments for the query spec. O-2 arguments may be provided depending on the context.

PROBQUERY

Types /

ProbQuery

Record Layout for Probability Queries. Also used for Causality Queries (i.e. Interventional, Counterfactual) General form for Probability Queries: - P(target | conditions) – e.g., P(Y=1 | X1=1.5, X2=-.3, .5 <= X3 <= 1.0) - E(target | conditions) - distr(target | conditions)

FIELD **id** ||| UNSIGNED8 — Unique id for each query, used to correlate results.

FIELD **target** ||| TABLE (ProbSpec) — The target of the query (e.g. 'Y', 'Y'=1, .5 <= 'Y' <= 1.0)

FIELD **conditions** ||| TABLE (ProbSpec) — The set of conditions to apply to the target query (e.g. ['X1', 'X2'=1, .5 <= 'X3' <= 1.0]). Defaults to empty set meaning no conditions.

FIELD **interventions** ||| TABLE (ProbSpec) — The set of interventions for causal (interventional) queries. These represent "do()" operations, and must be set to exact values (e.g. 'X1'=1.0).

FIELD **counterfacs** ||| TABLE (ProbSpec) — No Doc

HISTENTRY

Types /

HistEntry

Histogram Entry Represents one bin of a discretized probability histogram

FIELD **Min** ||| REAL8 — The minimum value for this bin

FIELD **Max** ||| REAL8 — The maximum value for this bin. Values within this bin fall into the interval [Min, Max). For discrete variables, Min and Max will both equal the discrete value.

FIELD **P** ||| REAL8 — The probability that the random variable will take on a value within this bin.

DISTRIBUTION

Types /

Distribution

Record to represent the Distribution of a single random variable Values are discretized. For discrete variables, there will be as many bins as the cardinality of the variable. For continuous variables, the number of bins is determined automatically based on the number of observations. Datasets with more observations are discretized more finely than smaller datasets.

FIELD **id** ||| UNSIGNED8 — Identifier for the given requested distribution. Matches the id of the corresponding request.

FIELD **query** ||| STRING — A representation of the query in (near) standard Pearl notation. Format: Distr<counterfactual>(target | conditions, do(interventions)) The fields: counterfactual, conditions and do(interventions) may or may not appear in any given query. Angle brackets <> are used in place of subscripting as in Pearl notation.</counterfactual>

FIELD **nSamples** ||| UNSIGNED8 — The number of samples upon which the distribution is based.

FIELD **isDiscrete** ||| BOOLEAN — Boolean is TRUE if this is a discrete variable, otherwise FALSE.

FIELD **minVal** ||| REAL8 — The minimum observed value of the variable.

FIELD **maxVal** ||| REAL8 — The maximum observed value of the variable.

FIELD **Mean** ||| REAL8 — The sample mean of the variable.

FIELD **StDev** ||| REAL8 — The sample standard deviation of the variable.

FIELD **Skew** ||| REAL8 — The sample skew of the variable.

FIELD **Kurtosis** ||| REAL8 — The sample excess kurtosis of the variable.

FIELD **Median** ||| REAL8 — The median sample value of the variable.

FIELD **Mode** ||| REAL8 — The most common value of the variable. For continuous variables, this is the midpoint of the bin containing the most samples.

FIELD **Histogram** ||| TABLE (HistEntry) — The set of discretized bins representing the distribution's PDF.

FIELD **Deciles** ||| TABLE (HistEntry) — The Deciles of the variable's distribution From 10 to 90.

DATTYPEENUM

Types /

DatTypeEnum

Enumeration for Random Variable Data Type (see RV below).

RETURN **UNSIGNED4** —

RV

Types /

RV

Random Variable Record type for causal model representation

FIELD **Name** ||| STRING — The name of the Random Variable.

FIELD **Parents** ||| SET (STRING) — A set of RV Names representing the causal parents of this variable.

FIELD **isObserved** ||| BOOLEAN — Boolean is TRUE if this variable has measurable data associated with it. Otherwise FALSE.

FIELD **DataType** ||| UNSIGNED4 — Enumeration of the data type associated with this variable. Currently only Numeric and Categorical are supported.

CMODEL

Types /

cModel

Causal Model Definition Record

FIELD **Name** ||| STRING — The name of the model.

FIELD **Nodes** ||| TABLE (RV) — The list of Random Variables that comprise the model. This must be in the order of variable in the dataset.

SEM

Types /

SEM

Record to represent a Structural Equation Model (SEM) See Synth/synthTest.ecl for details on use of fields.

FIELD **Init** ||| SET (STRING) — An ordered list of statements to be executed once to do any required variable initialization.

FIELD **VarNames** ||| SET (STRING) — An ordered set of variables representing the output of the SEM. The produced data will follow the order of variable in this set.

FIELD **EQ** ||| SET (STRING) — An ordered set of equations that will be executed to generate each observation of the generated dataset. Equations may refer to variables initialized during Init processing, or variables set by previous equations.

VALIDATIONREPORT

Types /

ValidationReport

Model Validation Report Shows result of a model validation test. Four types of tests are conducted: - Type 0: Verify all exogenous variables are independent of one another. - Type 1: Verify expected independencies. - Type 2: Verified expected dependencies. - Type 3: Verify causal direction.

FIELD **Confidence** ||| REAL8 — The confidence in the model between 0 and 1. 0 implies no confidence. 1 implies perfect confidence.

FIELD **NumTotalTests** ||| UNSIGNED8 — The total number of tests conducted.

FIELD **NumTestsByType** ||| — An array of four values indicating the number of tests of each type 0-3 conducted.

FIELD **NumErrsPerType** ||| SET (UNSIGNED8) — An array of four values indicating the number of errors detected for each test type 0-3.

FIELD **NumWarnsPerType** ||| SET (UNSIGNED8) — An array of four values indicating the number of warnings detected for each test type 0-3.

FIELD **Errors** ||| SET (STRING) — Array of strings describing each error that occurred.

FIELD **Warnings** ||| SET (STRING) — Array of strings describing each warning that occurred.

FIELD **numtestspertype** ||| SET (UNSIGNED8) — No Doc

CMETRICS

Types /

cMetrics

Record type for the results of a metrics query.

FIELD **id** ||| UNSIGNED8 — The id of the result corresponding to the original id in the query.

FIELD **query** ||| STRING — A representation of the original query e.g., Source -> Destination.

FIELD **AveCausalEffect** ||| REAL8 — The average causal effect (ACE) of the source variable on the destination variable.

FIELD **ContDirEffect** ||| — The controlled direct effect (CDE) of the source variable on the destination variable.

FIELD **IndirEffect** ||| REAL8 — The indirect effect (via other variables) of the source variable on the destination variable.

FIELD **contrdireffect** ||| REAL8 — No Doc

SETMEMBERS

Types /

SetMembers

Represents a named set along with its members

FIELD **Name** ||| STRING — The identifier of the set

FIELD **Members** ||| SET (STRING) — A list of unique set member identifiers

---

DISCOVERYREPORT

Types /

DiscoveryReport

Results of the DiscoverModel function. Provides the information about what was discovered from analyzing the dataset.

FIELD **Exos** ||| SET (STRING) — A list of exogenous variables.

FIELD **Clusters** ||| SET (STRING) — A list of all of the discovered data cluster names.

FIELD **ClustMembers** ||| TABLE (SetMembers) — A list of each cluster and its members.

FIELD **ClustGraph** ||| TABLE (SetMembers) — A list of clusters and the set of parent clusters for each, representing a Directed Acyclic Graph (DAG) of cluster-to-cluster relationships.

FIELD **VarGraph** ||| TABLE (SetMembers) — A list of variables and the set of parents for each, representing a Directed Acyclic Graph (DAG) of variable relationships.
