LogisticRegression

Go Up

Name	LogisticRegression
Version	1.0.0
Description	Logistic Regression implementation
License	http://www.apache.org/licenses/LICENSE-2.0
Copyright	Copyright (C) 2017 HPCC Systems
Authors	HPCCSystems
DependsOn	ML_Core, PBblas
Platform	6.2.0

Table of Contents

Extract the beta values form the model dataset
ExtractBeta_CI.ecl Extract the beta values form the model dataset
ExtractBeta_pval.ecl Extract the beta values form the model dataset
ExtractReport.ecl Extract Report records from model
LogitPredict.ecl Predict the category values with the logit function and the the supplied beta coefficients
LogitScore.ecl Calculate the score using the logit function and the the supplied beta coefficients
Model_Deviance.ecl Model Deviance
Null_Deviance.ecl Deviance for the null model, that is, a model with only an intercept
Types.ecl

LogisticRegression/ BinomialConfusion

Go Up

IMPORTS

ML_Core.Types | LogisticRegression | LogisticRegression.Types |

DESCRIPTIONS

FUNCTION BinomialConfusion

DATASET(Types.Binomial_Confusion_Summary)	**BinomialConfusion**
(DATASET(Core_Types.Confusion_Detail) d)	

Binomial confusion matrix. Work items with multinomial responses are ignored by this function. The higher value lexically is considered to be the positive indication.

PARAMETER **d** confusion detail for the work item and classifier

RETURN confusion matrix for a binomial classifier

LogisticRegression/

BinomialLogisticRegression

Go Up

IMPORTS

LogisticRegression | LogisticRegression.Constants | ML_Core.Interfaces | ML_Core.Types |

DESCRIPTIONS

MODULE BinomialLogisticRegression

BinomialLogisticRegression

```
(UNSIGNED max_iter=200, REAL8 epsilon=Constants.default_epsilon, REAL8
ridge=Constants.default_ridge)
```

Binomial logistic regression using iteratively re-weighted least squares.

PARAMETER **max_iter** maximum number of iterations to try

PARAMETER **epsilon** the minimum change in the Beta value estimate to continue

PARAMETER **ridge** a value to populate a diagonal matrix that is added to a matrix help assure that the matrix is invertible.

Children

1. GetModel : Calculate the model to fit the observation data to the observed classes

2. Classify : Classify the observations using a model
3. Report : Report the confusion matrix for the classifier and training data

---

FUNCTION GetModel

BinomialLogisticRegression \

DATASET(Types.Layout_Model)	**GetModel**
(DATASET(Types.NumericField) observations, DATASET(Types.DiscreteField) classifications)	

Calculate the model to fit the observation data to the observed classes.

PARAMETER **observations** the observed explanatory values

PARAMETER **classifications** the observed classification used to build the model

RETURN the encoded model

OVERRIDE True

FUNCTION Classify

BinomialLogisticRegression \

DATASET(Types.Classify_Result)	**Classify**
(DATASET(Types.Layout_Model) model, DATASET(Types.NumericField) new_observations)	

Classify the observations using a model.

PARAMETER **model** The model, which must be produced by a corresponding getModel function.

PARAMETER **new_observations** observations to be classified

RETURN Classification with a confidence value

OVERRIDE True

FUNCTION Report

BinomialLogisticRegression \

DATASET(Types.Confusion_Detail)	**Report**
(DATASET(Types.Layout_Model) model, DATASET(Types.NumericField) observations, DATASET(Types.DiscreteField) classifications)	

Report the confusion matrix for the classifier and training data.

PARAMETER **model** the encoded model

PARAMETER **observations** the explanatory values.

PARAMETER **classifications** the classifications associated with the observations

RETURN the confusion matrix showing correct and incorrect results

OVERRIDE True

LogisticRegression/

Confusion

Go Up

IMPORTS

ML_Core | ML_Core.Types | LogisticRegression | LogisticRegression.Types |

DESCRIPTIONS

FUNCTION Confusion

DATASET(Confusion_Detail)	**Confusion**
(DATASET(DiscreteField) dependents, DATASET(DiscreteField) predicts)	

Detail confusion records to compare actual versus predicted response variable values.

PARAMETER **dependents** the original response values

PARAMETER **predicts** the predicted responses

RETURN confusion counts by predicted and actual response values.

LogisticRegression/

Constants

Go Up

DESCRIPTIONS

MODULE Constants

	Constants

Children

1. limit_card
2. default_epsilon
3. default_ridge
4. local_cap
5. id_base
6. id_iters
7. id_delta
8. id_correct
9. id_incorrect
10. id_stat_set
11. id_betas
12. id_betas_coef
13. id_betas_SE
14. base_builder
15. base_max_iter

16. base_epsilon
17. base_ind_vars
18. base_dep_vars
19. base_obs
20. builder_irls_local
21. builder_irls_global
22. builder_softmax

ATTRIBUTE limit_card

Constants \

UNSIGNED2	**limit_card**

ATTRIBUTE default_epsilon

Constants \

REAL8	**default_epsilon**

ATTRIBUTE default_ridge

Constants \

REAL8	**default_ridge**

ATTRIBUTE local_cap

Constants \

UNSIGNED4	local_cap

ATTRIBUTE id_base

Constants \

	id_base

ATTRIBUTE id_iters

Constants \

	id_iters

ATTRIBUTE id_delta

Constants \

	id_delta

ATTRIBUTE id_correct

Constants \

	id_correct

ATTRIBUTE id_incorrect

Constants \

	id_incorrect

ATTRIBUTE id_stat_set

Constants \

	id_stat_set

ATTRIBUTE id_betas

Constants \

	id_betas

ATTRIBUTE id_betas_coef

Constants \

	id_betas_coef

ATTRIBUTE id_betas_SE

Constants \

id_betas_SE

ATTRIBUTE base_builder

Constants \

base_builder

ATTRIBUTE base_max_iter

Constants \

base_max_iter

ATTRIBUTE base_epsilon

Constants \

base_epsilon

ATTRIBUTE base_ind_vars

Constants \

	base_ind_vars

ATTRIBUTE base_dep_vars

Constants \

	base_dep_vars

ATTRIBUTE base_obs

Constants \

	base_obs

ATTRIBUTE builder_irls_local

Constants \

	builder_irls_local

ATTRIBUTE builder_irls_global

Constants \

	builder_irls_global

ATTRIBUTE builder_softmax

Constants \

	builder__softmax

LogisticRegression/
DataStats

Go Up

IMPORTS

ML_Core.Types | LogisticRegression | LogisticRegression.Types | LogisticRegression.Constants |

DESCRIPTIONS

FUNCTION DataStats

DATASET(Types.Data_Info)	**DataStats**
(DATASET(Core_Types.NumericField) indep, DATASET(Core_Types.DiscreteField) dep, BOOLEAN field_details=FALSE)	

Information about the datasets. Without details the range for the x and y (independent and dependent) columns. Note that a column of all zero values cannot be distinguished from a missing column. When details are requested, the cardinality, minimum, and maximum values are returned. A zero cardinality is returned when the field cardinality exceeds the Constants.limit_card value.

PARAMETER **indep** data set of independent variables

PARAMETER **dep** data set of dependent variables

PARAMETER **field_details** Boolean directive to provide field level info

LogisticRegression/
Deviance_Analysis

Go Up

IMPORTS

LogisticRegression | LogisticRegression.Types |

DESCRIPTIONS

FUNCTION Deviance_Analysis

DATASET(Types.AOD_Record)	**Deviance_Analysis**
(DATASET(Types.Deviance_Record) proposed, DATASET(Types.Deviance_Record) base)	

Compare deviance information for an analysis of deviance.

PARAMETER **proposed** the proposed model

PARAMETER **base** the base model for comparison

RETURN the comparison of the deviance between the models

LogisticRegression/

Deviance_Detail

Go Up

IMPORTS

ML_Core | ML_Core.Types | LogisticRegression | LogisticRegression.Types |

DESCRIPTIONS

FUNCTION Deviance_Detail

DATASET(Types.Observation_Deviance)	**Deviance_Detail**
(DATASET(Core_Types.DiscreteField) dependents, DATASET(Types.Raw_Prediction) predicts)	

Detail deviance for each observation.

PARAMETER **dependents** original dependent records for the model

PARAMETER **predicts** the predicted values of the response variable

RETURN the deviance information by observation and the log likelihood of the predicted result.

LogisticRegression/
dimm

Go Up

IMPORTS

std.blas | std.BLAS.Types |

DESCRIPTIONS

EMBED dimm

Types.matrix_t	**dimm**
(BOOLEAN transposeA, BOOLEAN transposeB, BOOLEAN diagonalA, BOOLEAN diagonalB, Types.dimension_t m, Types.dimension_t n, Types.dimension_t k, Types.value_t alpha, Types.matrix_t A, Types.matrix_t B, Types.value_t beta=0.0, Types.matrix_t C=[])	

Matrix multiply when either A or B is a diagonal and is passed as a vector. alpha*op(A) op(B) + beta*C where op() is transpose

PARAMETER **transposeA** true when transpose of A is used

PARAMETER **transposeB** true when transpose of B is used

PARAMETER **diagonalA** true when A is the diagonal matrix

PARAMETER **diagonalB** true when B is the diagonal matrix

PARAMETER **m** number of rows in product

PARAMETER **n** number of columns in product

PARAMETER **k** number of columns/rows for the multiplier/multiplicand

PARAMETER **alpha** scalar used on A

PARAMETER **A** matrix A

PARAMETER **B** matrix B

PARAMETER **beta** scalar for matrix C

PARAMETER **C** matrix C or empty

LogisticRegression/

Distributions

Go Up

IMPORTS

ML_Core.Constants | ML_Core.Math |

DESCRIPTIONS

MODULE Distributions

	Distributions

Children

1. Normal_CDF : Cumulative Distribution of the standard normal distribution, the probability that a normal random variable will be smaller than x standard deviations above or below the mean
2. Normal_PPF : Normal Distribution Percentage Point Function
3. T_CDF : Students t distribution integral evaluated between negative infinity and x
4. T_PPF : Percentage point function for the T distribution
5. Chi2_CDF : The cumulative distribution function for the Chi Square distribution
6. Chi2_PPF : The Chi Squared PPF function

FUNCTION Normal_CDF

Distributions \

REAL8	**Normal_CDF**
(REAL8 x)	

Cumulative Distribution of the standard normal distribution, the probability that a normal random variable will be smaller than x standard deviations above or below the mean. Taken from C/C++ Mathematical Algorithms for Scientists and Engineers, n. Shammas, McGraw-Hill, 1995

PARAMETER **x** the number of standard deviations

FUNCTION Normal_PPF

Distributions \

REAL8	**Normal_PPF**
(REAL8 x)	

Normal Distribution Percentage Point Function. Translated from C/C++ Mathematical Algorithms for Scientists and Engineers, N. Shammas, McGraw-Hill, 1995

PARAMETER **x** probability

FUNCTION T_CDF

Distributions \

REAL8	**T_CDF**
(REAL8 x, REAL8 df)	

Students t distribution integral evaluated between negative infinity and x. Translated from NIST SEL DATAPAC Fortran TCDF.f source

PARAMETER **x** value of the evaluation

PARAMETER **df** degrees of freedom

FUNCTION T_PPF

Distributions \

REAL8	**T_PPF**
(REAL8 x, REAL8 df)	

Percentage point function for the T distribution. Translated from NIST SEL DATAPAC Fortran TPPF.f source

---

FUNCTION Chi2_CDF

Distributions \

REAL8	**Chi2_CDF**
(REAL8 x, REAL8 df)	

The cumulative distribution function for the Chi Square distribution. the CDF for the specfied degrees of freedom. Translated from the NIST SEL DATAPAC Fortran subroutine CHSCDF.

FUNCTION Chi2_PPF

Distributions \

REAL8	**Chi2_PPF**
(REAL8 x, REAL8 df)	

The Chi Squared PPF function. Translated from the NIST SEL DATAPAC Fortran subroutine CHSPPF.

LogisticRegression/

ExtractBeta

Go Up

IMPORTS

ML_Core.Types | LogisticRegression | LogisticRegression.Types |

DESCRIPTIONS

FUNCTION ExtractBeta

	ExtractBeta
`(DATASET(Core_Types.Layout_Model) mod_ds)`	

Extract the beta values form the model dataset.

PARAMETER **mod_ds** the model dataset

RETURN a beta values as Model Coefficient records, zero as the constant term.

LogisticRegression/
ExtractBeta_CI

Go Up

IMPORTS

ML_Core.Types | LogisticRegression | LogisticRegression.Types |

DESCRIPTIONS

FUNCTION ExtractBeta_CI

DATASET(Types.Confidence_Model_Coef)	**ExtractBeta_CI**
(DATASET(Core_Types.Layout_Model) mod_ds, REAL8 level)	

Extract the beta values form the model dataset.

PARAMETER **mod_ds** the model dataset

PARAMETER **level** the significance value for the intervals

RETURN the beta values with confidence intervals term.

LogisticRegression/

ExtractBeta_pval

Go Up

IMPORTS

ML_Core.Types | LogisticRegression | LogisticRegression.Types |

DESCRIPTIONS

FUNCTION ExtractBeta_pval

DATASET(Types.pval_Model_Coef)	**ExtractBeta_pval**
(DATASET(Core_Types.Layout_Model) mod_ds)	

Extract the beta values form the model dataset.

PARAMETER **mod_ds** the model dataset

RETURN the beta values with p-values as Model Coefficient records, zero as the constant term.

LogisticRegression/

ExtractReport

Go Up

IMPORTS

ML_Core.Types | LogisticRegression | LogisticRegression.Types | LogisticRegression.Constants |

DESCRIPTIONS

FUNCTION ExtractReport

DATASET(Types.Model_Report)	**ExtractReport**
(DATASET(Core_Types.Layout_Model) mod_ds)	

Extract Report records from model

PARAMETER **mod_ds** the model dataset

RETURN the model report dataset

LogisticRegression/ LogitPredict

Go Up

IMPORTS

ML_Core.Types | LogisticRegression | LogisticRegression.Types |

DESCRIPTIONS

FUNCTION LogitPredict

`DATASET(Classify_Result)`	**LogitPredict**
`(DATASET(Model_Coef) coef, DATASET(NumericField) independents)`	

Predict the category values with the logit function and the the supplied beta coefficients.

PARAMETER **coef** the model beta coefficients

PARAMETER **independents** the observations

RETURN the predicted category values and a confidence score

LogisticRegression/

LogitScore

Go Up

IMPORTS

ML_Core.Types | LogisticRegression | LogisticRegression.Types |

DESCRIPTIONS

FUNCTION LogitScore

`DATASET(Raw_Prediction)`	**LogitScore**
`(DATASET(Model_Coef) coef, DATASET(NumericField) independents)`	

Calculate the score using the logit function and the the supplied beta coefficients.

PARAMETER **coef** the model beta coefficients

PARAMETER **independents** the observations

RETURN the raw prediction value

LogisticRegression/

Model_Deviance

Go Up

IMPORTS

LogisticRegression | LogisticRegression.Types |

DESCRIPTIONS

FUNCTION Model_Deviance

DATASET(Types.Deviance_Record)	**Model_Deviance**
(DATASET(Types.Observation_Deviance) od, DATASET(Types.Model_Coef) mod)	

Model Deviance.

PARAMETER **od** observation deviance record

PARAMETER **mod** model co-efficients

RETURN model deviance

LogisticRegression/ Null_Deviance

Go Up

IMPORTS

LogisticRegression | LogisticRegression.Types |

DESCRIPTIONS

FUNCTION Null_Deviance

DATASET(Types.Deviance_Record)	**Null_Deviance**
(DATASET(Types.Observation_Deviance) od)	

Deviance for the null model, that is, a model with only an intercept.

PARAMETER **od** Observation Deviance record set.

RETURN a data set of the null model deviances for each work item and classifier.

LogisticRegression/
Types

Go Up

IMPORTS

ML_Core.Types |

DESCRIPTIONS

MODULE Types

	Types

Children

1. t_Universe
2. Field_Desc
3. Data_Info
4. NumericField_U
5. DiscreteField_U
6. Layout_Column_Map
7. Classifier_Stats
8. Model_Report
9. Binomial_Confusion_Summary
10. Model_Coef

11. Confidence_Model_Coef
12. pval_Model_Coef
13. Raw_Prediction
14. Observation_Deviance
15. Deviance_Record
16. AOD_Record

ATTRIBUTE t_Universe

Types \

	t_Universe

RECORD Field_Desc

Types \

	Field_Desc

RECORD Data_Info

Types \

	Data_Info

RECORD NumericField_U

Types \

	NumericField_U

RECORD DiscreteField_U

Types \

	DiscreteField_U

RECORD Layout_Column_Map

Types \

	Layout_Column_Map

RECORD Classifier_Stats

Types \

	Classifier_Stats

RECORD Model_Report

Types \

	Model_Report

RECORD Binomial_Confusion_Summary

Types \

	Binomial_Confusion_Summary

RECORD Model_Coef

Types \

	Model_Coef

RECORD Confidence_Model_Coef

Types \

	Confidence_Model_Coef

RECORD pval_Model_Coef

Types \

	pval_Model_Coef

RECORD Raw_Prediction

Types \

	Raw_Prediction

RECORD Observation_Deviance

Types \

	Observation_Deviance

RECORD Deviance_Record

Types \

	Deviance_Record

RECORD AOD_Record

Types \

	AOD_Record