LogisticRegression

Module name: LogisticRegression
Description: Logistic regression using clinical data and .cls file
Author: Yujin Hoshida (Broad Institute) gp-help@broad.mit.edu
Date: 5/2/2006
Release: 1.0

Summary: Once a genomics marker or predictive model for clinical outcome is built, the next step is to test whether it really associates with outcome of new cohorts, or whether it has better predictive ability compared to already known clinical markers/predictive models. "Survival analysis" modules perform such analyses seamlessly after clustering/marker gene/prediction analyses.

Logistic regression is a commonly used to assess association of variable(s) of interest with binary clinical data (e.g., treatment response) [1-3]. For the LogisticRegression module, a .cls file identifies the phenotype label for each sample and a separate tab-delimited text file defines all other clinical variables. The module generates a result file similar to the one shown below. Interpreting the results requires basic knowledge of biostatistics and clinical data analysis.

```
# Logistic regression results
variable selection: none


Call:
glm(formula = cls.clinical ~ factor(cls), family = binomial(logit),
   data = clin.data)


Deviance Residuals:
  Min    1Q Median     3Q    Max
-1.641 -0.459 -0.459  0.776  2.146


Coefficients:
        Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.0460    0.3224  3.244 0.00118 **
factor(cls)2 -3.2432    0.5711 -5.679 1.35e-08 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)


   Null deviance: 136.058  on 99  degrees of freedom
Residual deviance:  89.814  on 98  degrees of freedom
AIC: 93.814


Number of Fisher Scoring iterations: 4


Deviance: 89.8139890522869


         Estimate Std. Error    Wald      OR lower.95%.CI
(Intercept)     1.045969    0.3224127    3.244192   2.84615385
1.51291723
factor(cls)2   -3.243193    0.5710828   -5.679025   0.03903904
0.01274629
         upper.95%.CI    P-value
(Intercept)    5.3542861 1.177844e-03
factor(cls)2   0.1195679 1.354649e-08
```

Reference:

1. Hastie TJ, Pregibon D. *Generalized linear models* in *Statistical Models in S* eds Chambers JM, Hastie TJ. 1992, Wadsworth&Brooks/Cole.
2. R Development Core Team. *Stats 2.2.1*, R package, http://www.r-project.org/
3. Venables WN, Ripley BD. *Generalized linear models* in *Modern Applied Statistics with S-PLUS*, 3rd Ed. 1999, Springer, New York.

Parameters:

Name	Description	Choices
input clin data filename	clinical data - .txt Format shown below.	
input cls filename	class information file -.cls Format described in GenePattern File Formats.	
output file	Name for the output file.	

response variable	Column name for the response/independent variable. Must be a column in the input clin data file.	
variable continous	Comma separated list of the continuous dependent variables (e.g. age) to use for the analysis. Each variable must be a column in the input clin data file.	
variable category	Comma separated list of the categorical dependent variables (e.g. gender) to use for the analysis. Each variable must be a column in the input clin data file or cls (to use the classes in the input cls file).	
variable interaction terms	Separate interaction variables with a colon (e.g. sex:cls). Separate sets of interaction variables with a comma (e.g. sex:cls, sex:grade). Each variable must be a column in the input clin data file or cls (to use the classes in the input cls file).	
input subgroup	Analyze a subgroup of the data. Identify the subgroup by specifying a variable followed by a colon and a value for the variable. For example: gender:M or cls:1).	
variable selection	Select the way dependent variables are entered into the model. Use the default value unless you are familiar with clinical data analysis.	Default: no Stepwise selection Forward selection Backward selection

Format of input clinical data file:
Tab-delimited text file. The first row contains column heads. The first column must be the sample identifier; the column head is your choice; the order of samples in this file

and the input cls file must match. Each subsequent column contains a clinical variable. For example:

name	time	censor	cls.clinical	labdata1	labdata1	age	sex	grade
case1	207	0	poor	0.9	430	72	1	4
case4	49	0	poor	2.7	220	54	1	4
case5	267	0	poor	28.7	500	53	1	4
case7	8	0	poor	2.4	1090	75	0	4
…	…	…	…	…	…	…	…	…

Identify continuous (e.g., age) or categorical (e.g., sex) dependent variables for the analysis by entering the column heads in the *variable continuous* or *variable category* fields, respectively. Identify the response/independent variable (e.g. cls.clinical) for the analysis by entering the column head in the *response variable* field.

Example input files: surv.txt, surv.cls
The example files contain mock data. To run an analysis using these files, enter the column head "cls.clinical" in the *response variable* field and "cls" in the *variable category* field.

Output files:
LogisticRegression_result.txt: summary of the results.

Platform dependencies:
      task type: Survival Analysis
      CPU type: any
      OS: any
      Language: R (version 2.2.1 or above)
      Suppot files: none