PNNXValidationOptimization Documentation

Module name: PNNXValidationOptimization
Description: PNN Leave-One-Out Cross-Validation Optimization
Author: Gad Getz (Broad Institute) gp-help@broad.mit.edu

Summary
Determines the optimal parameters to use when running PNN. The best model is determined by (i) the fewest number of leave-one-out errors on the training set, which include both false-positive and false-negative errors with the same weight, and (ii) among all conditions with the same error rate, the parameters that gave rise to the maximal mean log-likelihood of the training set were selected. The mean log-likelihood is defined as

$$L[\{\mathbf{x}_i\};M] = \frac{1}{\#\text{of training examples}} \sum_i \log(P_M(\mathrm{c_i} \mid \mathbf{x}_i))$$

where c_i is the true class of sample $\mathbf{x}_i$ and the probability is evaluated using the model *M*. The top n features are selected using chosen test statistic in a balanced manner; n/2 features with the top positive scores and n/2 features with most negative scores.

Parameters

Name	Description
data.filename	data file name - .gct
class.filename	class file for data - .cls
number.of.features	range of number of features to use (min:max or min:step:max)
test statistic used to select features	test.type
min.std	minimal standard deviation to use in variance-thresholded t-test
distance.type	distance measure used between samples
sigma.type	method to determine sigma scale
sigma	values of sigmas (min:max or min:step:max)
prior.type	method to determine class prior probablities
prior	prior probability to use for all classes (ignored if prior type is empirical)
optimization.results.file	name of output file for optimization results - .odf

References:

1. Lu, Getz, Miska, et al. “MicroRNA Expression Profiles Classify Human Cancers,” Nature 435, 834-838 (9 June 2005)

Return Value:

1. optimization results odf file that can be used as input to PNN

Platform dependencies:

Task type:	Prediction
CPU type:	x86
OS:	Windows
Language:	MATLAB