mlpy Documentation

Release 2.1.0

D. Albanese, G. Jurman, R. Visintainer

December 22, 2009

CONTENTS

Homepage: https://mlpy.fbk.eu *Platforms:* Linux, Mac OS X, Unix, Windows

Module author: mlpy Developers <albanese@fbk.eu>

Section author: Davide Albanese <albanese@fbk.eu>

`mlpy` is a high-performance Python package for predictive modeling. It makes extensive use of NumPy (http://scipy.org) to provide fast N-dimensional array manipulation and easy integration of C code. `mlpy` provides high level procedures that support, with few lines of code, the design of rich Data Analysis Protocols (DAPs) for preprocessing, clustering, predictive classification and feature selection. Methods are available for feature weighting and ranking, data resampling, error evaluation and experiment landscaping. The package includes tools to measure stability in sets of ranked feature lists.

`mlpy` is a project of the MPBA Research Unit at FBK, the Bruno Kessler Foundation in Trento, Italy (http://mpba.fbk.eu).

CHAPTER

ONE

TUTORIAL

1.1 A Simple Example

In this example the performance of SVM classifier is evaluated in a stratified k-fold resampling schema.

First, import NumPy and mlpy modules:

```
>>> import numpy as np
>>> import mlpy
```

Then, load a data file (*data.dat*) containing 30 samples described by 100 features (x) and labels (y):

```
>>> x, y = mlpy.data_fromfile('data.dat') # import data file
>>> x.shape
(30, 100)
```

Initialize SVM classifier, specifying kernel type (*linear*) and regularization parameter (C):

```
>>> classifier = mlpy.Svm(kernel = 'linear', C = 1.0)  # initialize the svm classifier
```

Define a stratified 10-fold resampling schema, where *idx* contains the sample indexes (list of train/test pairs):

```
>>> idx = mlpy.kfoldS(cl = y, sets = 10)
```

Actually build train and test data. Train the model on *xtr* and test it on *xts*. The performance is evaluated computing the average prediction error:

```
>>> pred_err = 0.0
>>> for idxtr, idxts in idx:
...     xtr, xts = x[idxtr], x[idxts]       # build training data
...     ytr, yts = y[idxtr], y[idxts]       # build test data
...     ret = classifier.compute(xtr, ytr)  # compute the model
...     pred = classifier.predict(xts)      # test the model on test data
...     pred_err += mlpy.err(yts, pred)          # compute the prediction error
>>> av_pred_err = pred_err / len(idx)       # compute the average prediction error
>>> av_pred_err
0.17499999999999999
```

CHAPTER

TWO

WAVELET TRANSFORM

2.1 Extend data

This function should be used in `dwt()` and `uwt()` to extend the length of data to power of two. `cwt()` use it as internal function.

extend(*x, method='reflection', length='powerof2'*)

Extend the 1D numpy array x beyond its original length.

Input

- *x* - [1D numpy array] data
- *method* - [string] indicates which extension method to use ('reflection', 'periodic', 'zeros')
- *length* - [string] indicates how to determinate the length of the extended data ('powerof2', 'double')

Output

- *xext* - [1D numpy array] extended version of x

Example

```
>>> import numpy as np
>>> import mlpy
>>> a = np.array([1,2,3,4,5])
>>> mlpy.extend(a, method='periodic', length='powerof2')
array([1, 2, 3, 4, 5, 1, 2, 3])
```

New in version 2.0.6.

2.2 Discrete Wavelet Transform

Discrete Wavelet Transform based on the GSL DWT [Gsldwt].

dwt(*x, wf, k*)

Discrete Wavelet Tranform

Input

- *x* - [1D numpy array float] data (the length is restricted to powers of two)
- *wf* - [string] wavelet type ('d': daubechies, 'h': haar, 'b': bspline)
- *k* - [integer] member of the wavelet family

–daubechies: k = 4, 6, ..., 20 with k even

–haar: the only valid choice of k is k = 2

–bspline: k = 103, 105, 202, 204, 206, 208, 301, 303, 305 307, 309

Output

•*X* - [1D numpy array float] discrete wavelet transform

idwt (*X, wf, k*)

Inverse Discrete Wavelet Tranform

Input

•*X* - [1D numpy array float] data

•*wf* - [string] wavelet type ('d': daubechies, 'h': haar, 'b': bspline)

•*k* - [integer] member of the wavelet family

–daubechies: k = 4, 6, ..., 20 with k even

–haar: the only valid choice of k is k = 2

–bspline: k = 103, 105, 202, 204, 206, 208, 301, 303, 305 307, 309

Output

•*x* - [1D numpy array float]

2.3 Undecimated Wavelet Transform

Undecimated Wavelet Transform based on the "wavelets" R package.

uwt (*x, wf, k, levels=0*)

Undecimated Wavelet Tranform

Input

•*x* - [1D numpy array float] data (the length is restricted to powers of two)

•*wf* - [string] wavelet type ('d': daubechies, 'h': haar, 'b': bspline)

•*k* - [integer] member of the wavelet family

–daubechies: k = 4, 6, ..., 20 with k even

–haar: the only valid choice of k is k = 2

–bspline: k = 103, 105, 202, 204, 206, 208, 301, 303, 305 307, 309

•***levels* - [integer] level of the decomposition (J).** If *levels* = 0 this is the value J such that the length of X is at least as great as the length of the level J wavelet filter, but less than the length of the level J+1 wavelet filter. Thus, j <= log_2((n-1)/(l-1)+1), where n is the length of x

Output

•*X* - [2D numpy array float] (2J * len(x)) undecimated wavelet transform

Data:

```
[[wavelet coefficients W_1]
 [wavelet coefficients W_2]
                :
 [wavelet coefficients W_J]
 [scaling coefficients V_1]
 [scaling coefficients V_2]
                :
 [scaling coefficients V_J]]
```

iuwt (*X, wf, k*)
Inverse Undecimated Wavelet Tranform

Input

- *X* - [2D numpy array float] data
- *wf* - [string] wavelet type ('d': daubechies, 'h': haar, 'b': bspline)
- *k* - [integer] member of the wavelet family
  - daubechies: k = 4, 6, ..., 20 with k even
  - haar: the only valid choice of k is k = 2
  - bspline: k = 103, 105, 202, 204, 206, 208, 301, 303, 305 307, 309

Output

- *x* - [1D numpy array float]

New in version 2.0.2.

2.4 Continuous Wavelet Transform

Continuous Wavelet Transform based on [Torrence98].

cwt (*x, dt, dj, wf='dog', p=2, extmethod='none', extlength='powerof2'*)
Continuous Wavelet Tranform.

Input

- *x* - [1D numpy array float] data
- *dt* - [float] time step
- *dj* - [float] scale resolution (smaller values of dj give finer resolution)
- *wf* - [string] wavelet function ('morlet', 'paul', 'dog')
- *p* - [float] wavelet function parameter
- *extmethod* - [string] indicates which extension method to use ('none', 'reflection', 'periodic', 'zeros')
- *extlength* - [string] indicates how to determinate the length of the extended data ('powerof2', 'double')

Output

- X, scales - (scales x angularfreq) [2D numpy array complex], scales [1D numpy array float]

icwt (*X, dt, dj, wf='dog', p=2, recf=True*)
Inverse Continuous Wavelet Tranform.

Input

- *X* - (scales x angularfreq) [2D numpy array complex]
- *dt* - [float] time step
- *dj* - [float] scale resolution (smaller values of dt give finer resolution)
- *wf* - [string] wavelet function ('morlet', 'paul', 'dog')
- *p* - [int] wavelet function parameter
  - morlet: 2, 4, 6
  - paul: 2, 4, 6
  - dog: 2, 6, 10
- *recf* - [bool] use the reconstruction factor (Cdelta*psi_0(0))

Output

- x - [1D numpy array float]

2.4.1 Other functions

See [Torrence98].

angularfreq(*N, dt*)
Compute angular frequencies.

Input

- *N* - [integer] number of data samples
- *dt* - [float] time step

Output

- *angular frequencies* - [1D numpy array float]

scales(*N, dj, dt, s0*)
Compute scales.

Input

- *N* - [integer] number of data samples
- *dj* - [float] scale resolution
- *dt* - [float] time step

Output

- *scales* - [1D numpy array float]

compute_s0(*dt, p, wf*)
Compute s0.

Input

- *dt* - [float] time step
- *p* - [float] omega0 ('morlet') or order ('paul', 'dog')
- *wf* - [string] wavelet function ('morlet', 'paul', 'dog')

Output

- *s0* - [float]

CHAPTER

THREE

IMPUTING

3.1 Purify

purify(*x, th0=0.10000000000000001, th1=0.10000000000000001*)

Purify.

Return the matrix x without rows and cols containing respectively more than th0 * x.shape[1] and th1 * x.shape[0] NaNs.

Output

- *xout, v0, v1*

where v0 are the valid index at dimension 0 and v1 are the valid index at dimension 1

Example:

```
>>> import numpy as np
>>> import mlpy
>>> x = np.array([[1,      4,      4     ],
...               [2,      9,      np.NaN],
...               [2,      5,      8     ],
...               [8,      np.NaN, np.NaN],
...               [np.NaN, 4,      4     ]])
>>> y = np.array([1, -1, 1, -1, -1])
>>> x, v0, v1 = mlpy.purify(x, 0.4, 0.4)
>>> x
array([[  1.,   4.,   4.],
       [  2.,   9.,  NaN],
       [  2.,   5.,   8.],
       [ NaN,   4.,   4.]])
>>> v0
array([0, 1, 2, 4])
>>> v1
array([0, 1, 2])
```

New in version 2.0.4.

3.2 KNN imputing

knn_imputing(*x, y, k, dist='e', method='mean', ldep=True*)

Knn imputing.

Input

- *x* - [2D numpy array float] (#sample x #feature) data to impute
- *y* - [1D numpy array integer/float] labels
- *k* - [integer] number of nearest neighbor
- *dist* - [string] adopted distance ('se' = SQUARED EUCLIDEAN, 'e' = EUCLIDEAN)
- *method* - [string] method to compute the missing values ('mean', 'median')
- *ldep* - [bool] label depended

Output

- *xout* - [2D numpy array float] (#sample x #feature) data imputed

New in version 2.0.4.

3.3 Examples

```
>>> import numpy as np
>>> import mlpy
>>> x = np.array([[1,      4,      4     ],
...               [2,      9,      np.NaN],
...               [2,      5,      8     ],
...               [8,      np.NaN, np.NaN],
...               [np.NaN, 4,      4     ]])
>>> y = np.array([1, -1, 1, -1, -1])
>>> x, v0, v1 = mlpy.purify(x, 0.4, 0.4)
>>> x
array([[  1.,   4.,   4.],
       [  2.,   9.,  NaN],
       [  2.,   5.,   8.],
       [ NaN,   4.,   4.]])
>>> v0
array([0, 1, 2, 4])
>>> v1
array([0, 1, 2])
>>> y = y[v0]
>>> x = mlpy.knn_imputing(x, y, 2, dist='e', method='mean', ldep=False)
>>> x
array([[ 1. ,  4. ,  4. ],
       [ 2. ,  9. ,  6. ],
       [ 2. ,  5. ,  8. ],
       [ 1.5,  4. ,  4. ]])
```

CHAPTER

FOUR

DISTANCE COMPUTATIONS

4.1 Dynamic Time Warping

Features:

- Naive and Derivative [Keogh01] DTW
- Symmetric, Asymmetric, Quasi-Symmetric implementation with Slope Constraint Condition P=0 [Sakoe78]
- Sakoe-Chiba window condition [Sakoe78] option
- Linear space-complexity implementation option

class `Dtw` (*derivative=False, startbc=True, steppattern='symmetric0', wincond='nowindow', r=0.0, onlydist=True*)

Input

- *derivative* - [bool] Derivative DTW (DDTW).
- *startbc* - [bool] (0, 0) boundary condition
- *steppattern* - [string] step pattern ('symmetric', 'asymmetric', 'quasisymmetric')
- *wincond* - [string] window condition ('nowindow', 'sakoechiba')
- *r* - [float] sakoe-chiba window length
- *onlydist* - [bool] linear space-complexity implementation. Only the current and previous columns are kept in memory.

New in version 2.0.7.

`compute` (*x, y*)

Input

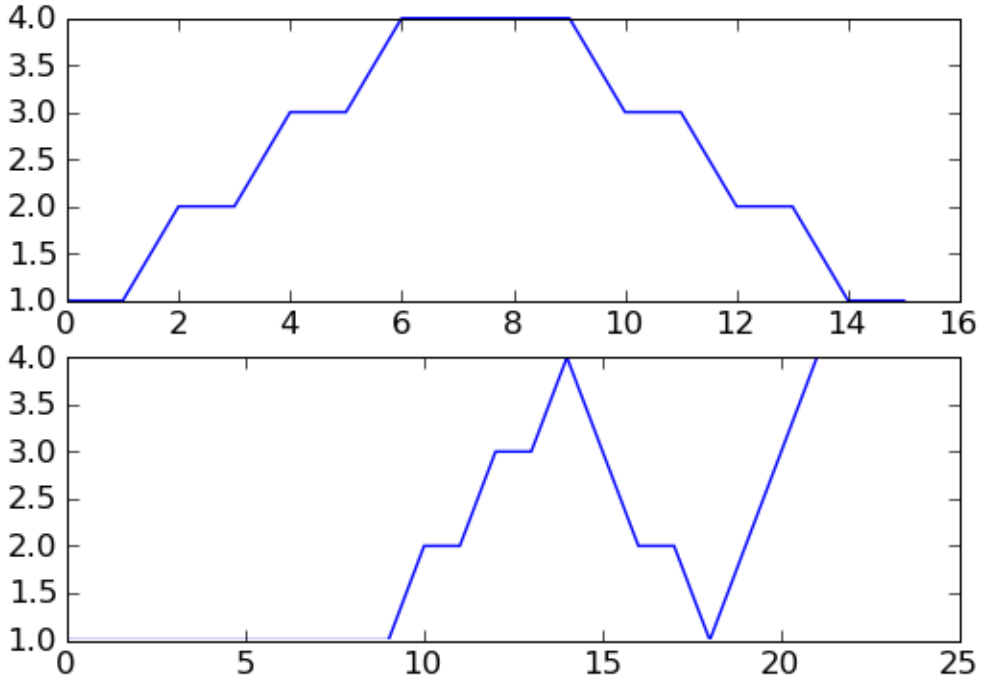
- *x* - [1D numpy array float / list] first time series
- *y* - [1D numpy array float / list] second time series

Output

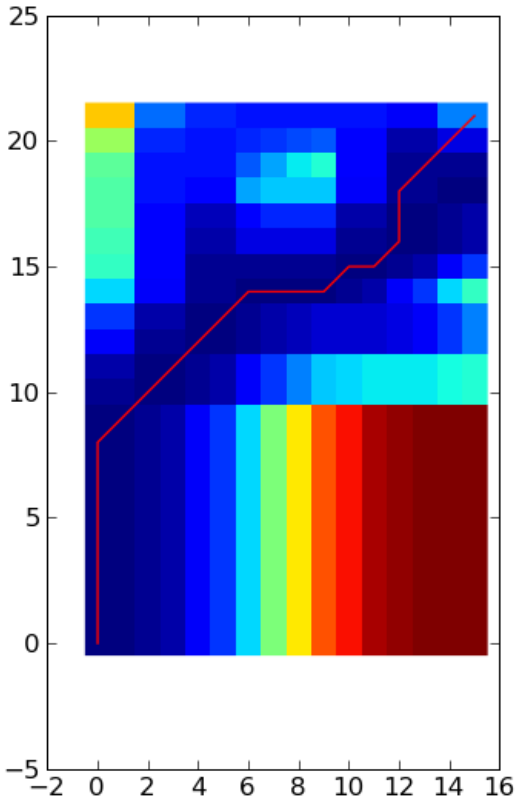
- *d* - [float] normalized distance
- *self.px* - [1D numpy array int] optimal warping path (for x time series) (for onlydist=False)
- *self.py* - [1D numpy array int] optimal warping path (for y time series) (for onlydist=False)
- *self.cost* - [2D numpy array float] cost matrix (for onlydist=False)

Example:

```
>>> import numpy as np
>>> import matplotlib.pyplot as plt
>>> import mlpy
>>> x = np.array([1,1,2,2,3,3,4,4,4,4,3,3,2,2,1,1])
>>> y = np.array([1,1,1,1,1,1,1,1,1,1,2,2,3,3,4,3,2,2,1,2,3,4])
>>> plt.figure(1)
>>> plt.subplot(211)
>>> plt.plot(x)
>>> plt.subplot(212)
>>> plt.plot(y)
>>> plt.show()
```




```
>>> mydtw = mlpy.Dtw()
>>> d = mydtw.compute(x, y)
>>> plt.figure(2)
>>> plt.imshow(mydtw.cost.T, interpolation='nearest', origin='lower')
>>> plt.plot(mydtw.px, mydtw.py, 'r')
>>> plt.show()
```



4.2 Minkowski Distance

class `Minkowski` (p)

Computes the Minkowski distance between two vectors `x` and `y`.

$$||x - y||_p = (\sum |x_i - y_i|^p)^{1/p}.$$

Initialize Minkowski class.

Parameters

p [float] The norm of the difference $||x - y||_p$

New in version 2.0.8.

`compute` (x, y)

Compute Minkowski distance

Parameters

x [ndarray] An 1-dimensional vector.

y [ndarray] An 1-dimensional vector.

Returns

d [float] The Minkowski distance between vectors `x` and `y`

CHAPTER

FIVE

CLUSTERING

5.1 Hierarchical Clustering

Hierarchical Clustering algorithm derived from the R package 'amap' [Amap].

class **HCluster** (*method='euclidean', link='complete'*)

Hierarchical Cluster.

Initialize Hierarchical Cluster.

Parameters

method [string ('euclidean')] the distance measure to be used

link [string ('single', 'complete', 'mcquitty', 'median')] the agglomeration method to be used

Example:

```
>>> import numpy as np
>>> import mlpy
>>> x = np.array([[ 1. ,  1.5],
...               [ 1.1,  1.8],
...               [ 2. ,  2.8],
...               [ 3.2,  3.1],
...               [ 3.4,  3.2]])
>>> hc = mlpy.HCluster()
>>> hc.compute(x)
>>> hc.ia
array([-4, -1, -3,  2])
>>> hc.ib
array([-5, -2,  1,  3])
>>> hc.heights
array([ 0.2236068 ,  0.31622776,  1.4560219 ,  2.94108844])
>>> hc.cut(0.5)
array([0, 0, 1, 2, 2])
```

compute (*x*)

Compute Hierarchical Cluster.

Parameters

x [ndarray] An 2-dimensional vector (sample x features).

Returns

self.ia [ndarray (1-dimensional vector)] merge

self.ib [ndarray (1-dimensional vector)] merge

self.heights [ndarray (1-dimensional vector)] a set of n-1 non-decreasing real values. The clustering height: that is, the value of the criterion associated with the clustering method for the particular agglomeration.

Element i of merge describes the merging of clusters at step i of the clustering. If an element j is negative, then observation -j was merged at this stage. If j is positive then the merge was with the cluster formed at the (earlier) stage j of the algorithm. Thus negative entries in merge indicate agglomerations of singletons, and positive entries indicate agglomerations of non-singletons.

cut (*ht*)

Cuts the tree into several groups by specifying the cut height.

Parameters

ht [float] height where the tree should be cut

Returns

cl [ndarray (1-dimensional vector)] group memberships. Groups are in 0, ..., N-1

5.2 k-medoids

class Kmedoids (*k, dist, maxloops=100, rs=0*)

k-medoids algorithm.

Initialize Kmedoids.

Parameters

k [int] Number of clusters/medoids

dist [class] class with a .compute(x, y) method which returns a distance

maxloops [int] maximum number of loops

rs [int] random seed

Example:

```
>>> import numpy as np
>>> import mlpy
>>> x = np.array([[ 1. ,  1.5],
...               [ 1.1,  1.8],
...               [ 2. ,  2.8],
...               [ 3.2,  3.1],
...               [ 3.4,  3.2]])
>>> dtw = mlpy.Dtw(onlydist=True)
>>> km = mlpy.Kmedoids(k=3, dist=dtw)
>>> km.compute(x)
(array([4, 0, 2]), array([3, 1]), array([0, 1]), 0.072499999999999981)
```

Samples 4, 0, 2 are medoids and represent cluster 0, 1, 2 respectively.

- cluster 0: samples 4 (medoid) and 3
- cluster 1: samples 0 (medoid) and 1
- cluster 2: sample 2 (medoid)

New in version 2.0.8.

compute(*x*)
Compute Kmedoids.

Parameters

x [ndarray] An 2-dimensional vector (sample x features).

Returns

m [ndarray (1-dimensional vector)] medoids indexes

n [ndarray (1-dimensional vector)] non-medoids indexes

cl [ndarray 1-dimensional vector)] cluster membership for non-medoids. Groups are in 0, ..., k-1

co [double] total cost of configuration

CHAPTER

SIX

CLASSIFICATION

Every classifier must be initialized with a specific set of parameters. Two distinct methods are deployed for the *training* and the *testing* phases. Whenever possible, the real valued prediction is stored in the *realpred* variable.

6.1 Binary Classification

6.1.1 Compute Model

compute (*x, y*)

x - training data [2D numpy array float]

- x.shape[0] number of samples
- x.shape[1] number of features

y - training classes (1 or -1) [1D numpy array integer]

- y.shape[0] number of samples

6.1.2 Test Model

predict (*p*)

p - test data [1D or 2D numpy array float]

- 1D: one sample
  - p.shape[0] number of features
- 2D: more than one sample
  - p.shape[0] number of samples
  - p.shape[1] number of features

6.2 Multiclass Classification

6.2.1 Compute Model

compute (*x, y*)

x - training data [2D float numpy array]

- x.shape[0] number of samples
- x.shape[1] number of features

y - training classes (1, ..., #classes) [1D integer numpy array]

- y.shape[0] number of samples

6.2.2 Test Model

predict(*p*)

p - test data [1D or 2D float numpy array]

- 1D: one sample
  - p.shape[0] number of features
- 2D: more than one sample
  - p.shape[0] number of samples
  - p.shape[1] number of features

6.3 Classifiers

6.3.1 Support Vector Machines (SVMs)

class Svm(*kernel='linear', kp=0.10000000000000001, C=1.0, tol=0.001, eps=0.001, maxloops=1000, cost=0.0, alpha_tversky=1.0, beta_tversky=1.0, opt_offset=True*)

Support Vector Machines (SVM).

Example

```
>>> import numpy as np
>>> import mlpy
>>> xtr = np.array([[1.0, 2.0, 3.0, 1.0],  # first sample
...                 [1.0, 2.0, 3.0, 2.0],  # second sample
...                 [1.0, 2.0, 3.0, 1.0]]) # third sample
>>> ytr = np.array([1, -1, 1])             # classes
>>> mysvm = mlpy.Svm()                     # initialize Svm class
>>> mysvm.compute(xtr, ytr)                # compute SVM
1
>>> mysvm.predict(xtr)                     # predict SVM model on training data
array([ 1, -1,  1])
>>> xts = np.array([4.0, 5.0, 6.0, 7.0])   # test point
>>> mysvm.predict(xts)                     # predict SVM model on test point
-1
>>> mysvm.realpred                         # real-valued prediction
-5.5
>>> mysvm.weights(xtr, ytr)                # compute weights on training data
array([ 0.,  0.,  0.,  1.])
```

Initialize the Svm class

Parameters

kernel [string ['linear', 'gaussian', 'polynomial', 'tr', 'tversky']] kernel

kp [float] kernel parameter (two sigma squared) for gaussian and polynomial kernel

C [float] regularization parameter

tol [float] tolerance for testing KKT conditions

eps [float] convergence parameter

maxloops [integer] maximum number of optimization loops

cost [float [-1.0, ..., 1.0]] for cost-sensitive classification

alpha_tversky [float] positive multiplicative parameter for the norm of the first vector

beta_tversky [float] positive multiplicative parameter for the norm of the second vector

opt_offset [bool] compute the optimal offset

compute(*x, y*)
Compute SVM model

Parameters

x [2d ndarray float (samples x feats)] training data

y [1d ndarray integer (-1 or 1)] classes

Returns

conv [integer] svm convergence (0: false, 1: true)

predict(*p*)
Predict svm model on a test point(s)

Parameters

p [1d or 2d ndarray float (samples x feats)] test point(s)training dataInput

Returns

cl [integer or 1d ndarray integer] class(es) predicted

Attributes

Svm.realpred [float or 1d ndarray float] real valued prediction

weights(*x, y*)
Return feature weights

Parameters

x [2d ndarray float (samples x feats)] training data

y [1d ndarray integer (-1 or 1)] classes

Returns

fw [1d ndarray float] feature weights

Note: For *tr* kernel (Terminated Ramp Kernel) see [Merler06].

6.3.2 K Nearest Neighbor (KNN)

class Knn(*k, dist='se'*)
k-Nearest Neighbor (KNN).

Initialize the Knn class.

Input

- *k* - [integer] number of NN
- *dist* - [string] adopted distance ('se' = SQUARED EUCLIDEAN, 'e' = EUCLIDEAN)

compute (*x, y*)

Store x and y data.

Input

- *x* - [2D numpy array float] (#sample x #feature) training data
- *y* - [1D numpy array integer] classes
  - -1 or 1 for binary classification
  - 1, ..., nclasses for multiclass classification

Output

- 1

predict (*p*)

Predict knn model on a test point(s).

Input

- *p* - [1D or 2D numpy array float] test point(s)

Output: the predicted value(s) on success:

- -1 or 1 for binary classification
- 1, ..., nclasses for multiclass classification
- 0 on succes with non unique classification
- -2 otherwise

6.3.3 Fisher Discriminant Analysis (FDA)

Described in [Mika01].

class Fda (*C=1*)

Fisher Discriminant Analysis.

Initialize Fda class.

Input

- *C* - [float] Regularization parameter

compute (*x, y*)

Compute fda model.

Input

- *x* - [2D numpy array float] (sample x feature) training data
- *y* - [1D numpy array integer] (two classes, 1 or -1) classes

Output

- 1

predict (*p*)

Predict fda model on test point(s).

Input

- *p* - [1D or 2D numpy array float] test point(s)

Output

- *cl* - [integer or 1D numpy array integer] class(es) predicted
- *self.realpred* - [float or 1D numpy array float] real valued prediction

weights (*x, y*)

Return feature weights.

Input

- *x* - [2D numpy array float] (sample x feature) training data
- *y* - [1D numpy array integer] (two classes, 1 or -1) classes

Output

- *fw* - [1D numpy array float] feature weights

6.3.4 Spectral Regression Discriminant Analysis (SRDA)

Described in [Cai08].

class Srda (*alpha=1.0*)

Spectral Regression Discriminant Analysis (SRDA).

Example:

```
>>> from numpy import *
>>> from mlpy import *
>>> xtr = array([[1.0, 2.0, 3.1, 1.0],  # first sample
...              [1.0, 2.0, 3.0, 2.0],  # second sample
...              [1.0, 2.0, 3.2, 1.0]]) # third sample
>>> ytr = array([1, -1, 1])             # classes
>>> mysrda = Srda()                     # initialize srda class
>>> mysrda.compute(xtr, ytr)            # compute srda
1
>>> mysrda.predict(xtr)                 # predict srda model on training data
array([ 1, -1,  1])
>>> xts = array([4.0, 5.0, 6.0, 7.0])   # test point
>>> mysrda.predict(xts)                 # predict srda model on test point
-1
>>> mysrda.realpred                     # real-valued prediction
-16.500000000001439
>>> mysrda.weights(xtr, ytr)            # compute weights on training data
array([  1.00000000e+00,   2.00000000e+00,   5.40012479e-13,
         4.50000000e+00])
```

Initialize the Srda class.

Input

- *alpha* - [float] (>=0.0) regularization parameter

compute (*x, y*)

Compute Srda model.

Initialize array of alphas a.

Input

- *x* - [2D numpy array float] (sample x feature) training data
- *y* - [1D numpy array integer] (two classes) classes

Output

- 1

predict (*p*)
Predict Srda model on test point(s).

Input

- *p* - test point(s) [1D or 2D numpy array float]

Output

- *cl* - [integer or 1D numpy array integer] class(es) predicted
- *self.realpred* - [float or 1D numpy array float] real valued prediction

weights (*x, y*)
Return feature weights.

Input

- *x* - [2D numpy array float] (sample x feature) training data
- *y* - [1D numpy array integer] (two classes) classes

Output

- *fw* - [1D numpy array float] feature weights

6.3.5 Penalized Discriminant Analysis (PDA)

Described in [Ghosh03].

class Pda (*Nreg=3*)
Penalized Discriminant Analysis (PDA).

Initialize Pda class.

Input

- *Nreg* - [integer] number of regressions

compute (*x, y*)
Compute Pda model.

Input

- *x* - [2D numpy array float] (sample x feature) training data
- *y* - [1D numpy array integer] (two classes, 1 or -1) classes

Output

- 1

predict (*p*)
Predict Pda model on test point(s).

Input

- *p* - [1D or 2D numpy array float] test point(s)

Output

- *cl* - [integer or 1D numpy array integer] class(es) predicted
- *self.realpred* - [float or 1D numpy array float] real valued prediction

weights (*x, y*)

Return feature weights.

Input

- *x* - [2D numpy array float] (sample x feature) training data
- *y* - [1D numpy array integer] (two classes, 1 or -1) classes

Output

- *fw* - [1D numpy array float] feature weights

6.3.6 Diagonal Linear Discriminant Analysis (DLDA)

class Dlda (*nf=0, tol=10, overview=False, bal=False*)

Diagonal Linear Discriminant Analysis.

Example:

```
>>> from numpy import *
>>> from mlpy import *
>>> xtr = array([[1.1, 2.4, 3.1, 1.0],  # first sample
...              [1.2, 2.3, 3.0, 2.0],  # second sample
...              [1.3, 2.2, 3.5, 1.0],  # third sample
...              [1.4, 2.1, 3.2, 2.0]]) # fourth sample
>>> ytr = array([1, -1, 1, -1])         # classes
>>> mydlda = Dlda()                     # initialize dlda class
>>> xtr_std = data_standardize(xtr)     # standardize the training dataset
>>> mydlda.compute(xtr_std, ytr)        # compute dlda
1
>>> mydlda.predict(xtr_std)             # predict dlda model on training data
array([ 1, -1,  1, -1])
>>> xts = array([4.0, 5.0, 6.0, 7.0])   # test point
>>> mydlda.predict(xts)                 # predict dlda model on test point
-1
>>> mydlda.realpred                     # real-valued prediction
-45.32292203746583
>>> mydlda.weights(xtr_std, ytr)        # compute weights on training data
array([  3.00000000e+00,   3.00000000e+00,   9.32587341e-15,
         2.61756029e+00])
```

Initialize Dlda class.

Input:

- *nf* - [integer] the number of the best features that we want to use in the model (1 <= nf >= #features). If nf = 0 the system stops at a number of features corresponding to a peak of accuracy
- *tol* - [integer] in case of nf = 0 it's the number of steps of classification to be calculated after the peak to avoid a local maximum
- *overview* - [bool] set True to print informations about the accuracy of the classifier at every step of the compute
- *bal* - [bool] set True if it's reasonable to consider the unbalancement of the test set similar to the one of the training set

compute (*x, y, mf=0*)
Compute Dlda model.

Initialize array of alphas a.

Input

- *x* - [2D numpy array float] (sample x feature) training data
- *y* - [1D numpy array integer] (two classes) classes
- *mf* - [integer] (More Features) number of classification steps to be calculated more on a model already computed

Output

- 1

predict (*p*)
Predict Dlda model on test point(s).

Input

- *p* - [1D or 2D numpy array float] test point(s)

Output

- *cl* - [integer or 1D numpy array integer] class(es) predicted
- *self.realpred* - [1D numpy array float] real valued prediction

weights (*x, y*)
Return feature weights.

Input

- *x* - [2D numpy array float] (sample x feature) training data
- *y* - [1D numpy array integer] (two classes, 1 or -1) classes

Output

- *fw* - [1D numpy array float] feature weights (they are gonna be > 0 for the features chosen for the classification and = 0 for all the others)

CHAPTER

SEVEN

FEATURE WEIGHTING

Algorithms for assessing the quality of features.

7.1 Compute Feature Weights

weights (*x, y*)

x - data [2D float numpy array]

- x.shape[0] number of samples
- x.shape[1] number of features

y - classes (-1 or 1) [1D integer numpy array]

- y.shape[0] number of samples

7.2 Methods

7.2.1 Classifier-derived

See *Classification*.

7.2.2 Classifier-independent

Iterative RELIEF (I-RELIEF)

class Irelief (*T=1000, sigma=1.0, theta=0.001*)

Iterative RELIEF for Feature Weighting.

Example:

```
>>> from numpy import *
>>> from mlpy import *
>>> x = array([[1.1, 2.1, 3.1, -1.0],  # first sample
...            [1.2, 2.2, 3.2, 1.0],   # second sample
...            [1.3, 2.3, 3.3, -1.0]]) # third sample
>>> y = array([1, 2, 1])               # classes
>>> myir = Irelief()                   # initialize irelief class
```

```
>>> myir.weights(x, y)                  # compute feature weights
array([ 0.,  0.,  0.,  1.])
```

Initialize the Irelief class.

Input

- *T* - [integer] (>0) max loops
- *sigma* - [float] (>0.0) kernel width
- *theta* - [float] (>0.0) convergence parameter

weights (*x, y*)
Return feature weights.

Input

- *x* - [2D numpy array float] (sample x feature) training data
- *y* - [1D numpy array integer] (two classes) classes

Output

- *fw* - [1D numpy array float] feature weights

exception SigmaError
Sigma Error

Sigma parameter is too small.

Feature Weighting/Selection Sun08

A feature weighting/selection algorithm described in [Sun08].

class FSSun (*T=1000, sigma=1.0, theta=0.001, lmbd=1.0, eps=0.001, alpha0=1.0, c=0.01, rho=0.5*)
Sun Algorithm for feature weighting/selection

Initialize the FSSun class

Parameters

T [int (> 0)] max loops

sigma [float (> 0.0)] kernel width

theta [float (> 0.0)] convergence parameter

lmbd [float] regularization parameter

eps [float (0 < eps << 1)] termination tolerance for steepest descent method

alpha0 [float (> 0.0)] initial step length (usually 1.0) for line search

c [float (0 < c < 1/2)] costant for line search

rho [flaot (0 < rho < 1)] alpha coefficient for line search

New in version 2.0.9.

weights (*x, y*)
Compute the feature weights

Parameters

x [2d ndarray float (samples x feats)] training data

y [1d ndarray integer (-1 or 1)] classes

Returns

fw [1d ndarray float] feature weights

Attributes

FSSun.loops [int] number of loops

Raises

ValueError if classes are not -1 or 1

SigmaError if sigma parameter is too small

Discrete Wavelet Transform (DWT)

class `Dwt` (*specdiff='rpv'*)

Discrete Wavelet Transform (DWT).

Initialize the Dwt class.

Input

- *specdiff* - [string] spectral difference method ('rpv', 'arpv', 'crpv')

`weights` (*x, y*)

Return ABSOLUTE feature weights.

Input

- *x* - [2D numpy array float] (sample x feature) training data
- *y* - [1D numpy array integer] (two classes, 1 and -1) classes

Output

- *fw* - [1D numpy array float] feature weights

CHAPTER

EIGHT

FEATURE RANKING

The feature weights are used for selecting and ranking purposes inside one of the implemented schemes:

- *Recursive Feature Elimination family* [Guyon02]: RFE, ERFE [Furlanello03], BISRFE, SQRTRFE
- *Recursive Forward Selection* family [Louw06]: RFS
- *One-step*

class **Ranking** (*method='rfe', lastsinglesteps=0*)

Ranking class based on Recursive Feature Elimination (RFE) and Recursive Forward Selection (RFS) methods.

Example:

```
>>> from numpy import *
>>> from mlpy import *
>>> x = array([[1.1, 2.1, 3.1, -1.0],  # first sample
...            [1.2, 2.2, 3.2, 1.0],   # second sample
...            [1.3, 2.3, 3.3, -1.0]]) # third sample
>>> y = array([1, -1, 1])              # classes
>>> myrank = Ranking()                 # initialize ranking class
>>> mysvm = Svm()                      # initialize svm class
>>> myrank.compute(x, y, mysvm)        # compute feature ranking
array([3, 1, 2, 0])
```

Initialize Ranking class.

Input

•*method* - [string] method ('onestep', 'rfe', 'bisrfe', 'sqrtrfe', 'erfe', 'rfs')

•*lastsinglesteps* - [integer] last single steps with 'rfe'

compute (*x, y, w, debug=False*)

Compute the feature ranking.

Input

•*x* - [2D numpy array float] (sample x feature) training data

•*y* - [1D numpy array integer] (1 or -1) classes

•*w* - object (e.g. classifier) with weights() method

•*debug* - [bool] show remaining number of feature at each step (True or False)

Output

•*feature ranking* - [1D numpy array integer] ranked feature indexes

CHAPTER

NINE

RESAMPLING METHODS

9.1 k-fold

kfold(*nsamples, sets, rseed=0, indexes=None*)
K-fold Resampling Method.

Input

- *nsamples* - [integer] number of samples
- *sets* - [integer] number of subsets (= number of tr/ts pairs)
- *rseed* - [integer] random seed
- *indexes* - [list integer] source indexes (None for [0, nsamples-1])

Output

- *idx* - list of *sets* tuples: ([training indexes], [test indexes])

kfoldS(*cl, sets, rseed=0, indexes=None*)
Stratified K-fold Resampling Method.

Input

- *cl* - [list (1 or -1)] class label
- *sets* - [integer] number of subsets (= number of tr/ts pairs)
- *rseed* - [integer] random seed
- *indexes* - [list integer] source indexes (None for [0, nsamples-1])

Output

- *idx* - list of *sets* tuples: ([training indexes], [test indexes])

9.2 Monte Carlo

montecarlo(*nsamples, pairs, sets, rseed=0, indexes=None*)
Monte Carlo Resampling Method.

Input

- *nsamples* - [integer] number of samples
- *pairs* - [integer] number of tr/ts pairs

- *sets* - [integer] 1/(fraction of data in test sets)
- *rseed* - [integer] random seed
- *indexes* - [list integer] source indexes (None for [0, nsamples-1])

Output

- *idx* - list of *pairs* tuples: ([training indexes], [test indexes])

montecarloS(*cl, pairs, sets, rseed=0, indexes=None*)
Stratified Monte Carlo Resampling Method.

Input

- *cl* - [list (1 or -1)] class label
- *pairs* - [integer] number of tr/ts pairs
- *sets* - [integer] 1/(fraction of data in test sets)
- *rseed* - [integer] random seed
- *indexes* - [list integer] source indexes (None for [0, nsamples-1])

Output

- *idx* - list of *pairs* tuples: ([training indexes], [test indexes])

9.3 Leave-one-out

leaveoneout(*nsamples, indexes=None*)
Leave-one-out Resampling Method.

Input

- *nsamples* - [integer] number of samples
- *indexes* - [list integer] source indexes (None for [0, nsamples-1])

Output

- *idx* - list of *nsamples* tuples: ([training indexes], [test indexes])

9.4 All Combinations

allcombinations(*cl, sets, indexes=None*)
All Combinations Resampling Method.

Input

- *cl* - [list (1 or -1)] class label
- *sets* - [integer] number of subset
- *indexes* - [list integer] source indexes (None for [0, nsamples-1])

Output

- *idx* - list of tuples: ([training indexes], [test indexes])

9.5 Manual Resampling

manresampling(*cl, pairs, trp, trn, tsp, tsn, rseed=0*)

Manual Resampling.

Input

- *cl* - [list (1 or -1)] class label
- *pairs* - [integer] number of tr/ts pairs
- *trp* - [integer] number of positive samples in training
- *trn* - [integer] number of negative samples in training
- *tsp* - [integer] number of positive samples in test
- *tsn* - [integer] number of negative samples in test

Output

- *idx* - list of *pairs* tuples: ([training indexes], [test indexes])

9.6 Resampling File

resamplingfile(*nsamples, file, sep='t'*)

Resampling file from file.

Returns a list of tuples: ([training indexes],[test indexes])

Read a file in the form:

```
[test indexes 'sep'-separated for the first  replicate]
[test indexes 'sep'-separated for the second replicate]
                          .
                          .
                          .
[test indexes 'sep'-separated for the last   replicate]
```

where indexes must be integers in [0, nsamples-1].

Input

- *file* - [string] test indexes file
- *nsamples* - [integer] number of samples

Output

- *idx* - list of tuples: ([training indexes],[test indexes])

CHAPTER

TEN

METRIC FUNCTIONS

Compute metrics for assessing the performance of binary classification models.

The Confusion Matrix:

Total Samples (ts)	Actual Positives (ap)	Actual Negatives (an)
Predicted Positives (pp)	True Positives (tp)	False Positives (fp)
Predicted Negatives (pn)	False Positives (fn)	True Negatives (tn)

10.1 Error

err (*y, p*)

Compute the Error.

error = (fp + fn) / ts

Input

- *y* - classes (two classes) [1D numpy array integer]
- *p* - prediction (two classes) [1D numpy array integer]

Output

- error

errp (*y, p*)

Compute the Error for positive samples.

errp = fp / ap

Input

- *y* - classes (two classes +1 and -1) [1D numpy array integer]
- *p* - prediction (two classes +1 and -1) [1D numpy array integer]

Output

- error for positive samples

errn (*y, p*)

Compute the Error for negative samples.

errn = fn / an

Input

- *y* - classes (two classes +1 and -1) [1D numpy array integer]

- *p* - prediction (two classes +1 and -1) [1D numpy array integer]

Output

- error for negative samples

10.2 Accuracy

acc (*y*, *p*)

Compute the Accuracy.

accuracy = (tp + tn) / ts

Input

- *y* - classes (two classes) [1D numpy array integer]
- *p* - prediction (two classes) [1D numpy array integer]

Output

- accuracy

10.3 Sensitivity and Specificity

sens (*y*, *p*)

Compute the Sensitivity.

sensitivity = tp / ap

Input

- *y* - classes (two classes +1 and -1) [1D numpy array integer]
- *p* - prediction (two classes +1 and -1) [1D numpy array integer]

Output

- sensitivity

spec (*y*, *p*)

Compute the Specificity.

specificity = tn / an

Input

- *y* - classes (two classes +1 and -1) [1D numpy array integer]
- *p* - prediction (two classes +1 and -1) [1D numpy array integer]

Output

- specificity

10.4 AUC

single_auc (*y*, *p*)

Compute the single AUC.

Input

- *y* - classes (two classes +1 and -1) [1D numpy array integer]
- *p* - prediction (two classes +1 and -1) [1D numpy array integer]

Output

- singleAUC

wmw_auc (*y, r*)

Compute the AUC by using the Wilcoxon-Mann-Whitney formula.

Input

- *y* - classes (two classes +1 and -1) [1D numpy array integer]
- *r* - real-valued prediction [1D numpy array float]

Output

- wmwAUC

10.5 Other

ppv (*y, p*)

Compute the Positive Predictive Value (PPV).

PPV = tp / pp

Input

- *y* - classes (two classes +1 and -1) [1D numpy array integer]
- *p* - prediction (two classes +1 and -1) [1D numpy array integer]

Output

- PPV

npv (*y, p*)

Compute the Negative Predictive Value (NPV).

NPV = tn / pn

Input

- *y* - classes (two classes +1 and -1) [1D numpy array integer]
- *p* - prediction (two classes +1 and -1) [1D numpy array integer]

Output

- NPV

mcc (*y, p*)

Compute the Matthews Correlation Coefficient (MCC).

MCC = ((tp*tn)-(fp*fn)) / sqrt((tp+fn)*(tp+fp)*(tn+fn)*(tn+fp))

Input

- *y* - classes (two classes +1 and -1) [1D numpy array integer]
- *p* - prediction (two classes +1 and -1) [1D numpy array integer]

Output

- MCC

CHAPTER

ELEVEN

FEATURE LIST ANALYSIS

11.1 Canberra Indicator

Canberra stability indicator on top-k positions [Jurman08]

canberra(*lists, k, dist=False, modules=None*)

Compute mean Canberra distance indicator on top-k sublists.

Input

- *lists* - [2D numpy array integer] position lists Positions must be in [0, #elems-1]
- *k* - [integer] top-k sublists
- *modules* - [list] modules (list of group indicies)
- *dist* - [bool] return partial distances (True or False)

Output

- *cd* - [float] canberra distance
- *i1* - [1D numpy array integer] index 1 (if dist == True)
- *i2* - [1D numpy array integer] index 2 (if dist == True)
- *pd* - [1D numpy array float] partial distances for index1 and index2 (if dist == True)

```
>>> from numpy import *
>>> from mlpy import *
>>> lists = array([[2,4,1,3,0],  # first positions list
...                [3,4,1,2,0],  # second positions list
...                [2,4,3,0,1],  # third positions list
...                [0,1,4,2,3]]) # fourth positions list
>>> canberra(lists, 3)
1.0861983059292479
```

canberraq(*lists, complete=True, normalize=False, dist=False*)

Compute mean Canberra distance indicator on generic lists.

Input

- *lists* - [2D numpy array integer] position lists Positions must be in [-1, #elems-1], where -1 indicates features not present in the list
- *complete* - [bool] complete
- *normalize* - [bool] normalize

•*dist* - [bool] return partial distances (True or False)

Output

•*cd* - [float] canberra distance

•*i1* - [1D numpy array integer] index 1 (if dist == True)

•*i2* - [1D numpy array integer] index 2 (if dist == True)

•*pd* - [1D numpy array float] partial distances for index1 and index2 (if dist == True)

```
>>> from numpy import *
>>> from mlpy import *
>>> lists = array([[2,-1,1,-1,0],  # first positions list
...                [3,4,1,2,0],    # second positions list
...                [2,-1,3,0,1],   # third positions list
...                [0,1,4,2,3]])   # fourth positions list
>>> canberraq(lists)
1.0628570368721744
```

normalizer(*lists*)

Compute the average length of the partial lists (nm) and the corresponding normalizing factor (nf) given by 1 - a / b where a is the exact value computed on the average length and b is the exact value computed on the whole set of features.

Inputs

•*lists* - [2D numpy array integer] position lists Positions must be in [-1, #elems-1], where -1 indicates features not present in the list

Output

•*(nm, nf)* - (float, float)

11.2 Borda Count, Extraction Indicator, Mean Position Indicator

Borda Count [Borda1781]

borda(*lists, k, modules=None*)

Compute the number of extractions on top-k sublists and the mean position on lists for each element. Sort the element ids with decreasing number of extractions, AND element ids with equal number of extractions should be sorted with increasing mean positions.

Input

•*lists* - [2D numpy array integer] ranked feature-id lists. Feature-id must be in [0, #elems-1].

•*k* - [integer] on top-k sublists

•*modules* - [list] modules (list of group indicies)

Output

•*borda* - (feature-id, number of extractions, mean positions)

Example:

```
>>> from numpy import *
>>> from mlpy import *
>>> lists = array([[2,4,1,3,0],  # first ranked feature-id list
```

```
...                 [3,4,1,2,0],  # second ranked feature-id list
...                 [2,4,3,0,1],  # third ranked feature-id list
...                 [0,1,4,2,3]]) # fourth ranked feature-id list
>>> borda(lists, 3)
(array([4, 1, 2, 3, 0]), array([4, 3, 2, 2, 1]), array([ 1.25      ,  1.66666667,  0.        ,
```

- Element 4 is in the first position with 4 extractions and mean position 1.25.
- Element 1 is in the first position with 3 extractions and mean position 1.67.
- Element 2 is in the first position with 2 extractions and mean position 0.00.
- Element 3 is in the first position with 2 extractions and mean position 1.00.
- Element 0 is in the first position with 1 extractions and mean position 0.00.

CHAPTER

TWELVE

DATA MANAGEMENT

12.1 Importing and exporting data

data_fromfile (*file, ytype=<type 'int'>*)

Read data file in the form:

```
x11 [TAB] x12 [TAB] ... x1n [TAB] y1
x21 [TAB] x22 [TAB] ... x2n [TAB] y2
 .         .         .    .         .
 .         .          .   .         .
 .         .           .  .         .
xm1 [TAB] xm2 [TAB] ... xmn [TAB] ym
```

where xij are float and yi are of type 'ytype' (numpy.int or numpy.float).

Input

- *file* - data file name
- *ytype* - numpy datatype for labels (numpy.int or numpy.float)

Output

- *x* - data [2D numpy array float]
- *y* - classes [1D numpy array int or float]

Example:

```
>>> from numpy import *
>>> from mlpy import *
>>> x, y = data_fromfile('data_example.dat')
>>> x
array([[ 1.1,  2. ,  5.3,  3.1],
...    [ 3.7,  1.4,  2.3,  4.5],
...    [ 1.4,  5.4,  3.1,  1.4]])
>>> y
array([ 1, -1,  1])
```

data_fromfile_wl (*file*)

Read data file in the form:

```
x11 [TAB] x12 [TAB] ... x1n [TAB]
x21 [TAB] x22 [TAB] ... x2n [TAB]
 .         .         .    .
```

```
 .          .          .    .
 .          .          .    .
xm1 [TAB] xm2 [TAB] ... xmn [TAB]
```

where xij are float.

Input

•*file* - data file name

Output

•*x* - data [2D numpy array float]

Example:

```
>>> from numpy import *
>>> from mlpy import *
>>> x, y = data_fromfile('data_example.dat')
>>> x
array([[ 1.1,  2. ,  5.3,  3.1],
...    [ 3.7,  1.4,  2.3,  4.5],
...    [ 1.4,  5.4,  3.1,  1.4]])
```

data_tofile (*file, x, y, sep='t'*)
Write data file in the form:

```
x11 [sep] x12 [sep] ... x1n [sep] y1
x21 [sep] x22 [sep] ... x2n [sep] y2
 .          .          .    .         .
 .          .          .    .         .
 .          .          .    .         .
xm1 [sep] xm2 [sep] ... xmn [sep] ym
```

where xij are float and yi are integer.

Input

•*file* - data file name

•*x* - data [2D numpy array float]

•*y* - classes [1D numpy array integer]

•*sep* - separator

data_tofile_wl (*file, x, sep='t'*)
Write data file in the form:

```
x11 [sep] x12 [sep] ... x1n [sep]
x21 [sep] x22 [sep] ... x2n [sep]
 .          .          .    .
 .          .          .    .
 .          .          .    .
xm1 [sep] xm2 [sep] ... xmn [sep]
```

where xij are float.

Input

•*file* - data file name

•*x* - data [2D numpy array float]

•*sep* - separator

12.2 Normalization and Standardization

data_normalize(*x*)

Normalize numpy array (2D) x.

Input

•*x* - data [2D numpy array float]

Output

•normalized data

Example:

```
>>> from numpy import *
>>> from mlpy import *
>>> x = array([[ 1.1,  2. ,  5.3,  3.1],
...            [ 3.7,  1.4,  2.3,  4.5],
...            [ 1.4,  5.4,  3.1,  1.4]])
>>> data_normalize(x)
array([[-0.9797065 , -0.48295391,  1.33847226,  0.12418815],
...    [ 0.52197912, -1.13395464, -0.48598056,  1.09795608],
...    [-0.75217354,  1.35919078,  0.1451563 , -0.75217354]])
```

data_standardize(*x*, *p=None*)

Standardize numpy array (2D) x and optionally standardize p using mean and std of x.

Input

•*x* - data [2D numpy array float]

•*p* - optional data [2D numpy array float]

Output

•standardized data

Example:

```
>>> from numpy import *
>>> from mlpy import *
>>> x = array([[ 1.1,  2. ,  5.3,  3.1],
...            [ 3.7,  1.4,  2.3,  4.5],
...            [ 1.4,  5.4,  3.1,  1.4]])
>>> data_standardize(x)
array([[-0.67958381, -0.43266792,  1.1157668 ,  0.06441566],
...    [ 1.1482623 , -0.71081158, -0.81536804,  0.96623494],
...    [-0.46867849,  1.1434795 , -0.30039875, -1.0306506 ]])
```

CHAPTER

THIRTEEN

MISCELLANEOUS

13.1 Confidence Interval

percentile_ci_median(*x, nboot=1000, alpha=0.025000000000000001, rseed=0*)

Percentile confidence interval for the median of a sample x and unknown distribution.

Input

- *x* - [1D numpy array] sample
- *nboot* - [integer] (>1) number of resamples
- *alpha* - [float] confidence level is 100*(1-2*alpha) (0.0<alpha<1.0)
- *rseed* - [integer] random seed

Output

- *ci* - (cimin, cimax) confidence interval

Example:

```
>>> from numpy import *
>>> from mlpy import *
>>> x = array([1,2,4,3,2,2,1,1,2,3,4,3,2])
>>> percentile_ci_median(x, nboot = 100)
(1.8461538461538463, 2.8461538461538463)
```

13.2 Peaks Detection

span_pd(*x, span*)

span peaks detection.

Input

- *x* - [1D numpy array float] data
- *span* - [odd int] span

Output

- *idx* - [1D numpy array integer] peaks indexes

New in version 2.0.7.

13.3 Functions from GSL

gamma (*x*)
Gamma Function.

Input

- *x* - [float] data

Output

- *gx* - [float] gamma(x)

fact (*x*)
Factorial x!. The factorial is related to the gamma function by x! = gamma(x+1)

Input

- *x* - [int] data

Output

- *fx* - [float] factorial x!

quantile (*x, f*)
Quantile value of sorted data. The elements of the array must be in ascending numerical order. The quantile is determined by the f, a fraction between 0 and 1. The quantile is found by interpolation, using the formula: quantile = (1 - delta) x_i + delta x_{i+1} where i is floor((n - 1)f) and delta is (n-1)f - i.

Input

- *x* - [1D numpy array float] sorted data
- *f* - [float] fraction between 0 and 1

Output

- *q* - [float] quantile

cdf_gaussian_P (*x, sigma*)
Cumulative Distribution Functions (CDF) P(x) for the Gaussian distribution.

Input

- *x* - [float] data
- *sigma* - [float] standard deviation

Output

- *p* - [float]

New in version 2.0.2.

13.4 Other

away (*a, b, d*)
Given numpy 1D array *a* and numpy 1D array *b* compute *c* = { bi : | bi - aj | > d for each i, j}

Input

- *a* - [1D numpy array float]
- *b* - [1D numpy array float]

- *d* - [double]

Output

- *c* - [1D numpy array float]

New in version 2.0.3.

is_power(*n, b*)
Return True if 'n' is power of 'b', False otherwise. New in version 2.0.6.

next_power(*n, b*)
Returns the smallest integer, greater than or equal to 'n' which can be obtained as power of 'b'. New in version 2.0.6.

CHAPTER

FOURTEEN

TOOLS

14.1 Landscaping and Parameter Tuning

`mlpy` includes executable scripts to be used off-the-shelf for landscaping and parameter tuning tasks. The classification and optionally feature ranking operations are organized in a sampling procedure (k-fold or Monte Carlo cross validation).

- **svm-landscape**: landscaping and regularization parameter (*C*) tuning
- **fda-landscape**: landscaping and regularization parameter (*C*) tuning
- **srda-landscape**: landscaping and alpha parameter (*alpha*) tuning
- **pda-landscape**: landscaping and number of regressions parameter (*Nreg*) tuning
- **dlda-landscape**
- **nn-landscape**: landscaping

Error (`mlpy.err()`), Matthews Correlation Coefficient (`mlpy.mcc()`) and optionally Canberra Distance (`mlpy.canberra()`) are retrieved at each parameter step.

`mlpy` includes executable scripts to be used exclusively for parameter tuning tasks:

- **irelief-sigma**: kernel width parameter (*sigma*) tuning

In order to print help message:

```
$ command --help
```

14.2 Other Tools

borda

Compute Borda Count, Extraction Indicator, Mean Position Indicator from a text file containing feature lists.

canberra

Compute mean Canberra distance indicator on top-k sublists from a text file containing feature lists and one containing the top-k positions.

In order to print help message:

```
$ command --help
```

14.2.1 The Feature Lists File

The feature lists file is a plain text TAB-separated file where each row is a feature ranking (a feature list).

Example:

```
feat6 [TAB] feat2 [TAB] ... [TAB] feat1
feat4 [TAB] feat1 [TAB] ... [TAB] feat7
feat4 [TAB] feat9 [TAB] ... [TAB] feat3
feat2 [TAB] feat3 [TAB] ... [TAB] feat9
feat8 [TAB] feat4 [TAB] ... [TAB] feat2
```

Index

BIBLIOGRAPHY

[Torrence98] C Torrence and G P Compo. Practical Guide to Wavelet Analysis

[Gsldwt] Gnu Scientific Library, http://www.gnu.org/software/gsl/

[Senin08] Pavel Senin. Dynamic Time Warping Algorithm Review

[Keogh01] Eamonn J. Keogh and Michael J. Pazzani. Derivative Dynamic Time Warping. First SIAM International Conference on Data Mining (SDM 2001), 2001.

[Sakoe78] Hiroaki Sakoe and Seibi Chiba. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing. Volume 26, 1978.

[Amap] amap: Another Multidimensional Analysis Package, http://cran.r-project.org/web/packages/amap/index.html

[Vapnik95] V Vapnik. The Nature of Statistical Learning Theory. Springer-Verlag, 1995.

[Cristianini] N Cristianini and J Shawe-Taylor. An introduction to support vector machines. Cambridge University Press.

[Merler06] S Merler and G Jurman. Terminated Ramp - Support Vector Machine: a nonparametric data dependent kernel. Neural Network, 19:1597-1611, 2006.

[Nasr09] 1. Nasr, S. Swamidass, and P. Baldi. Large scale study of multiplemolecule queries. Journal of Cheminformatics, vol. 1, no. 1, p. 7, 2009.

[Mika01] S Mika and A Smola and B Scholkopf. An improved training algorithm for kernel fisher discriminants. Proceedings AISTATS 2001, 2001.

[Cristianini02] N Cristianini, J Shawe-Taylor and A Elisseeff. On Kernel-Target Alignment. Advances in Neural Information Processing Systems, Volume 14, 2002.

[Cai08] D Cai, X He, J Han. SRDA: An Efficient Algorithm for Large-Scale Discriminant Analysis. Knowledge and Data Engineering, IEEE Transactions on Volume 20, Issue 1, Jan. 2008 Page(s):1 - 12.

[Ghosh03] D Ghosh. Penalized discriminant methods for the classification of tumors from gene expression data. Biometrics on Volume 59, Dec. 2003 Page(s):992 - 1000(9).

[Sun07] Yijun Sun. Iterative RELIEF for Feature Weighting: Algorithms, Theories, and Applications. IEEE Trans. Pattern Anal. Mach. Intell. 29(6): 1035-1051, 2007.

[Sun08] Yijun Sun, S. Todorovic, and S. Goodison. A Feature Selection Algorithm Capable of Handling Extremely Large Data Dimensionality. In Proc. 8th SIAM International Conference on Data Mining (SDM08), pp. 530-540, April 2008.

[Subramani06] P Subramani, R Sahu and S Verma. Feature selection using Haar wavelet power spectrum. In BMC Bioinformatics 2006, 7:432.

[Guyon02] Isabelle Guyon, Jason Weston, Stephen Barnhill, Vladimir Vapnik. Gene Selection for Cancer Classification using Support Vector Machines, Machine Learning, v.46 n.1-3, p.389-422, 2002.

[Furlanello03] C Furlanello, M Serafini, S Merler, and G Jurman. Advances in Neural Network Research: IJCNN 2003, chapter An accelerated procedure for recursive feature ranking on microarray data. Elsevier, 2003.

[Louw06] N Louw and S J Steel. Variable selection in kernel Fisher discriminant analysis by means of recursive feature elimination. Computational Statistics & Data Analysis, Volume 51 Issue 3 Pages 2043-2055, 2006.

[Jurman08] G Jurman, S Merler, A Barla, S Paoli, A Galea, and C Furlanello. Algebraic stability indicators for ranked lists in molecular profiling. Bioinformatics, 24(2):258-264, 2008.

[Borda1781] J C Borda. Mémoire sur les élections au scrutin. Histoire de l'Académie Royale des Sciences, 1781.

INDEX

N

P

Q

R

S

U

W