GURLS guide to `train` and `test`

September 23, 2014

A very common machine learning task consists in training a model on a given dataset and then to test it. Here the two steps are made simple:

```
model = train(Xtrain, ytrain);
ypredicted = test(Xtest);
```

In particular the functions `train, test` solve different kinds of supervised learning problems: regression, classification, multiclass classification. The functions automatically recognize the kind of problem, find the most suitable model and learning algorithm for the given dataset and select the appropriate metaparameters, if not specified in the options.

1 train

The function train has different signatures

```
model = train(Xtrain, ytrain);
model = train(Xtrain, ytrain, options);
model = train(Xtrain, ytrain, 'optname1', optval1, 'optname2', optvalue2, ...);
```

`Xtrain` can be a $n \times d$ matrix of the form `[x1; x2;... xn]` where `xi` are the n inputs of d dimensions. Otherwise it can be a precomputed Gram matrix (thus a $n \times n$ positive semidefinite matrix).

`ytrain` is a $n \times T$ matrix where the i-th row is the vector of labels associated to the i-th sample point `xi` in `Xtrain`. `ytrain` is a matrix of real numbers in the regression tasks, while a matrix of $-1, +1$ in the classification or multiclassification ones, where the `ytrain(i,j)` is 1 if `xi` belongs to the class j otherwise it is -1. This notation is general and covers the case of overlapping classes. Note that for the multiclass case another allowed format for `ytrain` is a $n \times 1$ vectors of numbers in $\{1, 2, .., T\}$, where `ytrain(i)=j` if `xi` belongs to the class j.

`options` is a list of options and is defined as follows

```
options = struct('optname1', optvalue1, 'optname2', optvale2, ...)
```

if `options` is defined in this way the second and the third signatures are equivalent.

`'optname1', optval1,...` these are options with the associated values that are passed to the train function. Note that all the options specified here will be passed to the underlying modules of gurls. The following is a list of options with the possible values.

- `'datatype'`: (automatically deduced, if not specified) can be `'vector'` when the input type is a matrix with the format specified for `Xtrain`, or `'kernel'` when `Xtrain` is a Gram matrix.

- `'problem'`: (automatically deduced, if not specified) can be `'regression'` or `'classification'`, see the definition of `'ytrain'`
- `'algorithm'`: it specifies the algorithm to use for solving the problem, it can be
  - `'lrls'`: (Only when the datatype is vector) it is linear regularized least squares,
  - `'krls'`: (default) it is kernel regularized least squares,
  - `'krlsrf'`: it is kernel regularized least squares with random features,
  - `'gpr'`: it is gaussian processes regression.
- `'filter'`: (automatically deduced, if not specified) it specifies the filter used by the algorithm, it can be
  - `'tikh'`: (default) tikhonov filter,
  - `'land'`: landweber iterative filter. To use this, specify the parameter `regrange` with the range of metaparameters to test (e.g. `1:100`),
  - `'nu'`: nu-method iterative filter. To use this, specify the parameter `regrange` with the range of metaparameters to test (e.g. `1:100`),
  - `'conjgrad'`: conjugate gradient filter,
  - `'randtikh'`: randomized tikhonov filter.
- `'kernelfun'`: it specifies the kernel function used by the algorithm. It applies for all the algorithms except for lrls that is linear by default. Possible values are
  - `'linear'`: linear kernel,
  - `'rbf'`: (default) gaussian kernel,
  - `'quasiperiodic'`: quasiperiodic kernel,
  - `'chisquared'`: chi-square kernel,
  - `'datatype'`: used when `'Xtrain'` is a precomputed Gram matrix.
- `'partuning'`: specifies the cross validation approach to be used for automatically selecting the metaparameters. It can be `'loo'` for leave one out, or `'ho'` for hold-out. The default is `ho`
- `'perfm'`: measure of the performances of the model. It can be
  - `'rmse'`: root mean square error (default when the problem is of regression type),
  - `'macroavg'`: macro average (default when the problem is of classification type),
  - `'precrec'`: precision recall,
  - `'gpregr'`: specific performance measure for the gaussian process regression (mandatory when the algorithm is gpr).
- `'pars'`: (automatically deduced, if not specified) it tells the system which metaparameters must be selected.
  - `'none'`: both the regularization and the kernel metaparameters are specified in the options,
  - `'reg'`: the regularization metaparameter has to be found by cross validation,
  - `'ker'`: the kernel metaparameter has to be found by cross validation,
  - `'all'`: both the regularization and the kernel metaparameter has to be found by cross validation.

- `'regrange'`: the range where the automatic cross validation will search for the best regularization parameter. It must be specified for the nu and land filters, otherwise it is automatically computed by the system.
- `'kerrange'`: the range where the automatic cross validation will search for the best kernel parameter. It is automatically computed by the system, if not specified.
- `'regpar'`: specific value for the regularization metaparameter.
- `'kerpar'`: specific value for the kernel metaparameter.

2 test

```
ypredicted = test(Xtest)
[ypredicted, accuracy] = test(Xtest, ytest)
[ypredicted, accuracy] = test(Xtest, ytest, perfmeas)
```

The first signature computes the predicted labels `ypredicted` associated to the test dataset. The second computes the predicted labels and their accuracy with respect to the test label `ytest`. In this case the accuracy is computed by the performance measure used in the training phase. In order to compute it with a different performance measure, use the third signature. Note that the format of `Xtest` and of `ytest` must be the same of `Xtrain` and `ytrain`.