User Guide of `Examiner.jar`

Kilho Shin
University of Hyogo

June 30, 2011

1 Example

You can run the command

```
java -cp Examiner.jar Examiner -svm c:\util\libsvm-3.0\windows
\ -f sample -sf -ratio 4 -round 5 -fold 10 -level 3 -p 4
```

in a directory that includes `Examiner.jar` and `sample.kernel`.

`sample.kernel` is assumed to include at least one Gram matrix of the tree kernel of `SF` type (known as `EF` type in [Shin et al. 2011]) for some dataset of trees. The tree kernel of `SF` type is defined by

$$K(X,Y) = \sum_{(X',Y')\in M^{\mathsf{SF}}_{X,Y}} \lambda^{|X'|},$$

where $M^{\mathsf{SF}}_{X,Y}$ is the entire set of isomorphic pairs of substructures of trees X and Y and λ denotes a variable that represents the decay factor. Hence, X' and Y' are substructures of X and Y, and are isomorphic with each other. An element of a Gram matrix specified in `sample.kernel` is a polynomial in λ.

The program `Examiner.jar` performs the following, and outputs the results to a file with extension `.result`

1. The program first gentrates 5 (specified by `-round`) pairs of a training dataset and a test dataset by dividing the tree dataset at random. The ratio of the size of the training dataset is 4 times as large as that of the test dataset as specified by `-ratio`.
2. For each pair of training and test datasets, the program executes the following.
   (a) The program performs a grid search, and finds *optimal* values for the decay factor λ and the regulation parameter C of C-SVM that shows the maximum Cross Validation Rate on the training dataset. As the Cross Validation Rate, AUC of ROC Curve is used, and the number of folds for each cross validation is 10 as specified by `-fold`.

The broadness of the grid search is determined by the option `-level`. The greater the value for the option is, the broader is the search space of the grid search.

   (b) The program trains the libSVM classifier with the training dataset with the obtained optimal parameters λ and C, and creates a model (hypothesis).

   (c) The program applies the obtained model to the relevant test dataset, and computes values for the four measures of Accuracy, Ballanced Accuracy, F-Score and AUC of ROC Curve.

3. The program finally outputs the ten quadruples of the measurements (Accuracy, Ballanced Accuracy, F-Score and AUC) for the ten pairs of training and test datasets.

2 Options

The following are a part of the options that the program supports.

-f To specify the name of `.kernel` without extension, which includes the Gram matrices for one or more tree kernels. The program automatically adds `.kernel` as an extension.

-svm A path to `svm-train.exe` and `svm-predict`. When specified, the last character must be \.

-norm When specified, the normalized kernel values are used for evaluation. Hence, for the value $K(X, Y)$ obtained from a Gram matrix in the specified `.kernel` file,

$$\bar{K}(X, Y) = \frac{K(X, Y)}{\sqrt{K(X, X)K(Y, Y)}}$$

is used for training and testing.

-D m.mm,M.MM Minimum (m.mm) and maximum (M.MM) of the search range for decay factor. When left out, `-D 1.0,10.0` is applied.

-logC m.mm,M.MM Minimum (m.mm) and maximum (M.MM) of the search range for log2 C. When left out, `-logC -3.0,5.0` is applied.

-p Prallelism for grid search. Default `-p 2`.

-level Number of iteration for grid search. Default `-level 2`.

The kernel type is determined by the following options. For definition of the types, please refer to [Shin et al. 2011].

The program supports more kernel types such as the tree kernel derived from Taï tree edit distance.

More than one options can be specified simultaneously.

-sf `SF` type, known as `EF` type in [Shin et al. 2011].

-st `ST` type, known as `ET` type in [Shin et al. 2011].

-sp `SP` type, known as `EP` type in [Shin et al. 2011].

-at `AT` type.

-cf `CF` type.

-ct `CT` type.

-cp `CP` type.

-cot `CoT` type.

3 Reference

[Shin et al. 2011] K. Shin, M. Cuturi and T. Kuboyama, "Mapping kernels for trees", ICML 2011