Example Session for Supervised Classification
Andreas Borg, Murat Sariyar
January 16, 2015

This document shows an example session for using supervised classification in the package RecordLinkage for deduplication of a single data set. Conducting linkage of two data sets differs only in the step of generating record pairs.

See also the vignette on Fellegi-Sunter deduplication for some general information on using the package.

1 Generating comparison patterns

In this session, a training set with 50 matches and 250 non-matches is generated from the included data set RLData10000. Record pairs from the set RLData500 are used to calibrate and subsequently evaluate the classifiers.

```r
> data(RLdata500)
> data(RLdata10000)
> train_pairs=compare.dedup(RLdata10000, identity=identity.RLdata10000,
+   n_match=500, n_non_match=500)
> eval_pairs=compare.dedup(RLdata500, identity=identity.RLdata500)
```

2 Training

trainSupv handles calibration of supervised classifiers which are selected through the argument method. In the following, a single decision tree (rpart), a bootstrap aggregation of decision trees (bagging) and a support vector machine are calibrated (svm).

```r
> model_rpart=trainSupv(train_pairs, method="rpart")
> model_bagging=trainSupv(train_pairs, method="bagging")
> model_svm=trainSupv(train_pairs, method="svm")
```

3 Classification

classifySupv handles classification for all supervised classifiers, taking as arguments the structure returned by trainSupv which contains the classification model and the set of record pairs which to classify.

```r
> result_rpart=classifySupv(model_rpart, eval_pairs)
> result_bagging=classifySupv(model_bagging, eval_pairs)
> result_svm=classifySupv(model_svm, eval_pairs)
```
4 Results

4.1 Rpart
alpha error 0.020000
beta error 0.012638
accuracy 0.987359

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>P</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>FALSE</td>
<td>123124</td>
<td>0</td>
<td>1576</td>
</tr>
<tr>
<td>TRUE</td>
<td>1</td>
<td>0</td>
<td>49</td>
</tr>
</tbody>
</table>

4.2 Bagging
alpha error 0.000000
beta error 0.002574
accuracy 0.997427

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>P</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>FALSE</td>
<td>124379</td>
<td>0</td>
<td>321</td>
</tr>
<tr>
<td>TRUE</td>
<td>0</td>
<td>0</td>
<td>50</td>
</tr>
</tbody>
</table>

4.3 SVM
alpha error 0.000000
beta error 0.003496
accuracy 0.996505

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>P</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>FALSE</td>
<td>124264</td>
<td>0</td>
<td>436</td>
</tr>
<tr>
<td>TRUE</td>
<td>0</td>
<td>0</td>
<td>50</td>
</tr>
</tbody>
</table>