Package ‘medSTC’

February 20, 2015

Type Package

Title A max-margin supervised Sparse Topical Coding Model

Version 1.0.0

Date 2013-01-15

Author Jun Zhu <junzhu@cs.cmu.edu>, Aykut FIRAT <aykut@mit.edu>

Maintainer Aykut FIRAT <aykut@crimsonhexagon.com>

Description This is a C++ implementation of Sparse Topical Coding (STC), a model of discrete data which is fully described in Zhu et al. (2011) (http://www.cs.cmu.edu/~junzhu/stc/stc.pdf). It can be used for multi-class classification and describing documents with underlying sparse topics.

Depends R (>= 2.10)

License file LICENSE

LazyLoad yes

Repository CRAN

Date/Publication 2013-01-24 08:24:00

NeedsCompilation yes

License_restricts_use yes

R topics documented:

medSTC-package ......................................................... 2
medSTC ........................................................................ 3
newsgroups ................................................................. 4
predict.medSTC ........................................................... 5

Index 7
Description

This package employs sparse topical coding models for multi-class classification developed by Jun Zhu and Eric P. Xing. The package uses a fast C implementation of SVM, SVMlight (http://svmlight.joachims.org/), developed by Thorsten Joachims <thorsten@joachims.org>.

Details

<table>
<thead>
<tr>
<th>Package</th>
<th>medSTC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>Package</td>
</tr>
<tr>
<td>Version</td>
<td>1.0.0</td>
</tr>
<tr>
<td>Date</td>
<td>2013-1-15</td>
</tr>
<tr>
<td>License</td>
<td>GPL</td>
</tr>
<tr>
<td>LazyLoad</td>
<td>yes</td>
</tr>
</tbody>
</table>

Author(s)

Jun Zhu (<junzhu@cs.cmu.edu>), Aykut Firat (<aykut@crimsonhexagon.com>)

References


See Also

Functions to fit models: medSTC
Functions for prediction: predict.medSTC
Included data sets: newsgroups

Examples

```r
## See the demo for the newsgroup example:
## Not run: demo(medSTC)
```
medSTC

A max-margin Sparse Topical Coding model (Med-STC) for multi-class classification

Description

MedSTC is a novel classification algorithm by Prof. Jun Zhu (http://www.ml-thu.net/~jun/).

Usage

medSTC(documents, mlabels, ntopics, initial_c=0.5, lambda=1, rho=0.01, delta_ell=3600, supervised=TRUE, primal_svm=1, var_max_iter=20, convergence=1e-4, em_max_iter=100, em_convergence=1e-4, svm_alg_type=2, output_dir=".")

Arguments

- **documents**: A list whose length is equal to the number of documents, D. Each element of documents is an integer matrix with two rows. Each column of documents[i] (i.e., document i) represents a word occurring in the document. documents[i][1, j] is a 0-indexed word identifier for the jth word in document i. documents[i][2, j] is an integer specifying the number of times that word appears in the document.
- **mlabels**: The training labels for the documents.
- **ntopics**: Number of topics to be used in modeling the corpus.
- **initial_c, lambda, rho**: These are positive-valued regularization constants. Default values are initial_c=0.5, lambda=0.1, rho=0.01.
- **delta_ell**: The parameter for the svm cost function, i.e., 0/(delta ell) loss. Only positive values are allowed. Default value is 3600.
- **supervised**: If the value is TRUE, the model is a supervised MedSTC; if FALSE, the model is the unsupervised STC.
- **primal_svm**: Only works when "supervised" is set at 1. If the value is 1, uses the loss-augmented prediction (i.e., sub-gradient) to update document codes; otherwise it uses the gradient with Lagrangian multipliers to update document codes.
- **var_max_iter**: The maximum number of iterations of coordinate descent for a single document.
- **convergence**: The convergence criteria for coordinate descent. Stop if (objective_old - objective) / abs(objective_old) is less than this value (or after the maximum number of iterations). Note that "objective" is the objective value for a single document.
- **em_max_iter**: The maximum number of iterations of hierarchical sparse coding, dictionary learning, and svm training (for supervised MedSTC).
- **em_convergence**: The convergence criteria for coordinate descent. Stop if (objective_old - objective) / abs(objective_old) is less than this value (or after the maximum number of iterations). Note that "objective" is the objective value for the whole corpus.
newsgroups

svm_alg_type
If set to 0 then the n-slack multi-class SVM is used. If set to 2, then the 1-slack multi-class SVM is used. In our testing, the 1-slack SVM is faster.

output_dir
A directory for writing intermediate results. Directory is removed after the calculation is done, but is needed during the run.

Value

model
A model object of the medSTC class, which has a state list with five elements: The first two list elements are for storing the model parameter state after the model completed training. The third list element is the LogProbabilityOfWordsForTopics, which can be used for word assignments to topics. The fourth and fifth model state list elements are Eta and Mu. (refer to paper) The model also stores the original parameter values.

Author(s)

Jun Zhu (<junzhu@cs.cmu.edu>), Aykut Firat (<aykutfirat@gmail.com>)

References


Examples

## Not run: demo(medSTC)

---

newsgroups

A shortened collection of newsgroup messages with the first 3 classes.

Description

The 20 Newsgroups data set is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. We use in this package only its first 3 classes for demonstration purposes.

Usage

data(newsgroup.train.documents)
data(newsgroup.test.documents)
data(newsgroup.train.labels)
data(newsgroup.test.labels)
data(newsgroup.vocab)
**Predict.medSTC**

**Format**

newsgroup.train.documents and newsgroup.test.documents comprise a corpus of 2731 newsgroup documents partitioned into 1633 training and 1098 test cases evenly distributed across 3 classes.

newsgroup.train.labels is a numeric vector of length 1633 which gives a class label from 1 to 3 for each training document in the corpus.

newsgroup.test.labels is a numeric vector of length 1098 which gives a class label from 1 to 3 for each test document in the corpus.

newsgroup.vocab is the vocabulary of the corpus.

stopwords English stopwords extracted from the tm package.

**Source**


**Examples**

```r
data(newsgroup.train.documents)
data(newsgroup.test.documents)
data(newsgroup.train.labels)
data(newsgroup.test.labels)
data(newsgroup.vocab)
data(stopwords)
```

**predict.medSTC**  
*Prediction for max-margin Sparse Topical Coding model (Med-STC) for multi-class classification*

**Description**

MedSTC is a novel classification algorithm by Prof. Jun Zhu ([http://www.ml-thu.net/~jun/](http://www.ml-thu.net/~jun/)). This function predicts test class labels using a medSTC model.

**Usage**

```r
## S3 method for class 'medSTC'
predict(object, documents,...)
```

**Arguments**

- `object`  
  A model object of the medSTC class, which is obtained after running `medSTC`.
documents A list whose length is equal to the number of documents, \( D \). Each element of \( \text{documents} \) is an integer matrix with two rows. Each column of \( \text{documents}[i] \) (i.e., document \( i \)) represents a word occurring in the document. \( \text{documents}[i][1, j] \) is a 0-indexed word identifier for the \( j \)th word in document \( i \). \( \text{documents}[i][2,j] \) is an integer specifying the number of times that word appears in the document.

... Currently not used.

Value

scores Score values for each label for each document.
assignments Predicted class labels for each document.

Author(s)

Jun Zhu (<junzhu@cs.cmu.edu>), Aykut Firat (<aykut@crimsonhexagon.com>)

References


Examples

```r
## Not run: demo(medSTC)
```
Index

*Topic datasets
newsgroups, 4

*Topic package
medSTC-package, 2

*Topic sparse topical coding models
medSTC, 3
predict.medSTC, 5

medSTC, 2, 3
medSTC-package, 2

newsgroup.test.documents (newsgroups), 4
newsgroup.test.labels (newsgroups), 4
newsgroup.train.documents (newsgroups), 4
newsgroup.train.labels (newsgroups), 4
newsgroup.vocab (newsgroups), 4
newsgroups, 2, 4

predict.medSTC, 2, 5

stopwords (newsgroups), 4