Machine Learning with R at LRZ: Introduction to mlr

Spam E-mail Database

Description

A data set collected at Hewlett-Packard Labs, that classifies 4601 e-mails as spam or non-spam. In addition to this class label there are 57 variables indicating the frequency of certain words and characters in the e-mail.

Format

A data frame with 4601 observations and 58 variables.

The first 48 variables contain the frequency of the variable name (e.g., business) in the e-mail. If the variable name starts with num (e.g., num650) the it indicates the frequency of the corresponding number (e.g., 650). The variables 49-54 indicate the frequency of the characters ';', '(', '[', '!', '\$', and '\#'. The variables 55-57 contain the average, longest and total run-length of capital letters. Variable 58 indicates the type of the mail and is either `"nonspam"` or `"spam"`, i.e. unsolicited commercial e-mail.

Details

The data set contains 2788 e-mails classified as `"nonspam"` and 1813 classified as `"spam"`.

The "spam" concept is diverse: advertisements for products/web sites, make money fast schemes, chain letters, pornography... This collection of spam e-mails came from the collectors' postmaster and individuals who had filed spam. The collection of non-spam e-mails came from filed work and personal e-mails, and hence the word 'george' and the area code '650' are indicators of non-spam. These are useful when constructing a personalized spam filter. One would either have to blind such non-spam indicators or get a very wide collection of non-spam to generate a general purpose spam filter.

Source

- Creators: Mark Hopkins, Erik Reeber, George Forman, Jaap Suermondt at Hewlett-Packard Labs, 1501 Page Mill Rd., Palo Alto, CA 94304
- Donor: George Forman (gforman at nospam hpl.hp.com) 650-857-7835

These data have been taken from the UCI Repository Of Machine Learning Databases at http://www.ics.uci.edu/~mlearn/MLRepository.html

References

T. Hastie, R. Tibshirani, J.H. Friedman. *The Elements of Statistical Learning.* Springer, 2001.

Exercise

a) Create a binary classifcation task from the spam data

```
library(mlr)
data(spam, package = "kernlab")

spam.task = ...
```

b) List all learners that could be trained on `spam.task`

c) Select a learner you like and create it. If you want to can change its hyperparameters

d) Create an index set of train and test indicies. The test set should have 1000 observations.

d*) Ensure that the fraction between `"spam"` and `"nonspam"` is the training and test set is the same as in the full dataset.

```
train.inds = ...
test.inds = ...
```

e) Train your model on the train subset of the spam data and predict on the test subset.

```
mod = ...
preds = ...
```

f) Evaluate the performance of your model based on accuracy and area under the curve.

```
perf = ...
```

g) Try to find a model with an AUC of at least 98%.

- Try different models
- Change hyperparameters
- Have a closer look at the feature and try to find transformations or combination of features that improve your model's performance