Data

KDD Cup 2009: Customer relationship prediction

Data Download

Training and test data matrices and practice target values

The large dataset archives are available since the onset of the challenge. The small dataset will be made available at the end of the fast challenge. Both training and test sets contain **50,000 examples**. The data are split similarly for the small and large versions, but the samples are ordered differently within the training and within the test sets. Both small and large datasets have numerical and categorical variables. For the large dataset, the first **14,740 variables are numerical** and the last **260 are categorical**. For the small dataset, the first **190 variables are numerical** and the last **40 are categorical**. Toy target values are available only for practice purpose. The prediction of the toy target values will not be part of the final evaluation.

Small version (230 var.):

- orange_small_train.data.zip (8.2 Mbytes)
- orange_small_test.data.zip (8.2 Mbytes)

Large version (15,000 var.):

- orange_large_train.data.chunk1.zip (52.7 Mbytes)
- orange_large_train.data.chunk2.zip (52.7 Mbytes)
- orange_large_train.data.chunk3.zip (52.6 Mbytes)
- orange_large_train.data.chunk4.zip (52.5 Mbytes)
- orange_large_train.data.chunk5.zip (52.6 Mbytes)

- orange_large_test.data.chunk1.zip (52.8 Mbytes)
- orange_large_test.data.chunk2.zip (52.5 Mbytes)
- orange_large_test.data.chunk3.zip (52.6 Mbytes)
- orange_large_test.data.chunk4.zip (52.6 Mbytes)
- orange_large_test.data.chunk5.zip (52.6 Mbytes)

Toy targets (large):

- orange_large_train_toy.labels

True task labels

Real binary targets (small):

- orange_small_train_appentency.labels

- orange_small_train_churn.labels
- orange_small_train_upselling.labels

Real binary targets (large):

- orange_large_train_appetency.labels
- orange_large_train_churn.labels
- orange_large_train_upselling.labels

Data Format

The datasets use a format similar as that of the text export format from relational databases:

- One header lines with the variables names
- One line per instance
- Separator tabulation between the values
- There are missing values (consecutive tabulations)

The large matrix results from appending the various chunks downloaded in their order number. The header line is present only in the first chunk.

The target values (.labels files) have one example per line in the same order as the corresponding data files. Note that churn, appetency, and up-selling are three separate binary classification problems. The target values are +1 or -1. We refer to examples having +1 (resp. -1) target values as positive (resp. negative) examples.

The Matlab matrices are numeric. When loaded, the data matrix is called X. The categorical variables are mapped to integers. Missing values are replaced by NaN for the original numeric variables while they are mapped to 0 for categorical variables.