kaggle   Search kaggle   🔍   **Competitions**   **Datasets**   **Kernels**   **Discussion**   **Jobs**

Cause-effect pairs

Given samples from a pair of variables A, B, find whether A is a cause of B.

$10,000 · 266 teams · 4 years ago

Overview   **Data**   Discussion   Leaderboard   Rules   Late Submission

---

Competition Data   Edit

⊞ basic_python_benchma...	**SUP1data_text.zip** 74.23 MB	⬇ Download
🗋 CEfinal_test_text.zi...		
🗋 CEfinal_train_text.z...		
🗋 CEfinal_valid_text.z...		
🗋 SUP1data_text.zip		
🗋 SUP2data_text.zip		
🗋 SUP3data_text.zip		

Data Description

This is the July 1, 2013 final data release.

The data provided on this page is in csv format, suitable to be read by:

- basic python benchmark
- Matlab sample code

Archived data and data in the split format (one pair per file) are also available.

The file CEfinal_basic_python_benchmark.csv provides a sample submission.

We released the final test data and an equivalent amount of training and validation data distributed similarly. **The test data is encrypted, the decryption key will be revealed at the end of the development phase.** The new validation set is replacing the old validation set on the leaderboard and all the scores are reset to 0.5, please re-submit results on the new validation set. The new data include pairs of variables generated in a similar way as those of SUP2data and pairs of real variables from various sources. The final data is different from the original training and validation data with respect to normalization and quantization of variables to address a problem of bias in the original data.

NEW: May-June 2013 supplementary data release:

We provide three additional training datasets artificially generated: SUP1data, SUP2data, and SUP3data. Those training datasets have normalized numerical variables and have balanced number of unique values across all classes. SUP1data includes ~6000 pairs of numerical variables. SUP2 includes ~6000 pairs of mixed variables (numerical, categorical, binary). SUP3 data includes 81 pairs of real cause-effect pairs and 81 control pairs A|B and A-B generated from the real pairs.

March 2013 data release: archived.

In CEfinal_xxx_text.zip, you will find the following files:

- CEdata_xx_pairs.csv: the actual data for the A, B pairs.
- CEdata_xx_publicinfo.csv: information on the type of variables "Numerical", "Binary" or "Categorical".
- CEdata_train_target.csv: The target values for training data. The second column contains trinary targets +1 for A->B, -1 for B->A, 0 otherwise.

Our Team   Terms   Privacy   Contact/Support