**Description for gina\_agnostic dataset**

Author: Isabelle Guyon

Source: Agnostic Learning vs. Prior Knowledge Challenge

Please cite: None

Dataset from the Agnostic Learning vs. Prior Knowledge Challenge (http://www.agnostic.inf.ethz.ch), which consisted of 5 different datasets (SYLVA, GINA, NOVA, HIVA, ADA). The purpose of the challenge was to check if the performance of domain-specific feature engineering (prior knowledge) can be met by algorithms that were trained on data without any domain-specific knowledge (agnostic). For the latter, the data was anonymised and preprocessed in a way that makes them uninterpretable.

Modified by TunedIT (converted to ARFF format)

**Topic**

The task of GINA is handwritten digit recognition. This is the agnostic version of a subset of the MNIST data set. We chose the problem of separating the odd numbers from even numbers. We use 2-digit numbers. Only the unit digit is informative for that task, therefore at least ½ of the features are distracters. This is a twoclass classification problem with sparse continuous input variables, in which each class is composed of several clusters. It is a problems with heterogeneous classes

**Source**

The data set was constructed from the MNIST data that is made available by Yann LeCun of the NEC Research Institute at http://yann.lecun.com/exdb/mnist/. The digits have been size-normalized and centered in a fixed-size image of dimension 28x28. Examples are shown in the documentation in chapter 3.

**Description**

To construct the “agnostic” dataset, we performed the following steps:

We removed the pixels that were 99% of the time white. This reduced the original feature set of 784 pixels to 485. The original resolution (256 gray levels) was kept.

In spite of the fact that the data are rather sparse (about 30% of the values are non-zero), we saved the data as a dense matrix because we found that it can be compressed better in this way (to 19 MB.)

The feature names are the (i,j) matrix coordinates of the pixels (in a 28x28 matrix).

We created 2 digit numbers by dividing the datasets into to parts and pairing the

. The task is to separate odd from even numbers. The digit of the tens being not digits at random informative, the features of that digit act as distracters.To construct the “prior” dataset, we went back to the original data and fetched the “informative” digit in its original representation. Therefore, this data representation consists in a vector of concatenating the lines of a 28x28 pixel map.

Data type: non-sparse

Number of features: 970

Number of examples and check-sums :

Pos\_ex Neg\_ex Tot\_ex Check\_sum

Train 1550 1603 3153 164947945.00

Valid 155 160 315 16688946.00

This dataset contains samples from both training and validation datasets