Data descriptions

In this appendix we provide a brief description of each dataset used in this study. When possible we have used the description of the data provided by the original authors.

2dplanes: This is an artificial data set described in Breiman et al. [1984], with variance 1 instead of 2. The 10 attributes are generated independently using:

$$P(X_1 = -1) = P(X_1 = 1) = 1/2$$
$$P(X_m = -1) = P(X_m = 0) = P(X_m = 1) = 1/3, m = 2, ..., 10$$

Obtain the value of the target variable Y using the rule:

$$\text{if } X_1 = 1 \text{ set } Y = 3 + 3X_2 + 2X_3 + X_4 + \sigma(0, 1)$$
$$\text{if } X_1 = -1 \text{ set } Y = -3 + 3X_5 + 2X_6 + X_7 + \sigma(0, 1)$$

abalone: The task is to predict the age of abalone from physical measurements. Features include the sex, dimensions, and weight.

ailerons: This data set addresses a control problem, namely flying a F16 aircraft. The attributes describe the status of the aeroplane, while the goal is to predict the control action on the ailerons of the aircraft.

airlines: The dataset consists of a large amount of records, containing flight arrival and departure details for all the commercial flights within the USA, in 2008. The task is to predict the delay of a flight based on attributes like distance traveled, plane age, and date/time features. Categorical features such as day of week were converted to a one-hot encoding.

bank32nh: A synthetically generated dataset from a simulation of how bank-customers choose their banks. Tasks are based on predicting the fraction of bank customers who leave the bank because of full queues.

calHousing: The task here is to predict the median house value based on features like median income and median age, collected from the 1990 California census.

cpu_act: A collection of computer systems activity measures. The task is to predict the portion of time that CPUs run in user mode, based on attributes such as number of reads/writes to the system, system calls, and page requests.

electricity_prices: From the ICON Challenge on Forecasting and Scheduling[1]. The task is to forecast the electricity prices for a time period, based on

[1] http://iconchallenge.insight-centre.org/challenge-energy

features like the national load, time, and weather. It has been pre-processed to impute missing values ($< 1\%$ of the data).

elevators: This data set is also obtained from the task of controlling a F16 aircraft, although the target variable and attributes are different from the ailerons domain. In this case the goal variable is related to an action taken on the elevators of the aircraft.

friedman: This is an artificial data set used in Friedman [1991]. The cases are generated using the following method: Generate the values of 10 attributes, $X_1, ..., X_{10}$ independently, each of which is uniformly distributed over $[0, 1]$. Obtain the value of the target variable Y using the equation:

$$Y = 10 * sin(\pi * X_1 * X_2) + 20 * (X_3 - 0.5)^2 + 10 * X_4 + 5 * X_5 + \sigma(0, 1)$$

house: Both the `house_8L` and `house_16H` datasets are concerned with predicting the median price of a house in a region based on demographic composition and a state of housing market. The number signifies the approximate difficulty of the task, L for low difficulty, H for high.

kin8nm: This is data set is concerned with the forward kinematics of an 8 link robot arm. Among the existing variants of this data set we have used the variant 8nm, which is known to be highly non-linear and medium noisy.

mv: This is an artificial dataset with dependencies between the attribute values.

newsPopularity: This dataset summarizes a heterogeneous set of features about articles published on Mashable[2] in a period of two years. The goal is to predict the number of shares in social networks (popularity) based on attributes such as the number of words, images, and videos in the article, publication time and other word-based derived features.

puma: This is a family of datasets synthetically generated from a realistic simulation of the dynamics of a Unimation Puma 560 robot arm. There are eight datastets in this family . The task in these datasets is to predict the angular accelaration of one of the robot arm's links. The inputs include angular positions, velocities and torques of the robot arm. The number in the name of the dataset indicates the number of features, 8 or 32.

qsar-chembl: This dataset contains QSAR data (from ChEMBL[3] version 17) showing activity values (unit is pseudo-pCI50) of several compounds on drug target ChEMBL_ID.

sulfur: These are measurements for the amount of sulfur recovered by recovery units that remove environmental pollutants from acid gas streams in industrial settings before they are released into the atmosphere. The features are gas and air flows.

yprop_4_1: A drug design dataset, used in Feng et al. [2003]. The task is to predict the toxic effects of a substance based on a number of chemical descriptors.

[2] https://www.mashable.com

[3] https://www.ebi.ac.uk/chembl/index.php/

References

Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and Regression Trees*. CRC press, 1984.

Jun Feng, Laura Lurati, Haojun Ouyang, Tracy Robinson, Yuanyuan Wang, Shenglan Yuan, and S. Stanley Young. Predictive Toxicology: Benchmarking Molecular Descriptors and Statistical Methods. *Journal of Chemical Information and Computer Sciences*, 43:1463–1470, 2003.

Jerome H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19:1–67, 1991.