Data description

August 11, 2020

The following document contain a short description of the datasets used for the data competitions described in the paper.

1 Melbourne properties price dataset

The dataset '*Melbourne properties price*' contains sales prices of different types of residential properties in Melbourne area along with other characteristics of the properties.

The variables and their descriptions prevented in table 1:

Variable	Description
price	Price house sold for in AUS dollars
suburb	Suburb: different areas of Melbourne
land_size	Area of the land in square meters
house_size	Area of the house or the apartment in square meters
nbeds	Number of bedrooms
nbaths	Number of bathrooms/toilets
ncars	How many cars fit in the carport/garage
result	S - property sold; SP - property sold prior; PI - property passed in; VB - vendor bid; SA - sold after auction
agent	Selling agent
property_type	h house; u appartment; t town-house or unit
nvisits	Number of visitors before the selling
rating	The (rounded) average rating by the visitors, 0 to 10, 0-"awful" to 10-"unbelivebly lovely"
year	YYYY Year sold
month	MM Month sold
day	DD day sold
id	Unique id for each house

Table 1: Melbourne properties price dataset: variable description.

The data were collected between Feb 2, 2013 and Dec 17, 2016. The training set and the test set contained 75,366 and 25,122 observations, respectively. The associated errors were calculated as absolute errors: $\frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$.

2 Spam dataset

The '*Spam*' data contains few thousands emails that were classified by their owners as either spam or non-spam (3066 in the training set and additional 1533 in the testing set). Each email message also includes many other characteristics.

The variables and their descriptions prevented in table 2:

Variable	Description
numRecipients	The number of people in the To: or Cc: lines
Domain	Domain name of the sender's email address
replyToSameAddress	TRUE/FALSE
Weekday	Day of the week
Hour	Hour of the day
percentLowerCase	% of lower case letters in the subject line
numDigits	Number of digits in the sender's email address
percentNonLetters	% of non-letters in the sender's email address
Size	Size of the email in kb
numLinks	Integer number of links included in the email
localSender	Indicator - was the sender in the local domain as the
receiver credit, porn, sucker, pharm, prescription, drugs, save, sex, dis- creet, free, sell, sale, asseenon, discount	Binary variables computed by the presence of certain key words
newsletter	Indicator - is this email basically a newsletter
Spam	Indicator - classification of the email as spam or not by the receiver
id	Unique id for each email

Table 2: Spam dataset: variable description.

The mean consequential error (MCE) was selected as the method for prediction accuracy measurement. The mean/average of the "consequential error", where all errors are equally bad is given by:

$$MCE = \frac{1}{N}\sum_{y_i \neq \hat{y}_i} 1$$