Machine Learning

Exercise "Data set description"

Gernot Steindl
gernot.steindl@tuwien.ac.at

Jürgen Pannosch
juergen.pannosch@tuwien.ac.at

Bernd Windholz
bernd.windholz@ait.ac.at

March 25, 2020

1 Introduction

In this exercise two data sets were chosen and evaluated regarding the statistical characteristics. The first data set is called *Communities and Crime Data Set* and is online available[1]. The second data set is called *Occupancy Detection* which is also online available[2]. In the following sections 2 and 3 there is a detailed description of the individual data sets.

2 Data set: Communities and Crime

The *Communities and Crime Data Set* was chosen due to its relatively small number of samples (1994), its relatively high number of dimensions (128), and because it has also missing values. The associated task of the data set is regression, which makes it a perfect candidate for exercise 0 "Data set description".

According to the attribute information on the website, the data set consists of 122 predictive, 5 non-predictive, and 1 goal attribute. The goal attribute is the total number of violent crimes per 100K population in communities within the United States. The data combines socio-economic data (1990 US Census), law enforcement data (1990 US LEMAS[3] survey), and crime data (1995 FBI UCR[4]).

Attributes are nominal (e.g. *US state*), ordinal (*LemasGangUnitDeploy*: gang unit deployed ([0, 0.5, 1] mean [no, part time, yes])), but mostly ratio quantities (e.g. the goal *ViolentCrimesPerPop*: total number of violent crimes per 100K population). The

[1] `http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime`
[2] `http://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+`
[3] Law Enforcement Management and Administrative Statistics
[4] Uniform Crime Reporting

distribution of the goal attribute is shown in Figure 1. For prediction purposes the distribution of the goal attribute in the training data should cover the whole range. I.e. missing ranges in the training data can lead to poor prediction results in this range.

There is only one categorial attribute in the data set, the ordinal attribute *LemasGangUnitDeploy* (gang unit deployed): [0, 0.5, 1] mean [no, part time, yes]. In the follow-up exercises, when applying algorithms to the data, it has to be considered, that the values can only be exactly one of these discrete categories (e.g. pre-processing by one-hot encoding).

The distribution of the 122 predictive attributes is shown in Figure 2. As can be seen in the histograms, the predictive and the goal attribute values are normalized to the range from zero to one. Compare the data set information on the website:

> All numeric data was normalized into the decimal range 0.00-1.00 using an unsupervised, equal-interval binning method. Attributes retain their distribution and skew (hence for example the population attribute has a mean value of 0.06 because most communities are small).
>
> ...
>
> all values more than 3 SD above the mean are normalized to 1.00; all values more than 3 SD below the mean are normalized to 0.00

Figure 3 shows the relative number of missing values in the attributes (non-predictive attributes are not shown). Since there are many missing values in 23 of the 122 predictive attributes (no missing values in the goal attribute), further pre-processing of the numeric values will be necessary (e.g. deletion of samples or variables, imputation).

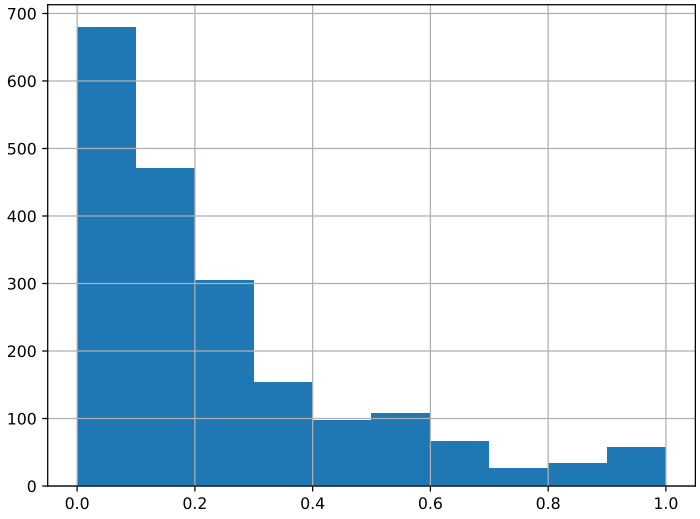

Figure 1: Histogram of the goal attribute of data set *Communities and Crime*

3 Data set: Occupancy Detection

The *Occupancy Detection* data set was chosen due to its relatively high number of samples (20560) and its small number of dimensions (7). There is one goal attribute (Occupancy). The data set can be used for binary classification of the occupancy variable. The available input features are:

- date - Timestamp (YYYY-mm-dd hh:MM:ss)
- Temperature - °C
- Humidity - %
- Light - Lux
- CO2 - ppm
- HumidityRatio - $\mathrm{kg}_{water-vapor}/\mathrm{kg}_{air}$

The binary occupancy variable is an ordinal quantity, the input features are numerical quantities. Of these, temperature and date are interval quantities, the other features are ratio quantities.

The whole data set is divided into the three parts Training, Test and Test2 (same features but different date ranges). Figure 4, Figure 5, and Figure 6 show the histograms of the training and the two test data sets. Table 1 shows the min, max, and mean values of the attributes in the three data sets. There are no missing values within each of the three data sets, i.e. for all available timestamps all other features are valid.

		Temperature °C	Humidity %	Light Lux	CO2 ppm	Humidity Ratio g/kg
Training (4.-10.2.2015)	min	19.00	16.74	0.00	412.75	2.67
	max	23.18	39.11	1546.33	2028.50	6.48
	mean	20.61	25.73	119.51	606.54	3.86
Test (2.-4.2.2015)	min	20.20	22.10	0.00	427.50	3.30
	max	24.40	31.47	1697.25	1402.25	5.38
	mean	21.43	25.35	193.22	717.90	4.03
Test2 (11.-18.2.2015)	min	19.50	21.86	0.00	484.66	3.27
	max	24.39	39.50	1581.00	2076.50	5.77
	mean	21.00	29.89	123.06	753.22	4.59

Table 1: Basic statistics of the *Occupancy Detection* data sets

Due to the very different ranges of the values, normalization (e.g. z-score standardisation or min-max scaling) will be required for algorithms that use distances (if this is not done by the algorithm itself).

The sampling rate of the time series is about one minute. Figure 7 shows the time series for the temperature attribute of the test data set.

Figure 2: Histograms of the 122 predictive attributes of data set *Communities and Crime*

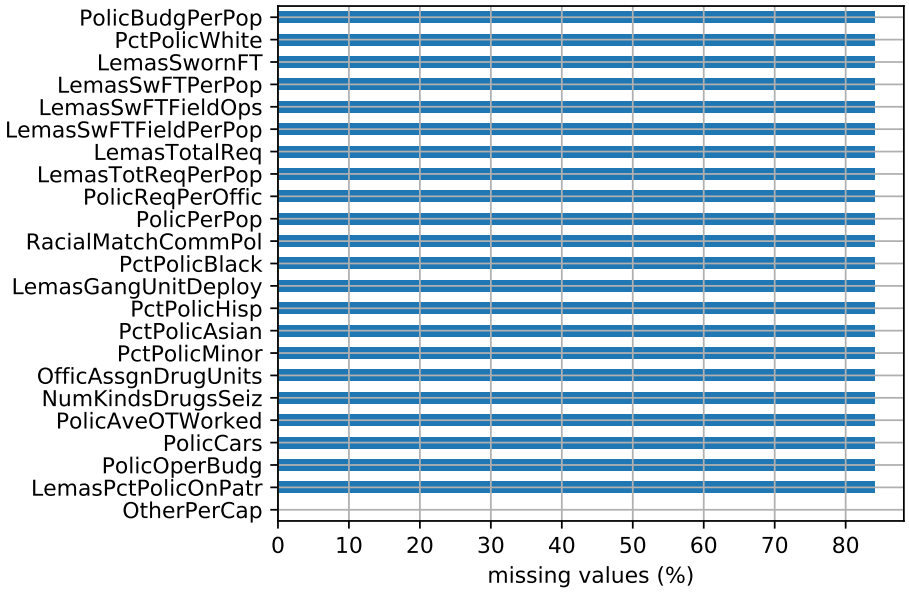

Figure 3: Relative number of missing values (23 of 122 predictive attributes) of data set *Communities and Crime*

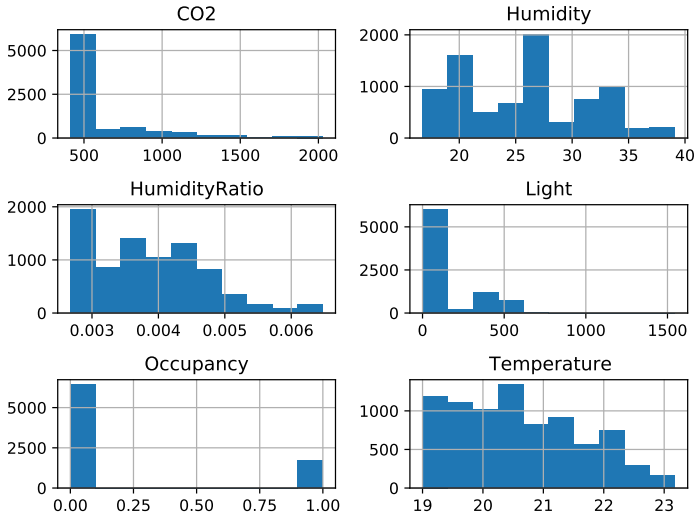

Figure 4: Histograms of the attributes of the *Occupancy Detection Training* data set

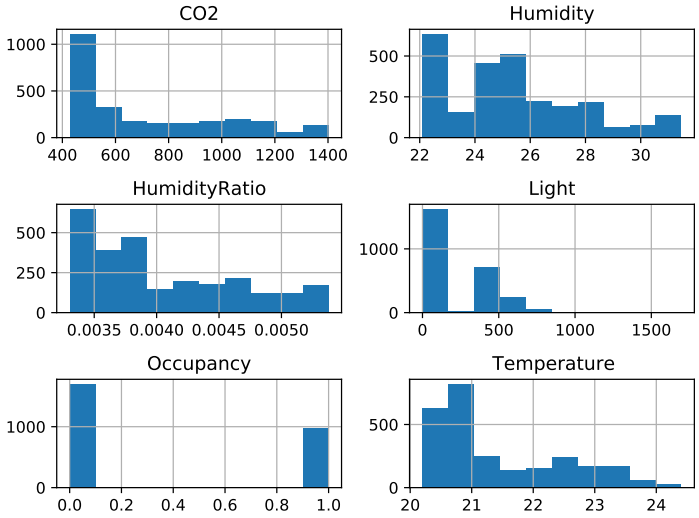

Figure 5: Histograms of the attributes of the *Occupancy Detection Test* data set

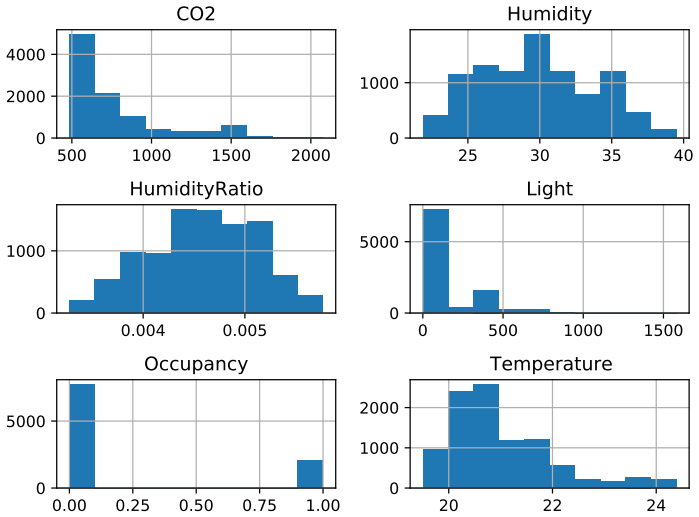

Figure 6: Histograms of the attributes of the *Occupancy Detection Test2* data set

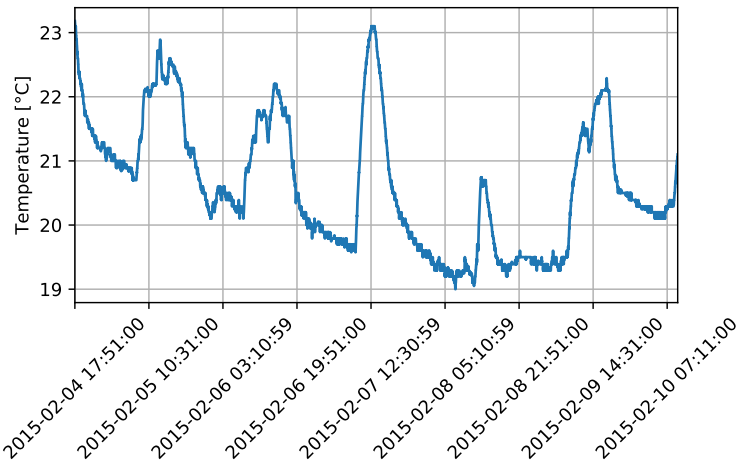

Figure 7: Time series of the temperature feature (2015-02-04 17:51:00 - 2015-02-10 09:33:00)