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 22, 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 of the source, 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 means NO, 1 means YES, 0.5 means part time), but mostly ratio (e.g. the goal *ViolentCrimesPerPop*: total number of violent crimes per 100K population).

[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

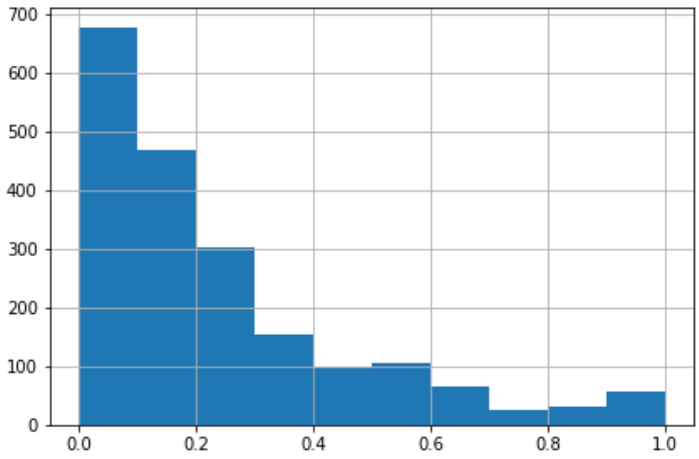

Figure 1: Histogram of the goal attribute of data set *Communities and Crime*

The distribution of the goal attribute shows 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 means NO, 1 means YES, 0.5 means part time. In the follow-up exercises it has to be considered, that the values can only be exactly one of these discrete categories.

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:

> 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

From today's point of view, further pre-processing of the numeric values is not necessary.

3 Data set: Occupancy Detection

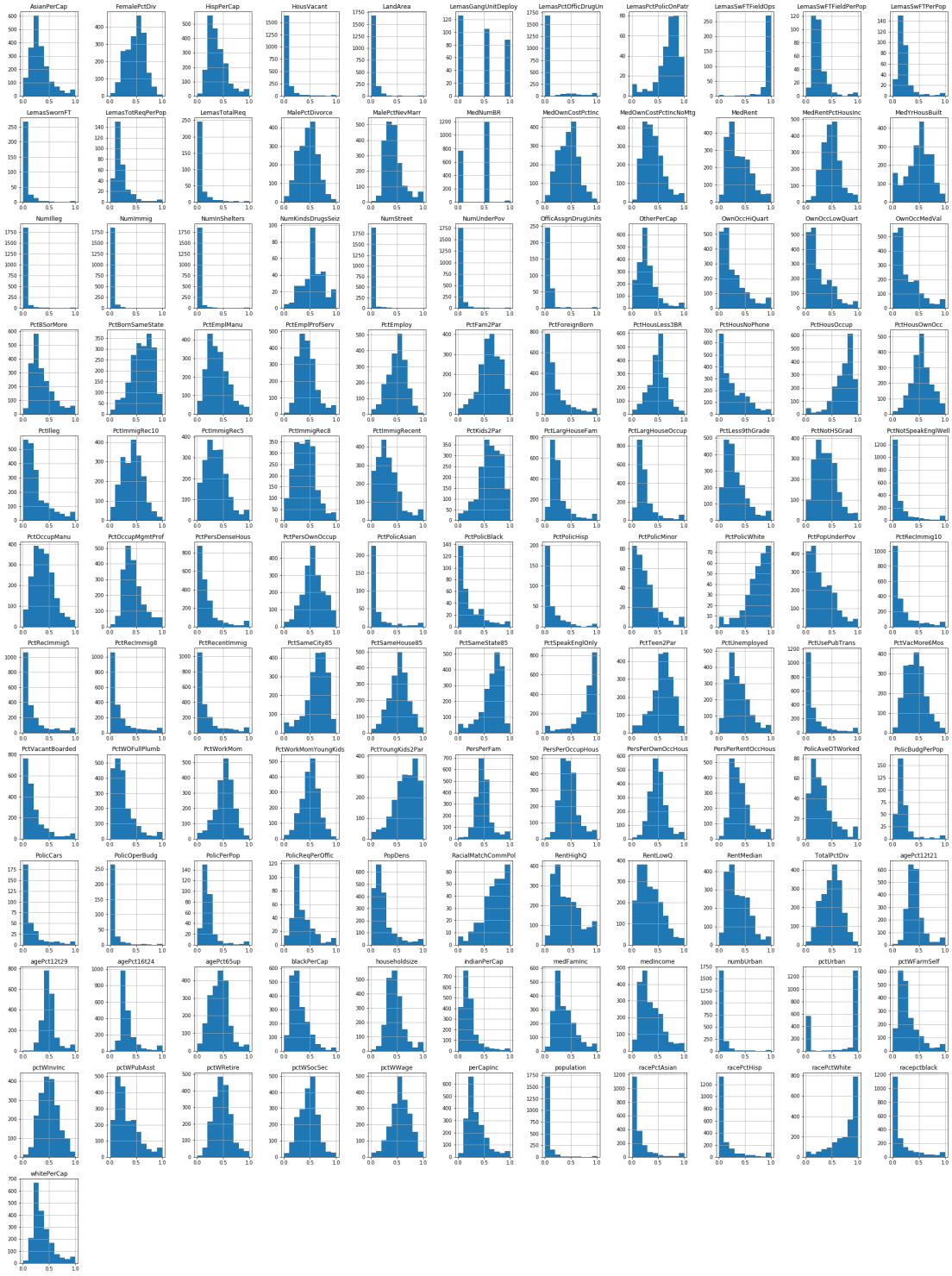

Figure 2: Histograms of the 122 predictive attributes of data set *Communities and Crime*