TECHNISCHE UNIVERSITÄT WIEN

PROJECT REPORT: 184.702 MACHINE LEARNING — 2022S, EXERCISE 0

Group 8: Beck Viktor, Simhandl Stefan, Wanecek Wilhelm

March 18, 2022

1 Choice of data sets

We've chosen following data sets:

- Communities Crime data set (regression)
  Source: `https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime`
- Kinematics Motion data set (classification)
  Source: `https://www.kaggle.com/yasserh/kinematics-motion-data`

Our main goal was to choose diverse data sets covering the given requirements: number of samples — small vs. large, number of dimensions — low vs. high dimensional and at least one data set shall come with missing values. Furthermore the data should be easy to understand, allowing us to focus on learning several machine learning models rather than intricate pre-processing procedures. For the Kinematics data-set we're interested in how precisely the models predict the activity state without any pre-computation of the accelerations with reference to the sensors 3D orientation.

2 Data exploration

To explore the data we've used a Jupyter notebook, the fundamentals of the analysis established using the module "Pandas Profiling". Pandas Profiling creates an HTML report which describes basic data set aspects. Further it plots the distribution of each feature, computes correlations between the features and provides some additional information. We refer to the HTML-documents for the distribution plots, rather than repeating them in this report.

3 Communities Crime data set (regression)

This data set consists of 128 attributes over 1994 instances of socioeconomic, law enforcement and crime data. It was collected in the United States in the years 1990 and 1995. The machine learning goal (target attribute) is to predict the number of violent crimes per population.

3.1 Data set description

After dropping non predictive attributes (state, county, community, communityname, fold) in a preprocessing step, the data set now only consists of numeric attributes (50 interval, 73 ratio). In contrast to the Kinematic data set, it has a fairly high number of missing values (39 202 cells, or 15.4 %) mostly due to some states not collecting statistics on rape, as well as on the employment statistics of police officers. Beside that, certain non-predictive columns included — e.g. 'county' stating the numeric code for a county — has an especially high ratio of missing values (in this case 58.9 %).

The target attribute (number of violent crimes, normalized against population) seems to have an exponential distribution, with the exception of a higher-than-expected number of values for the maximum value (1).

3.2 Notes on the data set and analysis

- The raw data provided by file "communities.DATA" and was already in a kind of csv format, the info about feature names manually extracted from file "communities_info.INFO". We finally merged those sources manually into "communities_crime.csv" file which feeds the data exploration.
- 5 of the 128 attributes were dismissed as they are non-predictable which was also stated in the data description.
- For those attributes which do not provide a ratio or descriptive statistics, we might have to think about some normalization like z-score as an additional pre-processing step if required for the ML model.
- We executed Pandas Profiling in "minimal mode" to reduce resource consumption respectively the size of the HTML report which only provides basic data set aspects and the variable distribution plots.

4 Kinematics Motion data set (classification)

The Kinematics motion data set contains over 88 000 sensor data samples, collected by an accelerator and gyroscope of a smartphone (iPhone 5c). Both sensors measure in 3D space, in addition the data set provides the wrist information — left or right — resulting in a total of 7 input attributes. The task is to classify which activity — walking or running — the user is performing based on the sensor data and which wrist the phone is attached to.

Given the well-formatted input, the pre-processing required was minimal. Indeed, the main step besides importing the csv was to drop the meta-data columns (`date, time,` and `name`).

4.1 Data set description

The data set covers 88 588 observations, each with 8 variables: `wrist`, `activity`, `acceleration_x`, `acceleration_y`, `acceleration_z`, `gyro_x`, `gyro_y`, and `gyro_z`. The acceleration and gyro attributes are interval numbers (see Table 1 for the range of values), with the unit for acceleration being the relative acceleration (with $1 = 9.81ms^{-2}$), and the gyro being reported in radians per second. The wrist information (left or right wrist) as well as the target attribute "activity" contains nominal data (each with two distinct values).

The data set is completely without missing values.

Attribute	Min	Mean	Max
acceleration_x	-5.35	-0.07	5.56
acceleration_y	-3.30	0.56	2.67
acceleration_z	-3.75	-0.31	1.64
gyro_x	-4.43	0	4.87
gyro_y	-7.46	0.04	8.50
gyro_z	-9.48	0.02	11.23

Table 1: Minimum, mean, and maximum of numeric attributes in Kinematic motion data set.