Machine Learning - SS 2021
Exercise 0

Federico Ambrogi, e1449911@student.tuwien.ac.at
Adam Höfler, e11847620@student.tuwien.ac.at
Matteo Panzieri 12039996@student.tuwien.ac.at
TU Wien

March 25, 2021

Contents

Introduction

In this report we describe and analyse the main features of the datasets *Drug consumption (quantified) Data Set* [1] and *Cuff-Less Blood Pressure Estimation Data Set* [2]. They differ in size of the data set, number of features, type of features, etc. to give us a variety of possible scenarios to work with.

1 Dataset: Cuff-Less Blood Pressure Estimation

The original data comes from the Multi-parameter Intelligent Monitoring in Intensive Care II (MIMIC II) online waveform database [3][4]. The data set was extracted from MIMIC II and reduced onto Photoplethysmograph (PPG), Electrocardiogram (ECG) and arterial blood pressure (ABP) waveform signals [5]. The data was collected between the years 2001 and 2008 from a variety of Intensive Care Units and were sampled at rate of 125 Hz with 8 bit accuracy. The preprocessing on the original included the smoothing of the waveforms and the removal blocks with a) unreasonable blood pressures, b) unreasonable heart rate, c) severe discontinuities and d) big differences in the PPG signal correlation between neighbouring blocks.

Dataset The data set contains 12,000 rows with 3 attributes each, i.e. the PPG, ECG and ABP. Due to the preprocessing, there should be no missing values or outliers. The features are of numeric type.

The signals and the distributions of their values, can be inspected in Fig. 1.

Task This data set is our choice for the regression part with ABP being the target variable. The most common approach to measure (BP) is by using a cuff, wrapped around the upper arm and inflated. A pressure sensor inserted in the cuff records the arterial pulsations during the cuff deflation, and the amplitudes of these pulsations are used to calculate systolic and diastolic BP. The limitations of the approach include that it relies on a set of empirical coefficients to map the pulse amplitudes to systolic

PPG signal	Ratio
ECG signal	Ratio
ABP signal	Ratio

Table 1: Features & Data type of the Life Expectancy data set

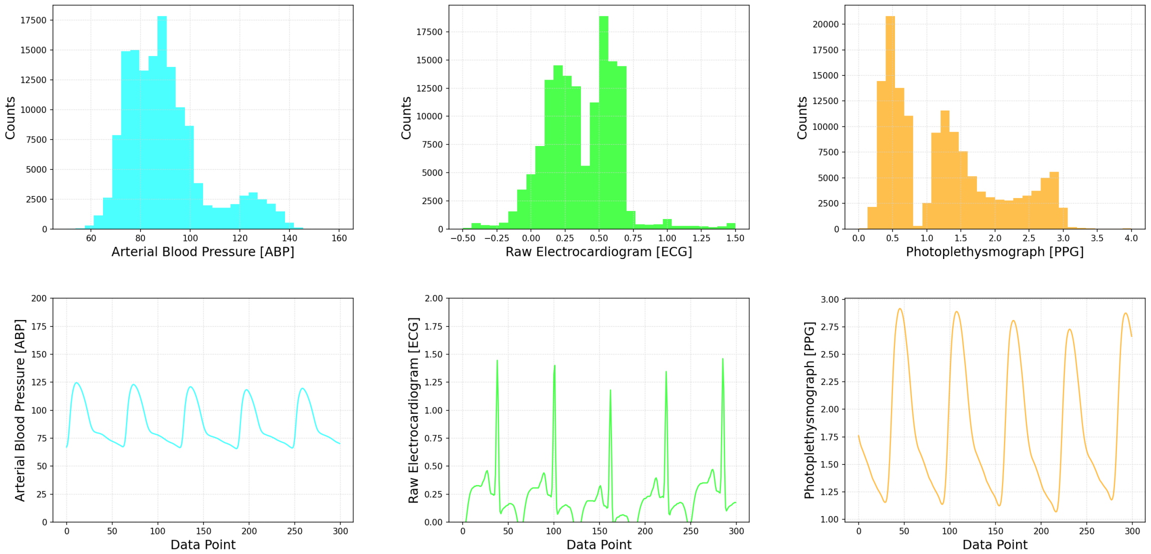

Figure 1: Distribution for the ABP, ECG and PPG signals.

and diastolic BP, specific to the used device. Furthemore, this kind of measurements in patients with atherosclerosis or obese patients, whose pulse amplitudes can be weak, are subject to large errors.

For these reasons, it is very interesting to study the possibility to study a regression model able to predict the BP by using the PPG and/or the ECG signals.

2 Drug consumption Data Set

This dataset [6] includes records of drug usage of 1885 respondents, together with attributes regarding personal information. For each respondent 12 attributes are known: Personality measurements which include NEO-FFI-R (neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness), BIS-11 (impulsivity), and ImpSS (sensation seeking), level of education, age, gender, country of residence and ethnicity.

All input attributes are originally categorical and are quantified. After quantification values of all input features can be considered as real-valued. In addition, participants were questioned concerning their use of 18 legal and illegal drugs (alcohol, amphetamines, amyl nitrite, benzodiazepine, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushrooms, nicotine and volatile substance abuse and one fictitious drug (Semeron) which was introduced to identify over-claimers. For each drug they have to select one of the answers: never used the drug, used it over a decade ago, or in the last decade, year, month, week, or day. Database contains 18 classification problems. Each of independent label variables contains seven classes: ”Never Used”, ”Used over a Decade Ago”, ”Used in Last Decade”, ”Used in Last Year”, ”Used in Last Month”, ”Used in Last Week”, and ”Used in Last Day”.

Dataset The data set contains xxx TO DO.

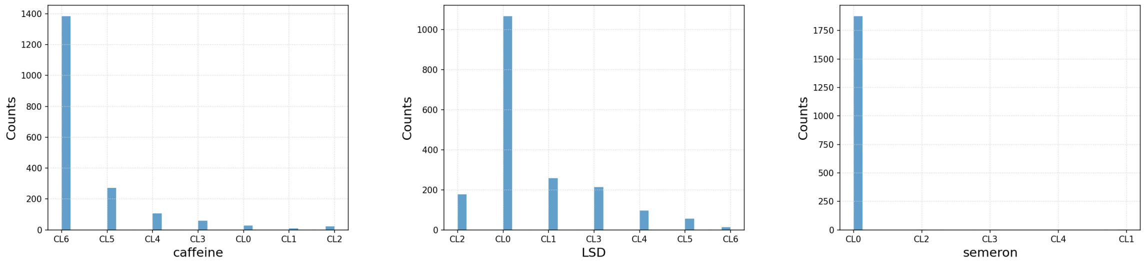

Figure 2: Distribution for drugs.

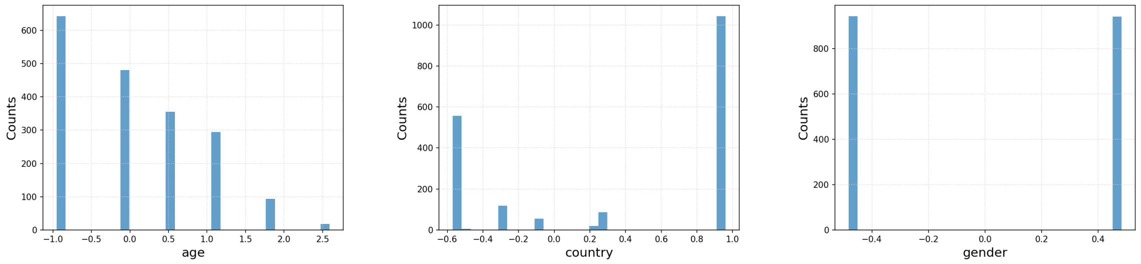

Figure 3: Distribution for drugs.

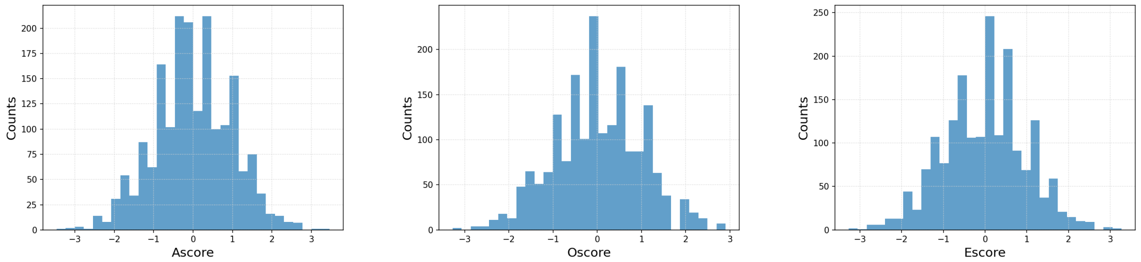

Figure 4: Distribution for drugs.

Task This data set is very appropriate for classification experiments, given the partition of the data attributes into different classes.

References

[1] Evgeny M. Mirkes et al. Uci - drug consumption (quantified) data set. https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29. Accessed: 25.03.2021.

[2] Mohamad Kachuee et al. Uci - cuff-less blood pressure estimation data set. https://archive.ics.uci.edu/ml/datasets/Cuff-Less+Blood+Pressure+Estimation. Accessed: 19.03.2021.

[3] Physionet - multiparameter intelligent monitoring in intensive care (mimic) ii. https://archive.physionet.org/mimic2/. Accessed: 20.03.2021.

[4] A. Goldberger, L. A. Amaral, L. Glass, Jeffrey M. Hausdorff, P. Ivanov, R. Mark, J. Mietus, G. Moody, C. Peng, and H. Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation*, 101 23:E215–20, 2000.

[5] Mohamad Kachuee, Mohammad Mahdi Kiani, Hoda Mohammadzade, and M. Shabany. Cuff-less high-accuracy calibration-free blood pressure estimation using pulse transit time. *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1006–1009, 2015.

[6] E. Fehrman, A. K. Muhammad, E. M. Mirkes, V. Egan, and A. N. Gorban. The five factor model of personality and evaluation of drug consumption risk. 2017.