Machine Learning

Spring Semester 2021
Prof. Dr. Peter Zaspel
Abhieshree Dhami, Kristijan Spirkoski, Kristian Sterjo

JACOBS UNIVERSITY

Assignment Sheet 3.

Submit on **Tuesday, March 2, 2021, 10:00**.

Excercise 1. (Modelling inputs / outputs)

In this exercise you work with two data sets:

- Statlog (Shuttle) Data Set
- Computer Hardware Data Set,

which are both available in the UCI Machine Learning Repository. For each of them, perform the follwing tasks:

a) Briefly describe the data set and all involved variables in your own words. If some information is missing on the UCI Repository site, do your own search for these details.

b) Model the data set via input and output random variables / vectors.

c) Formulate a question that can be solved using machine learning on this data set and give the type of machine learning (supervises / unsupervised/ regression / classification) that will allow to answer the question.

(8 Points)

Excercise 2. (SPAM e-mail representation)

The Spambase Data Set is a SPAM classification data set that has exactly the 57 input variables that are roughly described in Example 2.9 of the lecture.

a) In Example 2.9, we did not mention the specific words and characters that are used in the features. Use the information of the UCI Repository to make a complete description of these variables, i.e. give all the key words, etc.

b) Search the web for alternative features that can be used to describe (SPAM) emails. Pick one example feature set, cite the source, and describe these features.

(4 Points)

Programming Exercise 1. (SPAM e-mail representation)

Implement a code that reads a text file and extracts the 57 features discussed in the previous task. Apply the feature extractor to the provided three example emails.
Reference solutions will only be provided in Python. The submission format for Python is a Jupyter notebook. The submission format for C/C++ is standard source files.

(4 Points)