Final Course Assignment

Mark Niehues, Stefaan Hessmann
Mathematical Aspects in Machine Learning

13. Juli 2017

1 Introduction

In the past course we dealt with the broad mathematical foundations of machine learning. To get an idea of what the consequences of those mathematical theorems and approaches are and to get in touch with the standard Python tools, we have evaluated an comparatively easy data science example found on kaggle.com. The example dataset [1] consists of the historic passenger records of the disastrous Titanic maiden voyage in 1912.

Listing 1: Hello World

```
#   Copyright (C) 2017 Mark Niehues, Stefaan Hessmann
#
#   This program is free software: you can redistribute it and/or modify
#   it under the terms of the GNU General Public License as published by
#   the Free Software Foundation, either version 3 of the License, or
#   (at your option) any later version.
#
#   This program is distributed in the hope that it will be useful,
#   but WITHOUT ANY WARRANTY; without even the implied warranty of
#   MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
#   GNU General Public License for more details.
#
#   You should have received a copy of the GNU General Public License
#
```

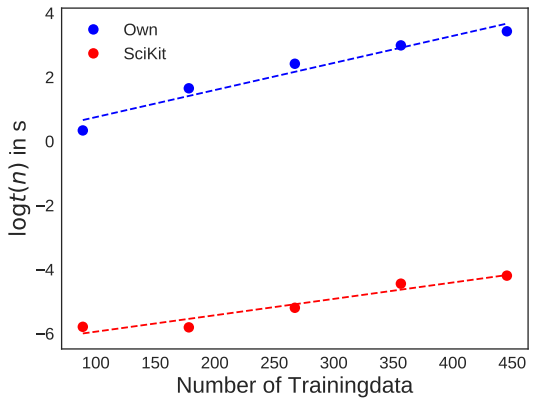

Abbildung 1: Benchmark

Feature	Description	Missing Data (%)
PassengerId	Unique ID for every passenger	0.0
Survived	Survived (1) or died (0)	0.0
Pclass	Passenger's class	0.0
Name	Passenger's name	0.0
Sex	Passenger's sex	0.0
Age	Passenger's age	19.87
SibSp	Number of siblings/spouses aboard	0.0
Parch	Number of parents/children aboard	0.0
Ticket	Ticket number	0.0
Fare	Ticket-price	0.0
Cabin	Number of the passenger's cabin	77.10
Embarked	Port of embarkation	0.22

Tabelle 1: Description of the dataset.

```python
import numpy as np


class Kernels:
    """
    Class that holds different Kernels
    """
    def __init__(self, gamma):
        self.gamma = gamma
        self.kernels = {
            "rbf" : self.kernel_rbf,
            "linear": self.kernel_lin}

    def get_kernel(self, kernel_name):
        return self.kernels[kernel_name]

    def kernel_lin(self, x, y):
        """
        Linear kernel
        """
        return x.dot(y)

    def kernel_rbf(self, x, y):
        """
        RBF Kernel

        """
        d = x - y
        return np.exp(-np.dot(d, d) * self.gamma)
```

2 Applying Machine Learning Methods on the Titanic Disaster

2.1 Dataset

The given dataset consists of a CSV-file containing data of 891 passengers. The dataset contains an ID for every passenger, a label if the passenger has survived the disaster and the features that are described in table 2.1. It can be noticed that some of the features are incomplete.

After loading the dataset, it is necessary to process the data for our learning machine. Therefore the different features will be investigated to select meaningful features and the missing data needs to be handled.

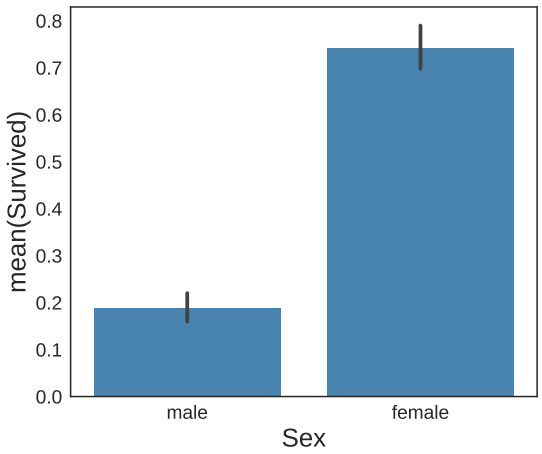

Abbildung 2: Distribution of survivors distributed by their sex.

2.2 Feature: Sex

3 Implementation of an easy SMO Algorithm

Literatur

[1] Kaggle. *Titanic: Machine Learning from Disaster*. 13. Juli 2017. URL: https://www.kaggle.com/c/titanic.