UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

A report on

Breast Cancer Wisconsin Data Set

Final Project

Heidelberg Collaboratory
for Image Processing

February 23, 2018

Submitted to:

Prof. Dr. Ulrich Köthe & Tutor: Lynton

Submitted by:

Jannik Lukas Kossen (JK)
Student ID: 3495228
M.Sc. Physics, ungraded

Ahmad Neishabouri (AN)
Student ID: 3436580
M.Sc. Scientific Computing, graded

Contents

1 Introduction (JK)

The following pages contain the report on our final project for the *Fundamentals of Machine Learning* lecture given by Dr. Köthe in the last semester. As the lecture focussed on the *fundamentals* of machine learning we see this project as an opportunity to explore and revise again some fundamental techniques of machine learning. Often, the application potential of these techniques on real-world tasks is rather small. However, they do constitute the foundations of more recent and better algorithms, which often embed or build upon these basic methods. The focus of this project shall therefore not be to *solve* a task using fundamental machine learning, but rather to carefully study the different algorithms and compare their strengths and shortcomings in comparison with each other. To narrow the focus of the project a bit, we have chosen to only compare *classification* methods. We work with the *Breast Cancer Wisconsin (Diagnostic) Data Set* (1), a binary classification task with the aim to identify whether a person does or does not have breast cancer based on a set of hand-crafted features. The data set should work very well for benchmarking algorithms, providing a nice balance between being well-behaved, i.e. solvable, and possessing enough depth and character to allow for some comparison between the algorithms, i.e. not being trivial to solve.

Structure. The structure of this report is as follows. Section 2 discusses the theoretical foundations of each of the compared methods. In section 3, unsupervised exploration techniques are employed to gain some understanding of the shape of our data set. The classification methods are applied and benchmarked. Section 4 then discusses these results in relation to peculiarities of each method and the data set.

As demanded by the faculty of Mathematics and Computer Science, each section or paragraph heading also contains a shorthand symbol, *(JK)* or *(AN)*, to indiciate whether a certain section was written by Jannik Kossen or Ahmad Neishabouri.

2 Theoretical Background

Preliminary Remarks (JK). Note that, unless marked otherwise, the theoretical derivation and description attempt follow the notation of the lecture (**?**) for clarity. This includes the convention to describe the data as $X = \left\{X_1, X_2, \ddots, X_N\right\}$, where N is the number of instances and each X_i is a D-dimensional feature (row) vector. The feature j of instance i is therefore given by X_{ij}. Index i is solely used to index along the N-dimensional instance axis, whereas j solely indexes the D-dimensional feature axis.

3 Experiments

3.1 Data Set – Breast Cancer Wisconsin (JK)

Before we

4 Discussion

5 Summary

6 Appendix

References

[1] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, “Nuclear feature extraction for breast tumor diagnosis,” in *Biomedical Image Processing and Biomedical Visualization*, vol. 1905, pp. 861–871, International Society for Optics and Photonics, 1993.