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 focused 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* (**?**), 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.

Data Set. The abovementioned data set was first published alongside a paper called *Nuclear Feature Extraction For Breast Cancer Diagnosis* (**?**) in the 1993 edition of the *International Symposium on Electronic Imaging: Science and Technology*, wherein a classification pipeline for breast cancer diagnosis is proposed. A small sample of skin tissue is extracted from the patients breast and analyzed using a camera mounted atop a microscope. Using an interactive tool the operator identifies individual cell nuclei, whose outlines are then semi-automatically determined. For each nucleus, ten features that are sought to describe the geometric shape and appearance of the nucleus are crafted. These features are the radius, perimeter, area, compactness, smoothness, concavity, concave points, symmetry, fractal dimension, and texture of the nucleus. For the sake of brevity, the reader is referred to the original publication (**?**) for a detailed description of how these features are calculated from the image and outline of a nucleus. The mean, max, and standard deviation values of these features is then calculated for the nuclei of a single sample, giving in total a 30-dimensional feature vector describing each cell nucleus. The hope here is obviously, that these features (or a subset thereof) capture the differentiating aspect of tumor-ridden cell nuclei from their healthy counterparts. The data set is explored further in **??**.

Structure. The structure of this report is as follows. **??** discusses the theoretical foundations of each of the compared methods. In **??**, unsupervised exploration techniques are employed to gain some understanding of the shape of our data set. The classification methods are applied and benchmarked. **??** 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 indicate 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 $X = X_1, X_2, \ldots, X_N$, 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. The labels or ground truth instances of supervised learning are given as Y_i.

(k-) Nearest Neighbor Classification (JK). One of the simplest approaches to classification is the the nearest neigbor (NN) classifier. Test set instances i are simply classified with the same class as their *nearest* neighbor from the training set (TS). Which neighbor is the *nearest* is given by some distance metric $d(X_i, X_{i'})$ between to instances X_i and $X_{i'}$. $d(X_i, X_{i'})$ is often chosen to be simply the Euclidean distance. In mathematical terms, the decision function for the NN classifier can be written as

$$f_{\text{NN}} = Y_i \quad \text{, where } i = \underset{i' \in \text{TS}}{\arg\min}\, d(X_i, X_{i'})\,.$$

The NN classifier requires the total memorization of the training set which becomes problematic for larger data sets. A simple variant of the NN classifier is the k-nearest neighbor algorithm, which classifies instances by taking a majority vote of the classes of the surrounding k-nearest neighbors of the instance that is to be classified. In contrast to the simple NN classifier, the k-NN classifier is consistent, i.e. the k-NN classifier converges to the optimal Bayes classifier as $N- > \infty$.

3 Experiments

3.1 Exploratory Analysis (JK)

Before the introduced classification methods are applied, it is crucial that some time is spent on exploring the *Breast Cancer Wisconsin Data Set*. The data set consists of $N = 569$ instances with $D = 30$ dimensions per instance. The data is somewhat imbalanced as 63 % of the instances are false, i.e. come from benign cells without breast cancer. This might have to be accounted for in the analysis. Data will be centered and standardized according to the training set distribution unless otherwise mentioned.

3.2 Performance Evaluation (JK)

In order to provide reliable performance analysis, 10-fold cross-validation as described in the lecture is employed for each algorithm. A strict separation of training and test set is ensured such that meaningful statements over the predictive power of the methods can be made. Whenever an algorithm has hyperparameters an exhaustive grid search is performed in order to find the best-performing set of parameters. *Performance* is given in terms of precision-recall graphs.

4 Discussion

5 Summary

6 Appendix