Mini-project: Computer Aided Diagnosis

- Group 6 -

R.J.P. van Bergen (1234032), N.N.H.C. Bodenstaff (1223152), S.H.A. Verschuren (1245272)

Technical University Eindhoven, Dept. of Biomedical Engineering

October 24, 2019

Contents

1 Introduction

About 1 in 8 women in the United States develops invasive breast cancer over the course of her lifetime. For men this statistic is slightly more optimistic; about 1 in 883 [1]. It is partly because of this relatively high incidence that there has been a vast amount of research dedicated to this disease. The war against cancer has produced steady progress over the past four decades, showing increasing survival rates, from 74.8% in 2003 to 90.3% in 2009 [2]. One challenge in increasing the survivability is being able to identify the cancer. One of the ways to do this is to judge cells of interest by their nucleus size. Nowadays, pathologists often make qualitative judgements on the size of these nuclei. Quantitative evaluation would be a better method, however it takes too much time to do manually. Computer programs that can automatically return the total size of the nuclei under review and classify them by either "small" or "large" will be necessary to effectively diagnose cancer. This project will attempt to create such a function, using computer learning techniques. Linear regression is used to calculate the area of nuclei. Logistic regression classifies the areas of the nuclei in two classes, being "small" and "large".

2 Methods

2.1 Area measurement

The dataset used in this project consists of around 40000 pictures of 24x24 pixels with a cell in the center [3]. These are split into a training and testing set of 20000 samples each. In these images, each pixel will be used as a feature. Since the pixels all have rgb-values, the number of features is tripled. This results in 1728 features. Furthermore, a 1 is added to the features to account for the intercept constant.

To calculate the nucleic area of a given microscopy image, a linear regression model is used. This fitting was done using four different methods. These methods were fitting linear and quadratic regression on the complete data-set and down-sampled data-set. For the quadratic variant another 1728 features were added describing the squared intensities of each pixel. For the down-sampling the training set is reduced to one fourth of its size.

The model will be used to automatically measure the size of previously unseen samples, without the need of manual measurement. To compute the parameters, the following formula is used.

$$\theta = \left(X^\top X\right)^{-1} X^\top y$$

Where X is the training dataset of features, theta are the weights assigned to the features and Y are the targets.

The different regression methods are then compared using there normalized square difference to their true labels, which acts as an error function. The method which yields the lowest error will be used in the classification of the nuclei.

2.2 Nuclei Classification

Classification of the nuclei is done through a logistic regression model. The model is trained to classify nuclei into a 'large' (1) and a 'small' (0) class. For the classification, a probability function which determines the probability that a certain sample belongs to the 'large' class is needed.

$$p(y=1|\mathbf{x}) = \sigma\left(\boldsymbol{\theta}^\top \mathbf{x}\right)$$

Here a Sigmoid function is used to compress the values from a linear regression model to values between 0 and 1. A negative log-likelihood function is used as loss function.

$$J(\theta) = -\sum_{i=1}^{N} y_i \log p\left(y=1|\mathbf{x}_i,\theta\right) + (1-y_i)\log\left\{1-p\left(y=1|\mathbf{x}_i,\theta\right)\right\}$$

The learning rate (μ) is lowered after every iteration (k), by using the following function.

$$\mu = \mu_0 \cdot e^{-5\cdot k/50}$$

An optimization code is run before the classification to find the best combination of batch size and initial learning rate (μ_0). These combinations are compared using the loss function. These values are then used in the training of the classification model for the nuclei. This will be done on the full training data-set, and on a down-sampled version of the training data-set (5%) for illustration and comparison. The training was done using a analytic expression of the loss function instead of using a numerical gradient. This was done to decrease the computing time of the script. Also note that the algorithm has an inbuilt stop-function, which lowers the computing time further and helps to counteract over-fitting.

3 Results

3.1 Area measurement

Figure 1 shows a scatter-plot for predicted versus the actual area. The "perfect" result is added for visual context in the form of a red dotted line. The upper models are trained on all training samples, the bottom models on a down-sampled training set. The errors for linear regression, in both the full trainset and down-sampled set are respectively, 374.91 and 624.45. For second order regression these errors are, 423.38 and 1698.51.

3.2 Nuclei Classification

Using the optimization code the optimal batch size was set to 200 and for an initial learning rate of 0.0002 was chosen. Figure 2 shows the logistic regression model trained on the full data-set. It yields a classification accuracy (naturally measured on the test data-set) of 0.781, after 37 iterations. The losses for the training and validation data are 0.507 and 0.492 respectively. It can also be seen that the overall fitting and stabilization of the loss function occurs after approximately 15 iterations.
The model trained on a vastly down-sampled training data-set, as is shown in figure 3, yields losses for the training and validation data of 0.637 and 0.762 respectively, after 47 iterations. It yields a classification accuracy of 0.486.

4 Discussion

4.1 Area measurement

As seen in figure 1, there is a large spread around the line y=x. This imperfection is due to the large data-set and the nature of linear regression. When taking a look at the errors of the prediction the conclusion can be drawn that the lowest error occurs when using linear regression trained with the full training set. The regression trained using the down-sampled training set results in both cases (linear and quadratic) in a higher error. This can easily be explained by the fact that the full data-set provides more information for determining the parameters. It is also logical that the quadratic model is more admissible to the growth in error due to a smaller data-set, since it over-fits more easily due to a higher number of dimensions.
Something that also stands out is that the quadratic prediction model yields a higher error than the linear models. This is possibly due to the fact that in the training phase the model is over-fit on the training data. This proven by the model trained with the down-sampled data-set. Specifically the error is four times the error of the model using the complete training set. This supports the theory that more over-fitting occurs when fewer data-points are presented resulting in a larger error.

4.2 Nuclei classification

As can be seen in figure 2, the training- and validation loss functions for the "best" model are rather similar, having respective end-values of 0.507 and 0.492. This would suggest that the model is not over-fitting in any dramatic way. It is also apparent that the loss functions descend into their respective minima after relatively few (approximately 15) iterations, most likely due to the variable learning rate function. A decent final classification accuracy of 0.781 supports these theories.
This is in stark contrast to the results of the model trained on a down-sampled training set, whose loss functions have respective end values of 0.637 and 0.762. Not only are these loss values higher than the ones described by the first model, but the difference between the two suggests a rather dramatic over-fit. A final classification accuracy of 0.486, which is lower than 0.5, suggests that this model does no good at classifying nucleus sizes at all.

In conclusion, this project has brought forth a first-order linear classification model with variable learning rate, a batch-size of 200, a maximum number of iterations of 50 and a stop condition which is able to classify cell nuclei into two classes, being 'large' and 'small', with an accuracy of approximately 78%.

References

[1] U.S. Breast Cancer Statistics — Breastcancer.org.

[2] Masaaki Kawai, Kathleen E. Malone, Mei Tzu C. Tang, and Christopher I. Li. Active smoking and the risk of estrogen receptor-positive and triple-negative breast cancer among women ages 20 to 44 years. *Cancer*, 120(7):1026–1034, 2014.

[3] Mitko Veta, Paul J. Van Diest, and Josien P.W. Pluim. Cutting out the middleman: Measuring nuclear area in histopathology slides without segmentation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9901 LNCS, pages 632–639. Springer Verlag, 2016.

Appendices

A Figures

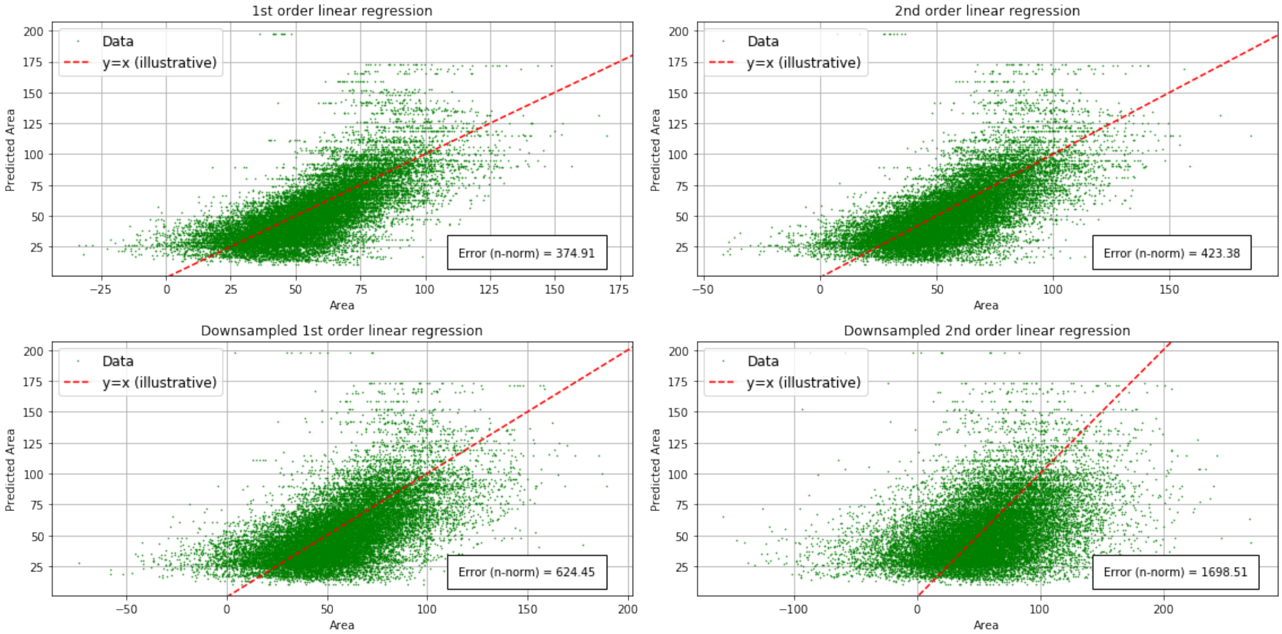
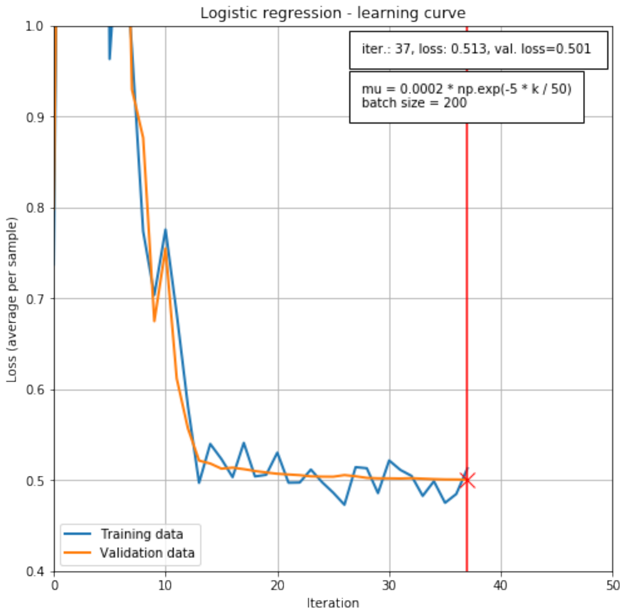


Figure 1: Predictions of the area of the test set plotted against the true area for linear and quadratic regression models trained with a normal and downsampled training datasets.



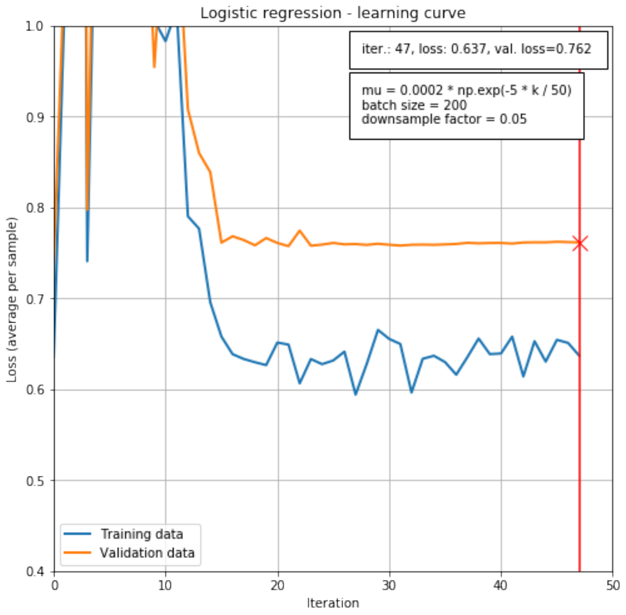


Figure 2: learning curve of the logistic regression on the full training set using μ_0= 0.0002 and batch size 200

Figure 3: learning curve of the logistic regression on 5 percent of training set using μ_0= 0.0002 and batch size 200